IIEEJ Transactions on

Image Electronics and Visual Computing

Special Issue on Journal Track Papers in IEVC2021 Part II

Special Issue on CG & Image Processing Technologies Supporting and Expanding Human Creativities

Vol. 10, No. 1 2022

The Institute of Image Electronics Engineers of Japan

Editorial Committee of IIEEJ

Contact Information

The Institute of Image Electronics Engineers of Japan (IIEEJ)
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan
Tel : +81-3-5615-2893　Fax : +81-3-5615-2894
E-mail : hensyu@iieej.org
http://www.iieej.org/ (in Japanese)
http://www.iieej.org/en/ (in English)
http://www.facebook.com/IIEEJ (in Japanese)
http://www.facebook.com/IIEEJ.E (in English)

IIEEJ Transactions on
Image Electronics and Visual Computing
Vol.10 No.1 June 2022

CONTENTS

Announcements

Guide for Authors

Published two times a year by the Institute of Image Electronics Engineers of Japan（IIEEJ）
3-35-4-101，Arakawa，Arakawa-ku，Tokyo 116-0002，Japan
Tel：+81-5615-2893　Fax：+81-5615-2894　E-mail：hensyu@iieej.org　http://www.iieej.org/

Upon the Special Issue on Journal Track Papers in IEVC2021 Part II

Editor: Osamu UCHIDA
(Tokai University)

The 7th International Conference on Image Electronics and Visual Computing (IEVC2021) was held on September 8-11, 2021, as the international academic event of Image Electronics Engineers of Japan (IIEEJ). It was based on the great success of the previous six workshops in 2007 (Cairns, Australia), 2010 (Nice, France), 2012 (Kuching, Malaysia), 2014 (Koh Samui, Thailand), 2017 (Da Nang, Vietnam), and in 2019 (Bali, Indonesia). The conference aims to bring together researchers, engineers, developers, and students from various fields in both academia and industry for discussing the latest studies, standards, developments, implementations, and application systems in all areas of image electronics and visual computing. The IEVC2021 Committee initially planned to hold IEVC2021 as a face-to-face conference in Shiretoko (Shari), Hokkaido, Japan, a World Natural Heritage Site. However, due to the worsening of the COVID-19 infection nationwide, the committee has decided to hold IEVC 2021 in a fully online format, prioritizing the safety of the participants.

As well as in IEVC2019, IEVC2021 has two paper categories: general paper and late-breaking paper (LBP). Moreover, the general paper category has two tracks: Journal track (JT) and Conference track (CT). Journal track provides authors with the advantage of having their papers published in the special issue of IIEEJ Transactions on Image Electronics and Visual Computing (IIEEJ Trans. IEVC) scheduled for the December 2021 issue (Part I), and this June 2022 issue (Part II).

This special issue Part II contains 9 papers that were accepted within the publication schedule of June issue. As the special issue Part I, published in December 2021, contained 4 papers, the number of accepted papers in Special Issues of JT papers in IEVC 2021 reached to 13 in total.

Last but not least, I would like to thank all the reviewers and editors for their contribution to improving the quality of papers. I would also like to express my deepest gratitude to the members of the editorial committee of IIEEJ and the staff at the IIEEJ office for various kinds of support.

IIEEJ **Paper** *-- Special Issue on Journal Track Papers in IEVC2021Part II --*

Robust Plane Detection in Terrestrial Laser Scanning Point Clouds using Efficient Multi-scale Sliding Voxels and Median Plane Weighted PCA

Jaime SANDOVAL† (*Member*), Kazuma UENISHI†† (*Member*),
Munetoshi IWAKIRI††† (*Member*), Kiyoshi TANAKA†† (*Fellow*)

† AB.do Corp., †† Shinshu University, ††† National Defense Academy of Japan

<**Summary**> Planes are simple shapes often found in urban environments. Their detection is crucial for several applications of 3D point clouds. Efforts have focused on detecting planes in low-cost and low-range scenarios. However, in Terrestrial Laser Scanning (TLS), the range is extended to a few hundred meters. This long-range adds huge variations in points' density that complicates defining proper thresholds for conventional methods. Therefore, we propose to employ the sliding voxel plane detector over multiple voxel sizes to estimate hypothetical planes in a coarse-to-fine way, followed by a merging Non-Maximum Suppression to detect planes robustly. According to experimental results, the proposed method shows superior precision, efficiency, and scalability in TLS point clouds.

Keywords: point couds, plane detection, ransac, hough transform, sliding voxel, PCA

1. Introduction

3D point clouds are sets of unorganized points that resemble the surface of objects. Recently, with the advent of low-cost sensors and photogrammetry, point cloud processing has gained the attention of researchers and industry. Planes are the most common geometric primitive in urban environments and man-made objects. Therefore, plane detection is an essential task in Simultaneous Localization And Mapping (SLAM)[1)–3)], reconstruction[4)], object detection[5)], keypoint detection[6)], among others.

Nonetheless, due to sensor artifacts, sensing patterns, noise, and other nuisances of point clouds, their detection is non-trivial. Particularly, large range sensors, such as those used in Terrestrial Laser Scanning (TLS), produce point clouds with high variations in density over dozens or even hundreds of meters. These nuisances complicate the decision of important plane detection parameters such as search radius or voxel size.

In the past, several approaches[7)–9)] tried to solve plane detection mostly for short range point clouds. However, they fail to detect most of the planes in TLS point clouds because they expect a uniform density[10)]. Another approach[11)] uses local robust statistics to detect planar patches, but expects accurate point-wise normal vectors for it to succeed. Therefore, to overcome conventional methods limitations, we propose a plane detection algorithm for TLS point clouds that tackles the sparsity problem by using multiple sliding voxel detectors over different scales, and maintains efficiency in large-scale TLS point clouds due to its coarse-to-fine approach.

Experiments results show that the proposed method is superior in accuracy and efficiency when compared to the state-of-the-art both in synthetic and real TLS point clouds. Furthermore, it is scalable to point clouds with dozens of millions of points without drastically affecting its performance.

The organization of this paper is as follows. In section 2 we brefly summarize the conventional methods. In section 3 we describe in detail the proposed method. Section 4 contains the experiments results and section 5 the conclusions.

2. Conventional Methods

2.1 Efficient RANSAC

Efficient RANSAC[7)](EFRANSAC) is a method that improves RANSAC[12)] efficiency in both hypothesis generation and model validation steps. To achieve higher performance, it uses an octree for localized sampling and scoring over disjoint random subsets.

It delimits the search space for hypothesis generation by selecting a random octree cell that contains a randomly drawn point. Then, it randomly selects the other 2 remaining samples from within that cell.

Once a hypothetical plane is drawn, it uses random disjoint subsets of the point cloud to extrapolate the model validation score (number of inliers) of one subset into the set of all points. After validating a model, inliers are refined by a connected component analysis in the projection of the shape to a plane.

However, as its authors stated, it can tolerate up to a 20% of outliers before returning false positives. Although it has optimizations in the hypothesis generation it is still expecting a degree of density and relatively high inliers ratio of the point cloud. Moreover, it does not compute point-wise normal vectors, which are frequently inaccurate and computationally expensive.

2.2 Randomized Hough transform

The Randomized Hough Transform(RHT)[8] for plane detection is a method that employs a conditioned random sampling to vote for hypothetical planes into an accumulator iteratively.

A random sample is valid only if the maximum distance between their points falls inside a user defined range. Once a plane gets enough votes, their inliers are removed from the point cloud and the accumulator is reseted. Then, it keeps iterating until there are not enough points left or if the failure rate of drawing random samples exceeds a user defined threshold.

As a result of removing planar points, in each iteration, the same-plane inliers ratio of the point cloud decreases, making increasingly difficult to draw good samples from the point cloud. Moreover, its random sampling mechanism expects some knowledge about the planes dimensions of the point cloud, and its finishing conditions have to be carefully defined. Otherwise, it will take increasingly longer times.

2.3 Robust statistics plane detector

The Robust Statistics Plane Detector (RSPD)[11] is a bottom-top approach for plane detection. It trasverses vertically an octree in the search for planar patches using planarity tests based on robust statistics. Each planar patch is grown over its coplanar neighbors and merged until stable to define a plane.

For an easier parameterization, they thresholded the planarity tests in terms of angles and defined them by robust statistics metrics such as Median Absolute Deviation(MAD).

This method requires normal vectors and a computationally intensive graph built from the neighborhood of each point estimated in advance. Moreover, as a bottom-top approach, it is more prone to fall into local noise, where noise levels could exceed the breakpoint of the median in the worst scenarios.

2.4 Sliding voxel

The Sliding Voxel (SV) algorithm[9] uses an octree to subdivide the point cloud into a set of voxel centroids and normal vectors V. A voxel and its 26 neighbors are called a sliding voxel. For each sliding voxel, it efficiently analyzes its neighborhood planarity and decides whether to grab a hypothetical plane or not.

It validates hypothetical planes P using a 2-step approach: (1) it counts inliers voxels of all hypothetical planes in V using point-to-plane distance and normal vectors deviation, and (2) it sorts P by their inliers count in descending order and extracts the inliers from V. P are pruned by their thickness

$$T = \left\{ \frac{\lambda_3}{\lambda_2} \in \mathbf{R} \mid \lambda_1 \geq \lambda_2 \geq \lambda_3 \right\} \tag{1}$$

estimated from PCA over the points distribution of each inlier voxel of V. Finally, the planes are refitted using the centroid and normal vector from PCA.

This method is simple, robust, and highly efficient in short-range point clouds. But in TLS point clouds, the high variations in density and noise make it difficult to define a suitable voxel size for every surface. At a larger voxel size, it can detect sparser planes but it can dismiss, merge or miscalculate smaller planes. On the other hand, at a smaller voxel size it is more affected by noise and tends to dismiss sparser planes. Therefore, the user has to decide the tradeoff balance by setting the most appropriate voxel size for each case.

3. Proposed Method

3.1 Overall routine

To solve previous methods issues in TLS point clouds, we propose to employ SV not as a detector but as a hypothetical planes estimator. The key idea is to detect bigger or sparser planes with a bigger voxel size and smaller or denser planes with a small voxel size in a coarse-to-fine way.

While the sliding voxel itself is a bottom-top approach, the proposed method is a top-bottom. Therefore, as opposed to RSPD, the proposed method is a combination of the best of both approaches.

The proposed method flow is depicted in **Fig. 1**. It starts by defining the voxel size of the octree levels defined by the user. Consequently, it employs a modified slid-

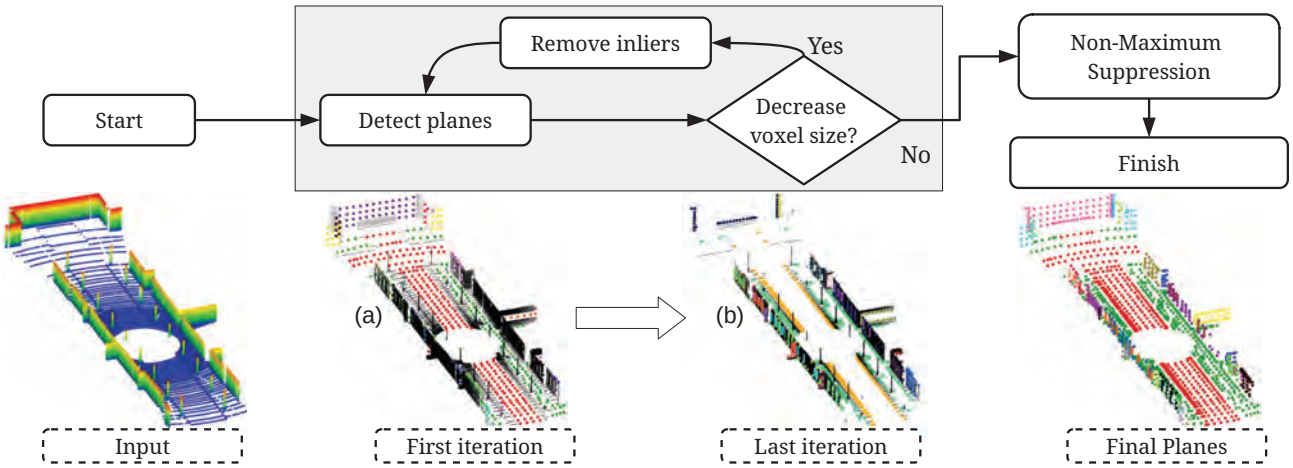


Fig. 1 Proposed method flow

ing voxel algorithm to collect hypothetical planes. This modification lies on filtering the points of V by thresholding their voxels thickness T. When T of Eq. (1) is obtained from a point distribution sampled from a surface, it represents both its noise level and curvature as T increases proportionally to these characteristics. Therefore, the proposed method can omit regions with high noise and outliers, especially when the voxel size is composed of more than 1 surface where a smaller voxel size is more suitable.

The parameters of the sliding voxel plane detector are fixed for all scales. The planarity threshold p_{th}, inliers threshold ϵ_{th}, and angular threshold θ_{th} are shared between the proposed method and SV.

Figure 1(a) and (b) shows the planes detected at the initial voxel size and final voxel size, respectively. In Figure 1(a), sliding voxel inliers are shown as big and colored dots, where each color represents a different plane, and the non-planar points are black. We can notice that only the biggest planes were detected, while the small planes were omitted by the detector. These non-planar points will be the input of a detector with a smaller voxel size.

Figure 1(b) shows a hollowed cloud due to the effect of removing planar points from previous scales. Here, we can observe that more detailed and smaller planes were detected. However, some sliding voxel inliers correspond to planes from previous scales. These inliers are near the border of other surfaces and need to be merged to avoid duplicated detections.

Therefore, after finishing the detection of hypothetical planes, similar planes are clustered by their angle $\hat{n} = \{a, b, c\}$ and offset d components of its cannonical form $ax + by + cz + d = 0$ using θ_{th} and ϵ_{th} thresholds. A plane with normal vector $\hat{n}_p$ belongs to a cluster when there is a plane in the cluster with normal vector $\hat{n}_c$ that satisfy

$$\frac{2}{\pi} \arccos(|\hat{n}_p \cdot \hat{n}_c|) \leq \theta_{th} \tag{2}$$

and

$$||d_p| - |d_c|| \leq \epsilon_{th}, \tag{3}$$

where $||$ is the absolute value.

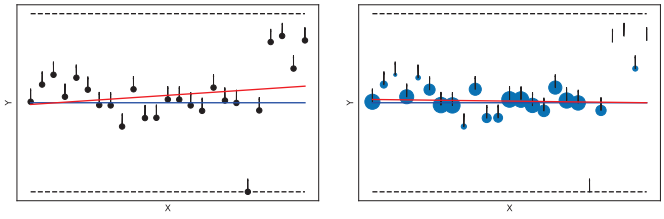
Then, a merging Non-Maximum Suppression (NMS) is applied iteratively until the resulting number of clusters is not reduced. The difference from a classic NMS is that, instead of only suppressing similar hypothetical planes, we merge those whose nearest points are up to 1 voxel size apart from the biggest plane in the cluster and create a new cluster otherwise. Here, the voxel size considered is the smallest used by the detectors. This merging mechanism allows us to maximize the true positives without throwing away planes that are likely to be part of a different structure. Furthermore, when merging clustered planes, we concatenate their sliding voxel inliers and re-estimate each cluster plane.

Finally, the centroids and normals point clouds used in each detector are combined for their use in pruning and validation of the resulting planes in the same way as SV.

3.2 Median Plane Weighted PCA

During the global validation of the sliding voxel algorithm and the merging NMS of the proposed method, a small percentage of errors can arise from merging sliding voxels with higher thresholds. Therefore, to extract only highly reliable planes coefficients, we apply a special case of median filter on the plane coefficients based on the sliding voxel inliers.

To achieve this, we estimate weights for every sliding voxel using a Radial Basis Function (RBF) over the



(a) PCA model (b) MPWPCA model

Fig. 2 Models estimated using PCA and MPWPCA

signed distance to the median plane, then we apply a weighted PCA to extract the coefficients, hence, we call it Median Plane Weighted PCA (MPWPCA).

Figure 2 depicts a 2D example of this technique applied in a similar condition. Fig. 2(a) shows a set of points and normal vectors, the blue line is the ground truth and the red line model was obtained using PCA over the points distribution. The upper and bottom dashed lines define a distance threshold between the model and the data points. Next, the median line model is estimated from the points positions and normal vectors orientations.

Fig. 2(b) shows the resulting model in red after applying a Weighted PCA with a strict threshold over the distance to their median line. To understand the contribution of each point to the final model, the blue points were resized according to the weights estimated from the distance to their median plane.

The proposed method performs MPWPCA as follows. For every plane p from the sliding voxel plane detector we have n pairs of centroids $C \in \mathbf{R}^{n\times 3}$ and their corresponding normal vectors $N \in \mathbf{R}^{n\times 3}$. We flip every sliding voxel normal if their deviation from p is more than 90[deg]. Then, let

$$\mathrm{med}(N) = \frac{\{\mathrm{med}(N_x), \mathrm{med}(N_y), \mathrm{med}(N_z)\}}{||\{\mathrm{med}(N_x), \mathrm{med}(N_y), \mathrm{med}(N_z)\}||_2} \quad (4)$$

be the normalized median of the x,y, and z components of the normal vectors N, and

$$\mathrm{med}(C) = \{\mathrm{med}(C_x), \mathrm{med}(C_y), \mathrm{med}(C_z)\} \quad (5)$$

be the centroids median values the x, y, and z components of C, hence the median plane is defined as

$$p_m = \{\mathrm{med}(N), -\mathrm{med}(N) \cdot \mathrm{med}(C)\} \quad (6)$$

where med stands for "median" and $||\cdot||_2$ is the L^2-norm.

We use p_m to estimate sliding-voxel-wise weights $w_i \in W$ based on the distance of every $\mathbf{c}_i$ to p_m as follows,

$$\forall \mathbf{c}_i \in C, w_i = \mathrm{e}^{-\frac{\delta(\mathbf{c}_i, p_m)^2}{\epsilon_{th}}} \quad (7)$$

Table 1 Point clouds information

Point Cloud	Points[#]	BBDD[m]	Planes[#]
S	129,621	200.9812	38
S noisy	129,621	201.2574	38
B	131,021	160.6900	74
B noisy	131,021	160.8506	74
N	33,789,031	347.0569	96

where ϵ_{th} is both the inliers threshold and clustering offset threshold of the proposed method and δ is the point-to-plane signed distance function. The inliers count I_c of the plane is also weighted using W

$$I_c = \sum_{i=1}^{n} w_i \Omega_i, \quad (8)$$

where Ω_i is the number of points inside a sliding voxel center. Planes that reduced their inliers by more than the median breakdown point (50%) are discarded.

Consequently, we compute a weighted mean

$$\mu = \frac{1}{\sum w_i} \sum w_i \mathbf{c}_i \quad (9)$$

and covariance matrix

$$\Sigma_\mu = \frac{1}{\sum w_i} (C-\mu)^T \mathrm{diag}(W)(C-\mu) \quad (10)$$

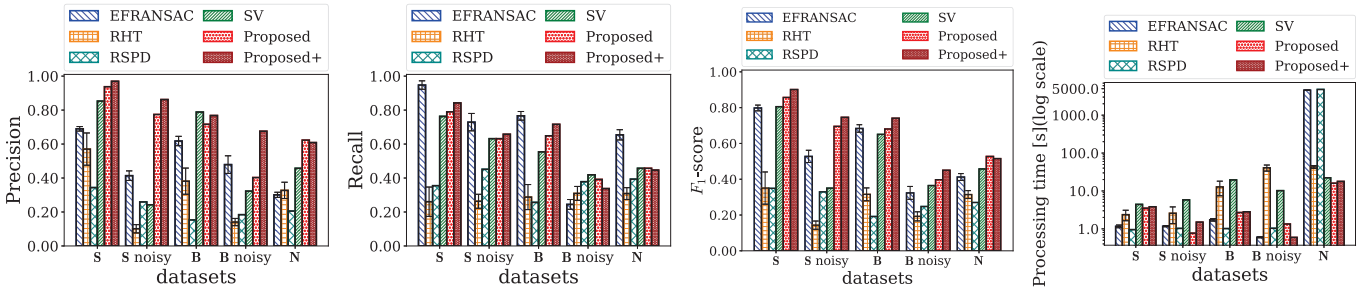
where $\mathrm{diag}(W)$ is a diagonal matrix containing the weights and $(C-\mu)$ represents the centroids cloud matrix with points in their its rows and centered to the median μ. Then, we estimate the robust plane from the smallest eigenvector of Σ_μ and $\mathrm{med}(C)$ as in Eq. (6).

4. Experiments and Results

Table 1 shows the basic information of the point clouds used for the experiments, where the range is expressed in Bounding Box Diagonal Distance (BBDD). The **S** and **S** noisy point clouds were obtained from a Velodyne HDL-64E simulation in Blensor[13]. **S** noisy was generated by setting Gaussian noise on with 0-mean and 0.01 variance.

The **B** and **B** noisy point clouds were obtained in the same setting of **S** and **S** noisy , the model used for simulations comes from the Bacelona Robot Lab Dataset[14]. **N** is the G_1 point cloud from a sphere detection study[15]. It was obtained by a high-precision FARO® LiDAR scanner and roughly 80% of the points are planar.

For comparison experiments, we carefully generated ground truth planes by manually selecting planar surface points and estimated plane coefficients with PCA using a tool developed specifically for this purpose. EFRANSAC and RHT are non-deterministic, therefore, we averaged the data of 10 executions for all the evaluation metrics.



(a) Precision, higher is better (b) Recall, higher is better (c) F_1-score, higher is better (d) Processing time[s](log scale), lower is better

Fig. 3 Numerical evaluation results

The source code of the proposed method and SV was implemented in C++ using the PCL 1.11.1 routines on Windows 10 compiled with MSVC 16.77 in release mode. We used the implementation of EFRANSAC from the CGAL library[16]. For RHT, and RSPD we used the implementations made public by their authors. The computer for experiments has an Intel® Xeon® E-2224G CPU at 3.5GHz with 64GB of RAM.

EFRANSAC and RSPD require point-wise normal vectors. Therefore, to avoid bias in the detection results, we used the same input normal vectors for both RSPD and EFRANSAC estimated with the PCL command line tool as follows. For the **N** point cloud we used a radius of 0.15[m], and for the synthetic datasets we used a neighborhood size (k) of 30.

We validated the performance of the proposed method in terms of processing time and the well-known precision, recall, and F_1-score from the field of information retrieval[17], against synthetic point clouds with realistic noise simulation. For this purpose, a detected plane was matched to a ground truth plane if their angle is less than 15[deg] and its offset absolute difference is less than 0.2.

Furthermore, we manually adjusted the parameters of the evaluated methods to maximize their F_1-score, and provided a visual assessment of the detection results by extracting the inliers of their resulting plane coefficients in a consistent way for all the evaluated methods.

Figure 3 shows the numerical results of the experiments. Each bar represents the result of executing a method over a dataset and all bars are grouped by their datasets. Among the evaluated methods, "Proposed" stands for the proposed method and "Proposed+" stands for the proposed method with its median filter (MPW-PCA). For the non-deterministic methods, we show the variations in their results using the standard deviation from their mean value.

Fig. 3(a) shows the precision (correctness) results. The proposed method is clearly more precise than the conventional methods, particularly, when median filter is turned on. Fig. 3(b) shows the recall (completeness) results. EFRANSAC has the highest recall in all the synthetic cases, except for the **B** noisy point cloud. On the other hand, the proposed method is always 2nd or 3rd place in all cases.

Fig. 3(c) shows the F_1-score results. Due to the superior precision and competitive recall, Proposed+ is clearly superior to the conventional methods.

In Fig. 3(d) we observe that the proposed method is one of the most efficient since it removes planar points adaptively with respect to the voxel size. Furthermore, it escalates well with the size of the point cloud as observed in the **N** results.

On the other hand, EFRANSAC and RSPD are of the most efficient methods for the synthetic datasets but their performance do not escalate well in massive point clouds. Although both require computationally expensive point-wise normal vectors, even if we subtract the estimation time from EFRANSAC and RSPD, Proposed+ is still 2.48 times faster than EFRANSAC and 15.02 times faster than RSPD in the **N** point cloud.

To assess the results visually we selected only the most accurate results. For each point cloud, we selected the proposed method option with the highest F_1-score (Proposed+) and the next 2 highest from the conventional methods which turned out to be EFRANSAC and SV.

To avoid bias in visual assessment, we segmented the planes by counting their inliers in a consistent way for each point cloud. The segmentation parameters are shown in **Tab. 2**, ϵ is the threshold for the absolute point-to-plane distance, and θ is the threshold for the maximum angle deviation between the inlier and the plane normal

(a) Ground truth (b) EFRANSAC (c) SV (d) Proposed+

Fig. 4 Visual results for **N**

Table 2 Segmentation parameters

Point Cloud	Neighborhood	ϵ[m]	θ[deg]
S	30[#]	0.050	13.5
S noisy	30[#]	0.070	30.0
B	30[#]	0.050	45.0
B noisy	30[#]	0.100	45.0
N	0.15[m]	0.025	45.0

vector. The neighborhood size in number of points or in radius size was estimated as the minimum possible to estimate normal vectors with correct orientations.

Each plane was assigned with a different color from the Glasbey LUT color scheme of the PCL library. They were sorted sorted by their number of inliers in descending order for their extraction. Plane inliers are shown in their respective color and outliers in black. Noticeably, not all plane inliers can be selected due to noise and high deviation of normal vectors from their estimation.

Figure 4 shows the visual results for the **N** point cloud. The biggest planes were correctly detected by most methods. The high recall of EFRANSAC can also be confirmed easily. However, its precision decreases when it detects false positives on slightly curved regions such as the ones shown in **Fig. 5**(a). On the other hand, both the proposed method and SV had a low false positives rate on these surfaces because they test for planarity before accepting a hypothetical plane.

Figure 6 shows the visual results for the **S** point cloud. The selected methods performed visually good and also had similar segmentation patterns when compared with their ground truth. On the other hand, in **Fig. 7**, the results **S** noisy point cloud indicates a lack of robustness of SV in broad planes and of EFRANSAC in less dense regions. This is due to insufficient voxel size of SV and the model validation of EFRANSAC which extrapolates the inliers count for model validation resulting in a lack of robustness in less dense regions.

(a) EFRANSAC (b) Proposed+

Fig. 5 False positives comparison in the **N** point cloud

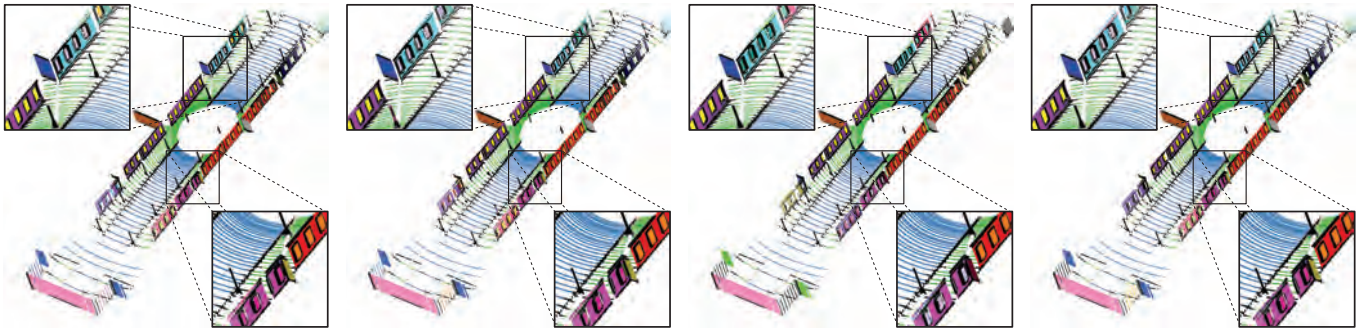
Figure 8 shows the visual results for the **B** point cloud. The selected methods performed well in the noiseless dataset with negligible errors. However, in **Fig. 9** we can observe a lack of robustness of the selected methods. While the proposed method detected more accurately bigger planes, it had difficulties in detecting some of the smaller planes due to an aggressive filtering. On the other hand, EFRANSAC and SV detected more smaller planes at the cost of showing several false positives that decreased drastically their precision.

5. Conclusions

In this paper, we addressed the problem of detecting planes in typical TLS point clouds obtained from the simulation of a cheaper and less precise rotating-head LiDAR and a high-precision LiDAR scanner with a range of up to roughly 300[m].

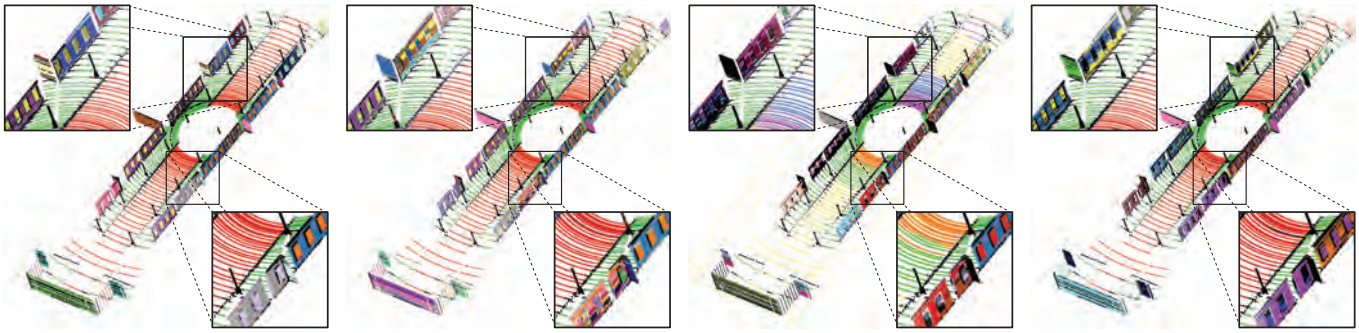
Conventional plane detection methods were designed to address the short-range scenarios and many of them suffer from difficulties of setting appropriate thresholds for TLS point clouds. Also, their efficiency do not escalate well with the number of points.

SV is the best performing on short-range point clouds, but the sparsity of TLS increase the amount of computations it needs. Our experiments indicate that



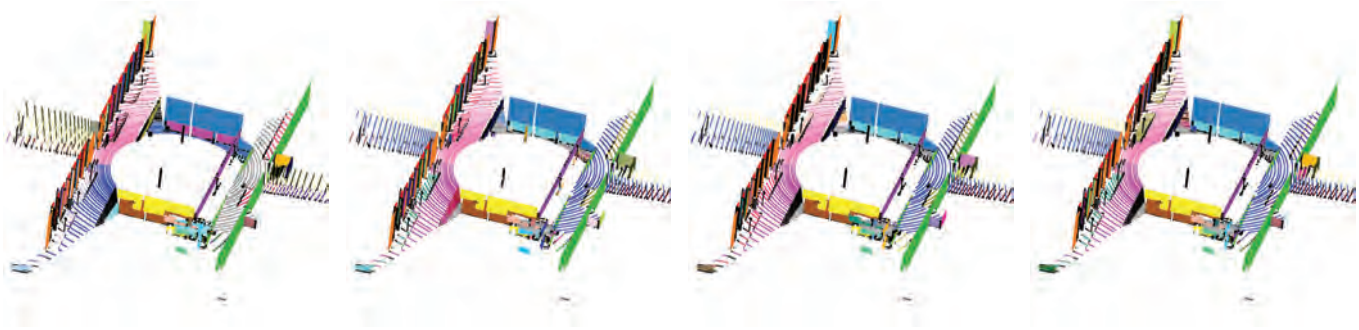
(a) Ground truth (b) EFRANSAC (c) SV (d) Proposed+

Fig. 6 Visual results for **S**



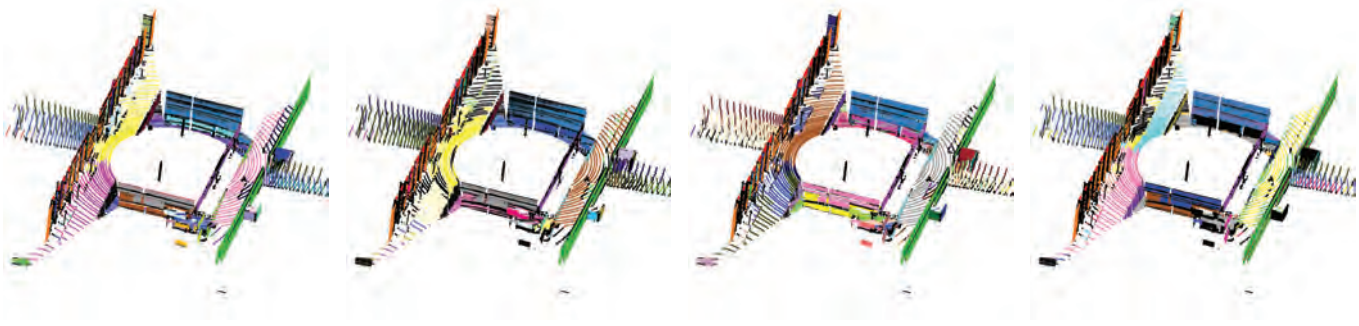
(a) Ground truth (b) EFRANSAC (c) SV (d) Proposed+

Fig. 7 Visual results for **S** noisy



(a) Ground truth (b) EFRANSAC (c) SV (d) Proposed+

Fig. 8 Visual results for **B**



(a) Ground truth (b) EFRANSAC (c) SV (d) Proposed+

Fig. 9 Visual results for **B** noisy

EFRANSAC can be drastically inefficient in most TLS point clouds. However, it is a good option for planar approximation of slightly curved surfaces rather than plane detection when performance is not important. We found that both RSPD and EFRANSAC efficiency can be competitive only when the point-wise normal vectors are given or easy to compute. The latter being the case of sparse TLS point clouds but its performance drops drastically when density is increased.

Therefore, to overcome these limitations, we designed a more robust approach for detecting planes in TLS point clouds by leveraging the best features of the sliding voxel plane detector and use it as a hypothetical plane estimator for different voxel sizes. The proposed method can also can handle point clouds with dozens of millions of points efficiently due to its coarse-to-fine approach.

According to the experiments results, the proposed method with the median filtering option is drastically more precise than the conventional methods and has the best equilibrium between precision and recall. We demonstrated that a hierarchical coarse-to-fine method for plane detection is highly efficient in TLS point clouds and provides state-of-the-art precision in the evaluated datasets.

For future work, we are considering improving the recall of the proposed method efficiently by conducting experiments on other planarity estimation metrics and a better quantization method of the 3D space. Moreover, we believe that the proposed method can be extended to simultaneously detect other parametric shapes such as spheres or cylinders.

References

1) K. Pathak, A. Birk, N. Vaskevicius, M. Pfingsthorn, S. Schwertfeger, J. Poppinga: "Online Three-Dimensional SLAM by Registration of Large Planar Surface Segments and Closed-Form Pose-Graph Relaxation", Journal of Field Robotics, Vol. 27, No. 1, pp. 52–84 (2010).
2) K. Pathak, A. Birk, N. Vaskevicius, M. Pfingsthorn, J. Poppinga: "Fast Registration Based on Noisy Planes with Unknown Correspondences for 3-D Mapping", IEEE Transactions on Robotics, Vol. 26, No. 3, pp. 424-441 (2010).
3) K. Lenac, A. Kitanov, R. Cupec, I. Petrović: "Fast Planar Surface 3D SLAM Using LIDAR", Robotics and Autonomous Systems, Vol. 92, pp. 197–220 (2017).
4) Y. Zhang, W. Xu, Y. Tong, K. Zhou: "Online Structure Analysis for Real-Time Indoor Scene Reconstruction", ACM Transactions on Graphics (TOG), Vol. 34, No. 5, pp. 159 (2015).
5) D. Lin, S. Fidler, R. Urtasun: "Holistic Scene Understanding for 3D Object Detection with RGBD Cameras", Proc. of the IEEE International Conference on Computer Vision, pp. 1417–1424 (2013).
6) K. Uenishi, J. Sandoval, I. Munetoshi, K. Tanaka: "VKOP: 3D Virtual Keypoint Detector Adapted to Geometric Structures and its Feature Descriptor", The journal of the Institute of Image Electronics Engineers of Japan: Visual Computing, Devices & Communications, Vol. 46, No. 2, pp. 283–297 (In Japanese)(2017).
7) R. Schnabel, R. Wahl, R. Klein: "Efficient RANSAC for Point-Cloud Shape Detection", Proc. of Computer Graphics Forum, Vol. 26, No. 2, pp. 214–226 (2007).
8) D. Borrmann, J. Elseberg, K. Lingemann, A. Nüchter: "The 3D Hough Transform for Plane Detection in Point Clouds: A Review and a New Accumulator Design", 3D Research, Vol. 2, Article 3 (2011).
9) J. Sandoval, K. Uenishi, M. Iwakiri, K. Tanaka: "Robust, Efficient and Deterministic Planes Detection in Unorganized Point Clouds Based on Sliding Voxels", IIEEJ Trans. on Image Electronics and Visual Computing, Vol. 7, No. 2, pp. 67–77 (2019).
10) S. Xia, D. Chen, R. Wang, J. Li, X. Zhang: "Geometric Primitives in LiDAR Point Clouds: A Review", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 13, pp. 685–707 (2020).
11) A.M.C. Araújo, M.M. Oliveira: "A Robust Statistics Approach for Plane Detection in Unorganized Point Clouds", Pattern Recognition, Vol. 100, Article 107115, pp. 1–12 (2020).
12) M. Fischler, R. Bolles: "Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Communications of the ACM, Vol. 24, No. 6, pp. 381–395 (1981).
13) M. Gschwandtner, R. Kwitt, A. Uhl, W. Pree: "BlenSor: Blender Sensor Simulation Toolbox", Proc. of the International Symposium on Visual Computing, pp. 199–208 (2011).
14) Barcelona Robot Lab Dataset, `http://www.iri.upc.edu/research/webprojects/pau/datasets/BRL/index.php` (2021).
15) J. Sandoval, K. Uenishi, M. Iwakiri, K. Tanaka: "Robust Sphere Detection in Unorganized 3D Point Clouds Using an Efficient Hough Voting Scheme Based on Sliding Voxels", IIEEJ Trans. on Image Electronics and Visual Computing, Vol. 8, No. 2, pp. 121–135 (2020).
16) CGAL, Computational Geometry Algorithms Library, https://www.cgal.org (2021).
17) C. Manning, P. Raghavan, H. Schütze: "Introduction to Information Retrieval", Cambridge University Press (2009).

(Received May 28, 2021)
(Revised September 24, 2021)

Jaime SANDOVAL (*Member*)
He received his B.E. degree in Computer Systems Engineering from the Universidad del Valle del Fuerte (Mexico) in 2009, and his M.E. degree in Electrical and Electronic Engineering from Shinshu University in 2017. In 2020 he received the Dr. Eng. degree from Shinshu University, and in the same year he joined AB.do Corporation. His research interests are 3D Point Clouds Processing, Computer Vision and Image Processing.

Kazuma UENISHI (*Member*)
He received his B.E. degree in Computer Science in 2008, and his M.E. degree in Mathematics and Computer Science in 2014 from the National Defense Academy of Japan. Currently, he is enrolled as a Ph.D. student in the Interdisciplinary Graduate School of Science and Technology of Shinshu University with a major in Systems Development Engineering. He is pursuing research in 3D Point Clouds Processing.

Munetoshi IWAKIRI (*Member*)
He received his B. E. degree in Computer Science in 1993, and received his M. E. degree in Mathematics and Computer Science from National Defense Academy of Japan in 1998. In 1999, he joined Department of Computer Science, National Defense Academy of Japan, as a Research Associate. In 2002, he received Dr. Eng. degree from Keio University, Tokyo, Japan. In 2005 he became Lecturer and in 2015 he became Associate Professor in the same institution. He is pursuing research related to Multimedia Processing and Information Security. He is a member of the Information Processing Society of Japan.

Kiyoshi TANAKA (*Fellow*)
He received his B.S and M.S. degrees in Electrical Engineering and Operations Research from National Defense Academy, Yokosuka, Japan, in 1984 and 1989, respectively. In 1992, he received Dr. Eng. degree from Keio University, Tokyo, Japan. In 1995, he joined the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University, Nagano, Japan, and currently he is a full professor in the academic assembly (Institute of Engineering) of Shinshu University. He is the Vice-President of Shinshu University as well as the director of Global Education Center (GEC) of Shinshu University. His research interests include image and video processing, 3D point cloud processing, information hiding, human visual perception, evolutionary computation, multi-objective optimization, smart grid, and their applications. He is a project leader of JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation entitled Global Research on the Framework of Evolutionary Solution Search to Accelerate Innovation started from 2013. He received IEVC2010 Best Paper Award from IIEEJ, iFAN2010 Best Paper Award from SICE, GECCO2011 Best Paper Award and GECCO2015 Best Paper Award from ACM-SIGEVO, ISPACS2011 Best Paper Award from IEEE, Excellent Journal Paper Award from IIEEJ two times, in 2012 and in 2014, and Best Journal Paper Award from JSEC in 2012. He is a fellow of IIEEJ. He is a member of IEEE, IEICE, IPSJ and JSEC. He is the former editor in chief of Journal of the Institute of Image Electronics Engineers Japan as well as IIEEJ Transactions on Image Electronics and Visual Computing.

IIEEJ Paper -- Special Issue on Journal Track Papers in IEVC2021Part II --

Aggregative Input Convolution for Large-Scale Point Cloud Semantic Segmentation

Kana KURATA†, Yasuhiro YAO†, Shingo ANDO (*Member*)†, Naoki ITO†, Jun SHIMAMURA†

†NTT Human Informatics Laboratories

<Summary> We propose an efficient semantic segmentation method for a large-scale point cloud. Previous point-based semantic segmentation methods to large-scale point clouds have been difficult. This is because those methods infer semantic labels to all the points used for feature extraction, and large-scale point clouds easily exceed their capacity. To solve this problem, we propose a novel point-based approach that predicts class labels for a downsampled point cloud and expands the labels to the whole point cloud by nearest-neighbor interpolation. The key idea of our approach is to give local features derived from the whole point cloud to each sample point by the newly developed Aggregative Input Convolution (AIC) and convert those features into wider context features by a point-based model for small-scale point clouds. AIC was experimentally confirmed to improve semantic segmentation accuracy on a large-scale dataset.

Keywords: point cloud, deep neural network, semantic segmentation

1. Introduction

Recent advances in sensor and resolution enhancement techniques[1] have enabled easier acquisition of high-resolution and spatially large-scale point clouds (10^4-10^7 points per scene). Spatial information for an object (e.g., location and size) can also be acquired by detecting the object from an obtained point cloud.

One successful technique for extracting objects from point clouds is semantic segmentation using a deep neural network (DNN). DNNs for point clouds have been actively researched since Pointnet[2], which predicts semantic labels without converting point clouds to other structures (e. g., images or voxels). Methods that can directly process point clouds are called point-based methods. These give a class label for each target point depending on features calculated from coordinates. To classify each point based on its local shape, methods based on the relative coordinates between each inference target point and its neighbor points have been developed [3, 4, 5, 6, 7]. However, these methods have high redundancy and high computational cost for searching neighbor points and calculating features using a DNN for all points in large-scale point clouds. Thus, the targets of most point-based semantic segmentation models are limited to small-scale (less than 10^4 points) point clouds.

To solve this problem, we propose a new point-based approach that predicts class labels for downsampled point clouds and expands the labels to whole point clouds by nearest-neighbor interpolation. This paper introduces a new point-based deep neural network (DNN) module named Aggregative Input Convolution (AIC), which applies local features derived from a whole point cloud to each inference target point. AIC was designed to preserve useful features in high-resolution point clouds through downsampling. The local features extracted by the AIC module are converted to wider context features by a conventional point-based DNN module for small-scale point clouds. Also, we evaluate the performance of nearest-neighbor interpolation under several sampling densities to predict semantic labels for an entire point cloud.

In summary, our contributions are as follows:

・We introduced AIC as the DNN module that passes useful features in high-resolution point clouds to small-scale convolutional DNNs.

・We proposed a semantic segmentation method for large-scale point clouds by combination of AIC, Parametric Continuous Convolution Neural Networks (PCCN) and color-aware nearest-neighbors.

・We experimentally evaluated the semantic segmentation method on S3DIS dataset.

2. Related Work

Point cloud semantic segmentation is a computer vision task that divides point clouds into subsets according to semantic meanings. The output of most point cloud segmentation methods is usually a semantic label for each point. Recent DNN-based methods can be categorized as either methods that rely on intermediate representations or point-based methods.

The former methods include projection and voxel-based architecture. Projection-based architectures[8, 9, 10] are derived from many works on 2D convolutional neural networks (CNNs). Additionally, in voxel based methods, methods to make feature extraction process more efficient have been studied[11, 12]. OctNet[11] reduced computational cost by efficiently handling empty areas

using hybrid grid-octree structure[13]. These methods with data transformation are based on previous research on graph neural networks (GNNs) and 2D CNNs. However, due to the transformation of the data structure, they cannot directly handle the detailed shapes represented by continuous coordinates, or translation and rotation in 3D space, etc.

The latter, point-based DNNs, have been developed to directly handle continuous coordinate point clouds without conversion to regular grid structures (e.g. 2D image and 3D voxels). Recent point-based methods tend to score highly in semantic segmentation tasks. Because of the difference of data structures, convolution operations used in standard convolutional neural networks (CNNs) cannot be applied to point clouds. Therefore, various types of feature extraction operations for non-grid data have been developed.

One way to process continuous coordinates is to use graph neural networks (GNNs). GNNs consume graph structured data that consists of inference target nodes and edges defining relationships between the nodes[14, 15, 16]. For example, SPG[15] applies gated recurrent units (GRU) [17] to graphs that consist of clusters of primitive shapes and edges between clusters.

PointNet[2], the pioneer point-based method, directly encodes coordinates of each point by pointwise multi-layer perceptron (MLP). In PointNet++[3], DNN architecture was improved to capture local shapes by encoding relative coordinates of neighbor points around each reference point using pointwise MLP on relative coordinates to reference points. As well as PointNet++, KCNet[4], PCNN[5], PCCN[6] and A-CNN[7] also extract features based on relative coordinates between target points and its neighbor points. In particular, PCNN[5] and PCCN[6] build CNN-like architectures for non-grid data by developing ways to give convolutional weights. PCNN[5] learns convolution weights according to the order of distance, and then PCCN[6] calculates the weights from relative coordinates. On the other hand, A-CNN[7] proposed annular convolution operators to better capture local neighborhood geometry in point clouds.

In conventional point-based methods such as those described above, it is common to give predictions to all inference target points based on a directed graph between every target point and their neighbor points. Since searching for neighbor points and learning features for a large number of points is computationally expensive, downsampling is often used for preprocessing. However, high-resolution shape features are lost by coarse downsampling.

3. Method

3.1 Overview

Our model is designed to improve efficiency by reducing the number of inference points without losing useful features in high-resolution point clouds. Our method is composed of semantic label prediction by DNN that includes an Aggregative Input Convolution (AIC) network and PCCN[6], followed by color-aware nearest-neighbor interpolation. We introduce AIC, which is a DNN module, to aggregate information for detail point distribution as features of a few inference target points. PCCN is a convolutional neural network that requires higher computation cost than discrete convolution because it calculates convolution weights from neighbor point coordinates. Generally, the computation cost of convolution operation in PCCN linearly decreases as the number of input points decreases. In this study, the input points of PCCN are restricted to a small scale using AIC and downsampling to reduce computation cost while preserving high-resolution features, and reducing classification errors around object boundaries by color-aware nearest-neighbor interpolation. **Figure 1** shows the entire DNN architecture.

3.2 Aggregative Input Convolution (AIC) network

AIC is designed to utilize the rich shape information of a high-resolution point cloud with minimum computational cost by reducing the computationally expensive process of determining neighbor points for all points of large-scale point cloud.

AIC is a DNN module that aggregates information for detailed point distribution as features of a few inference target points. It is used in combination with a semantic segmentation DNN. AIC input is the relative coordinates and colors of neighbor points of a few inference target points.

AIC searches the neighbor points for inference target points from high-resolution point clouds before downsampling. In this paper, inference points are chosen by random sampling. In Fig. 1, “sparse indexing” corresponds to this operation, and includes neighbor points search and indexing of coordinates and features of neighbor points for inference points. We incorporated dilation into sparse indexing, inspired by the dilated convolution method[18]. The dilation factor is a coefficient that determines the distance between points. For example, if the dilation factor is 3, neighboring points are selected from the closest point every three points. “Concatenate” is a function that combines the elements of multiple vectors in the feature channel axis. The number of neighbor points per target point is set to be larger than a segmentation network, to cover the wide sphere of the entire point cloud. **Figure 2** visualizes the AIC input. The blue points in Fig. 2 are inference target points which are given features by AIC, while the other points are their supporting neighbor points. Additionally, to adapt the method to various data, the number of AIC channels for shape and color features are designed to be adjusted according to the characteristics of the target data. AIC extracts shape and color features separately and combines them into AIC output features as shown in Fig. 1.

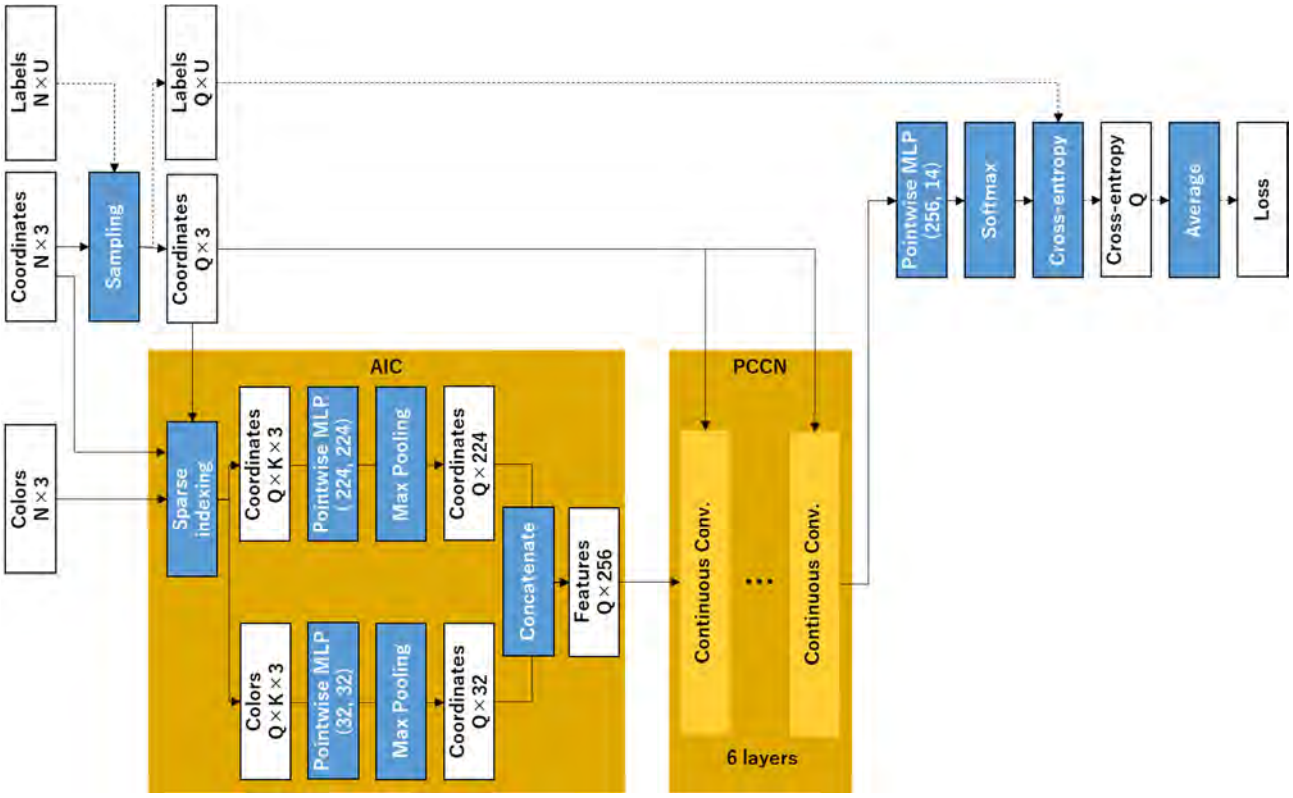


Fig. 1 Proposed DNN architecture consists of AIC and semantic segmentation networks. N, Q, K mean the total number of input points, inference points and neighbor points. Numbers in bracket are layer sizes. Labels' size "U" is the number of recognition target classes. White boxes mean input or intermediate data, blues are processing, and dotted lines show data flow only in the training phase.

The coordinates of a target point and its neighbor points, and the RGB vector of each neighbor point are transformed by pointwise MLPs and are aggregated around each inference point by a pooling layer. This operation encodes local distribution of points by MLP as well as PointNet++[3] but differs in that it is applied to a high-resolution point cloud before downsampling, and only representative points with the derived features are given.

In our implementation, layer sizes were set as shown in Fig. 1, and pooling layers were max pooling. Since indoor objects come in a variety of colors, the number of channels for shape and color features was manually set to 224 and 32 respectively to cover shape variations preferentially. Such separation of shape and color enables adjustment of the number of channels for shape and color features. The appropriate number of neighbor points and the dilation factor are experimentally evaluated in Section 4.4.

3.3 DNN architecture

AIC and PCCN are connected as shown in Fig. 1 and trained together. PCCN, composed of three continuous convolution blocks with residual connections, receives AIC output feature and outputs a class label prediction. Our method consumes lower computational cost than methods that process all points equally since only a small subset (less than 10^4 points per scene) of the entire points are targeted in the nearest neighbor search and AIC feature extraction processes. This network is trained with cross-entropy loss between PCCN output and true label shown as “Labels” in Fig. 1 of inference target points using an Adam optimizer.

Fig. 2 AIC input at the number of inference target points 2048 and neighbor points 128

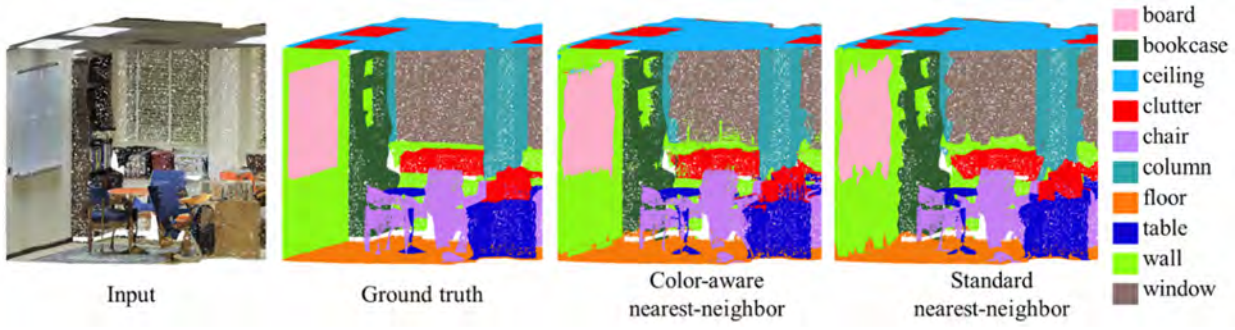


Fig. 3 Results for nearest-neighbor interpolation and color-aware nearest-neighbor interpolation under the 1/256 sampling rate.

3.4 Color-aware nearest-neighbor interpolation

In our method, semantic predictions by DNN of downsampled point clouds are expanded by color-aware nearest-neighbor interpolation. Nearest-neighbor interpolation requires low computational cost to expand prediction from a few points and thus is often used in point cloud processing. However, interpolation errors around object boundaries become larger as the number of points having labels decreases. To avoid such errors, we adopt color-aware nearest-neighbor interpolation. This method determines the closest point based on the Euclidean distance of concatenated vectors of coordinates and normalized RGB as follows.

$$R' = (Red - \min(Red))/\max(Red) \quad (1)$$

$$G' = (Green - \min(Green))/\max(Green) \quad (2)$$

$$B' = (Blue - \min(Blue))/\max(Blue) \quad (3)$$

$$Distance = \sqrt{x^2 + y^2 + z^2 + R'^2 + G'^2 + B'^2} \quad (4)$$

where *Red*, *Green*, and *Blue* are each channel of RGB, and *x*, *y*, *z* are values of three-dimensional coordinates. The functions “max” and “min” take the maximum and the minimum values from all input points.

4. Experiments

In this Chapter, we report the results of three experiments: accuracy measurement of color-aware nearest-neighbor interpolation, verification of accuracy improvement by AIC, and parameter study of AIC. All experiments are conducted on an NVIDIA RTX Titan GPU using an S3DIS dataset. In experiments described in Sections 4.3 and 4.4, DNN models were trained for 300 epochs and the batch size for training was 4.

4.1 Dataset

S3DIS is a large-scale dataset of indoor space that contains 3D scans of 271 rooms in 6 areas that contain multiple types of rooms (hallway, conference room, water closet, storage, and office). Each point has an RGB attribute and a semantic label from 13 categories of indoor objects. This paper identified these 13 classes and an additional category, "Other," to avoid exceptions when there are points with no labels. We used all scans of S3DIS for evaluation in Section 4.2 and assigned scans excluded Area 5 for training and scans of Area-5 for the test in Sections 4.3 and 4.4, following Wang et al.[6]. As S3DIS contains various sizes of data (from 85 thousand to 7 million points), we split scans of large rooms while limiting the max number of points to 524,288.

Table 1 Accuracy of color-aware and standard nearest-neighbor interpolation.

Sampling rate	OA (standard)	OA (color-aware)
1/512	91.89%	**93.02%**
1/256	94.39%	**95.05%**
1/128	96.14%	**96.43%**

4.2 Interpolation accuracy using color-aware nearest-neighbor

First, we tested color-aware nearest-neighbor interpolation of multiple sampling density in all scans of S3DIS dataset. This interpolation is non-learning and it is independent from the semantic segmentation by the DNN. In this experiment, the ground truth labels are interpolated from randomly sampled points to the whole points by several methods including ours. And we evaluated the interpolation methods by the label accuracies. **Table 1** shows the overall accuracy of standard and color-aware nearest-neighbor interpolation. Sampling rate means a reduction ratio parameter of AIC. Overall accuracy (OA) increased using the color-aware method instead of the standard method. Also, **Fig. 3** shows that interpolation errors around object boundaries were reduced overall by color-aware interpolation. In detail, interpolation accuracy around the boundary between a wall (yellow-green) and a bookcase (dark green) is significantly improved by the color-aware method. However, interpolation errors still remain around some boundaries,

for example, between a wall and a window (gray). Thus, it is necessary to adjust the strength of color awareness by adjusting the RGB value normalization range.

4.3 Semantic segmentation accuracy

We evaluated the effect of AIC by comparing the performance of our architecture with two baselines. The first is PCCN with downsampling in preprocessing phase (pre-sampling), using a DNN with the same structure as the semantic segmentation network of our model ("PCCN w/ pre-sampling" in **Table 2**). The second is our model with pre-sampling ("Ours w/ pre-sampling" in Table 2), the same structure as our model but with downsampled points input to AIC. This paper adopted AIC sampling rate of 1/256 to balance efficiency and accuracy according to the results of Section 4.2. Also, the pre-sampling rate was 1/256 to set the number of PCCN input points in each model to the same value. We employed this experimental setting to evaluate the effectiveness of extracting features from high-density points by AIC. In the case of the two baselines, the number of all the input points in Fig. 1. "N" equals "Q." Note that this evaluation is independent from the interpolation evaluation in Section 4.2.

Table 2 is the result for the downsampled points before interpolation. In evaluation with the downsampled DNN inference target points, our model without sampling before AIC ("Ours w/o pre-sampling" in Table 2) achieved a higher score in overall accuracy (OA), mean class IoU (mIoU) and mean class accuracy (mAcc) than PCCN with downsampled points and Ours w/ pre-sampling (Table 2).

The only difference between our model w/o pre-sampling and w/ pre-sampling is whether the neighbor points in AIC are selected from a high-resolution point cloud or from a pre-sampled point cloud. Therefore, the method of assigning features from the high-density point cloud is confirmed to improve the accuracy.

In the case of predicting labels to an entire point cloud, our method with color-aware nearest-neighbor interpolation outperformed the result of original PCCN in Wang et al.[6] (**Table 3**). Table 3 listed the score of the original PCCN paper that shows mAcc and mIoU only. Unlike the experiment in Section 4.2, this experiment operated interpolation based on inferred labels. In this experiment, we followed the training and testing procedure used in Wang et al.[6]. However, we used a different preprocessing method of data cutting. While Wang et al.[6] split data into cubes of a specific size in preprocessing, we split data by picking a fixed number of points in order of closeness from a randomly determined point to avoid extra sampling in scenes with high-density points. In contrast to original PCCN, which has neighbor points for all input points, our model has neighbor points for only a small number of inference target points. Therefore, our method reduces the maximum array size of internal data stored in RAM to 1/256 (AIC downsampling rate) of that of directly applied original PCCN while improving accuracy.

Table 2 Quantitative results of our model and baselines for downsampled points

Method	OA	mAcc	mIoU
PCCN w/ pre-sampling	75.31%	61.41%	52.59%
Ours w/ pre-sampling	76.18%	69.81%	60.00%
Ours w/o pre-sampling	**81.62%**	**71.14%**	**62.16%**

Table 3 Accuracy for all points

	OA	mAcc	mIoU
PCCN[6]	-	67.01%	58.27%
Ours w/o pre-sampling (interpolated)	79.75%	**68.79%**	**59.09%**

Table 4 Evaluation results of AIC parameters and accuracy

The number of neighbor points and dilation	OA	mAcc	mIoU
64, 2	80.84%	71.32%	61.93%
64, 3	78.18%	67.04%	57.38%
128, 1	80.29%	70.09%	60.48%
128, 2	74.78%	67.12%	57.59%
128, 3	**81.62%**	**71.14%**	**62.16%**

4.4 Parameter study of AIC

Next, we evaluated the approximate number of AIC neighbor points and dilation (**Table 4**). For a sampling rate of 1/256, the wall hanging paintings on right side of **Fig. 4 (a)** were identified by our method, although they were not identified on the left in Fig. 4 (a) or the right in Fig. 4 (b). Particularly, some areas around the painting on the right in Fig. 4 (a) were classified as a whiteboard. These results indicate that our architecture was capable of capturing texture features, but texture variations were not sufficiently learned. Hence, it is necessary to study methods for learning enough features of objects with many texture variations and objects with small amounts of data. Similar to the results described in Section 4.2, some boundaries could not be separated correctly by color-aware nearest neighbor due to similarity in color between objects.

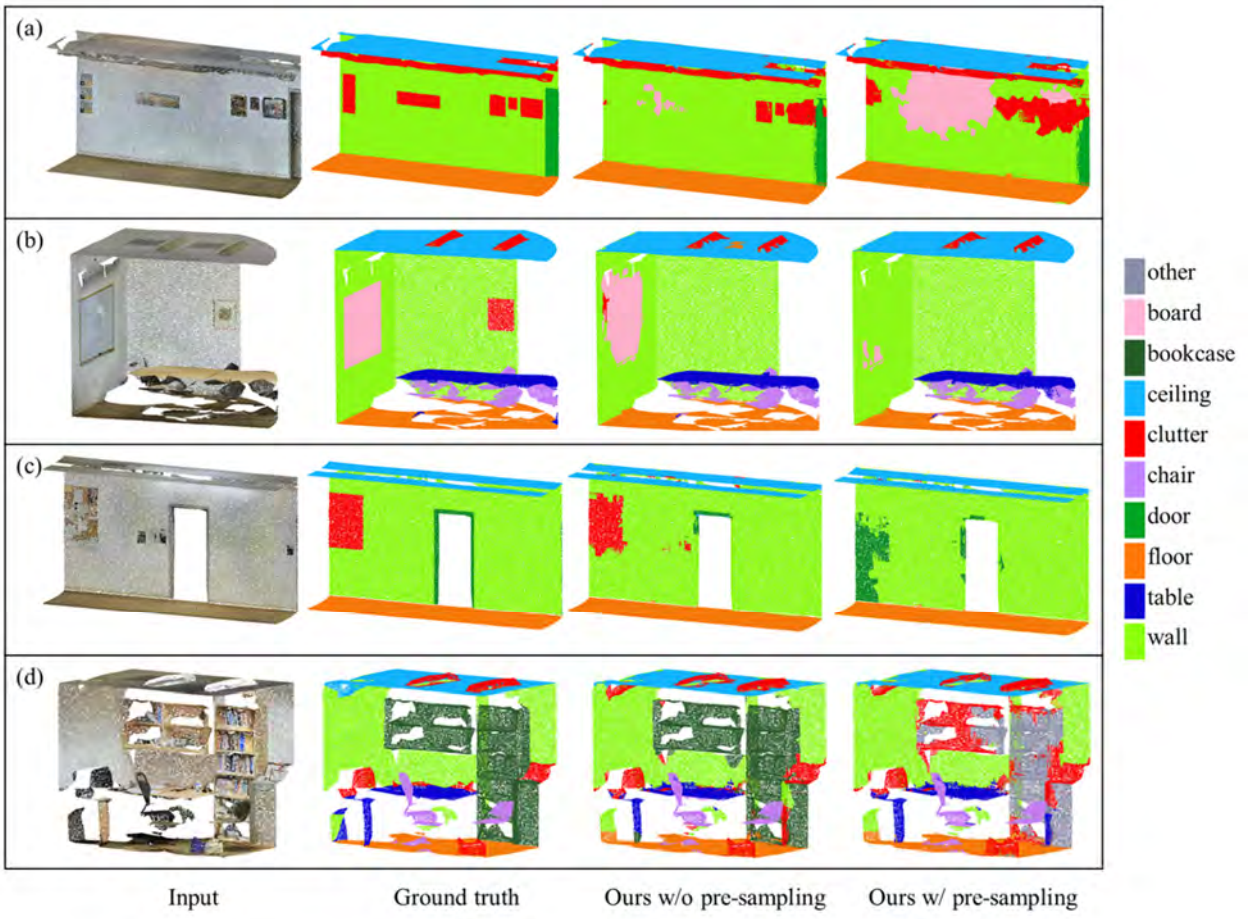


Fig. 4 Interpolated semantic segmentation result at sampling rate 1/256, 128 AIC neighbor points and AIC dilation 3

5. Conclusion

We presented DNN architecture with a novel AIC module that aggregates information of detailed point distribution as features of a few inference target points. Through the experiments on S3DIS dataset, we find out that high resolution features inputs captured by AIC improve semantic segmentation accuracy. In terms of the overall method, our method reduces the maximum array size of internal data to 1/256 (AIC downsampling rate) of that of directly applied PCCN[6] while improving the accuracy. Experimentations on other indoor/outdoor large-scale datasets and on the combination of AIC with other DNNs instead of PCCN are future issues to quantitatively evaluate the performance and efficiency of our model. Also, a future research topic is to develop methods for learning enough features of objects with many texture variations and objects with small amounts of data.

References

1) Y. Yao, M. Roxas, R. Ishikawa, S. Ando, J. Shimamura, T. Oishi: “Discontinuous and Smooth Depth Completion with Binary Anisotropic Diffusion Tensor”, IEEE Robotics and Automation Letters, Vol. 5, No. 4, pp. 5128–5135 (2020).

2) C. R. Qi, H. Su, K. Mo, L. J. Guibas: “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 652–660 (2017).

3) C. R. Qi, L. Yi, H. Su, L. J. Guibas: “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”, Proc of Advances in Neural Information Processing Systems 30 (NIPS 2017) (2017).

4) Yiru Shen, Chen Feng, Yaoqing Yang, Dong Tian: “Mining Point Cloud Local Structures by Kernel Correlation and Graph Pooling”, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), pp. 4548–4557 (2018).

5) Y. Li, R. Bu, M. Sun, W. Wu, X. Di, B. Chen: “PointCNN: Convolution on X-transformed Points”, Proc. of Advances in Neural Information Processing Systems 31 (NeurIPS 2018) (2018).

6) S. Wang, S. Suo, W.-C. Ma, A. Pokrovsky, R. Urtasun: “Deep Parametric Continuous Convolutional Neural Networks”, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), pp. 2589–2597 (2018).

7) A. Komarichev, Z. Zhong, J. Hua: “A-CNN: Annularly Convolutional Neural Networks on Point Clouds”, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), pp. 7421–

7430 (2019).

8) F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, M. Felsberg: “Deep Projective 3D Semantic Segmentation”, Proc. of 17th international Conference on Computer Analysis of Images and Patterns (CAIP) (2017).
9) A. Boulch, B. Le Saux, N. Audebert: “Unstructured Point Cloud Semantic Labeling Using Deep Segmentation Networks”, Eurographics Workshop on 3D Object Retrieval (3DOR 2017) (2017).
10) M. Tatarchenko, J. Park, V. Koltun, Q.-Y. Zhou: “Tangent Convolutions for Dense Prediction in 3D”, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), pp. 3887–3896 (2018).
11) L. Tchapmi, C. Choy, I. Armeni, J. Gwak, S. Savarese: “SEGCloud: Semantic Segmentation of 3D Point Clouds”, Proc. of 2017 International Conference on 3D Vision (3DV), pp. 537–547 (2017).
12) G. Riegler, A. O. Ulusoy, A. Geiger: "OctNet: Learning Deep 3D Representations at High Resolutions", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 3577–3586 (2017).
13) A. Miller, V. Jain, J. L. Mundy: “Real-time Rendering and Dynamic Updating of 3-d Volumetric Data”, Proc. of the Workshop on General Purpose Processing on Graphics Processing Units (2011).
14) Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, J. M. Solomon: “Dynamic Graph CNN for Learning on Point Clouds”, ACM Trans. Graph Vol. 38, No. 5 (2019).
15) L. Landrieu, M. Simonovsky: “Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs”, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), pp. 4558–4567 (2018).
16) L. Landrieu, M. Boussaha: “Point Cloud Oversegmentation with Graph-Structured Deep Metric Learning”, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), pp. 7440–7449 (2019).
17) K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio: “Learning Phrase Representations Using RNN Encoder–decoder for Statistical Machine Translation”, Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1724–1734 (2014).
18) F. Yu and V. Koltun: “Multi-Scale Context Aggregation by Dilated Convolutions”, Proc. of International Conference on Learning Representations (ICLR) (2016).

Appendix

APP1. DNN Structure of PCCN

PCCN[6] is a convolutional neural network with residual connection. Convolution weight is calculated as a continuous value from the relative coordinates of neighbor points.

In our method, PCCN is used to give predictions to small-scale point clouds that are downsampled in AIC rather than to the entire point cloud.

The original PCCN model used Wang et al.[6] includes eight convolution layers with residual connections and has a global pooling layer connected after the last convolutional block (**Fig. APP1**). Wang et al[6] adopts separable convolution.

DNN structure of PCCN used in this paper is composed of six continuous convolution layers with dropout layer. The global pooling layer was excluded from our model because of the effects of accuracy improvement were not to be experimentally confirmed.

We set the number of neighbor points to 32 following the original paper, feature channels to 256 based on the memory cap of VRAM, and dropout rate to 0.3 experimentally.

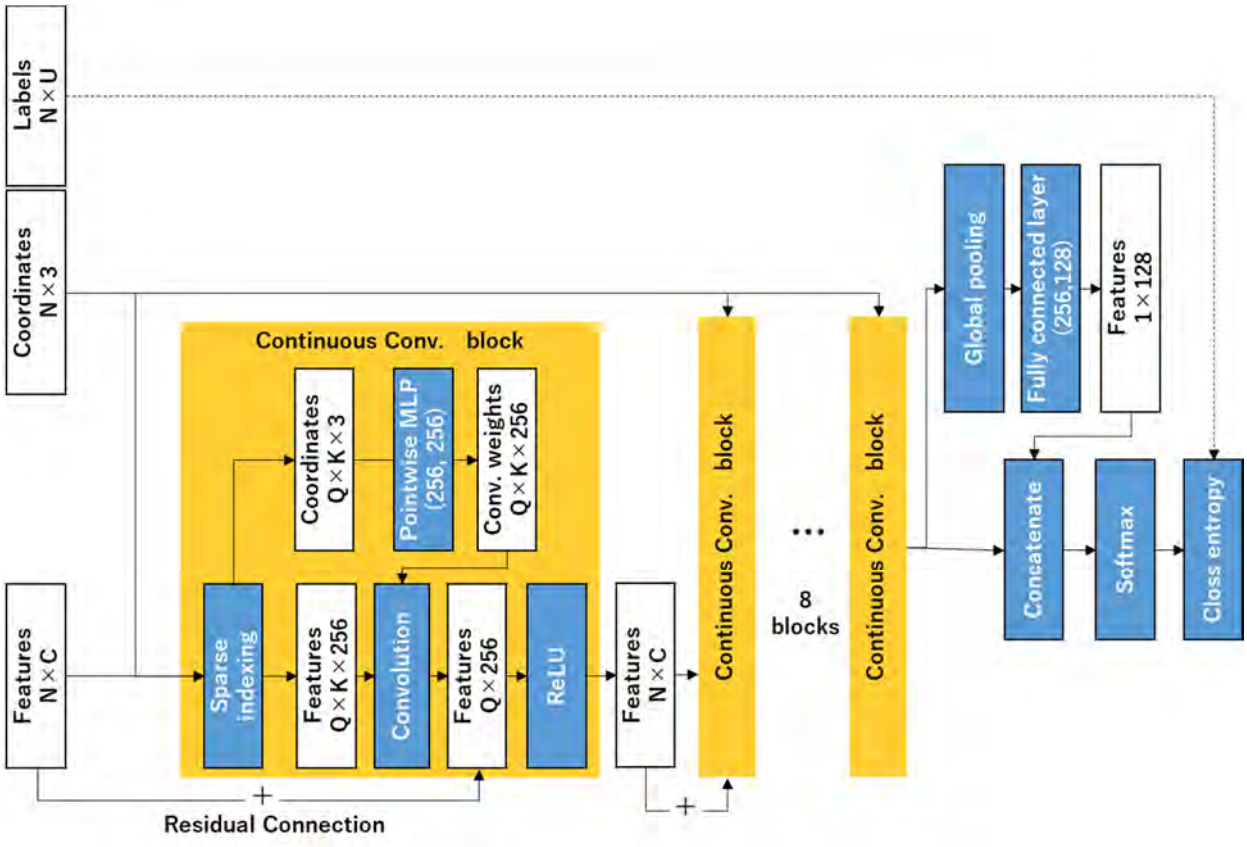


Fig. APP1 DNN architecture consists of PCCN.

(Received May 28, 2021)
(Revised Mar. 3, 2022)

Kana KURATA

She received a B.S. in earth and planetary sciences from Nagoya University in 2016 and an M.E. in environmental studies from Nagoya University in 2018. She joined NTT in 2018, where she has been engaged in research in computer vision and pattern recognition.

Yasuhiro YAO

He received the B.S. and M.E. degrees from The University of Tokyo, Japan, in 2007 and 2010, respectively, where he is currently pursuing the Ph.D. degree in information studies. In 2010, he joined NTT as a Researcher. From 2013 to 2016, he was a Cloud Solution Architect at Dimension Data APAC, Singapore. He is currently a Senior Research Engineer at NTT Human Informatics Laboratories. His research interests include computer vision and sensor fusion.

Shingo ANDO (*Member*)

He received the B.E. degree in electrical engineering from Keio University in 1998. He received the Ph.D. degree in engineering from Keio University in 2003. In 2003, he joined NTT. He has been engaged in research and practical application development in the fields of image processing, pattern recognition, and digital watermarks. He is a member of IEICE, ITE, and IIEEJ. He is currently an associate professor at Shonan Institute of Technology.

Naoki ITO

He received the M.E. from Toyohashi University of Technology, Aichi, in 2001. He joined NTT in 2001 and engaged in research on character recognition. He moved to NTT EAST in 2004 and engaged in the development of security systems. He moved to NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2008. He has been working on the development of real-world digitalization technologies.

Jun SHIMAMURA

He received a B.E. in engineering science from Osaka University in 1998 and an M.E. and Ph.D. from Nara Institute of Science and Technology in 2000 and 2006. He joined NTT Cyber Space Laboratories in 2000. He is currently senior research engineer, supervisor of scene analysis technology at NTT Human Informatics Laboratories, Japan. His research interests include computer vision and mixed reality. He is a member of IEICE.

IIEEJ Paper -- Special Issue on Journal Track Papers in IEVC2021Part II --

Registration of Histopathological Heterogeneous Stained Images Utilizing GAN Based Domain Adaptation Technique

Tanwi BISWAS†, Hiroyuki SUZUKI††, Masahiro ISHIKAWA††† (*Member*), Naoki KOBAYASHI††† (*Fellow*), Takashi OBI†

†Tokyo Institute of Technology, ††Gunma University, †††Saitama Medical University

<Summary> Registration of histopathological images obtained from different staining techniques is very challenging because of much difference of their color information. In this study, we propose a promising image registration method that can overcome the color difference of H&E and EVG stained images by means of GAN-based color conversion. Our proposed method consists of two main parts: one is GAN based unsupervised domain adaptation network for converting H&E stained image to EVG stained image which has similar distribution with the original EVG stained image and the other is SURF feature based registration framework which provides the registered EVG stained image leveraging the generated EVG stained image obtained from the domain adaptation network. The experimental result shows that our proposed method is able to provide better registration result than the conventional method where domain adaptation technique is not incorporated.

Keywords: image registration, histopathological image processing, unsupervised domain adaptation, generative adversarial network (GAN)

1. Introduction

Histopathological staining is an indispensable preprocessing step of preparing tissue slides for observing under the microscope where coloring reagent is used to visualize important features of tissue samples. Different types of histological stains are used to particularly highlight different tissue components. For the purpose of disease analysis and diagnosis, it is required to compare these tissue slides with different stains visually. Though these differently stained histopathological images have identical morphological information, they are neither aligned nor scale adjusted because of being captured by different device in different time. Registration makes these similar anatomically structured images aligned together which corresponds to each other and facilitates easy visual comparison of these tissue slides[1]. Also, for digital stain conversion or segmentation of particular tissue components with supervised learning, registration is considered to be a very important preprocessing step that makes the image data ready for further processing[2]. But, registration of these different types of stained images is very challenging because of having possible deformations during the preparation of tissue, repetitive pattern in texture, large image size and most importantly difference in appearances with diverse color information. This study aims to provide an efficient method for the automatic registration of these heterogeneous stained images.

Previous studies have introduced different non-learning and learning based approaches for addressing the problem of registration of differently stained images[1),3-11)]. The non-learning based approaches can be further categorized as intensity-based[3)], feature-based[1)] and both feature and intensity-based[4)] registration methods. Comparing the performances of different non-learning based methods, the hybrid method combining both intensity and features has been observed to provide better result at the cost of large processing time[5)]. Also, there is still significant gap between the obtained result and the result expected by the human annotator. Recently M. Wodzinski and A. Skalski have proposed an effective method for the registration of histology samples obtained from multiple stain types[6)]. Their proposed methodology follows multiple steps including preprocessing of input images, feature-based initial alignment, exhaustive rotation search, refinement of initial affine transformation and final calculation of deformation field. Though their proposed method provides better registration result, the method is too complicated and time consuming to be applied in practical applications. Also, according to the authors, the key component of their successful registration lies in the accuracy of initial alignment where our proposed method can be a great substitute. Nowadays, different learning-based approaches leveraging machine learning or deep learning, have also proved their effectiveness to be applied for the purpose of medical image registration[7-11)]. But, most of these learning

based methods are highly dependent on having annotated segmentation data[7-8] or prior spatially aligned corresponding images[10-11] as training data which is time-consuming, expensive and labor-intensive to collect. In our approach, we have also leveraged deep learning model for the adaptation of the domain of differently stained images in an unsupervised manner to aid the registration.

In the field of machine learning or deep learning, most of the analyses are performed using the dataset where training and testing data are from same distribution having similar features and are able to provide outstanding result. However, in real life applications the testing data may vary abruptly from the training data distribution because of the lighting condition, environmental issue etc. and in these cases the trained deep learning models are not generalized well and seem to provide poor performance[12-13]. Unsupervised domain adaptation technique can be a very useful solution in this regard which tackles the domain shift problem between the training (source domain) and testing (target domain) data without requiring labeled data from the target domain and is able to bring significant improvement to the performance of deep learning models in real life applications. Numerous researches have been performed to develop different methods of domain adaptation which have proved to provide improved results in various computer vision cases such as classification[14], object detection[15] and segmentation[16]. Also, in medical image analysis, either supervised or unsupervised domain adaptation technique seems to be very useful for different objectives such as cancer classification[17], tumor segmentation[18], skin disease classification[19], retina segmentation[20] and so on. However, fewer works have been observed where domain adaptation has been leveraged for the purpose of medical image registration[21]. Specially, for the registration of multiple stain types, to the best of our knowledge, no prior work has been found where domain adaptation has been utilized. Whereas we think of domain adaptation as a promising technique of reducing the gap in the appearances of the differently stained images thus aiding to find similar correspondences more easily. At the same time, we also believe that, adapting the domain of the histological images is a very challenging task due to the significant gap between the color information of two different domains.

In this study, we have proposed an automatic image registration method where GAN based unsupervised domain adaptation model has been incorporated to address the domain shift problem between two heterogeneous stained images and hand crafted feature detection technique like SURF feature[22] has been adopted to perform registration without requiring prior spatially corresponding training data. The details of the proposed method will be discussed in the next section and the other parts of this paper are organized as follows: section 3 describes the experimental details including data acquisition and preprocessing, section 4 represents experimental result, section 5 represents discussion and finally section 6 concludes our work.

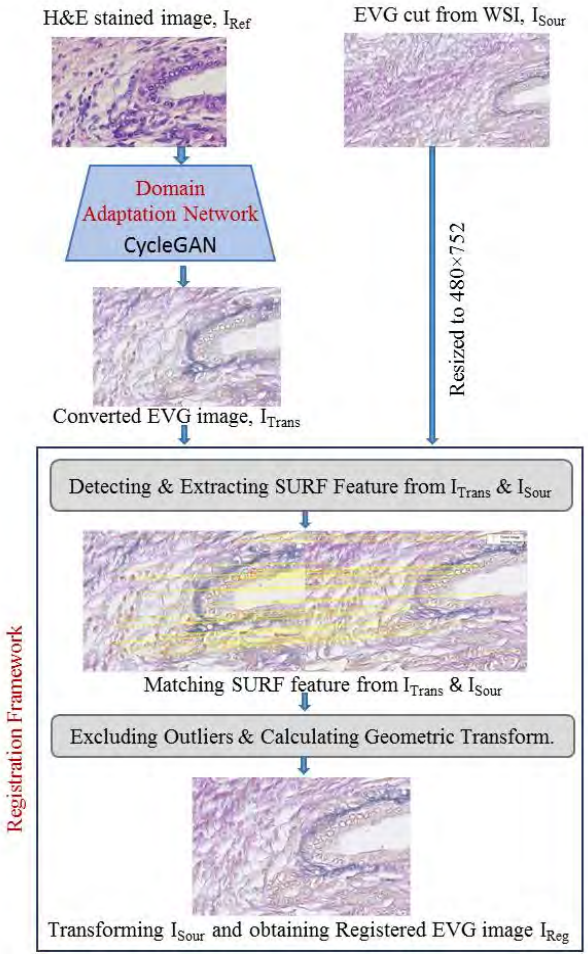


Fig.1 Proposed registration method

2. Proposed Method

This study aims to provide a registration framework which is able to overcome the problem of detecting and matching features from images of different stains because of the huge difference in the color information. For the implementation of our proposed method, two different types of stained images such as H&E (Hematoxylin and Eosin) stained and EVG (Verhoeff-Van Gieson) stained images are deployed in this research. The proposed method is shown in **Fig. 1**. It contains two stages including domain adaptation as a preprocessing step for registration and the registration itself. The domain adaptation part addresses the domain shift problem because of

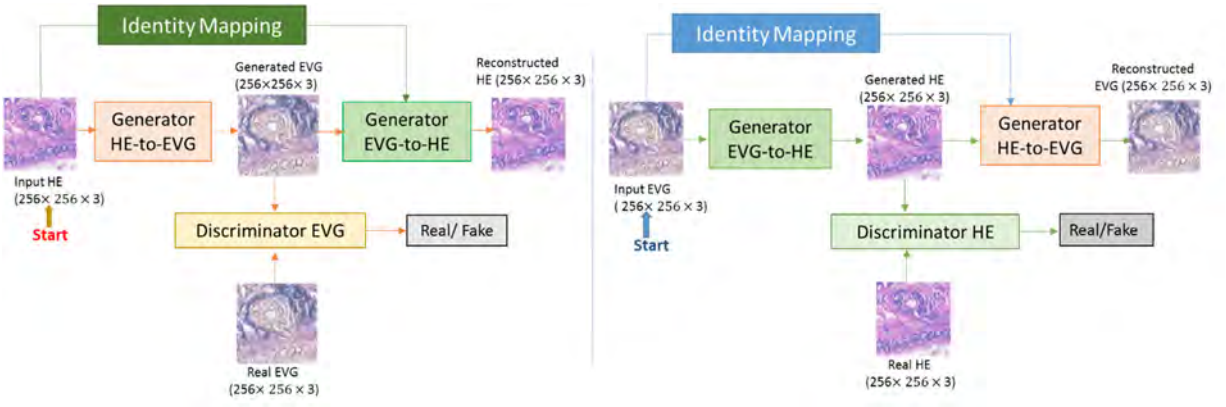


Fig.2 Methodology of adopting the domain of H&E to the domain EVG

color variation between differently stained images. This is done by adapting the domain of H&E to its equivalent domain EVG using GAN based translation. It reduces the distance of the distribution of color information between differently stained images which enables our registration framework to detect and match greater number of similar features between them. The other part detects and matches local features such as SURF between the original and translated EVG stained image and produces the registered EVG stained image by transforming the original EVG stained image.

2.1 Domain adaptation network

The objective of the domain adaptation network is to address the domain shift problem between H&E and EVG stained images by converting the images from H&E domain to its equivalent EVG domain so that the similarity in the color information between the H&E and EVG stained images are increased. This will contribute greatly to the detection and matching of larger number of local features by the registration framework in the later part. Our proposed domain adaptation network is based on CycleGAN model[23]. The methodology of adapting the domain of H&E stain to the domain of EVG stain has been shown in **Fig. 2**. The network consists of two generators and two discriminators. The 'Generator HE-to-EVG' takes the H&E stained image as input and generates the corresponding EVG stained image as output and the 'Generator EVG-to-HE' takes the EVG stained image as input and produces the equivalent H&E stained image as output. The 'Discriminator HE' tries to identify if the input image is coming from the original dataset of H&E domain or from the output of the 'Generator EVG-to-HE' and the 'Discriminator EVG' tries to identify if the input image is coming from the original dataset of EVG domain or from the output of the 'Generator HE-to-EVG'. The generators are designed following the ResNet architecture[24] and the discriminators are based on PatchGAN architecture[25]. The full objective function to train the model is described in Eq.(1).

$$\mathcal{L}_{Total} = \mathcal{L}_{adv.} + \lambda\mathcal{L}_{cycle-consis.} + \gamma\mathcal{L}_{idt.} \quad (1)$$

Where, $\mathcal{L}_{adv.}$, $\mathcal{L}_{cycle-consis.}$, $\mathcal{L}_{idt.}$ and $\mathcal{L}_{Total}$ represent adversarial loss, cycle-consistency loss, identity loss and the total loss to train the model respectively. Also, λ and γ are two scalar values which have been multiplied with cycle-consistency loss and identity loss respectively. Adversarial loss contributes to generate realistic H&E and EVG stained images which have similar distribution as the original dataset. The cycle-consistency loss tries to ensure that the anatomical structure of the input source domain and generated target domain remains unchanged. The identity loss contributes to imitate the exact color of the target domain. The adversarial loss has been calculated in terms of Mean Square Error (MSE) and the cycle-consistency and identity loss are calculated in terms of Mean Absolute Error (MAE). CycleGAN has been deployed in this study because of its ability to perform domain adaptation in a fully unsupervised way where prior paired (aligned and scale adjusted) training data is not required. Also, it facilitates the translation of color information from one domain to another while remaining their structural information unchanged.

2.2 Registration framework

Original EVG stained image cropped from the WSI, I_{Sour} and the converted EVG stained image, I_{Trans} obtained from the original H&E stained image, I_{Ref} has been used as the input of the registration framework. For the registration purpose, style transferred H&E stained image, I_{Trans} has been considered as target image according to which the registered image will be obtained. On the other hand, cropped EVG stained images

have been considered as the source image, I_{Sour} which is needed to be transformed. To provide lower computational complexity and higher effectiveness, hand crafted feature such as SURF feature[22] has been leveraged in our registration framework. Among different types of hand crafted features, we have utilized SURF feature because of its faster performance, robustness and scale and rotation invariant property. SURF feature is detected and extracted both from I_{Trans} and I_{Sour}. The features obtained from these images are then matched by calculating the distance between the feature sets. Outliers in the matched features are excluded and geometric transformation of type affine is calculated where a matrix contains the information of translation (horizontal or vertical shifting), rotation, scaling and shearing for determining the mapping of the position of each pixel from the source image to the corresponding position of that pixel in the target image. This is done using the M-estimator SAmple Consensus (MSAC) algorithm[26] which is a variant of popular RANSAC (Random Sample Consensus) algorithm and able to provide improved result than RANSAC by incorporating a better cost function. Though sometimes MLESAC (Maximum Likelihood Estimation Sample Consensus) provides better accuracy than MSAC, it takes more computational complexity with longer processing time. Considering the balance between processing time and performance, we chose to use MSAC. Finally, the original EVG stained image is warped according to the information of geometric transformation and the registered EVG stained image is obtained.

3. Experiment

3.1 Experimental data

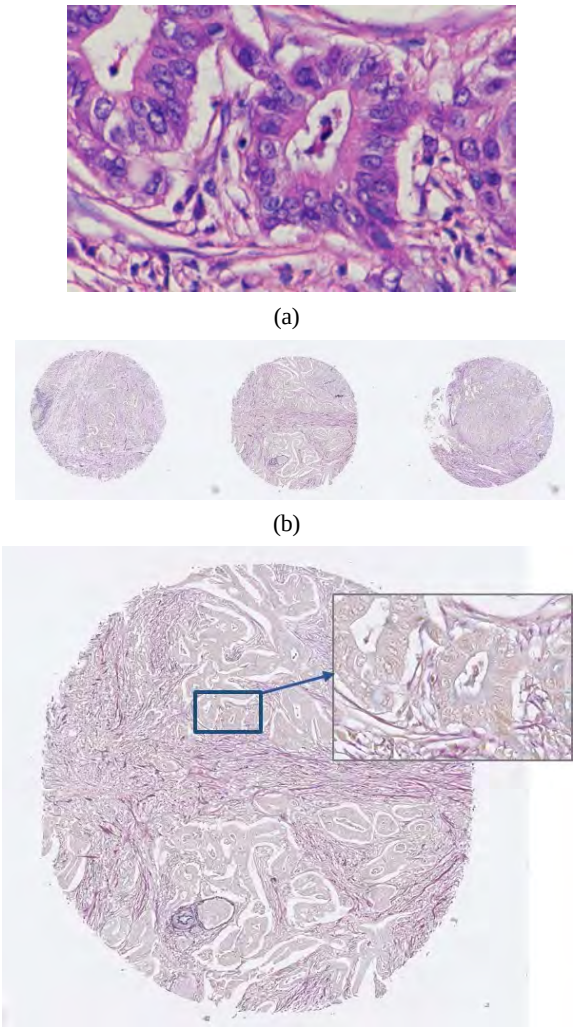
This study involves two different types of stained images: H&E stained image and EVG stained image. Though H&E stain is the most commonly used staining technique, elastic and collagen fiber cannot be differentiated using it because both of them are shown in pink color. On the other hand, EVG shows elastic in deep blue color and collagen in orchid color thus enables the easy discrimination of these two tissue components. For visual comparison of these two types of stains images or for digital conversion of these stain types and segmentation of particular tissue component, proper registration of these images plays an important role for further analysis or processing[2].

The H&E and EVG stained images used in this experiment are of human pancreatic tissue specimens which have been obtained from Biomax Inc. H&E stained images are hyperspectral images captured with hyperspectral camera of NH3 EBA JAPAN CO. LTD with which an optical microscope BX-53 by Olympus Corp. and a white LED are attached. Hyperspectral H&E stained images are obtained using the transmittance information of the tissue corresponding to the wavelength 350-1100nm with 5nm interval (151 channels) having the image size 480×752. After removing redundant information, we considered the hyperspectral image within the visible wavelength from 420 to 720 nm of 61 channels. EVG stained images are RGB Whole Slide Image (WSI) obtained using Hamamatsu photonics K.K. For the sake of retaining lower computational complexity, H&E stained hyperspectral images have been converted to sRGB image. The hyperspectral image is at first converted to X, Y and Z tristimulus values using illumination spectrum D65 which is a CIE standard illuminant corresponding to daylight and 1931 CIE XYZ color matching functions[27]. X, Y and Z tristimulus values are then converted to sRGB image using the XYZ to sRGB transformation matrix.

H&E and EVG stained images have a consistent anatomical structure because EVG staining has been performed after bleaching the tissue where H&E staining was imposed prior. Same area as H&E stained image has been cropped manually from EVG stained WSI to prepare the dataset with a set of H&E and EVG image which are unaligned and not scale adjusted (unpaired). Then, EVG stained image is resized so as to be 480×752. An example of preparing a set of H&E and EVG stained images is shown in **Fig. 3**.

3.2 Experimental details

For the implementation of the first stage of our proposed methodology, the domain adaptation CycleGAN network has been trained with images from two different domains: H&E and EVG. These training images have been prepared by cropping the overlapping area of size 256×256 from the set of H&E and EVG stained images which have been prepared previously. After applying the image augmentation techniques of rotation, flipping and brightness change on these images, a total of 9952 images in each domain has been finalized to train the model. Among the training images, H&E and EVG stained images may have similar biological structures but they do not spatially correspond to each other (unpaired). These training data for the domain adaptation network have been collected from two different tissue samples. While training the domain adaptation network, empirically it has been found that



(a)

(b)

(c)

Fig.3 Preparing dataset; (a) H&E stained image, (b) EVG Whole Slide Images (WSIs), and (c) EVG stained image with same biological structure as H&E is cropped from WSI

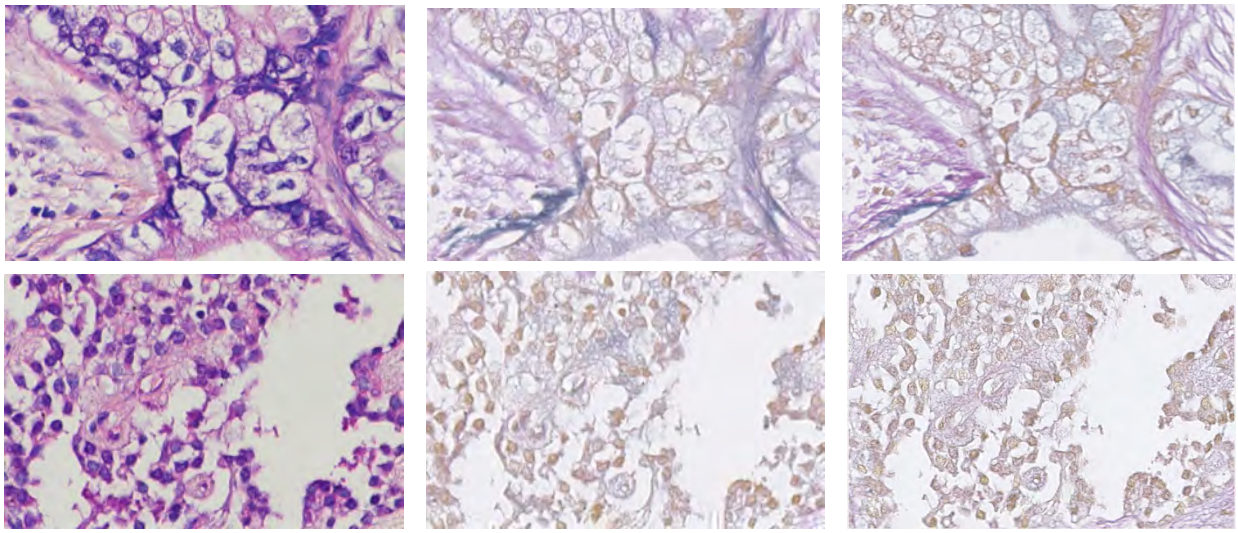
the model is generalized better for setting the value of λ and γ of Eq. 1 to 5 and 0.5 respectively. The CycleGAN domain adaptation model is trained on the GPU Quadro RTX 6000 for 81 epochs with training image batch size 1.

For the registration purpose, 85 image pairs of H&E stained and EVG stained (cropped from WSI) from four different tissue samples have been prepared. H&E stained image, I_{Ref} of size 480×752 has been converted to its corresponding EVG stained image, I_{Trans} using the trained domain adaptation CycleGAN model. This domain adapted H&E stained image, I_{Trans} and original EVG stained image, I_{Sour} of size 480×752 has been used as the input of registration framework. According to the information of calculated geometrical transformation, I_{Sour} has been transformed to obtain the registered EVG stained image. The domain adaptation part of this work has been implemented in Python using Pytorch toolkit and registration part has been implemented in Matlab using Computer Vision and Image Processing toolbox.

4. Result

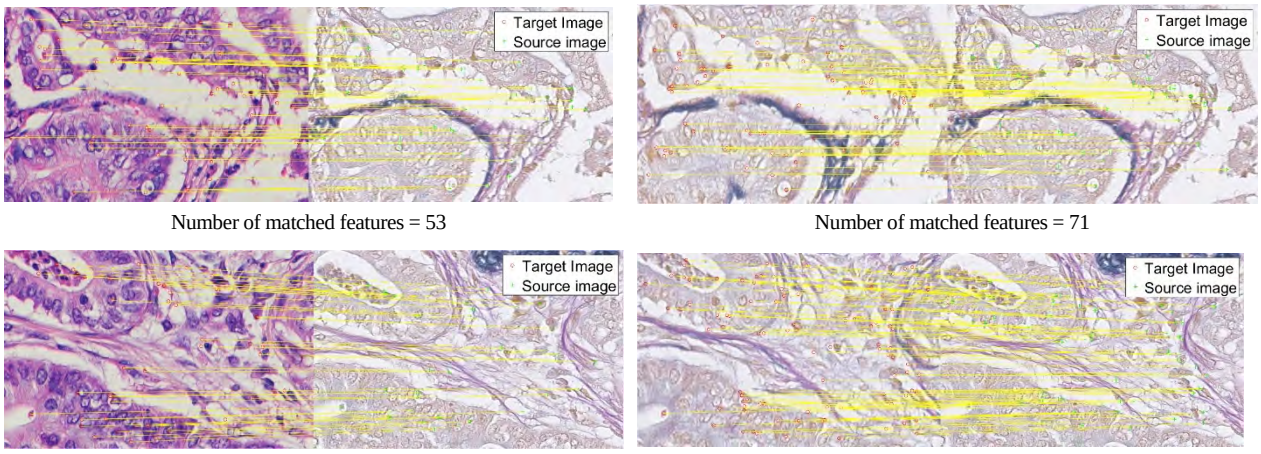
Our proposed domain adaptation network doesn't change the morphological information of the tissue. It only changes the color information of the tissue components to adapt the target domain (EVG stained) from its source domain (H&E stained). **Figure 4** represents the result of our domain adaptation network from which we can see that the converted EVG stained image has been adapted well to the original EVG stained image domain in terms of color information while retaining the morphological information same as the input H&E stained image. Now, the color information of the converted EVG stained image and the original EVG stained image has more similarity than that of original H&E and original EVG which has enabled the registration framework to obtain greater number of correctly matched features from the first pair than the later one. This is shown in **Fig. 5** where (a) represents the correctly matched features between the original EVG and original H&E without domain adaptation and (b) represents the correctly matched features between the original EVG and converted EVG using the domain adaptation network. By comparing these two, one can observe that larger number of matched features have been found for the pair (b) than that of (a). For the visual evaluation of the registration result, overlapping view of H&E stained image and EVG stained image before and after registration with and without domain adaptation network has been represented in **Fig. 6**. The magenta color represents H&E image and the green color represents EVG image. From Fig. 6 (a), we can see that, though H&E stained and EVG stained image have similar biological structure, they are not spatially aligned. Fig. 6 (b) and (c) represents the superimposed view of original H&E stained image and registered EVG stained image without and with domain adaptation network respectively. In the superimposed image, where H&E and EVG stained images are not aligned properly are looking blurred with dominating green color. From the zoomed view, it can be said that H&E and EVG stained images have been aligned (registered) more accurately with our proposed domain adaptation based registration method than the method where domain adaptation is not applied.

Apart from visual evaluation, we have also determined the registration result in terms of structural similarity (SSIM) and Mutual Information (MI). As the name indicates, SSIM



(a) (b) (c)

Fig.4 Result of domain adaptation network; (a) Input H&E, (b) Converted EVG, and (c) Original EVG image



Number of matched features = 53 Number of matched features = 71

Number of matched features = 50 Number of matched features = 96

(a) (b)

Fig.5 Matched feature between (a) Original H&E and original EVG, and (b) Converted EVG and original EVG

determines the similarity in the structural information between two images whereas MI indicates how well signal intensity of one image can be predicted when another is known. Both SSIM and MI are considered as good matching metrics of images which measures the spatial alignment between two images. Because of the non-deterministic characteristics of MSAC algorithm, there are a little variation in the result among different runs of the registration framework. To address this problem, we run each program for the registration of a single pair of images 3 times. In order to perform a fair comparison between our proposed and baseline method by considering the best possible performance provided by each method, among these three consecutive runs we have recorded the maximum resultant value for each pair of images. Then the average and standard deviation of the registration result for all the test images of four different tissue samples (T.S.1-4) have been recorded separately. **Table 1** summarizes this registration result in terms of structural similarity (SSIM) and **Table 2** summarizes the registration result in terms of MI comparing our proposed method (Prop. Meth.) with the baseline SURF feature-based registration without domain adaptation network (W/O DA) i.e., registration without stain conversion. In each cell of the table, upper and lower values represent averages and standard deviations respectively. For both cases of SSIM and MI, higher value indicates better registration. From Table 1 and Table 2, we can see that for every tissue sample, our proposed method has provided better registration result than the baseline method.

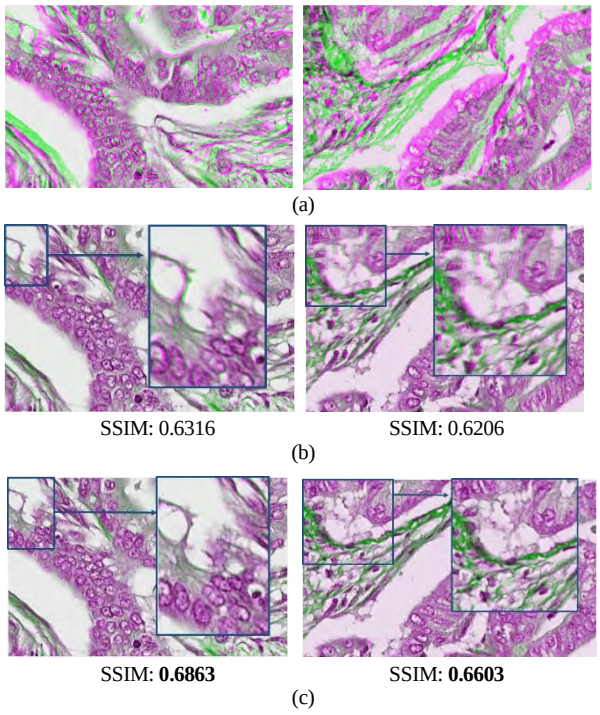


Fig.6 Visual evaluation of registration result: Overlapping view of H&E and EVG image; (a) before registration, (b) after registration without domain adaptation, and (c) after registration with proposed method

Table 1 Registration quality determination using SSIM (upper and lower values represent averages and standard deviations respectively)

Method	T.S.1	T.S.2	T.S.3	T.S.4	Avg.
W/O DA	0.7053	0.6322	0.6581	0.6785	0.6685
	0.0446	0.0424	0.0606	0.0452	0.0482
Prop. Meth.	**0.7315**	**0.6655**	**0.6919**	**0.7178**	**0.7017**
	0.0415	0.0456	0.0578	0.0413	0.0465

Table 2 Registration quality determination using MI (upper and lower values represent averages and standard deviations respectively)

Method	T.S.1	T.S.2	T.S.3	T.S.4	Avg.
W/O DA	0.8680	0.8101	0.8019	0.8621	0.8355
	0.1646	0.1437	0.1426	0.2313	0.1706
Prop. Meth.	**0.9253**	**0.8714**	**0.8693**	**0.9448**	**0.9027**
	0.1637	0.1565	0.1481	0.2182	0.1716

5. Discussion

In this work, the domain adaptation network is designed to change only the color information of the source domain (H&E) to the target domain (EVG) while remaining the morphological information of the source domain unchanged. This has enabled our domain adaptation network to produce an intermediate domain image which has exact same structural information as original H&E stained image and same color information as original EVG stained image. Because of the adversarial training nature of GAN, sometimes it may generate very smooth noiseless image but which has less similarity with the target domain color. On the other hand, it may generate image with very close adaptation of target domain color but introducing noises into the image. So, careful consideration is required while choosing the trained domain adaptation model for conversion purpose. Here, we have concentrated on two main factors at the same time while converting H&E stained image to EVG stained image: one is the generation of noiseless image and another is the perfect imitation to the color of original EVG stained image. Through the experiment, we have found that for ensuring larger number of matched features it is very important for the converted EVG stained image to be noiseless and perfectly adapted with target domain. As elastic and collagen fiber have very similar features in the H&E stained image and very different feature in the EVG stained image, the differentiation of the elastic and collagen fiber is not so perfect in the converted EVG stained image. But this didn't hamper the extraction and matching features because all other tissue components have been adapted properly to the color of original EVG stained image and also the area where elastic and collagen is not identified properly is less. Moreover, many CycleGAN based studies haven't considered identity loss during training[21),28)], but in our case considering identity loss has been very effective to imitate the exact color of EVG stained image and to differentiate the elastic and collagen fiber in the converted EVG stained image. The generation of perfectly adapted and noiseless EVG stained image has enabled the matching of larger number of features with the source image distributed all over the image and not concentrated on a single area of image which has contributed greatly to the better estimation of geometric transformation information and finally the improved registration result. It also takes very low processing time to convert an image from H&E to EVG. Once the domain adaptation model is trained it takes around 3s to convert the domain. The registration framework takes less than 1s (around 0.7 – 0.9s) for the registration of a single pair of images. So, the overall processing time of our proposed method is very low which is around 4s only.

6. Conclusion

In this research, we have proposed a promising method for the registration of histopathological images obtained from heterogeneous staining techniques such as H&E and EVG. For reducing the gap of color information between these two types of stained images, we have proposed to use a CycleGAN based domain adaptation network which has adapted the source domain images of H&E stained to the target domain images of EVG stained so that the registration framework is able to detect and match more features which has contributed greatly to the significant improvement of registered EVG stained image. As our proposed method leverages an unsupervised domain adaptation technique like CycleGAN and also uses the handcrafted feature such as SURF feature for obtaining the registered image, the major advantage is that without requiring prior aligned training data, the method is able to provide promising registration result while retaining briefness, simplicity and lower computational complexity in its implementation.

Acknowledgement

The authors thank to Dr. Takaya Ichimura from Saitama Medical University for his valuable comment on this work. The authors are also grateful to Otsuka Toshimi Foundation for the financial support.

References

1) L. Cooper,O. Sertel, J. Kong, G. Lozanski, K. Huang, M. Gurcana: “Feature-based Registration of Histopathology Images with Different Stains: an Application for Computerized Follicular Lymphoma Prognosis”, Computer Methods and Programs in Biomedicine, Vol. 96, No. 3, pp. 182–92 (2009).
2) L. Septiana, H. Suzuki, M. Ishikawa, T. Obi, N. Kobayashi, N. Ohyama, T. Ichimura, A. Sasaki, E. Wihardjo, D. Andiani: “Elastic and Collagen Fibers Discriminant Analysis Using H&E Stained Hyperspectral Images”, Optical Review, Vol.26, pp. 369–379 (2019).
3) I. Arganda-Carreras, C. Sorzano, R. Marabini, J. M. Carazo1, C. Ortiz-de-Solorzano, J. Kybic: “Consistent and Elastic Registration of Histological Sections Using Vector-Spline Regularization”, Computer Vision Approaches to Medical Image Analysis, Vol. 4241, pp. 85–95 (2006).
4) J. Borovec, J. Kybic, M. Bušta, C. Ortiz-de-Solórzano, A. Muñoz-Barrutia: “Registration of Multiple Stained Histological Sections”, Proc. of IEEE International Symposium on Biomedical Imaging, pp. 1034–1037 (2013).
5) J. Borovec, A. Munoz-Barrutia, J. Kybic: "Benchmarking of Image Registration Methods for Differently Stained Histological Slides", Proc. of 25th IEEE International Conference on Image Processing (ICIP 2018), pp. 3368–3372 (2018).
6) M. Wodzinski, A. Skalski: "Multistep, Automatic and Nonrigid Image Registration Method for Histology Samples Acquired Using Multiple Stains", Physics in Medicine & Biology, Vol. 66, No. 2 (025006), (2021).
7) J. Kybic, J. Borovec, "Automatic Simultaneous Segmentation and Fast Registration of Histological Images", Proc. of IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 774–777 (2014).
8) D. Mahapatra, Y. Sun: “Integrating Segmentation Information for Improved MRF-Based Elastic Image Registration”, IEEE Trans. on Image Processing, Vol.21, No.1, pp. 170–183 (2012).
9) D. Mahapatra, Z. Ge: "Training Data Independent Image Registration with Gans Using Transfer Learning and Segmentation Information", Proc. of IEEE 16th International Symposium on Biomedical Imaging, pp. 709–713 (2019).
10) H. Sokooti, B. de Vos, F. Berendsen, B. P. F. Lelieveldt, I. Isgum, M. Staring: "Non-Rigid Image Registration Using Multi-scale 3D Convolutional Neural Networks", Proc. of Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 232–239 (2017).
11) S. Miao, Z.J. Wang, R. Liao: "A CNN Regression Approach for Real-time 2D/3D Registration", IEEE Trans. on Medical Imaging, Vol.35, No.5, pp. 1352–1363 (2016).
12) A. Torralba, A. A. Efros: “Unbiased Look at Dataset Bias”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1521–1528 (2011).
13) E. A. Albadawy, A. Saha, M. A. Mazurowski: “Deep Learning for Segmentation of Brain Tumors: Impact of Cross-Institutional Training and Testing”, Medical Physics, Vol. 45, No. 3, pp. 1150–1158 (2018).
14) M. Long, H. Zhu, J. Wang, M. I. Jordan: “Deep Transfer Learning with Joint Adaptation Networks”, Proc. of the 34th International Conference on Machine Learning (ICML 2017), pp. 2208–2217 (2017).
15) Y. Chen, W. Li, C. Sakaridis, D. Dai, L. V. Gool: “Domain Adaptive Faster R-CNN for Object Detection in the Wild”, Proc. of Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3339–3348 (2018).
16) Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, M. Chandraker: “Learning to Adapt Structured Output Space for Semantic Segmentation”, Proc. of Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7472–7481 (2018).
17) T. Wollmann, C. S. Eijkman, K. Rohr: “Adversarial Domain Adaptation to Improve Automatic Breast Cancer Grading in Lymph Nodes”, Proc. of IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 582–585 (2018).
18) A. Gholami, Subramanian, V. S., Shenoy, N. Himthani, X. Yue, S. Zhao, P. Jin, G. Biros, K. Keutzer: “A Novel Domain Adaptation Framework for Medical Image Segmentation”, Proc. of International Medical Image Computing and Computer Assisted Intervention (MICCAI) Brainlesion Workshop, pp. 289–298 (2018).
19) E. Ahn, A. Kumar, M. Fulham, D. Feng, J. Kim: “Unsupervised Domain

Adaptation to Classify Medical Images Using Zero-Bias Convolutional Auto-Encoders and Context-Based Feature Augmentation", IEEE Trans. on Medical Imaging, Vol.39, No.7, pp. 2385–2394 (2020).

20) M. Javanmardi, T. Tasdizen: "Domain Adaptation for Biomedical Image Segmentation using Adversarial Training", Proc. of 15th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 554–558 (2018).
21) D. Mahapatra, Z. Ge: "Training Data Independent Image Registration using Generative Adversarial Networks and Domain Adaptation", Pattern Recognition, Vol.100 (2020).
22) H. Bay, A. Ess, T. Tuytelaars, L. Van Gool: "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol.110, No.3, pp. 1–10 (2008).
23) J. Zhu, T. Park, P. Isola, A. A. Efros: "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks", Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017).
24) K. He, X. Zhang, S. Ren, J. Sun: "Deep Residual Learning for Image Recognition", In Proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016).
25) P. Isola, J. Zhu, T. Zhou, A. A. Efros: "Image-to-Image Translation with Conditional Adversarial Networks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15967–5976 (2017).
26) P.H.S. Torr, A. Zisserman: "MLESAC: A New Robust Estimator with Application to Estimating Image Geometry", Computer Vision and Image Understanding, Vol.78, No.1, pp. 138–156 (2000).
27) N. P. Jacobson, M. R. Gupta: "Design Goals and Solutions for Display of Hyperspectral Images", IEEE Trans. on Geosci. And Remote Sensing Vol.43, No.11, pp. 2684–2692 (2005).
28) M. T. Shaban, C. Baur, N. Navab, S. Albarqouni: "Staingan: Stain Style Transfer for Digital Histological Images", Proc. of IEEE 16th International Symposium on Biomedical Imaging (ISBI), pp. 953–956 (2019).

(Received May 28, 2021)
(Revised December 12, 2021)

Tanwi BISWAS

She received B.Eng. degree on Electronics and Communication Engineering from Khulna University of Engineering and Technology, Bangladesh in 2016. She completed M.Eng. on Information and Communications in 2018 and currently pursuing PhD on Human Centered Science and Biomedical Engineering major in Tokyo Institute of Technology, Japan. She is working on medical image processing using deep learning, hyperspectral imaging, domain adaptation and generative adversarial network (GAN).

Hiroyuki SUZUKI

He received B.E., M.E. and Ph.D. degrees from Tokyo Institute of Technology in 1998, 2000 and 2006, respectively. From 2003 to 2004, he was a Researcher with Tokyo Institute of Technology. From 2004 to 2020, he was an Assistant Professor with Tokyo Institute of Technology. Since 2020, he has been working as an Associate Professor with Center for Mathematics and Data Science, Gunma University. His research interests include optical information security, holograms, biometric authentication, medical and healthcare information systems, and medical image processing. He is a member of IEICE, JSAP, and OSJ.

Masahiro ISHIKAWA (*Member*)

He is an assistant professor at the Faculty of Health and Medical Care, Saitama Medical University. He received his MS and PhD degrees in engineering from Niigata University, Niigata Japan, in 2004 and 2007, respectively. His research interests are medical image processing, pathology image recognition, and computer-aided diagnosis.

Naoki KOBAYASHI *(Fellow)*

He received his B.Sc. and M.E. degree from Tokyo Institute of Technology, Tokyo, Japan, in 1979 and 1981, respectively, and his Ph.D. from Niigata University, Niigata, Japan, in 2000. He worked for Laboratories of Nippon Telegraph and Telephone Corp. from 1981 and 2008. He has been a professor at the School of Biomedical Engineering, Faculty of Health and Medical Care of Saitama Medical University since 2003. His research interests include medical image proceeding, image compression and biosignal processing. He is a member of IEICE, IIEEJ, JSMBE and IEEE.

Takashi OBI

He earned B.S. degree in Physics, and M.S. and Ph.D degree in Information Physics from Tokyo Inst. of Tech, Japan in 1990, 1992 and 1996, respectively. Currently, he is an Associate Professor of Laboratory for Future Interdisciplinary Research of Science and Technology in Institute of Innovative Research, Tokyo Inst. of Tech. His research focuses on medical image processing, medical informatics, information system and security, etc. He is a member of IEICE, JAMIT, JSAP, JSNM, JSMP and IEEE.

IIEEJ **Paper** *-- Special Issue on Journal Track Papers in IEVC2021Part II --*

Spectral Super-Resolution Using CNN Decomposing a Color Image into Luminance and Chrominance Components

Masahiro SAKAMOTO† , Kazufumi KANEDA†(*Member*) , Bisser RAYTCHEV†

† Hiroshima University, Graduate School of Advanced Science and Engineering

<**Summary**> Hyper-spectral images are used in a wide range of fields such as industry, medicine, remote sensing, and so on. They are also used in computer graphics as light probe images and textures in spectral rendering. The acquisition of spectral images is, however, costly in terms of equipment and time, which hinders its acquisition and use. Conventional spectral super-resolutions using deep learning have been adopting a direct end-to-end learning method to RGB and hyper-spectral images. In contrast, we focus on the fact that hyper-spectral images are decomposed into luminance and chrominance components, and we propose a novel spectral super-resolution using a deep learning to estimate each component separately. Finally, in the proposed method, a hyper-spectral image is reconstructed by combining the estimated luminance and chrominance components.

Keywords: spectral super-resolution, hyper-spectral image, convolutional neural network, deep learning, luminance and chrominance

1. Introduction

RGB images use red, green, and blue to represent colors based on human visual perception, while hyper-spectral images are composed of many wavelength bands with dozens or more bands and represent physical quantities. Since materials in nature have unique spectral reflectance, hyper-spectral images show the characteristics of materials in detail. Hyper-spectral images are used in a wide range of fields such as industry, medicine, agriculture, and remote sensing[1)2)]because their abundant spectral information enables sophisticated discrimination that is difficult to achieve with RGB images. Hyper-spectral images are also used in computer graphics as light probe images and textures in spectral rendering. Physically-based rendering requires the use of hyper-spectral images to simulate physically accurate light transport for generating images.

Commonly used cameras capture RGB images where each light receiving element of the cameras is affected by a wide range of wavelength, while hyper-spectral cameras can identify wavelengths in a narrow range and acquire wavelength information from dozens or hundreds of bands. Capturing images in narrower bandwidth provide more spectral information, but it proportionally reduces the number of photons reaching a sensor element. As a result, it becomes difficult to maintain the signal-to-noise ratio without reducing the resolution of the image[3)]. Therefore, the trade-off between spectral and spatial resolutions is an important issue for the acquisition of hyper-spectral images. In addition, the high cost of hyper-spectral cameras is also a problem because of the special sensors and complex processing.

For this reason, research on super-resolution for hyper-spectral images has been conducted. Super-resolution for hyper-spectral images can be classified into two types: spatial super-resolution and spectral super-resolution. We focus on the spectral super-resolution from RGB images, which deals with the reconstruction of hyper-spectral images with many bands and with the same spatial resolution as the input RGB image. Since this super-resolution can use RGB images captured by general cameras, the acquisition of spectral images is a lower cost than using expensive hyper-spectral cameras.

Conventional spectral super-resolution using deep learning generates a hyper-spectral image by directly estimating the spectral intensities. Taking advantage of the fact that hyper-spectral images are decomposed into luminance and chrominance components, we propose a novel spectral super-resolution using a deep learning to estimate each component separately. Rather than learning spectral intensities directly end-to-end, a convolu-

tional neural network (CNN) can estimate the nonlinear relationship between RGB and hyper-spectral images with better accuracy when learning each component separately. Finally, in the proposed method, a hyper-spectral image is reconstructed by combining the estimated luminance and chrominance components.

The hyper-spectral images reconstructed by the proposed method can be used for spectral rendering in computer graphics. Although the proposed method cannot solve the metamerism problem, which is an important issue of estimating spectrum, it may provide some solution to the problem, because the spectral super-resolution using CNN estimates the spectral intensity not only from the tristimulus values at the pixel, but also from those of the surrounding pixels. This may lead to some avoidance of metamerism, and a spectrum close to the real could be estimated.

2. Related Work

2.1 Upsampling methods for rendering

Although three stimulus value representation of an RGB image may be sufficient for a color display, it is not exactly enough for an accurate simulation of light transport. Physically-based rendering that allows for realistic representations requires densely sampled wavelengths over the visible light region. However, we don't have many hyper-spectral images that can be used as textures or light probes. To solve the problem, research on obtaining spectral distributions from RGB tristimulus values has been developed.

Spectral rendering can generate physically accurate images, but it has the problem that the processing time and memory usage becomes large because the number of sampled wavelengths increases. It is desirable to express the spectral distribution in a compact manner. In addition, spectral distribution normally tends to be smooth. Considering these things, methods for reconstructing the spectral distribution using basis functions and few coefficients have been proposed[4),5)].

2.2 Spectral super-resolution

Most of the existing spectral super-resolution methods recover spatial resolution, while few methods attempt to recover spectral information.

To improve the spatial resolution of hyper-spectral image, methods for combining a high resolution RGB image and a low resolution hyper-spectral image have been developed. These methods require accurate alignment between the images, but it is not easy to capture two different images in the same position. To address the problem, some methods proposed the correction for spatial mis-alignment by using the spectral response of RGB camera[6)].

Spectral super-resolution, which creates a hyper-spectral image from an RGB image without using any auxiliary low-resolution hyper-spectral image, is most related to our research. In contrast to spatial super-resolution, spectral super-resolution recovers high-frequency components in spectral domain without reducing spatial resolution. The automatic colorization of a grayscale image to an RGB color image also belongs to a kind of spectral super-resolution. Recently a colorization method[7)]has been improved by separating the image into luminance and chrominance components. The method estimates only the chrominance component using CNN.

In reconstructing a hyper-spectral image from an RGB image, several methods have been proposed, such as a method using K-SVD to learn a sparse dictionary of hyper-spectral images[8)]and a method using an RBF (Radial Basis Function) network to learn a nonlinear mapping function between images[9)]. Recently, deep learning methods based on CNN have been used for spectral super-resolution. Inspired by a network model used for semantic segmentation[10)], Galliani et al.[11)]applied the model to spectral super-resolution. He et al.[12)]proposed a method optimizing the network model by utilizing an auxiliary spectral response function. They employed the loss function that combines L1 loss and a spectral angle mapper to achieve quick convergence. In the same way, Li et al.[13)] proposed a novel adaptive weighted attention network using a spectral response function. They incorporated the discrepancies between the original RGB image and the recovered RGB image from the hyper-spectral image as a finer constraint for accurate reconstruction.

3. Proposed Method

The overview of the proposed method is shown in **Fig. 1**. The proposed method learns luminance and chrominance components of hyper-spectral images separately, instead of directly learning spectral intensities. Luminance and chrominance components are estimated by the respective CNNs. The idea of the proposed method is based on the fact that the shape of the spectral distributions is not affected by the intensity of the incident light. Since the proposed method does not require one network to learn extra information, it can be

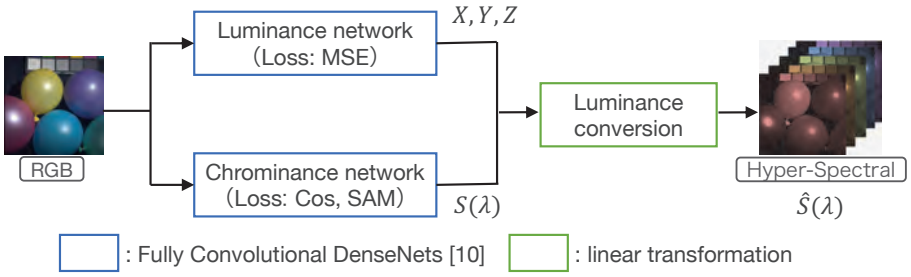


Fig. 1 Overview of the proposed method

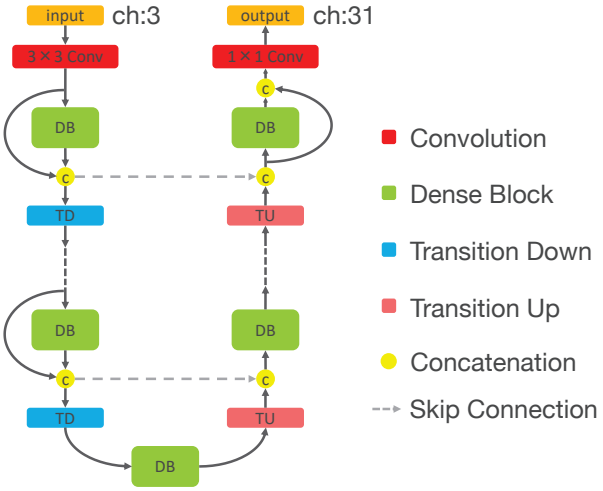


Fig. 2 Diagram of Fully Convolutional DenseNets

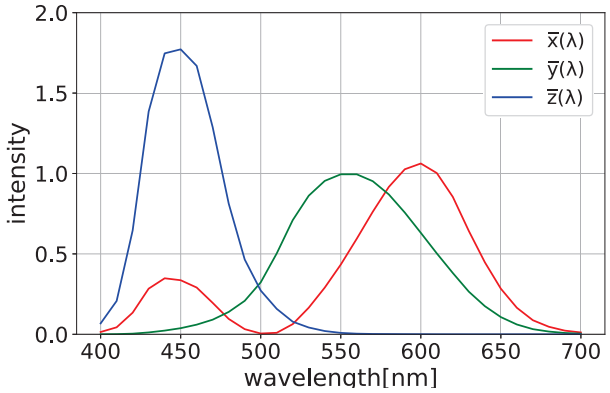


Fig. 3 The CIE XYZ color matching functions

expected to reduce the error of each component. Luminance has intensity information, and chrominance is represented by a nomalized spectral distribution. Using the luminance and chrominance output from the respective CNNs, a hyper-spectral image is reconstructed by the luminance conversion in which the normalized chrominance components are scaled to match the estimated luminance components.

In both networks for luminance and chrominance components, we employed Fully Convolutional DenseNets[10] (FC-DenseNets) shown in **Fig. 2**. We train the networks for luminance and chrominance components separately. Galliani et al.[11]applied the FC-DenseNets to spectral super-resolution, since the network for semantic segmentation outputs an image of the same size as the input image. The FC-DenseNets has been frequently followed in related works, and it is the standard method for spectral super-resolution using deep learning. Thus, we use the network in our experiments as well.

3.1 Luminance components

In the proposed method, the spectral magnitude of a hyper-spectral image is determined based on X, Y, and Z values in CIE XYZ color space. Here, we consider "spectral magnitude" to correspond to the intensity of the incident light. Section 3.3 describes the method for calculating the spectral magnitude in detail.

The conversion from a hyper-spectral image $S(\lambda)$ to (X, Y, Z) values is expressed by the following equations:

$$X = \frac{1}{k}\int_{\Lambda} S(\lambda)\bar{x}(\lambda)d\lambda, \tag{1}$$

$$Y = \frac{1}{k}\int_{\Lambda} S(\lambda)\bar{y}(\lambda)d\lambda, \tag{2}$$

$$Z = \frac{1}{k}\int_{\Lambda} S(\lambda)\bar{z}(\lambda)d\lambda, \tag{3}$$

$$k = \int_{\Lambda} \bar{y}(\lambda)d\lambda, \tag{4}$$

where λ is the wavelength, Λ is wavelength range, $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ are the color matching functions shown in **Fig. 3**. We use MSE Loss as a loss function for the network to estimate the spectral magnitude, and the network learns the relationship between (R, G, B) values of an RGB image and (X, Y, Z) values of a hyper-spectral image.

This allows the luminance network to learn the nonlinear relationship of tristimulus values between RGB and hyper-spectral images. Because of the nonlinear characteristics of the camera response function, RGB values and the spectral intensities normally have a nonlinear relationship. Although the camera response function is specific to each camera, it can be approximated with gamma correction for $\gamma = 2.2$ in most cases. Therefore, in our experiments, we use gamma correction $\gamma = 2.2$ to re-

generate RGB images from hyper-spectral images as a representation of a standard camera response function.

3.2 Chrominance componets

Suppose an N-bands hyper-spectral image, spectral values for each pixel of the image can be considered as an N-dimensional vector. For examples, just as (X, Y, Z) values in the color space can be considered as a three-dimensional vector. The direction of the vector corresponds to normalized chrominance components, and the magnitude of the vector corresponds to luminance. In other words, chrominance components of N-spectral bands can be expressed as the direction of an N-dimensional vector.

Cosine similarity is used to measure how close the directions of two vectors are. We apply a cosine similarity shown in the following equation to the loss function of the CNN to estimate the chrominance. The chrominance network learns the similarity of spectral distributions to the ground truth and estimates chrominance components of hyper-spectral images.

$$\mathrm{Cos} = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{S_c(i)\cdot S(i)}{\|S_c(i)\|_2\|S(i)\|_2}\right), \tag{5}$$

where n is the total number of pixels in a image, $S_c(i)$ is reconstructed spectral values in i-th pixel, $S(i)$ is the ground truth, $\cdot$ is the inner product of vectors, and $||\quad||_2$ is the Euclidean norm.

Spectral angle mapper (SAM) is also used to measure the similarity of spectral values, which is shown as follows:

$$\mathrm{SAM} = \frac{1}{n}\sum_{i=1}^{n}\cos^{-1}\left(\frac{S_c(i)\cdot S(i)}{\|S_c(i)\|_2\|S(i)\|_2}\right). \tag{6}$$

SAM represents the angle between two vectors; the smaller the spectral angle, the higher the similarity. In our experiments, we examine Cos and SAM as a loss function to learn chrominance components.

The input to the chrominance network is an RGB image, and the output is an N-bands hyper-spectral image without luminance information.

3.3 Luminance conversion

The final hyper-spectral image $\widehat{S}$ is reconstructed using the estimated (X_l, Y_l, Z_l) values. Pixel values of the normalized hyper-spectral image S_c is scaled by the following linear transform:

$$\widehat{S} = \frac{L_l}{L_c + \delta}S_c, \tag{7}$$

where the spectral magnitude L_l is calculated from the estimated (X_l, Y_l, Z_l) values, spectral magnitude L_c is calculated from the (X_c, Y_c, Z_c) values of the estimated chrominance components, and δ is a small value to avoid the singularity where L_c is zero.

Generally, Y in the CIE XYZ color space is assumed to be the luminance. However, the magnitude of spectral intensities is not only measured in Y. In our experiment, we also use X and Z in addition to Y. This allows $\widehat{S}$ to accurately represent the features of ground truth, even if the spectral distribution of S_c deviates significantly from the wavelength range covered in Y. L_l and L_c denote the following equations.

- Sum of tristimulus values:

$$L_l = X_l + Y_l + Z_l, \quad L_c = X_c + Y_c + Z_c. \tag{8}$$

  The spectral magnitude of a hyper-spectral image is converted considering the entire wavelength range, by adopting the sum of the tristimulus values for the spectral magnitude L.

- Maximum of tristimulus values:

$$L_l = \max(X_l, Y_l, Z_l), \quad L_c = \max(X_c, Y_c, Z_c). \tag{9}$$

  The tristimulus values are compared for each pixel, and the maximum value is used as the spectral magnitude L. The maximum value is selected, since it is representative of the most influential wavelength region. We select the same value for L_l and L_c from (X, Y, Z) values. For example, when $L_l = \max(X_l, Y_l, Z_l) = X_l$, then $L_c = X_c$ regardless of the maximum value of (X_c, Y_c, Z_c). The results may be different depending on which maximum value is chosen as the criterion. We examine the both criteria in our experiments.

4. Experimental Results

4.1 Datasets

We used two datasets, the CAVE dataset[14] and the ICVL dataset[8]. Both datasets provide hyper-spectral images as well as corresponding RGB images. To ensure accurate quantitative evaluation, input RGB images were synthetically re-generated in our experiment. We used the sRGB color space and approximated the nonlinear characteristics of the camera with gamma correction ($\gamma = 2.2$).

4.1.1 CAVE dataset

CAVE dataset is one of the most widely used hyper-spectral datasets. The dataset consists of 32 images with a size of 512×512. Each hyper-spectral image is sampled

in the wavelength range from 400 nm to 700 nm with a bandwidth of 10 nm and contains 31 bands. The images are divided into five sections, containing various objects such as faces, food, and paintings. Since the total number of images is small at 32, we use 2-fold cross-validation to evaluate the results.

4.1.2 ICVL dataset

The ICVL dataset consists of 201 images taken outdoors, with a size of 1300×1392. The dataset provides hyper-spectral images of 519 bands sampled from 400 nm to 1000 nm with a bandwidth of 1.25 nm, and 31 bands downsampled from 400 nm to 700 nm with a bandwidth of 10 nm. In our experiment, we employed hyper-spectral images downsampled to 31 bands. Note that the re-generated RGB images are different in color from the RGB images in the dataset, where many of the RGB images included in the dataset are underexposed or overexposed. Such RGB images are not suitable for the main purpose of this research. Therefore, we re-generated the RGB images according to the method described in the instruction attached to the dataset.

4.2 Evaluation metrics

We use root mean square error ($RMSE$) to evaluate the error in spectral intensity values, and also use *Similarity* and SAM to evaluate the similarity between the spectral distributions.

$RMSE$ is shown as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{S}(i) - S(i)\right)^2}. \tag{10}$$

Similarity is shown as follows:

$$Similarity = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{S}(i)\cdot S(i)}{\|\widehat{S}(i)\|_2\|S(i)\|_2}\right). \tag{11}$$

4.3 Quantitative evaluation

We also compare a conventional end-to-end method using MSE Loss.

4.3.1 Each component

Quantitative evaluations for each dataset is shown in **Tables 1**, **2**, **3**, and **4** for luminance and chrominance components. The tables show the mean and standard deviation (mean ± std).

Tables 3 and 4 show that the proposed method, which focuses on (X, Y, Z) values of hyper-spectral images, improves the accuracy of the spectral magnitude compared to the end-to-end method, which learns whole hyper-spectral images with MSE Loss for each dataset.

In Table 1, the accuracy of the chrominance components of our method is improved by using Cos or SAM as the loss function for each dataset. On the other hand, in the results of chrominance components on ICVL dataset shown in Table 2, the proposed method using Cos is worse than the end-to-end method in terms of SAM. This may be because the accuracy of the chrominance components is sufficiently high even using MSE Loss. The spectral distributions on ICVL dataset contain many similar distributions, and it is difficult to improve *Similarity* further by learning with Cos as the loss function. By learning chrominance components with SAM as a loss function, we can improve the accuracy of chrominance components on ICVL dataset compared to the end-to-end method.

4.3.2 Reconstructed hyper-spectral images

Quantitative evaluation of reconstructed hyper-spectral images of each dataset is shown in **Tables 5** and **6**.

Table 1 Quantitative evaluation of chrominance components on CAVE dataset

Method	Loss	*Similarity* ↑	SAM ↓
End-to-End	MSE	0.942 ± 0.043	16.089 ± 5.736
Propoesd	Cos	$\mathbf{0.981} \pm 0.011$	$\mathbf{9.065} \pm 2.915$
	SAM	0.980 ± 0.011	9.351 ± 2.880

Table 2 Quantitative evaluation of chrominance components on ICVL dataset

Method	Loss	*Similarity* ↑	SAM ↓
End-to-End	MSE	$\mathbf{0.999} \pm 0.001$	2.391 ± 0.900
Proposed	Cos	$\mathbf{0.999} \pm 0.001$	2.400 ± 0.779
	SAM	$\mathbf{0.999} \pm 0.001$	$\mathbf{2.216} \pm 0.835$

Table 3 Quantitative evaluation of luminance components on CAVE dataset

Method	Loss	$RMSE_{XYZ}$ ↓	$RMSE_X$ ↓	$RMSE_Y$ ↓	$RMSE_Z$ ↓
End-to-End	MSE(Spectral)	5.446 ± 3.375	4.980 ± 3.638	5.090 ± 3.899	5.901 ± 3.885
Proposed	MSE(XYZ)	$\mathbf{5.180} \pm 3.038$	$\mathbf{4.812} \pm 3.247$	$\mathbf{4.894} \pm 3.333$	$\mathbf{5.454} \pm 2.903$

Table 4 Quantitative evaluation of luminance components on ICVL dataset

Method	Loss	$RMSE_{XYZ}$ ↓	$RMSE_X$ ↓	$RMSE_Y$ ↓	$RMSE_Z$ ↓
End-to-End	MSE(Spectral)	0.399 ± 0.067	0.337 ± 0.052	0.399 ± 0.058	0.410 ± 0.119
Proposed	MSE(XYZ)	$\mathbf{0.278} \pm 0.041$	$\mathbf{0.230} \pm 0.049$	$\mathbf{0.275} \pm 0.029$	$\mathbf{0.320} \pm 0.057$

Table 5 Quantitative evaluation of reconstructed hyper-spectral images on CAVE dataset

Method	L. Loss	C. Loss	L. conversion	$RMSE_{Spectral} \downarrow$
End-to-End	MSE(Spectral)		-	8.168 ± 4.047
Proposed	MSE (XYZ)	Cos	Y	8.679 ± 3.756
			$X+Y+Z$	$\mathbf{8.085} \pm 3.566$
			$\max(X_l, Y_l, Z_l)$	8.524 ± 4.083
			$\max(X_c, Y_c, Z_c)$	8.288 ± 3.789
	MSE (XYZ)	SAM	Y	8.784 ± 3.667
			$X+Y+Z$	8.125 ± 3.441
			$\max(X_l, Y_l, Z_l)$	8.563 ± 3.924
			$\max(X_c, Y_c, Z_c)$	8.320 ± 3.671

Table 6 Quantitative evaluation of reconstructed hyper-spectral images on ICVL dataset

Method	L. Loss	C. Loss	L. conversion	$RMSE_{Spectral} \downarrow$
End-to-End	MSE(Spectral)		-	3.631 ± 1.629
Proposed	MSE (XYZ)	Cos	Y	3.868 ± 1.542
			$X+Y+Z$	3.856 ± 1.572
			$\max(X_l, Y_l, Z_l)$	3.860 ± 1.542
			$\max(X_c, Y_c, Z_c)$	3.860 ± 1.541
	MSE (XYZ)	SAM	Y	3.546 ± 1.570
			$X+Y+Z$	$\mathbf{3.527} \pm 1.582$
			$\max(X_l, Y_l, Z_l)$	3.535 ± 1.578
			$\max(X_c, Y_c, Z_c)$	3.534 ± 1.578

In the evaluation on CAVE dataset in Table 5, the proposed conversion method adopting $(X+Y+Z)$ as spectral magnitude achieved better results than the other conversion methods. This conversion method using all of the tristimulus values outperformed the end-to-end method (MSE(Spectral)). In addition, the conversion method adopting $\max(X_c, Y_c, Z_c)$ outperformed the conversion method adopting $\max(X_l, Y_l, Z_l)$. Therefore, it seems that the former is more accurate for calculating the maximum value of (X, Y, Z) values.

The evaluation on ICVL dataset in Table 6 shows similar results as the evaluation on CAVE dataset, although the best loss function is different. The proposed method learning chrominance components with SAM as a loss function achieved better results for reconstructed hyper-spectral images on ICVL dataset.

4.4 Qualitative evaluation

The reconstructed spectral intensities of each method on CAVE dataset are shown in **Fig. 4**. In the upper figure, spectral intensity of the proposed method is successfully reconstructed to the same spectral magnitude as the ground truth (GT). In the lower figure, the proposed method achieved better results compared to the end-to-end method (MSE(Spectral)) in the wavelength range of 400 to 600 [nm].

The reconstructed spectral intensities of "CD" images are shown in **Fig. 5.** It is difficult to reconstruct the hyper-spectral image because the image contains bright line spectra with high peak values. The upper and lower figures show that the proposed method properly recognizes the peak wavelength, although the spectral intensities are lower than the ground truth.

The reconstructed spectral intensities of each method on ICVL dataset are shown in **Fig. 6**. Because ICVL dataset consists of images taken outdoors, many images contain the sky. The error in the reconstructed spectral intensities of the sky is small as shown the lower figure. On the contrary, the reconstructed spectral intensities have some difference from the ground truth in the area containing fine patterns as shown in the upper figure. In both figures, there is no significant difference between the proposed method and the end-to-end method.

The common feature in Fig. 4, 5 and 6 is that the reconstructed spectral intensities have larger errors on the long wavelength side. One of the reasons is that the proposed method does not consider the response characteristics of the color matching functions. The loss function used in our experiments uniformly deals with the wavelength range of the spectral intensities, and the network is optimized by averaging the error at each wavelength. It is difficult to estimate the spectral intensities on the long wavelength side accurately because the color matching functions has low response characteristics on the long wavelength side. This is a common problem with the spectral super-resolution and is one of the future work.

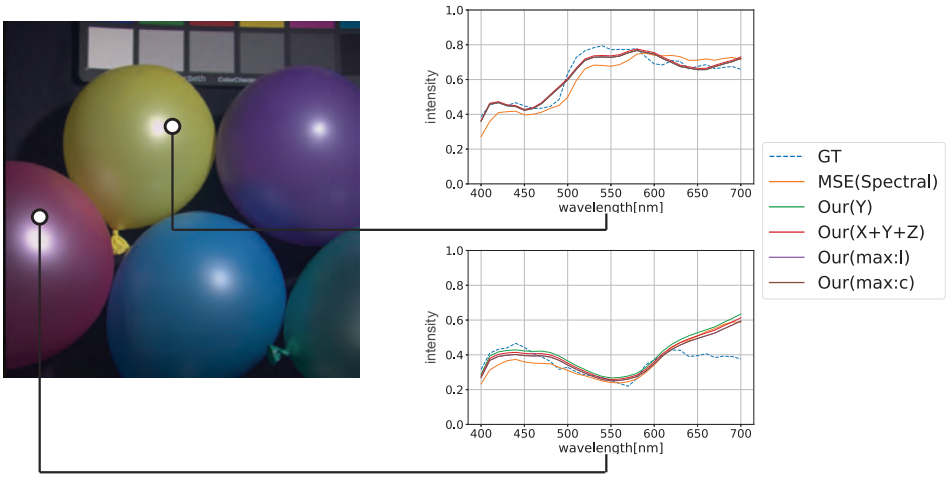


Fig. 4 Qualitative comparison of each method on "Balloon" images from CAVE dataset where "Our" is the proposed method that performed luminance conversion using Cos for chrominance components

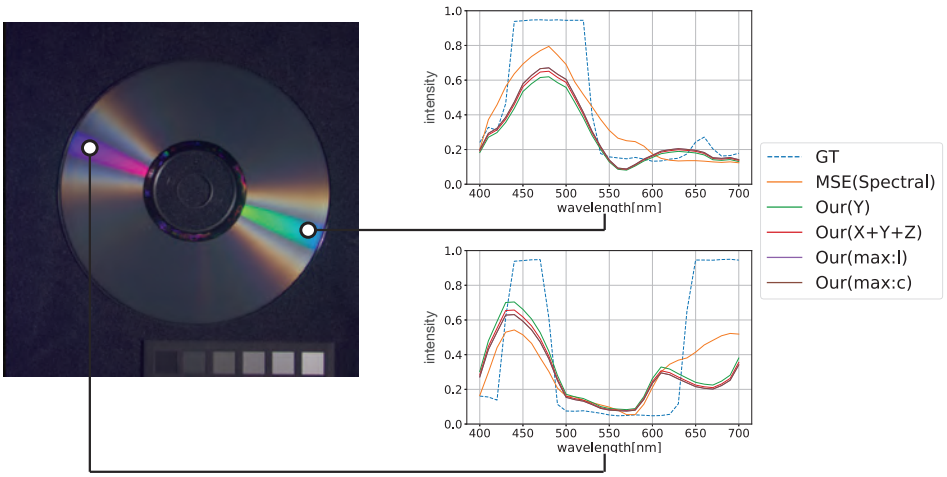


Fig. 5 Qualitative comparison of each method on "CD" images from CAVE dataset where "Our" is the proposed method that performed luminance conversion using Cos for chrominance components

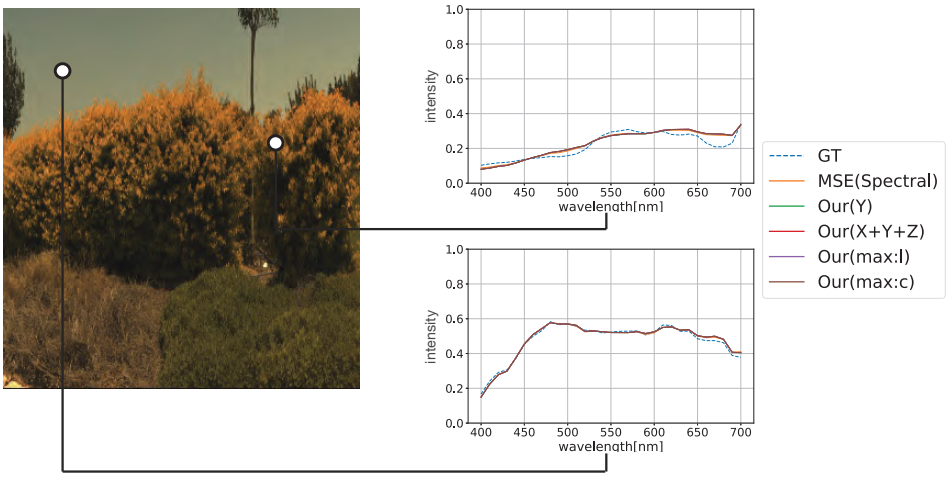


Fig. 6 Qualitative comparison of each method on ICVL dataset where "Our" is the proposed method that performed luminance conversion using SAM for chrominance components

5. Conclusion

We proposed a novel spectral super-resolution using a deep learning to estimate luminance and chrominance components separately. In the proposed method, the hyper-spectral image is finally reconstructed from the estimated chrominance component through a linear transformation using the estimated luminance component. Us-

ing all of three tristimulus values, the proposed method converts the spectral magnitude considering the entire wavelength range. Compared with the conventional end-to-end method that learns spectral intensities directly, the proposed method reduces the error of each component and achieves superior results for reconstructing hyperspectral images.

Future work includes the selection of appropriate networks for learning each component and the extension of our method to HDR images to reproduce a spectral image with higher dynamic range than LDR. Another future work includes experiment with a larger number of images to verify the further usefulness of the proposed method. The challenge of solving the metamerism problem is also one of the important issues of spectral super-resulution.

References

1) M. Moroni, A. Mei, A. Leonardi, E. Lupo, F. L. Marca: "PET and PVC Separation with Hyperspectral Imagery", Sensors, Vol. 15, No. 1, pp. 2205–2227 (2015).
2) E. Belluco, M. Camuffo, S. Ferrani, L. Modenese, S. Silvestri, A. Marani: "Mapping Salt-Marsh Vegetation by Multispectral and Hyperspectral Remote Sensing", Remote Sensing of Environment, Vol. 105, No. 1, pp. 54–67 (2006).
3) O. Yilmaz, O. Selimoglu, F. Turk, M. S. Kirik: "SNR Analysis of a Spaceborne Hyperspectral Imager", Proc. of the 2013 6th International Conference on Recent Advances in Space Technologies (RAST), pp. 601-606 (2013).
4) H. Otsu, M. Yamamoto, T. Hachisuka: "Reproducing Spectral Reflectances from Tristimulus Colours", Computer Graphics Forum, Vol. 37, No. 6, pp. 370–381 (2018).
5) W. Jakob, J. Hanika: "A Low Dimensional Function Space for Efficient Spectral Upsampling", Computer Graphics Forum, Vol. 38, No. 2, pp. 147–155 (2019).
6) R. Kawakami, J. Wright, Y. -W. Tai, M. B. Ezra, K. Ikeuchi, Y. Matsushita: "High Resolution Hyperspectral Imaging via Matrix Factorization", Proc. of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2329–2336 (2011).
7) S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, K. Murphy: "PixColor: Pixel Recursive Colorization", Proc. of the British Machine Vision Conference (BMVC), pp. 112.1-112.13 (2017).
8) B. Arad, O. B. Shahar: "Sparse Recovery of Hyperspectral Signal from Natural RGB Images", Proc. of the European Conference on Computer Vision (ECCV), pp. 19–34 (2016).
9) M. H. Nguyen, D. K. Prasad, M. S. Brown: "Training-Based Spectral Reconstruction from a Single RGB Image", Proc. of the European Conference on Computer Vision (ECCV), pp. 186–201 (2014).
10) S. Jégou, M. Drozdzal, D. Vázquez, A. Romero, Y. Bengio: "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation", Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1175-1183 (2017).
11) S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, K. Schindler: "Learned Spectral Super-Resolution", arXiv preprint arXiv:1703.09470, (2017).
12) J. He, J. Li, Q. Yuan, H. Shen, L. Zhang: "Spectral Response Function Guided Deep Optimization-Driven Network for Spectral Super-Resolution", IEEE Trans. on Neural Networks and Learning Systems, pp. 1-15 (2021).
13) J. Li, C. Wu, R. Song, Y. Li, F. Liu: "Adaptive Weighted Attention Network with Camera Spectral Sensitivity Prior for Spectral Reconstruction from RGB Images", Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1894-1903 (2020).
14) F. Yasuma, T. Mitsunaga, D. Iso and S. K. Nayar: "Generalized Assorted Pixel Camera: Postcapture Control of Resolution, Dynamic Range, and Spectrum", IEEE Trans. on Image Processing, Vol. 19, No. 9, pp. 2241–2253 (2010).

(Received June 4, 2021)
(Revised January 17, 2022)

Masahiro SAKAMOTO
He received his B.E. and M.E in information engineering from Hiroshima University, Japan, in 2020 and 2022. His research interests include spectral super-resolution.

Kazufumi KANEDA (*Member*)
He received his D.E. in system engineering from Hiroshima University, Japan, in 1991. Currently, he is a professor at Graduate School of Advanced Science and Engineering, Hiroshima University, Japan. His research interests include computer graphics, data visualization and medical graphics.

Bisser RAYTCHEV
He received his Ph.D. in Informatics from Tsukuba University, Japan, in 2000. After being a research associate at NTT Communication Science Labs and AIST, he is presently an associate professor in the Department of Information Engineering, Hiroshima University, Japan. His current research interests include machine learning, computer vision, natural language processing and brain-inspired computing.

IIEEJ Paper -- Special Issue on Journal Track Papers in IEVC2021Part II --

Evaluating YOLOv3 for Identification and Classification of Functional and Sclerosed Glomeruli

Thalita Munique COSTA† (*Student Member),* Lourenço BARBOSA ††, Yoko USAMI ‡, Mai IWAYA ‡, Kiyoshi TANAKA ‡ ‡ (*Fellow*), Fabio SCHNEIDER††,

†Shinshu University Graduate School of Medicine, Science and Technology, ††Federal University of Technology-Parana/CPGEI, ‡Department of Laboratory Medicine Shinshu University Hospital, ‡ ‡Academic Assembly (Institute of Engineering), Shinshu University

<Summary> In Digital Pathology, histological slides digitized allow the use of techniques for automatic procedures in histopathology, permitting the automatic quantification of the rate of sclerosed glomeruli to order the biopsy slides so that the most serious cases can be identified more quickly. In this work, we evaluate the YOLOv3 as deep neural network to identify glomeruli in WSI and classify in functional and sclerosed glomeruli. This work used the framework YOLOv3, with 53-layers convolutional neural network, and 30 complete slides from the Bio-Atlas repository (Pennsylvania State University), which resulted in 2448 images of 1024x1024 pixels with one or more glomeruli, used for training and performance evaluation. A total of 585 sclerosed glomeruli and 3383 functional glomeruli were labeled. Through the experiments, we achieve high performance in identification and classification of glomeruli (e.g., recall of 96.8%, precision of 95.9%, accuracy of 98.1%, and an F1 score of 96.3%). The method is capable to identify and report the location of the glomeruli on the slide, classify the glomeruli in functional and sclerosed, and precisely provide the percentage of sclerosed glomeruli, allowing support for a histopathological study of kidney diseases in the medical routine.

Keywords: Digital Pathology, Glomerular Sclerosis, Glomerular Hyalinization, YOLOv3, Deep Learning, Whole Slide Image (WSI).

1. Introduction

The functional unit of the kidneys responsible for the formation of urine is called Nephron. Initially, a healthy kidney has between 800,000 and 1,000,000 nephrons and has no regenerative capacity[1]. This number gradually reduces with aging and the reduction may be associated with several diseases such as diabetes, systemic lupus erythematosus, ingestion of drugs, and infections[2]. Studies have scored and analyzed recurrent cases of acute kidney injury caused by Coronavirus infection [3)-5)]. It is estimated that 6.7% of patients with Coronavirus acquire acute kidney injury and among these 91.7% cannot survive[5]. Each Nephron contains a glomerulus, which is formed by a tangle of capillaries, called glomerular capillary tuft, wrapped by Bowman's capsule[1]. They filter a large amount of liquid and other elements from the blood. The filtered liquid passes through a sequence of tubular structures, where selective resorption of water and ions occurs, and the residual elements, including excess water, are eliminated in the form of urine[1].

Each component of renal tissue (e.g., glomeruli, tubules, interstices and blood vessels) is usually affected differently for each disease[2]. The glomeruli may be affected resulting in a terminal lesion called glomerular sclerosis or hyalinization[6),2)], when plasma proteins are deposited in large quantities inside the glomeruli and the accumulation of material hyaline occurs, causing permanent loss of its function[6]. These alterations cause a chain reaction. Healthy glomeruli suffer hypertrophy to compensate for the loss of other nephrons seeking to maintain renal function through changes in structure, resulting in its sclerosis[2]. Renal disease is in progress to its final stage when the destruction of nephrons reduces the glomerular filtration rate to less than 50% of normal[2].

For the analysis of renal tissue, a renal biopsy is performed, an exam indicated when there is a renal failure of rapid progression, nephrotic syndrome, systemic diseases, or unexplained renal transplantation dysfunction[6]. Stained histological sections are evaluated microscopically to identify the nature of the disease[6]. One of the important items is the evaluation of the degree of glomerular healing (i.e., sclerosis). In these situations, it is sought to evaluate the percentage of affected glomeruli, whether the disease is focal (affects less than half of the glomeruli) or diffuse (affects more than half of the glomeruli)[6]. Typical images of functional and sclerosed glomeruli on a renal biopsy slide are shown in **Fig. 1**.

In the medical routine, there may be many slides to be analyzed at the same time. It becomes interesting to use a tool that identifies slides that, for not having a minimum number of glomeruli, should be discarded, while orders the slides by those that appear to have a more serious diagnosis, which can also be

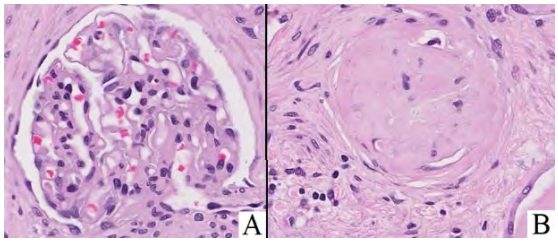


Fig. 1 Glomeruli found in renal biopsies. A: Functional glomerulus; B: Sclerosed glomerulus[1)]

applied to identify kidneys that are eligible for transplantation[7)]. The glomerular sclerosis rate can be a good parameter to make this classification. Fully digitized slide images (i.e., WSIs) are being widely used in digital pathology. WSIs make the histopathological evaluation possible in a virtual environment and allows the biopsy result to be shared with other specialists. Several laboratories use them for diagnostic routine. With the use of stains (e.g., hematoxylin-eosin, Schiff's periodic acid), performed to highlight and identify specific histological elements, in different colors and shades, there is the possibility of using the WSI for the automation of the analysis of histopathological alterations[8)].

The human vision and processing capacity in a trained and rested subject are accurate and capable of identifying an object instantly[9)]. On the other hand, automation allows benefits such as the reduction of the effects of human subjectivity in diagnosis and the ability to identify minimal variations in histological findings that would not be perceived by humans[8)]. In this context, neural networks emerged to try to model the human thought process using logical rules[10)]. Convolutional Neural Networks (i.e. CNN), using, for example, Deep Learning algorithms can help in classifications that are traditionally done manually[11)].

Deep learning algorithms use a large amount of training data in conjunction with multi-layered neural network architectures[12)]. In many histopathological analyses, one of the important threads is the classification of the evaluated components. In these situations, deep learning can present interesting results in the sense of adequate classification[13)]. In this work, it is proposed to evaluate the performance of the deep learning architecture YOLOv3 in the identification and classification of glomeruli into functional and sclerosed, to obtain the percentage of sclerosed glomeruli. YOLOv3 architecture was selected for this work because it presents many of the desired steps (training of convolutional neural network layers, prediction of bounding boxes, calculation of box probability and class probability) in a single architecture while presenting a superior performance for several applications when compared to some other architectures[9)].

2. Proposed Method

2.1 Entire flow

In this work we proposed to use a neural network architecture capable of identifying and classifying the items at the same time, using a prepared train set (i.e., teacher database) for supervised training of the neural network with images of 1024×1024 pixels of 26 varied WSIs. Also, we use early stopping to avoid overfitting, set a threshold to increase the accuracy of the system, reconstruct the WSI with the labels, treat the data to eliminate the counting of duplicate labels and bring the total number of glomeruli and the percentage of sclerosed glomeruli. We show the entire workflow of the proposed method in **Fig.2**

The first step consists of preparing the train set, selecting images, from the Bio-Atlas repository[14)], which have a good number of glomeruli, different shades and stains, and various diagnoses. Then the smaller images (i.e., 256×256 pixels) are concatenated creating bigger images with the size chosen (i.e., 1024×1024 pixels). Some of these images are separated in another file (i.e., test set) for performance evaluation. After that, text files were created with the coordinates and classes of the glomeruli using the YBAT software[15)]. The files were created with a supervision of a pathologist.

The next step is configuring the neural network by setting the hyperparameters of the neural network used according to the specifications of the computer used, the data augmentation characteristics that you want to apply and resizing done in the first layer of the neural network (i.e., 512×512 pixels, a resize that maintains the shape of the image with a simple scaling factor). Next, the training can start and should be stopped as soon as the accuracy value is stabilized to avoid overfitting.

After training, the threshold value for class confidence must be configured. Only items found with a confidence value greater than the threshold will be considered. With the trained neural network, it is possible to deliver slide images cut into pieces of 1024 ×1024 pixels, never seen before by the neural network, to be labeled (i.e., images separated for performance evaluation in the test set). The labeled images of 1024×1024 pixels are reconstituted in the WSI while the coordinates of each glomerulus found are calculated, duplicate tags are eliminated from the count, and a text file is generated with the information of the number of glomeruli found, number of sclerosed glomeruli, location of each glomerulus and percentage of sclerosed glomeruli in the slide.

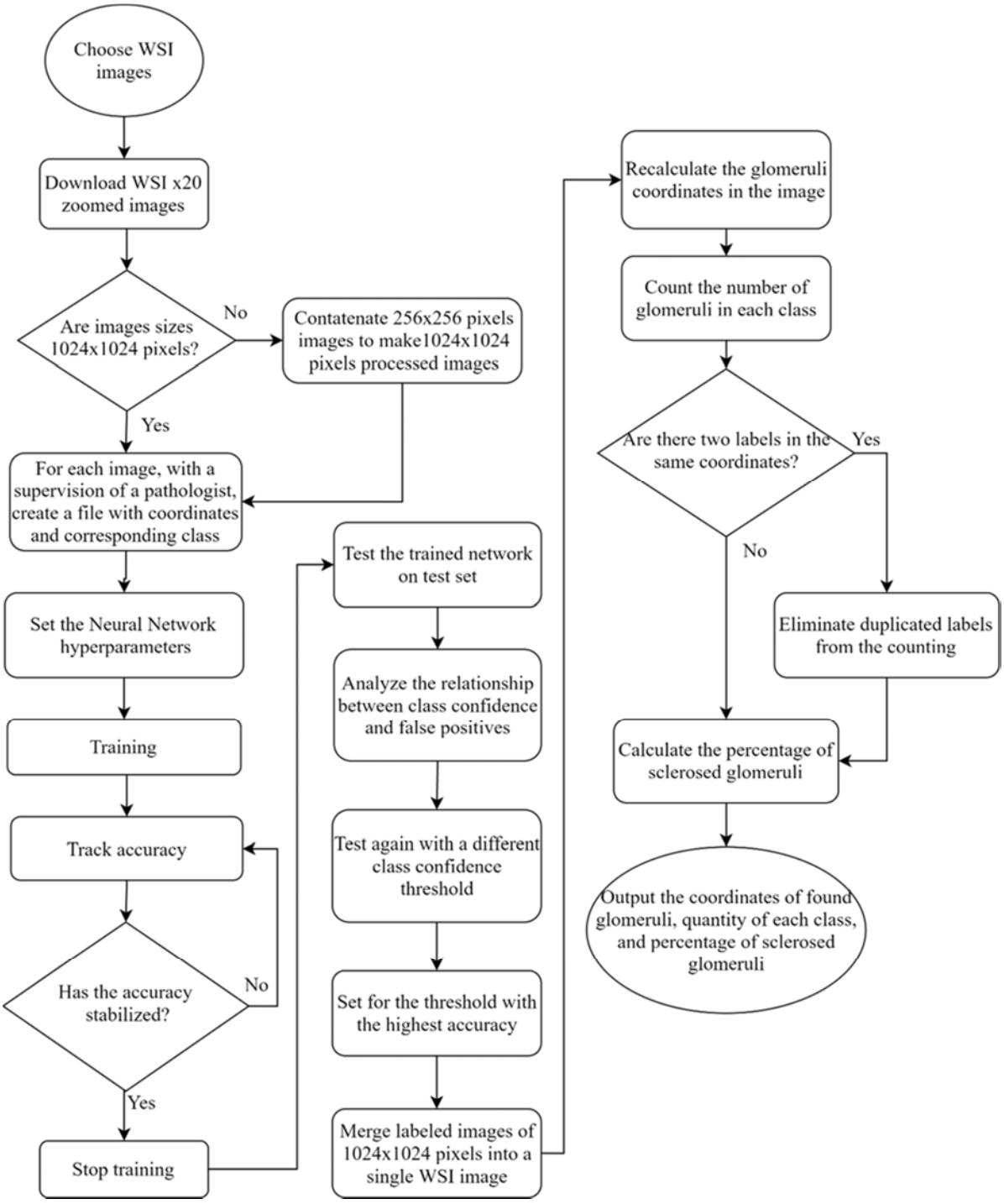


Fig. 2 Entire workflow of the proposed method

2.2 Deep neural network

The deep network architecture used in this work was YOLOv3[16], which is an evolution of YOLO (i.e. You Only Look Once). YOLO is a framework that brings a fast solution for the identification of objects in images, defining bounding boxes and class probabilities. YOLOv3 architecture has 53 layers (i.e., Darknet-53) as shown in **Fig. 3** and **Table 1**. The network has convolutional layers with varied numbers of 3×3 filters, pooling for image resizing, residual layers, a layer entirely connected at the output and a SoftMax layer, and an avgpool layer that despite being part of the network are not used in this version of YOLO[16]. The open-source YOLOv3 architecture can detect and classify the object found among more than one class and considers in the decisions the context in which the object is involved[16].

2.3 Database, training, and validation

The repository used (i.e., Bio-Atlas[14]) in the process to build the database (i.e., training set and test set) is composed of pieces of 256×256 pixels of WSI. This image size is not appropriate to the training once that does not show an entire glomerulus and its context in the image. To attend to these requirements while still maintaining the shape and a simple scaling factor in the first layer of the neural network, the image size that was chosen to be used is 1024×1024 pixels. A script was developed to download zoomed ×20 images, of size 256×256 pixels, from 26 human kidney WSIs of the Bio-Atlas repository[14], and concatenate the images building 2,759 images of 1024×1024 pixels. Also, for each image, a text file was built with the class and coordinates of the glomeruli presented on them.

The kidneys on the images are healthy and diagnosed with septic infarct, microscopic polyarteritis, malignant hypertension, tuberous sclerosis, chronic obstructive pyelonephritis, amyloidosis, interstitial nephritis, acute and chronic pyelonephritis, atheroembolism, chronic transplant rejection, myeloma nephropathy, focal segmental glomerulosclerosis, renal infarct, renal cell carcinoma, acute tubular necrosis from ethylene glycol, nephrocalcinosis, eclampsia, arteriolosclerosis, acute tubular necrosis, systemic lupus erythematosus and, polycystic kidney disease. The WSIs in the Bio-Atlas[14] repository are from laboratories from University of Kansas, University of New South Wales, University of Iowa, Medical College of Ohio Toledo, Universitas Basiliensis, Switzerland and from AACR Workshop. 26 WSIs were selected to compose the training set and 4 slides to compose the test set. The training set has 25 H&E-stained WSIs and one PAS-stained WSI. All WSIs in the test set are H&E-stained.

In the WSIs, 585 sclerosed and 3,383 functional glomeruli were located and classified. We used data

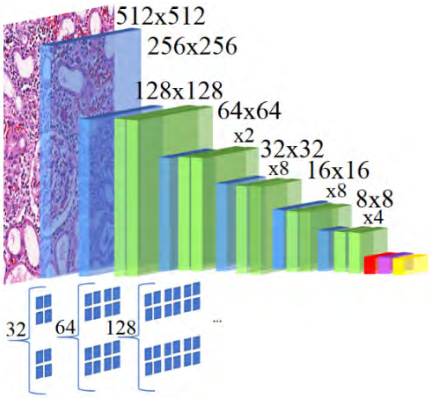


Fig. 3 Representation of YOLOv3 network architecture.

Table 1 Layers of neural network Darknet-53

Layer	No. of Layers	Type	Filters	Size of filters	Out
1º	1	Convolutional	32	3 x 3	256x256
2º	1	Convolutional	64	3 x 3 / 2	128x128
3º- 4º	1	Convolutional	32	1 x 1	
		Convolutional	64	3 x 3	
		Residual			128x128
5°	1	Convolutional	128	3 x 3 / 2	64x64
6º- 9º	2	Convolutional	64	1 x 1	
		Convolutional	128	3 x 3	
		Residual			64x64
10º	1	Convolutional	256	3 x 3 / 2	32x32
11º- 26º	8	Convolutional	128	1 x 1	
		Convolutional	256	3 x 3	
		Residual			32x32
27º	1	Convolutional	512	3 x 3 / 2	16x16
28º- 43º	8	Convolutional	256	1 x 1	
		Convolutional	512	3 x 3	
		Residual			16x16
44º	1	Convolutional	1024	3 x 3 / 2	8x8
52º	4	Convolutional	512	1 x 1	
		Convolutional	1024	3 x 3	
		Residual			8x8
		Avgpool		Global	
53º		Connected		1000	
		Softmax			

augmentation techniques, with rotation and HSV color map variations, generating 803,200 images in the set.

For the localization and classification of the glomeruli, the YBAT[15] software was used. In this application it's possible to build a file with the classification and coordinates of the objects in each image in the structure necessary for network training. In this procedure, besides the class of the object, two coordinates are recorded that indicate the displacement from the left superior corner of the image (i.e., x and y) as well as the width and height (i.e., w and h) of the

Table 2 Network architecture parameters configured in the experiments performed.

Parameter	Value
Percentage for training	80%
Percentage for error analysis	20%
Batch	16
Subdivision	8
Width	512
Height	512
Channel	3
Momentum	0.9
Decay	5
Angle	0
Saturation	1.5
Exposure	1.5
Hue	0.1

rectangle containing the classified object.

For the training, the images were randomly separated into two groups. The first, containing about 80% of the images, was intended for training, and the second, containing about 20% of the images, was used by the neural network itself to calculate the loss average during training. **Table 2** shows the hyperparameters used for the experiments. For other parameters, the initial settings of the YOLOv3 architecture were maintained. Batch and sub-division influence the speed of training, which tends to be higher with larger batch processing. The parameters had to be adjusted according to the processing capacity of the hardware used. In this work, a 9th generation Intel Core I7 processor, 16GB RAM and Nvidia GeForce RTX2060 6GB GDDR6 VRAM card were used. The values entered in width and height are used to resize the images in the input level of the neural network. Images were resized to 512×512 pixels. The channel used was 3, because the input items are RGB images. The parameters angle, saturation, exposure, and hue define how additional images will be generated for training.

3. Performance Validation

3.1 Experimental setup

For performance validation of the proposed method, 4 WSI containing sclerosed and functional glomeruli (i.e., 153 sclerosed and 540 functional glomeruli) were used as a test set, additionally to those 26 used for training. The 4 chosen slides have different shades as shown in **Fig. 4**, and different proportions of functional and sclerosed glomeruli. The slides in the test set were chosen to analyze the response to different diseases and different shades, once one of the main challenges in the computational interpretation of WSIs is the color variations due to the slide preparation, and scanning. The algorithms used in digital pathology have to be

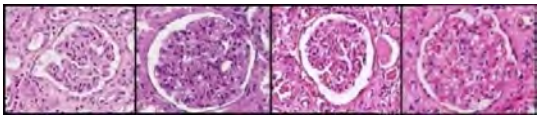


Fig. 4 Fragments of the 4 slides used in performance validation[1)]

resilient to these variations[12)]. The WSIs in the test set were detected with renal cell carcinoma, renal infarct, systemic sclerosis and disseminated intravascular coagulation, microthrombi microinfarcts. In the medical routine, is expected that the number of functional glomeruli found will be much higher than the number of sclerosed since the glomeruli may be compromised by several other diseases that are not in the hyalinization stage and even a kidney that is undergoing a hyalinization process will not have all its glomeruli compromised at the biopsy time. One of the validation slides had a region with failed fixation and unfocused. The images that presented this flaw were disregarded.

3.2 Evaluation metrics

For the evaluation of results, was considered the number of *True Positives* TP (i.e., glomeruli found and correctly labeled), *False Positives* FP (i.e., areas with no glomerulus labeled as a glomerulus), *False Negatives* FN (i.e., areas with glomerulus not labeled) and *True Negatives* TN (i.e., regions with no glomerulus not labeled as a glomerulus). The region considered for the True Negatives is each region of 512×512 pixels not labeled in the image that does not have a glomerulus. A glomerulus can be found in a wide range of sizes, but for x20 zoomed images, the size 512×512 pixels comprises most glomeruli. The true positive is the glomerulus in the image correctly labeled as a glomerulus. For these results, the images were verified manually with a supervision of a pathologist. The metrics evaluate how well the method can report the quantities and the percentage of the functional and sclerosed glomeruli. For this purpose, the bounding box of the detected glomeruli doesn't need to precisely have its size. The bounding box can have different sizes as long it identifies the glomerulus correctly and is a rectangle including just one correct glomerulus. A mark should be penalized (considered as a marking error) if the bounding box is much larger than the glomerulus (i.e., more than 50% bigger than the glomerulus and which the area is bigger than 262,144 pixels, which is the area of 512×512 pixels square). Because the WSI was cut into smaller pieces to be analyzed, when the WSI was reconstructed, there are glomeruli with more than one bounding box. The data passed to the post-processing analysis algorithm where all duplicate marks were eliminated from the count before the report was constructed.

We used as notations *Recall* R (i.e. sensitivity) in Eq.(1), *Precision* P in Eq.(2), *Specificity* S in Eq.(3),

True Negative Rate RTN in Eq.(4), *Accuracy* A in Eq.(5), *Jaccard coefficient* J in Eq.(6), *False Positive Rate* RFP in Eq.(7), *False Negative Rate* RFN in Eq.(8), *F1-Score* F1 in Eq.(9) and *Matthews Correlation Coefficient* M in Eq.(10)[17),18)]. Among these metrics, accuracy stands out as it is commonly discussed and analyzed in clinical practice. To evaluate our method in comparison with other similar works, we also used the same metrics used in recent similar research (i.e., Precision, Recall, F1-Score and Accuracy).

$$R = \frac{TP}{(TP+FN)} \quad (1)$$

$$P = \frac{TP}{(TP+FP)} \quad (2)$$

$$S = \frac{TN}{(TN+FP)} \quad (3)$$

$$RTN = \frac{TN}{(TN+FN)} \quad (4)$$

$$A = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (5)$$

$$J = \frac{TP}{(TP+FP+FN)} \quad (6)$$

$$RFP = \frac{FP}{(FP+TN)} \quad (7)$$

$$RFN = \frac{FN}{(FN+TN)} \quad (8)$$

$$F1 = \frac{2*P*R}{(P+R)} \quad (9)$$

$$M = \frac{TN*FP}{\sqrt{(TN+FN)(TP+FP)(FP+TN)(TP+FN)}} \quad (10)$$

3.3 Threshold analysis

When an image is run in the trained YOLOv3 architecture, the result presents the probability of a given object labeled of belonging to the identified class. It was observed that the number of false positives could be reduced by properly selecting the minimum probability that should be considered for assuming that a glomerulus was found. The performance evaluation test with the threshold of 0.20 also presented 13 marks with class error (i.e., object was correctly found but wrong classified), where 12 were items labeled twice, with the correct and with the wrong classification. It was observed that this problem could also be avoided with the correct threshold.

To analyze how the class probabilities can be used to improve the metrics, 490 marks (i.e., about 70% of the marks) were classified into FP and TP and their probabilities were plotted in **Fig. 5** (spread vertically), so it would be possible to understand how a change in the threshold value can change the number of TPs and FPs. In this graph, we chose 4 thresholds (0.20, 0.32, 0.46 and 0.53) that exclude a higher number of false positives compared to true positives represented with a vertical line. The test set were labeled again by the neural network, considering just images with confidence higher than each threshold.

3.4 Results and discussion

The results of the threshold analysis for the experiments are shown in **Table 3** and **Fig. 6**.

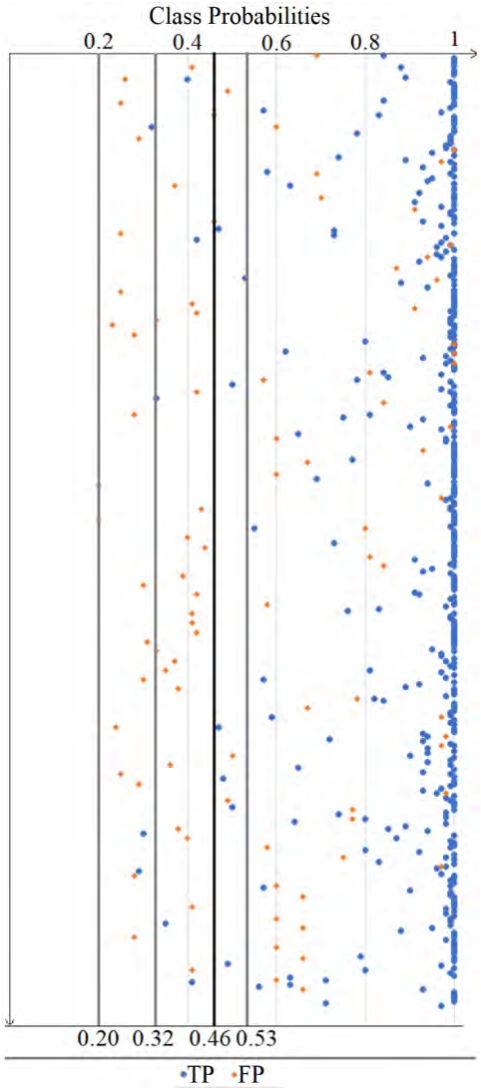


Fig. 5 Graph of the class probabilities of the markings made by the system

Table 3 Results of threshold experiments.

Threshold	Recall	Precision	Accuracy	F1-Score
0.20	**0.9730**	0.8739	0.9578	0.9202
0.32	0.9711	0.9161	0.9732	0.9427
0.46	0.9682	**0.9588**	**0.9810**	**0.9634**
0.53	0.9655	0.9324	0.9721	0.9535

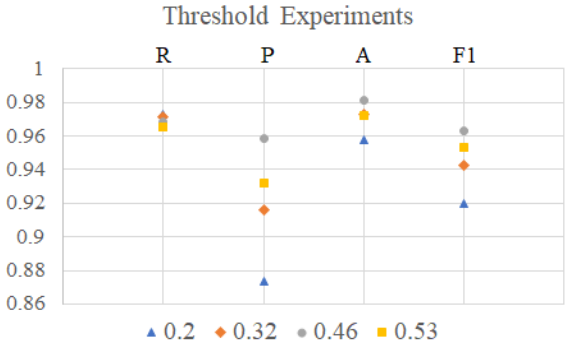


Fig. 6 Graph of Threshold experiments. R: Recall; P: Precision; A: Accuracy; F1: F1-Score

The experiment with the threshold of 0.46 was chosen to be better discussed once it presented the best results for precision, accuracy, and F1-Score. As expected, comparing this and the experiment with the standard value (i.e., 0.20), there was a decrease in recall and an increase in other parameters. While the Recall decreased 0.0048 percentual points, Precision increased 0.0849, Accuracy 0.0232 and F1-Score 0.0432 percentual points. This means that the evaluated method has lost some sensitivity in finding the right items but has become significantly less likely to identify wrong objects. The class error marks were reduced to 1 with no item labeled twice.

The results of the experiment separated by WSI are present in **Table 4**. The slides were analyzed separately so it becomes easier to identify the characteristics that may interfere with the results. Slide 3 for example, doesn't have sclerosed glomeruli and has the best results for recall and accuracy, so these parameters can be expected to be higher in samples with a smaller number of sclerosed glomeruli.

As a reference, to verify how the PAS-stained WSI present in the training set influences the results, experiments were performed by removing this WSI. **Table 5** shows the results for the same thresholds used in the previous experiments. The results were a bit deteriorated for some metrics by comparing with **Table 3** because the number of images used for the training was considerably reduced by removing the PAS-stained WSI. Therefore, to improve the results, we should increase the training data stained by PAS, and this issue is one of our future works.

3.5 Analysis of the results

Our method identifies the location of sclerosed and functional glomeruli and returns the quantities and positions of the items found. The method returns an image of a WSI with the found glomeruli marked by a frame (i.e., rectangle box) and with their class confidence value, as shown in **Fig. 7** and **Fig. 8**, and reports the percentage of sclerosed glomeruli, the total of glomeruli found, the number of these glomeruli that

Table 4 Results obtained in Experiment considering both classes using 0.46 as cut-off threshold.

Metric	S1	S2	S3	S4	**Total**
R	0.9695	0.9735	1	0.9300	0.9682
P	0.9408	0.9607	0.9848	0.9489	0.9588
S	0.9820	0.9921	1	0.9960	0.9925
R_{TN}	0.9950	0.9960	1	0.9969	0.9970
A	0.9670	0.9773	0.9972	0.9826	0.9810
J	0.9138	0.9363	0.9849	0.8857	0.9301
R_{FP}	0.0100	0.0060	0.0009	0.0022	0.0048
R_{FN}	0.0300	0.0260	0	0.0700	0.0314
F_1	0.9550	0.9671	0.9924	0.9390	0.9634
M	0.9475	0.9622	0.9920	0.9368	0.9596

Table 5 Results of experiments with the training set of 25 H&E-stained WSI by removing the PAS-stained WSI.

Threshold	R	P	A	F1
0.20	**0.9258**	0.8949	**0.9779**	**0.9072**
0.32	0.8921	0.9056	0.9764	0.8985
0.46	0.8639	0.9201	0.9757	0.8910
0.53	0.8607	**0.9224**	0.9757	0.8903

are sclerosed and the coordinates in the WSI where the glomeruli are located, as shown in **Fig 9**, where the location of the items is composed by the coordinates of the position of each glomerulus found and the size of the box that is marking each one.

Unlike some other related works, this work not just presents the WSI reconstructed with the glomeruli and the class confidence values labeled, but also the information about the number of glomeruli found, the percentage of them that are sclerosed, and the position of each glomerulus is reported, facilitating the use in medical routine. The duplicate glomeruli are avoided in the count checking the proximity between markings. The results still present false positives (i.e., objects B and D in **Fig. 10**), which generally occur because of the similarity of other kidney units with the glomeruli, especially with sclerosed. Figure 10 also shows examples of correct labels and a false negative area.

High results were obtained, mainly in terms of accuracy, true negative rate, and sensitivity. The three metrics consider the value of true negative, so it is concluded that the method has a good ability to not identify as a glomerulus an area where there is no glomerulus. The metric that brought the lowest result was the Jaccard coefficient, a metric that does not consider the values of true negatives, and to increase this value, the number of false positives and false negatives need to be reduced. Still, the experiment obtained good results for all metrics, indicating the method's ability to find and classify items correctly and the ability to not make wrong marks.

In the test data, there were also images smaller than

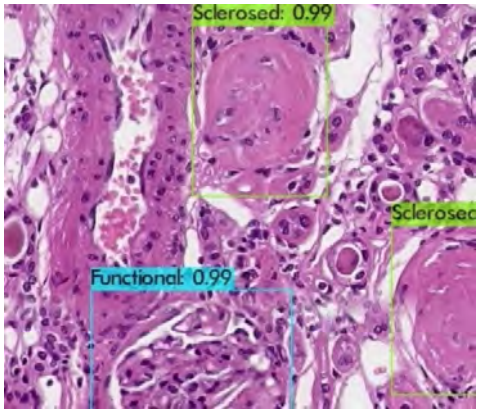


Fig. 7 Example of image labeled by the proposed method

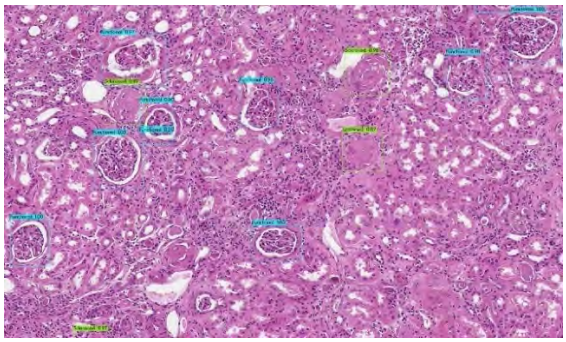


Fig. 8 WSI labeled by the evaluated method

```
File  Edit  Format  View  Help
Percentage of sclerosed glomeruli: 28.695652173913043%
Total of glomeruli: 115
Number of sclerosed glomeruli: 33
Location of items found:
[10664, 7093, '228', '137']
[21286, 5945, '283', '345']
[15452, 12010, '181', '314']
[20005, 11681, '377', '490']
```

Fig 9 Percentage of sclerosed glomeruli, total number of glomeruli, number of sclerosed glomeruli, the location of the items and the size of the box marked in the output file

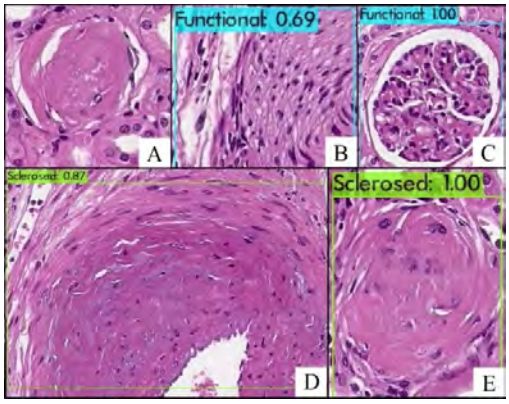


Fig. 10 A: False Negative of Sclerosed Glomerulus; B: False Positive of functional glomerulus; C: True Positive of Functional Glomerulus; D: False Positive of Sclerosed Glomerulus E: True Positive of Sclerosed Glomerulus

Table 6 Results obtained in experiment considering only functional glomeruli.

Metric	S1	S2	S3	S4	Total
R	0.9912	0.9916	1	0.9863	0.9923
P	0.9912	0.9915	0.9848	0.9863	0.9885
S	0.9981	0.9981	1	0.9987	0.9987
R_{TN}	0.9990	0.9990	1	0.9996	0.9994
A	0.9950	0.9951	0.9973	0.9970	0.9961
J	0.9826	0.9833	0.9848	0.9730	0.9809
R_{FP}	0.0009	0.0010	0.0009	0.0004	0.0008
R_{FN}	0.0088	0.0084	0	0.0137	0.0077
F_1	0.9912	0.9916	0.9924	0.9863	0.9904
M	0.9683	0.9708	0.9807	0.9358	0.9640

Table 7 Results obtained in experiment considering only sclerosed glomeruli.

Metric	S1	S2	S3	S4	Total
R	0.9200	0.9062	-	0.7778	0.8680
P	0.8363	0.8530	-	0.8400	0.8431
S	0.9856	0.9947	1	0.9975	0.9926
R_{TN}	0.9964	0.9973	1	0.9974	0.9978
A	0.9618	0.9751	-	0.9838	0.9736
J	0.7797	0.7838	-	0.6774	0.7470
R_{FP}	0.0081	0.0044	0	0.0017	0.0047
R_{FN}	0.0800	0.0937	-	0.2222	0.1320
F_1	0.8761	0.8788	-	0.8077	0.8542
M	0.8631	0.9288	-	0.7547	0.8489

those used in the training, corresponding to the edges of the slide. These images were analyzed separately and had an unsatisfactory result, showing the importance of maintaining the dimensions of the images used in the network.

3.6 Class analysis

Analyzing the results of each item separately, in **Table 6** for functional and **Table 7** for sclerosed glomeruli, the results were worse for the sclerosed, which can be explained by the smaller number of images of sclerosed glomeruli used in the training compared to functional, by the variety of the item searched and similarity with other elements shown in WSI. Blood vessels undergoing sclerosis are still commonly mistaken for sclerosed glomeruli, thus, as future work, traditional segmentation techniques can be used to avoid these misclassifications.

3.7 Comparison with other approaches

The methods used in works that have similar goals of using neural networks to identify and classify sclerosed and non-sclerosed glomeruli (i.e. Marsh et al.[7], Gallego et al.[19], and Bueno et al.[20]) are presented in

Table 8 Configurations of similar research[7),19),20)]

	Proposed	**Marsh et al.**[7)]	**Gallego et al.**[19)]	**Bueno et al.**[20)]
No. of WSI	26	48	51	47
Method Used	Identification and classification of bounding boxes in a single CNN architecture	CNN with semantic segmentation, Patch-Based model + fully convolutional model	CNN with semantic segmentation + post-processing	CNN with semantic segmentation
No. Sclerosed	585	870	564	303
No. Functional	3383	2997	1865	942
Image's size	1024×1024	448×448 and 1024×1024	256×256	2000×2000
Total of Images after DA	803200	471396	1126733	25320
CNN	YOLOv3	VGG16	U-Net	SegNet-VGG19

Table 9 Results from similar research[7),19),20)]

Author	**Precision**	**Recall**	**F1-score**	**Accuracy**
Proposed	0.9588	0.9682	0.9634	0.9810
Marsh et al.	0.7099	0.7914	0.7483	-
Gallego et al.	0.9630	0.8610	-	-
Bueno et al.	0.7499	0.9205	0.8191	0.9667

Table 10 Results in similar research separated by class [7),19),20)]

Metric	**Marsh et al.**		**Gallego et al.**		**Proposed**	
	Func.	Scle.	Func.	Scle.	Func.	Scle.
P	0.8128	0.6070	0.9760	0.9060	0.9885	0.9431
R	0.8852	0.6977	0.9170	0.6670	0.9923	0.8680
F1	0.8475	0.6492	0.9450	0.7680	0.9904	0.8542

Table 8 and the results are shown on **Table 9**.

Each author used a different method and different convolutional neural networks, although they all used semantic segmentation. The problem and method discussed in this work can't be exactly compared with these other works, once that they aim to define which regions of the WSI belong to the health and sclerosed glomeruli, while this work aims to identify the location and classification of the glomeruli and identify the quantities and percentage of glomeruli that are sclerosed. We understand that when comparing the metrics with the ones obtained in similar works, we are not looking for the best method between them, but rather considering that each one could be better in different situations.

Marsh et al.[7)] used Patch-Based and Fully Convolutional Model in the training. Gallego et al.[19)] used size analysis as post-processing. Bueno et al.[20)] used two methods in the paper, i.e., one with segmentation problem similar to the proposed method (i.e., non-glomerular, normal glomerular, and sclerosed structures) using convolutional neural networks in cascade, and another method using one unique neural network for 3 class segmentation problem similar to the proposed method (i.e., non-glomerular, normal glomerular, and sclerosed structures). For a fair comparison, in this paper, only the results involving 3 classes and a single CNN will be analyzed. As an extra information, Bueno et al. used just images stained with PAS.

Marsh et al.[7)] used H&E-stained frozen samples images. Gallego et al.[19)] used images stained with PAS and H&E. Another difference between the methods was the size of the images used. While Marsh et al.[7)] used 1024×1024 pixels image sets and isolated glomeruli in 448×448 pixels images, Gallego et al.[19)] used 256×256 pixels isolated glomerulus images and Bueno et al.[20)] used isolated glomeruli images of 2000×2000 pixels. Observing that the best results for the metrics analyzed were obtained in the proposed method, it is concluded that the analysis of the context where the object is

inserted is important for neural network learning. Some factors could explain why the proposed experiments had higher results compared to those presented by Bueno et al.[20]. Among them is the least scaling factor, as it is known that in the scaling process the image may lose relevant information. In addition, even though Bueno's research used a higher number of slides, the total number of glomeruli tagged in the present work is higher.

In **Table 10**, we can see that our method achieves significantly higher results in the identification and classification of sclerosed glomeruli compared to these other methods. Another observation is that the performance for sclerosed glomeruli detection is lower than for functional glomeruli in these methods. This is because the smaller number of sclerosed glomeruli in the database and blood vessels being wrongly labeled as sclerosed glomeruli by the neural network. This issue should be solved in the future.

4. Conclusion

In this paper, we evaluated the performance of YOLOv3 as deep neural network to identify the position and classify glomeruli, into functional and sclerosed, in WSIs from kidney biopsies. Also identify the total number of glomeruli and the rate of sclerosed glomeruli, to select WSIs that must be discarded for not having the minimum number of glomeruli for evaluation and to order the WSIs by damage.

Our method employs YOLOv3 with 53 layers neural network, to identify and classify glomerulus between functional and sclerosed. We prepared a database with images from Bio-Atlas repository, properly concatenated and classified, with a supervision of a pathologist, for training the neural network.

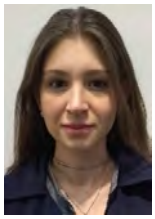
Our method achieved a high-performance with recall of 96.8%, precision of 95.9%, accuracy of 98.1%, and an F1 score of 96.3%. The results obtained were significant compared from those obtained in similar research. Also, we can show the WSIs reconstructed and with the glomeruli labeled, and a file containing the number of glomeruli found, the coordinates of each glomerulus on the slide and the percentage of sclerosed glomeruli, being useful in the medical routine.

As future works, we should increase the database with PAS-stained WSI and verify the results also with this kind of stain. Also, we should investigate another approach, as neural network cascade and semantic segmentation. Besides that, we are planning to extend our method to other specific diseases in medical routine.

References

1) A. Guyton, J. Hall: Tratado De Fisiología Medica. (2017).
2) V. Kumar, A. Abbas, K. Aster, J. Robins: Patologia Básica. (2013). doi:10.1016/b978-0-7020-3369-8.00001-x.
3) B. Henry, G. Lippi: "Chronic Kidney Disease is Associated with Severe Coronavirus Disease 2019 (COVID-19) Infection". Int. Urology and Nephrology. **52**, pp.1193–1194 (2020).
4) J. Ng, Y. Luo, K. Phua, A. Choong: "Acute Kidney Injury in Hospitalized Patients with Coronavirus Disease 2019 (COVID-19): A Meta-Analysis". J. Infect. - Lett. to Ed. **81**, pp. 661–664 (2020).
5) F. Zhang, Y. Liang: "Potential Risk of the Kidney Vulnerable To Novel Coronavirus 2019 Infection". Am J Physiol Ren. Physiol **318**, pp. 1136–1137 (2020).
6) R. Bonegio, D. Salant: "Glomerular Diseases". DECKER Intellect. Prop. INC- ACP Med. 2011;1-27 (2011).
7) J. Marsh, M. Matlock, S. Kudose, T. Liu, T. Stappenbeck, J. Gaut , J. Swamidass: "Deep Learning Global Glomerulosclerosis in Transplant Kidney Frozen Sections". IEEE Trans. on Med. Imaging (2018) doi:10.1109/TMI.2018.2851150.
8) Y. V. Eycke, J. Allard, I. Salmon, O. Debeir, C. Decaestecker: "Image Processing in Digital Pathology: An Opportunity to Solve Inter-Batch Variability of Immunohistochemical Staining". Sci. Rep. (2017). doi:10.1038/srep42964.
9) J. Redmon, S. Divvala, R. Girshick, A. Farhadi: "You Only Look Once: Unified, Real-Time Object Detection". Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016). doi:10.1109/CVPR.2016.91.
10) S. Skansi: Introduction to Deep Learning. From Logical Calculus to Artificial Intelligence, Springer (2018).
11) J. Ferreira, J. Raniery, N. Correia, P. Marques: "Aprendizado De Máquina Na Atenção À Saúde Humana". Computação Brasil - Revista da Sociedade Brasileira de Computação (SBC) 39ºed pp.37–39 (2019).
12) A. Madabhushi, G. Lee: "Image Analysis and Machine Learning in Digital Pathology: Challenges And Opportunities". Med. Image Anal. **33**, pp.170–175 (2016).
13) M. Najafabadi, F. Villanustre, T. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic: "Deep Learning Applications and Challenges in Big Data Analytics". J. Big Data. doi:10.1186/s40537-014-0007-7 (2015).
14) Bio-Atlas, http://zfatlas.psu.edu/, NIH grant 5R24 RR01744, Jake Gittlen Cancer Research Foundation, and PA Tobacco Settlement Fund. Slides 569, 603, 651, 657, 671, 704, 717, 832, 837, 842, 849, 868, 992, 1060, 1118, 1119, 1150, 1154, 1289, 1292, 1298, 1307, 1313, 1322, 1330, 1379, 1434, 1555, 1571, 1575, http://bio-atlas.psu.edu/view.php?s =603*.
15) S. Draining: Ybat - YOLO BBox Annotation Tool. https://github.com/drainingsun/ybat
16) J. Redmon, A. Farhadi: "YOLOv3: An Incremental Improvement". Tech Rep. (2018).
17) D. Powers: "Evaluation: From Precision, Recall , F-Factor To ROC, Informedness, Markdness & Correlation". J. Mach. Learn. Technol. **2**, pp.37–63 (2011).
18) D. Chicco, G. Jurman: "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation". BMC Genomics, doi:10.1186/s12864-019-6413-7 (2020).
19) J. Gallego, Z. Swiderska-Chadaj, T. Markiewicz, M. Yamashitae, A. Gabaldong, A. Gertych: "A U-Net Based Framework to Quantify Glomerulosclerosis in Digitized PAS and H&E Stained Human Tissues". Comput. Med. Imaging Graph. **89**, (2021).
20) G. Bueno, M. Fernandez-Carrobles, L. Gonzalez-Lopez, O. Deniz: "Glomerulosclerosis Identification in Whole Slide Images Using Semantic Segmentation". Comput. Methods Programs Biomed. **184**, 105273 (2020).

(Received May. 27, 2021)
(Revised Jan. 31, 2022)



Thalita Munique COSTA
(Student Member)

She received her B.S. and M.E. degrees in Electronic Engineering, and in Science, Area of Concentration: Biomedical Engineering, from Federal University of Technology - Paraná (UTFPR), Brazil in 2018 and 2020, respectively. She is now a Dr. Eng. Candidate in Telecommunication System at Shinshu University, Nagano, Japan. Her research interests include Digital Pathology, Artificial Intelligence, and Image Processing.

Lourenço BARBOSA

He graduated as an Electronics Technician at Federal Technological Center of Rio Grande do Sul (CEFET) in 2003, bachelor's in Information Systems at Paraná Society of Education and Informatics (SPEI) in 2010. In 2020, he received the title of Master of Science from the Federal University of Technology - Paraná (UTFPR), He is a Technical Leader in Software Evolution and Supports at NTT DATA Business Solutions Brazil. His research for the medical field encompasses the topics of image/video processing, deep learning, and machine learning. Currently, as a candidate for a Doctorate at UTFPR.

Yoko USAMI

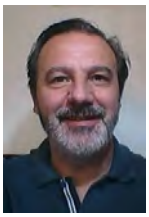
She received her Ph.D. degree from Tokyo Medical and Dental University, Tokyo, Japan in 2013. She joined the Department of Laboratory Medicine at Shinshu University Hospital, Japan where she is currently a vice tech supervisor. Her research interests include Laboratory Medicine, Clinical Chemistry, and Clinical Laboratory Immunology.

Mai IWAYA

She received her M.D. degree from Shinshu University in 2005. She completed anatomical pathology residency at Shinshu University Hospital in 2011 and completed gastrointestinal pathology fellowship at University of Toronto, Canada in 2018. She received the Ph.D. degree from Shinshu University in 2019. She joined the department of laboratory medicine at Shinshu University in 2018 and currently, works as an associate professor/lecturer. Her research interests include artificial intelligence in the pathology field.

Kiyoshi TANAKA (*Fellow*)

He received his B.S and M.S. degrees in Electrical Engineering and Operations Research from the National Defense Academy, Yokosuka, Japan, in 1984 and 1989, respectively. In 1992, he received the Dr. Eng. Degree from Keio University, Tokyo, Japan. In 1995, he joined the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University, Nagano, Japan, and currently, he is a full professor in the academic assembly (Institute of Engineering) of Shinshu University. He is the former Vice-president of Shinshu University in charge of international affairs and the former director of the Center for Global Education and Collaboration of Shinshu University. His research interests include image and video processing, 3D point cloud processing, information hiding, human visual perception, evolutionary computation, multi-objective optimization, smart grid, and their applications. Currently, he is the president of IIEEJ, a fellow of IIEEJ, a member of IEEE, IEICE, IPSJ, JSEC, and so on.



Fabio SCHNEIDER

He received his B.S and M.S. degrees in Electrical Engineering and Biomedical Engineering from the Federal University of Technology - Paraná (UTFPR), Brazil, in 1989 and 1995, respectively. In 2006, he received the Ph.D. degree from University of Washington, Seattle, U.S.A. In 1995, he joined the Department of Electronics and Telecommunication at UTFPR, Brazil where he is currently a full professor. He is a former Vice President for Research and Post Graduate School of UTFPR. His research interests include image and video processing, Artificial Intelligence and Deep Learning development and application, ultrasound imaging, computational algorithms, and architectures.

IIEEJ **Paper** *-- Special Issue on Journal Track Papers in IEVC2021Part II --*

Measurement of Eye Size Illusion Caused by Thickness of Eyeliner on Double-Eyelid Eyes

Rena OKURI† , Shuhei KODAMA†† , Tokiichiro TAKAHASHI†† (*Fellow*)

† Tokyo Denki University , †† ASTRODESIGN, Inc.

<**Summary**> Eyeliner is one of the makeup methods that enhances the attractiveness of female by a geometric illusion that makes the eyes appear larger. Eyeliner has no fixed shape, and its thickness and length can be freely adjusted. In this paper, we experimentally verify the relationship between thickness of eyeliner and the perceived eye size. In addition, by examining the sense of incongruity of the eyeliner, we clarify the optimal thickness of eyeliner as makeup to make the eyes appear larger. The results showed that the thicker eyeliner, the larger the eyes were perceived to be. However, the excessive thickness of eyeliner increased the sense of incongruity and reduced the illusion effect. There were gender differences in the results of these experiments. For male, thickness of eyeliner had a significant effect on the perception of eye size, while for female it had no significant effect.

Keywords: makeup, eyeliner, face, eyes, geometric illusion

1. Introduction

In recent years, with the growing interest in facial beauty and attractiveness, facial imaging studies have been actively conducted in many fields. According to experimental studies on facial morphology, large eyes have one of the features that increase the attractiveness of a female face[1] because they give the sense of childishness and femininity[2),3]. Makeup is a means of improving the attractiveness of a female face[4),5]. In particular, eye makeup improves the attractiveness of a female face by making the eyes appear larger and emphasizing them[5].

Matsushita et al.[6] revealed that eyeliner, a makeup method in which a line is drawn around the eyelash line, has the effect of making the eyes appear larger by the Delboeuf illusion. They measured how much the eyes appear to be larger due to the illusion (called the amount of illusion) by adjusting the color and range (upper eyelid only, upper eyelid + lower eyelid) of eyeliner.

In this paper, we assume that thickness of eyeliner affects the amount of the illusion in addition to the color and range of eyeliner as Matsushita et al.[6] focused on. We measure the amount of illusion between thickness of eyeliner and the size of the double-eyelid eyes to clarify the relationship between thickness of eyeliner and the amount of illusion using a psychophysical measurement method. Moreover, we investigate the appropriate thickness of eyeliner as eye makeup by conducting a subjective evaluation questionnaire of the sense of incongruity for the eyes with the eyeliner of each thickness. These results are used to verify the best thickness of eyeliner to make the eyes appear larger. Since it is known that there are gender differences in the perception of attractiveness with makeup[5], we also verify the gender differences in the results.

2. Related Work

Jones et al.[7] tested the hypothesis that the apparent size of facial features (eyebrows, eyes, nose, and mouth) changed by makeup. They found that the eyes appeared to be larger after applying makeup both by the participants themselves and by professionals. Furthermore, they compared whether the change in apparent feature size due to cosmetics was caused by the enhancement of fine details or by the coarse changes in the entire region by filtering the spatial frequency of the face image. As a result, the eyes appeared larger with makeup regardless of the spatial frequency, indicating that the effect of cosmetics on apparent feature size is not limited to whether the spatial frequency is low or high. These results support the idea that modification of the apparent size of facial features by makeup can enhance facial attractiveness.

Several studies focus on the eyes among the facial features. Morikawa et al.[8] found that the eyes with

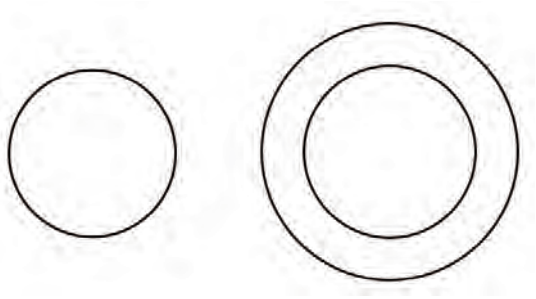


Fig. 1 Delboeuf illusion

eye shadow appear larger in their experiment using psychophysical methods. They showed that the perceived size of the eyes with eye shadow is related to the assimilation effect between the eye and the eye shadow, and the assimilation effect between the eyebrow and the eye when the eye shadow fills the area between the eye and the eyebrow. In addition, the perceived the eyes with eye shadow is affected by the viewing distance. They also experimentally showed that the eye size illusion caused by eye shadow and eyebrows is based on the same psychological effect as the Delboeuf illusion. The Delboeuf illusion is the illusion that the inner circle of a double circle appears to be larger than that of a single circle (the inner circle is the same size as the single circle), as shown in **Fig. 1**. This illusion is caused by the assimilation of the outer and inner circles of the double circles.

Matsushita et al.[6] experimentally showed that eyeliner has the effect of making the eyes look larger based on the Delboeuf illusion and that the perceived eye size changes depending on the color and range of eyeliner. They verified the effect of color and range of eyeliner on the eye size illusion, but they did not verify the effect of eyeliner thickness on the illusion.

In this paper, we experimentally examine the effect of eyeliner thickness on the perceived eye size. Furthermore, we discuss the results focusing on gender differences, which have not been verified in the previous studies[6),8)].

3. Illusion Measurement Experiment

3.1 Experimental environment and apparatus

Since it was difficult to gather all participants for the experiment due to the COVID-19 pandemic, we distributed the experimental program and instructed each participant to prepare the execution environment of the experimental program before conducting the experiment. During the measurement, we instructed that cold lighting should be used, the brightness should be constant, and the direction of the lighting should be fixed. The position

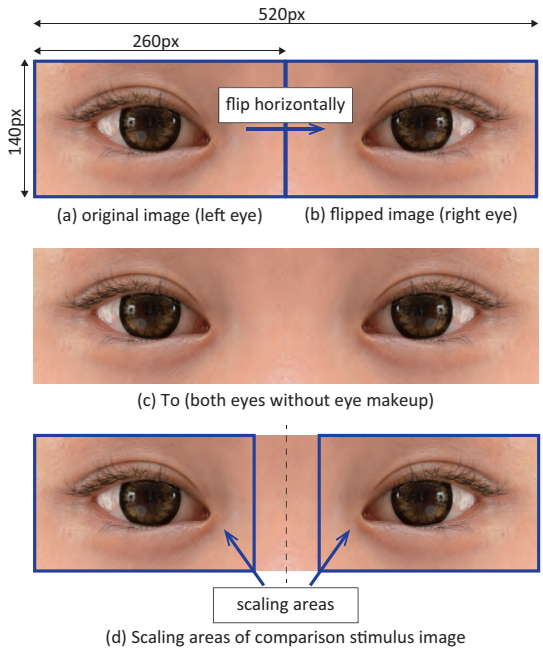


Fig. 2 Procedure for generating T_o which is both eyes image without eye makeup and the comparison stimulus image generated by scaling T_o image

and orientation of the display should be fixed, and the luminance should be measured with a color luminance meter (CS-100A, Konica Minolta) to be approximately 250 cd/m^2. The aspect ratio of all the displays used in the experiment was 16:9, and their vertical sizes H of the displays were between 14.9 cm and 33.6 cm. The observation distances between the subjects and the displays were set to $3H$. All participants were proficient in IT literacy, and they prepared the experimental environment as instructed and followed the experimental procedures.

3.2 Standard stimulus images and comparison stimulus images

The comparison stimulus images and the standard stimulus images used in the experiment were generated as follows.

The pixels resolution of both comparison and standard stimulus images was 520 px$\times$ 140 px, and they were displayed on the screen so that 1 px of the image was 0.2 mm (rounded to the second decimal place).

In this experiment, we generate and use symmetrical eyes images as stimulus images. To generate symmetrical eyes images, first, we prepare an image of the eye to the left of the root of the nose (left side of the screen) of a Japanese woman with double-eyelid and no makeup (**Fig. 2**(a)), as an original image. The width of each eye is set to about 30.0 mm, based on the average interocular breadth of 36.6 mm and biectocanthion breadth of 90.0 mm, which are typical for Japanese young adult female.[9)]

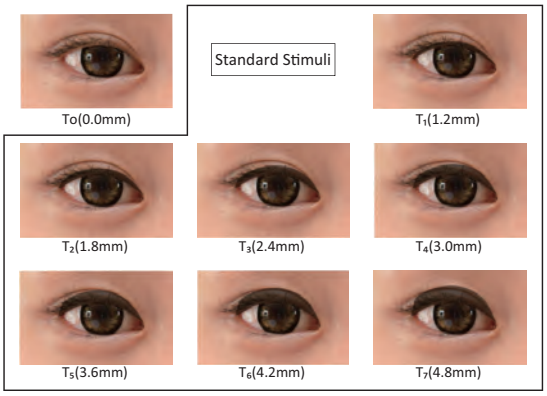


Fig. 3 Original eye image without eye makeup (T_o) and the standard stimulus images (T_1 to T_7)

Fig. 4 Comparison stimulus images (88%, 100%, 116%)

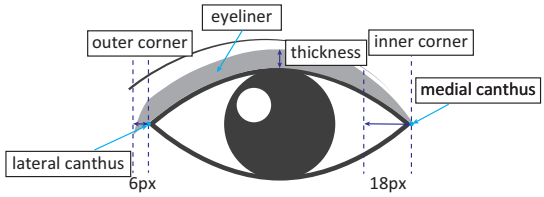


Fig. 5 Definition of eyes and eyeliner

Next, we generate an image of the right eye (right side of the screen) by flipping the original image horizontally (Fig. 2(b)). Finally, we generate a symmetrical image of both eyes T_o (thickness original) as shown in Fig. 2(c), based on the original image and the right eye image.

As standard stimulus images, we generate seven images T_1, T_2, T_3, T_4, T_6, and T_7 with eyeliner of thickness 6, 9, 12, 15, 18, 21, and 24 px (1.2, 1.8, 2.4, 3.0, 3.6, 4.2 and 4.8 mm in real size) drawn with paint software on T_o which is the image of both eyes without eye makeup. The ith image T_i is an image with eyeliner of thickness $(i \times 3 + 3)$ px or $(i \times 0.6 + 0.6)$ mm drawn by paint software. Only the left eye part of T_o and the standard stimulus images (T_1 to T_7) are shown in **Fig. 3**.

As comparison stimulus images, we generate images in which the area around each eye of T_o (Fig. 2(d)) is scaled by 2% from 88% to 116% without changing the aspect ratio of the original image in the horizontal and vertical directions. Here, T_o is the original image of the comparison stimulus image, scaled to 100%. For gaps in the image caused by scaling, we blur so that the merged area would not seem unnatural. **Figure 4** shows only the left eye part of the comparison stimulus images (88%, 100%, 116%). It is unknown to what extent thickness of eyeliner affects the apparent eye size both vertically and horizontally. We conducted the experiment assuming that the apparent eyes are isotropically scaled in horizontal and vertical directions. It would be a future task to verify the scaling ratios of the apparent eyes are either isotropic or anisotropic.

When the eyeliner is drawn in a smooth curve that follows the contour of the upper eyelid margin, the thickness of eyeliner is defined as the distance extended vertically upward from the curve. The eyeliner gradually becomes thinner toward the inner and outer corners of the eye to follow the contour of the eye. In this experiment, the inner corner of the eye is set between the medial canthus and the point 18 px inward from the medial canthus. The outer corner of the eye is defined as the point between the lateral canthus and the point 6 px outward from the lateral canthus. **Figure 5** shows the definitions of eyes and eyeliner.

3.3 Experimental procedure

For each standard stimulus, we obtain an equivalent comparison stimulus, *i.e.*, the points of subjective equivalence (PSE), and define the scaling rate of the comparison stimulus as the amount of illusion. The procedure of the experiment is shown in **Fig. 6**.

We used the staircase method to measure the amount of illusion. First, Image 1 of Fig. 6 (a gray scale image, R=G=B=153) is displayed for 7000 ms. Then, Image 2 of Fig. 6, in which standard and comparison stimulus are aligned on the left and right sides of the image, is displayed on a white background for 3000 ms. The participants compare both stimulus images, then answer in three ways: "I feel that the eyes on the right side of the image are larger," "I feel that the eyes on the left side of the image are larger," or "equal/not sure." Next, the Image 1 is displayed again for 4000 ms. Then, the task is repeated to answer the question by comparing the eye sizes of the stimulus images displayed on both the left and right sides of Image 2 and to watch Image 1 on the display.

The comparison stimulus is changed by the participant's judgment in each trial. When the participants judge that the standard stimulus is larger, the staircase direction of the comparison stimulus is upward (88% to 116%), while when they judge that the comparison stimulus is larger, the staircase direction of the comparison stimulus is downward (116% to 88%). The measurement

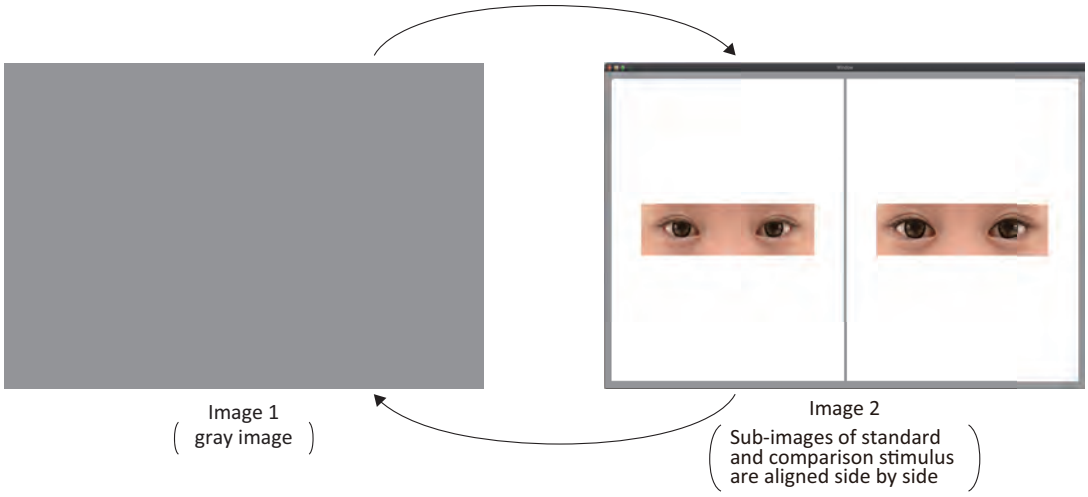


Fig. 6 Experimental procedure for measuring the amount of illusion

is completed when the direction of ascending or descending of the comparison stimulus is reversed 8 times, referring to the experimental method of Matsushita et al.[6]. Then, the mean of the stimulus intensity (%) of 8 times of the comparison stimuli is used as the amount of illusion under the condition of that standard stimulus.

The standard stimulus and the comparison stimulus are displayed on the left or right side of the screen at random. A break time of at least one minute is provided before changing the standard stimulus image and moving to the next measurement.

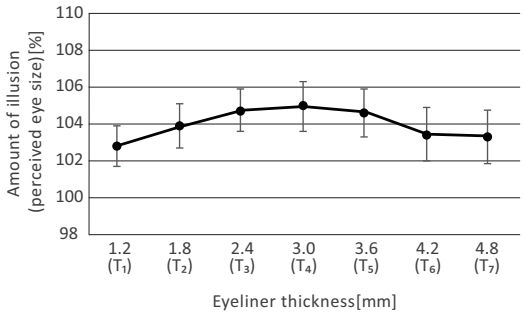


Fig. 7 Thickness of eyeliner and the amount of illusion (Markers indicate mean values of the amount of illusion and error bars indicate 95% confidence intervals)

4. Experimental Results and Statistical Analysis

4.1 Participants

Thirty-three participants (mean age 22.7 years, $SD = 1.48$, 17 males and 16 females) with normal visual acuity and color vision participated. We explained the contents of our experiment in verbal and written, then obtained the participants' consent before conducting the experiment.

4.2 Experimental results

Figure 7 shows the results of the illusion measurement experiment. The amount of illusion in Fig. 7 indicates the mean amount of illusion of the participants. In the standard stimulus images T_1 to T_7, the amount of illusions (perceived eye size) were about 102.8%, 103.9%, 104.7%, 104.9%, 104.6%, 103.4%, 103.3%, respectively (about 105.7%, 107.9%, 109.6%, 110.1%, 109.4%, 107.0%, 106.8% in area, respectively). For example, the amount of illusion of 105% makes the eyes appear to be 5% larger than their real size. The amount of illusion tended to increase when thickness of eyeliner was 1.2 mm to 3.0 mm, to reach the maximum at 3.0 mm, and then to decrease from 3.0 mm to 4.8 mm. We refer to the detailed analysis of the measurement results in section 4.3.

4.3 Statistical analysis

We analyzed whether the eyeliner thickness factor affects the amount of illusion using the analysis of variance and the effect size for the results of the illusion measurement experiment. Furthermore, we compared between groups using the multiple comparison method, 95% confidence intervals, and effect sizes. The results are described below.

4.3.1 Effect of eyeliner thickness factor

To analyze the effect of eyeliner thickness on the double-eyelid eye size illusion, we checked the normality of the experimental data and the equality of variance between groups and used the one-way ANOVA of within-subjects factors. The results of the analysis of variance showed that the effect of eyeliner thickness was statistically significant ($F(6, 32) = 3.80$, $p = .00 < .05$). According to the effect size η^2, which is an indicator of the real effect size regardless of the sample size, the strength of the relationship between the amount of illusion and thickness of eyeliner was slightly medium[10] ($\eta^2 = 0.04$). In other words, thickness of eyeliner affected the perception of the size of double-eyelid eyes.

4.3.2 Intergroup comparison

By comparing each group of T_1 to T_7, we can grasp the trend of the change in the amount of illusion. We compared the differences in the means between the groups using the multiple comparison method (Tukey-Kramer method) and found that T_3, T_4, and T_5 were significantly larger than T_1. Therefore, it can be said that the illusion increased gradually as thickness of eyeliner increased from T_1 to T_4, and was the largest around T_4. On the other hand, there were no statistically significant differences between the other groups. However, looking at the 95% confidence interval, which is an indicator of significant difference by not including 0, T_6 for T_4 (mean difference -1.52, 95%CI[-3.28, 0.24]) and T_7 for T_4 (mean difference -1.61, 95%CI[-3.36, 0.15]) slightly included 0. This indicates that the amount of illusion tends to decrease gradually with the eyeliner thickness after T_4 (3.0 mm or more). This suggests that above a certain thickness, the eyes do not appear to be as large as expected even if the eyeliner is thicker.

We then measured the magnitude of the experimental effect between each group using the effect size dD (the effect size that takes the correlation between the groups into account). The effect sizes of T_1 for T_2 and T_2 for T_3 were both slightly medium[10] ($dD = 0.35$ and $dD = 0.34$, respectively), which supports the monotonic increase of the amount of illusion from T_1 to T_3. In addition, the effect size of T_5 to T_6 was nearly medium[10] ($dD = 0.43$), suggesting that the amount of illusion decreases from T_5 to T_6. Moreover, the effect sizes of T_3 to T_4, T_4 to T_5, and T_6 to T_7 were nearly small to extremely small[10] ($dD = 0.07$, $dD = 0.13$ and $dD = 0.03$, respectively), suggesting that the change in the amount of illusion from T_3 to T_5 and from T_6 to T_7 are slight.

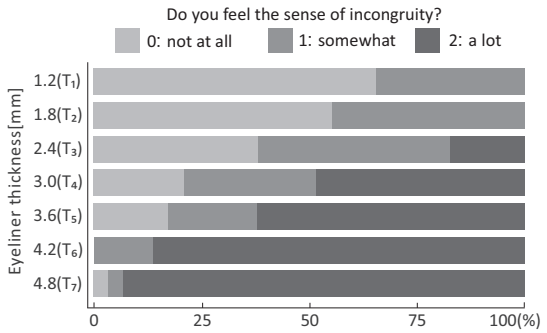


Fig. 8 Subjective evaluation questionnaire results

5. Subjective Evaluation Questionnaire

Although there is no rule for thickness of eyeliner, thick eyeliner is not appropriate as makeup if it causes the sense of incongruity. Thus, we investigated the sense of incongruity for each thickness of the eyeliner in the illusion measurement experiment using a subjective evaluation questionnaire. Twenty-nine participants (17 males and 12 females, aged 20-28 years) answered the questionnaire.

5.1 Contents

We conducted a questionnaire to investigate the sense of incongruity of the standard stimulus images T_1 to T_7 used in the illusion measurement experiment as makeup. One of the standard stimulus images T_1 through T_7 was proposed in the questionnaire, and participants were required to answer the question "Do you feel the sense of incongruity with the following image?" by selecting from "0. not at all," "1. somewhat," or "2. a lot." This task was performed for all standard stimulus images T_1 to T_7. In the field of psychiatry, the three-option Brief Grief Questionnaire[11] is often used for negative impressions such as anxiety and distress, so the three-point scale was also used in this experiment to investigate negative impressions of the sense of incongruity.

5.2 Response results and discussion

As shown in **Fig. 8**, it can be found that the thicker the eyeliner from T_1 to T_7, the more sense of incongruity the participants had. The number of participants who had the strong sense of incongruity gradually increased from T_3, reaching nearly 50% in T_4, and more than 50% in T_5, T_6, and T_7, which suggests that they are not appropriate as makeup. In T_1, some participants answered that they felt somewhat the sense of incongruity because they felt the sense of incongruity that the eyeliner was not actually applied but drawn with paint software and that the shape of the corner of the eyes.

6. Discussion

6.1 Effect of eyeliner of different thicknesses

From the experimental results on measuring the amount of illusion with various thicknesses of eyeliner, it was shown that the amount of illusion increased monotonically from 1.2 mm to 2.4 mm, and slightly from 2.4 mm to 3.0 mm, and that the eyes were perceived to be the largest around 3.0 mm. Furthermore, there was a slight decreasing trend from 3.0 mm to 3.6 mm, a decreasing trend from 3.6 mm to 4.2 mm, and a very small decreasing trend from 4.2 mm to 4.8 mm. A questionnaire for subjective evaluation suggested that the eyeliner with thickness of 3.0 mm or more gave the strong sense of incongruity as makeup. Therefore, thickness of eyeliner that can make the eyes appear the largest with the least the sense of incongruity as makeup is about 2.4 mm.

The reason why the illusion increases with thickness of eyeliner is thought to be that eyeliner area is misidentified as the eye area. It is also thought that the thicker eyeliner, the easier it is to perceive that the eye is surrounded by eyeliner and the Delboeuf illusion is more likely to occur.

However, the eyes with excessively thick eyeliner make the boundary between the face and the eyes clearer, thus making the eyes appear to float away from the face. Then, the eyes appear to be smaller in contrast to the size of face, which is called "size contrast illusion." This illusion is also seen with eye shadow[6]. In addition, the sense of incongruity of stimulus images with excessively thick eyeliner may affect the perception of the illusion and reduce the amount of the illusion. The measurement of the amount of illusion with the strong sense of incongruity makes it difficult for the participants to perceive the shape in a short time and to obtain the illusion effect. In addition, there is a possibility of large individual differences in the results because participants are hesitant to judge.

Therefore, the eyeliner thickness illusion is not merely an application of the Delboeuf illusion, but probably an interrelationship of various illusions. Further experiments are needed to clarify the relationship between the various illusion effects.

6.2 Differences between male and female

The results of the illusion measurement experiment and the subjective evaluation questionnaire showed a large differences between male and female. **Figure 9** shows the

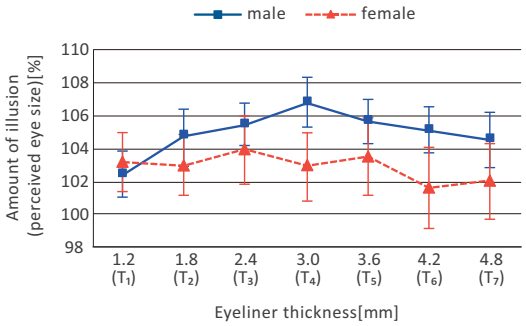


Fig. 9 Thickness of eyeliner and the amount of illusion by gender (Markers indicate mean values of the amount of illusion and error bars indicate 95% confidence intervals)

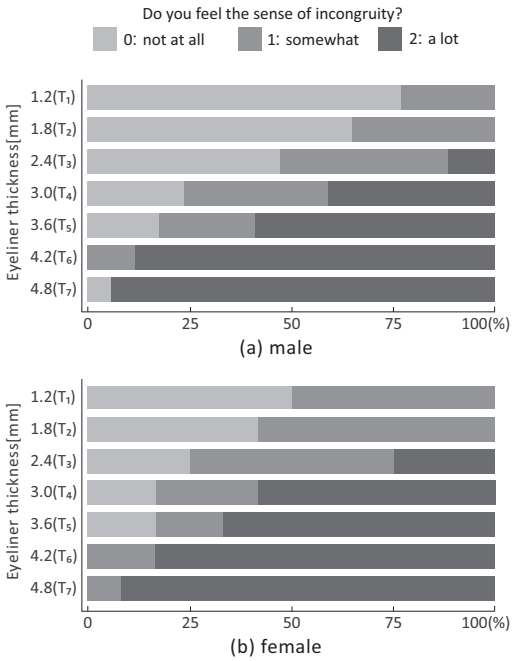


Fig. 10 Subjective evaluation questionnaire results by gender

results of the illusion measurement experiment by gender. **Figure 10** shows the results of the subjective evaluation questionnaire by gender.

The results of the illusion measurement experiment for male showed that the amount of illusion tended to increase when thickness of eyeliner was 1.2 mm to 3.0 mm, to reach the maximum at 3.0 mm, and then to decrease from 3.0 mm to 4.8 mm (Fig. 9). This trend is similar to that of all the participants (Fig. 7). For male, thickness of eyeliner had a significant effect on the amount of illusion ($F(6, 16) = 6.00$, $p < .001$). While for female, the variance was larger and thickness of eyeliner had no significant effect on the amount of illusion ($F(6, 15) = 1.90$, $p = .09$). This means the illusion effect of eyeliner

thickness in male was more marked than in female.

In the subjective evaluation questionnaire, about 50% of the female participants already felt a little sense of incongruity with the minimum thickness of eyeliner (1.2 mm), suggesting that the sense of incongruity affected the measurement of the amount of illusion more than the male participants. Since female has more opportunities to consciously look at eyeliner, they might be more likely to feel the sense of incongruity about the shape and position of the eyeliner.

Coren and Porac[12] showed the correlation among individual differences in spatial ability and the size of illusion for each type of illusion. They proved that the effect of illusion is larger in people with higher spatial ability for illusions related to the perception of area and direction, such as the Delboeuf illusion. Although male are generally said to be better than female in spatial ability[13], their study did not show clear and consistent sex differences in the correlations among the spatial abilities measures and the visual-geometric illusion scores. However, the eyeliner illusion, which is not a mere geometric illusion, is caused by the interaction of many complex perceptual processes. It is also possible that male with higher spatial ability is more likely to show this illusion effect. We suspect that these differences between male and female in the amount of illusion and the sense of incongruity on the eyes with makeup are one of the reasons why there are gender differences in makeup preferences.

7. Conclusion

In this paper, we measured the apparent size of double-eyelid eyes depending on thickness of eyeliner, *i.e.*, the amount of illusion. The results of the measurement experiment showed that the illusion increased monotonically to a certain level as thickness of eyeliner increased, and then decreased. This result indicates that thick eyeliner causes misperception of the eye area and makes the eyes appear larger, while too thick eyeliner has the opposite effect. We measured the amount of illusion by assuming that the thickness of eyeliner is isotropically large in the horizontal and vertical directions relative to the apparent eye size. Verifying different aspect ratios in the horizontal and vertical directions is a future work. Moreover, further experiments are needed to clarify whether the illusion effect of eyeliner thickness is affected by the texture of the actual eyeliner, such as liquid or pencil type.

We also revealed that gender differences within the results were also apparent. Thick eyeliner had a greater illusion effect only on male. In this paper, since we conducted on one type of eye sample, it is necessary to conduct validation on multiple types of eye samples that take the shape of the eyelid such as single-eyelid eyes into consideration. In the future, based on the verification with multiple types of eye samples, we will investigate the thickness model of eyeliner that makes the eyes appear larger according to the shape of individual eyes.

Acknowledgments

We would like to thank Dr. Hiroto Inouye, assistant professor at Tokyo Denki University, for his comments and suggestions in carrying out this paper. We would also like to thank Mr. Shigeaki Suzuki and ASTRODESIGN, Inc. for their kind support.

References

1) J. Y. Baudouin, G. Tiberghien: "Symmetry, Averageness, and Feature Size in the Facial Attractiveness of Women", Acta Psychologica, Vol.117, No.3, pp.313–332 (2004).
2) D. Jones, C. L. Brace, W. Jankowiak, K. N. Laland, L. E. Musselman, J. H. Langlois, L. A. Roggman, D. Pérusse, B. Schweder, D. Symons: "Sexual Selection, Physical Attractiveness and Facial Neoteny: Cross-Cultural Evidence and Implications", Current Anthropology, Vol.36, No.5, pp.723–748 (1995).
3) D. I. Perrett, K. J. Lee, I. Penton-Voak, D. Rowland, S. Yoshikawa, D. M. Burt, S. P. Henzi, D. L. Castles, S. Akamatsu: "Effects of Sexual Dimorphism on Facial Attractiveness", Nature, Vol.394, No.6696, pp.884–887 (1998).
4) K. Tagai, H. Ohtaka, H. Nittono: "Faces with Light Makeup Are Better Recognized than Faces with Heavy Makeup", Frontiers in Psychology, Vol.7, No.226 (2016).
5) R. Mulhern, G. Fieldman, T. Hussey, J. L. Lévêque, P. Pineau: "Do Cosmetics Enhance Female Caucasian Facial Attractiveness?", International Journal of Cosmetic Science, Vol.25, No.4, pp.199–205 (2003).
6) S. Matsushita, K. Morikawa, H. Yamanami: "Measurement of Eye Size Illusion Caused by Eyeliner, Mascara, and Eye Shadow", Journal of Cosmetic Science, Vol.66, No.3, pp.161–174 (2015).
7) A. L. Jones, A. Porcheron, R. Russell: "Makeup Changes the Apparent Size of Facial Features", Psychology of Aesthetics, Creativity, and the Arts, Vol.12, No.3, pp.359–368 (2018).
8) K. Morikawa, S. Matsushita, A. Tomita, H. Yamanami: "A Real-Life Illusion of Assimilation in the Human Face: Eye Size Illusion Caused by Eyebrows and Eye Shadow", Frontiers in Human Neuroscience, Vol.9, No.139, pp.1–9 (2015).
9) M. Kouchi, M. Mochimaru: "Anthropometric Database of Japanese Head 2001", National Institute of Advanced Industrial Science and Technology, H16PRO-212 (2008).
10) J. Cohen, Statistical Power Analysis for the Behavioral Sciences (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum (1988).
11) M. Ito, S. Nakajima, D. Fujisawa, M. Miyashita, Y. Kim, M. K. Shear, A. Ghesquiere, M. M. Wall: "Brief Measure for Screening Complicated Grief: Reliability and Discriminant Validity", PLoS One, Vol.7, No.2, e31209 (2012).
12) S. Coren, C. Porac: "Individual Differences in Visual-

Geometric Illusions: Predictions From Measures of Spatial Cognitive Abilities", Perception & Psychophysics, Vol.41, No.3, pp.211–219 (1987).

13) M. C. Linn, A. C. Petersen: "Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis", Child Development, Vol.56, No.6, pp.1479–1498 (1985).

(Received May 31, 2021)
(Revised April 30, 2022)

Rena OKURI
She received her B.E. degree from Tokyo Denki University in 2020. She is currently a Master's student at Tokyo Denki University Graduate School. Her research interests include facial image processing and cognitive science.

Shuhei KODAMA
He received his Ph.D. (Engineering) from Tokyo Denki University in 2019. He had been a visiting research student at Université de Montréal from March until June, 2018. He had been an engineer at ASTRODESIGN, Inc. from 2019 until 2021. He is currently an assistant professor at Tokyo Denki University. His research fields are NPR and HCI.

Tokiichiro TAKAHASHI (*Fellow*)
He received his B.E. degrees from Niigata Uiversity in 1977, and Ph. D. from the University of Tokyo in 2005, respectively. He had been working at Nippon Telegraph and Telephone Corporation and engaged in Pattern Recognition, Computer Graphics, and Learning Science since 1977. Since 2003, he has been a professor at Tokyo Denki University. He has been a visiting researcher at ASTRODESIGN, Inc. since 2017. His research interests include computer graphics, image processing and visual computing.

IIEEJ **Paper** *-- Special Issue on Journal Track Papers in IEVC2021Part II --*

3D Distance Field-Based Apparel Modeling

Masanori NAKAYAMA† (*Member*), Takami YAMAMOTO†† (*Member*),
Issei FUJISHIRO† (*Honorary Member/Fellow*)

† Keio University, †† Wayo Women's University

<**Summary**> This study aims to design an apparel item by converting a polygon model of the human body obtained through a 3D scan into a volume model. From the model, we can quickly generate a curved surface that retains a constant distance from the surface of the body, thus being able to design an apparel item that ensures accurate allowance by defining the distance field for the surface of the body. The curved surface, which would be an ideal garment that can envelop the body, can be generated by tweaking the threshold value for an isosurface from the surface of the body. Furthermore, by drawing design lines on the curved surface, the pattern required for actual sewing can be designed in 3D space. The curved surface trimmed by the design lines is converted into a polygon mesh. A dynamics simulation is applied to smooth the curvatures so that the desired flat surface pattern can be obtained. As a result of actually sewing the virtually designed pattern and trying it on, it was shown that we can generate clothes with the appropriate allowance and proportions.

Keywords: 3D scanner, volume model, 3D distance field, apparel items

1. Introduction

In conventional apparel design, polygon models have commonly been used to explicitly define fabric based on a methodology that mounts 2D patterns on a 3D human body[1)–3)]. The apparel industry has a long history of drafting patterns by hand. Over the course of digitalization, many kinds of pattern drafting software have been developed based on 2D industrial computer-aided design (CAD). Indeed, the culture of 2D pattern drafting has continued for many years, and the mainstream format in industrial design became the mainstream format in drafting. However, there is still no fundamental change in the conventional methodology of enveloping a 3D human body with a 2D cloth. The 2D design must be forcibly extended and drawn, especially when developing garments of multiple sizes (i.e., grading), to ensure an appropriate allowance. Therefore, major size changes cannot be made, and any allowance must be based on empirical evaluations. A typical work using this methodology is given by Umetani et al[4)], who developed a system that allows the user to draw a design and carry out the fitting in real time. However, the amount and position of the darts and shape of the sleeves are not practical in terms of clothing construction and have limited relevance to the human body and clothing.

With this in mind, we propose a reverse pattern design method of cutting a 3D curved surface and then expanding this into 2D space. In other words, this method first constructs a curved surface that expresses the ideal piece of clothing, cuts it directly in 3D space, and then converts it into a 2D pattern. This creates a seamless curved surface of the garment that would cover the human model serving as the base; consequently, the 3D consistency at the time of sewing can always be maintained, regardless of body shape. Furthermore, various sizes and appropriate allowances can be freely set by enlarging or reducing the size of the base human model.

In the current study, we have focused on the degree of freedom and accuracy of the human body model. A human body model measured with a 3D scanner includes body shape distortions and uneven noise. Therefore, most tradtional methods copy individual body shape information by fitting a simplified template model prepared with an actual model. In other words, the accuracy of these other methods is inferior because the extremely complex human body shape is simplified into a few parameters, such as the bust or waist. This accuracy issue is particularly significant in fields where "form-fitting clothing" is emphasized. We have adopted implicit function modeling to generate a curved surface that envelops the

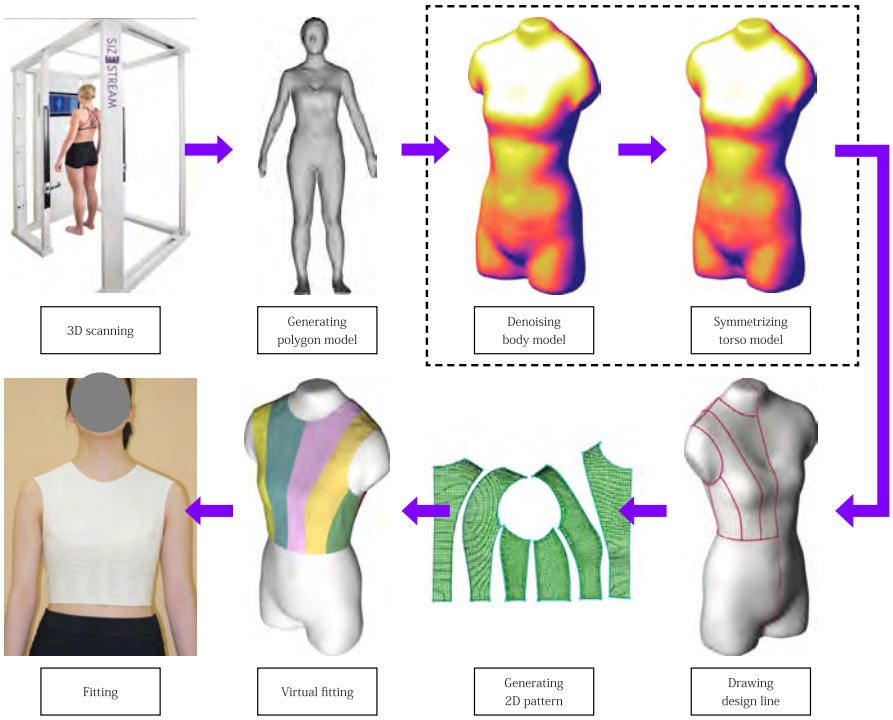


Fig. 1 Pattern development flow based on 3D distance field

human body. In other words, the human body model is implicitly defined as a distance field based on the surface of the body, thereby facilitating the generation of a 3D curved surface along the surface of the body. Specifically, the human body model is defined as a "volume model", which is done by using voxels.

The most prominent characteristic of the volume model is that the curved surface retaining a constant distance from the surface of the body can be simply generated by establishing the distance field. In other words, designing a garment with an appropriately defined allowance is possible because the measured human body can be made thinner or fatter just by changing the threshold value for isosurfacing.

When the 3D scanned form is used without any changes being made, an improved fit to the body is achieved, whereas the aesthetics of a well-proportioned garment decreases. To address this issue, a method to negate the left and right disparity of the body is developed, here by utilizing the characteristics of the voxel model, which simplifies the topological operations while maintaining personal body shape characteristics[5].

The flat surface pattern needed for actual sewing is obtained by trimming the ideal garment surface defined by the isosurface. Of course, the trimmed surface has inherent 3D curvatures. However, by converting the trimmed surface to a polygon mesh again, cloth simulation with a mass-spring model is applied. In other words, a 2D pattern is obtained while controlling distortion to a minimal degree by reducing the 3D curve of the fabric to 2D stretch expansion.

Furthermore, to evaluate this approach, we created a basic pattern of the upper body for both women and men. We also created a vest for men as another real apparel item. Our original intent was to provide a comprehensive framework for the design of apparel items, yet in the present study, we focus only on tight clothing because this type of clothing is the most sensitive to individual differences and, thus, is difficult to design accurately. Traditionally, apparel designers have only been able to make 3D cuts using a standard torso. In our study, by strictly maintaining the allowance, we establish a new design method that emphasizes the individual body shape.

2. Volumetric Apparel Design

The pattern development flow is shown in **Fig. 1**. First, a polygon model of the human body is obtained using a 3D scanner and is converted into a voxel model to define the distance field. Next, the left–right difference is removed by detecting the sagittal plane, and then, a torso with a generous margin is generated by changing the threshold for isosurfacing.

Generating a curved surface with a constant distance to an arbitrary curved surface is very difficult because special care must be taken when it comes to the possible changes in the topology. Computing the distance from

the curved surface based on a polygon model would be very expensive, and there would be a high possibility of generating incorrect data structures because of the self-intersection of polygons. On the other hand, if we rely on a volumetric model, it can always generate the correct surface, no matter how much allowance is specified.

After doing this, design lines are drawn to create the basic pattern for the upper body on the torso surface, and curved pattern surfaces are generated as a polygon mesh. Furthermore, a cut-out pattern of the curved surfaces is developed into a flat surface pattern. Finally, the fabric is cut using thick sheeting, sewn, and tried on while confirming any fabric distortion through a virtual fitting; finally, the design quality is evaluated.

3. Volume Model Generation

The core of the current research is a volume model of the human body. We can parametrically change its size by representing the surface of the body as an isosurface from the distance field. Conventional torso models that can be physically resized have been used. However, such a torso model only divides the surface of the body into multiple parts and deforms them mechanically. Our virtual parametric torso can deform while also maintaining a smooth and continuous surface. The volume model is generated based on a real human body that has been 3D scanned. Moreover, the process can be broadly divided into fairing the model as a highly accurate human body model and fairing it as a torso model that is useful for apparel design.

3.1 Body model generation

Although a 3D body scanner digitizes a person's body, the scanned data are generally output as a polygon model. Therefore, the polygon data must be converted into a volume model.

A human body scan is conducted, with the subject somewhat spreading both legs and hands so as not to hide their armpits and the space between their legs. However, human posture is always tilted, and there are also cases where an accurate horizontal reference may not be guaranteed, depending on the measurement method, such as when using photogrammetry. Therefore, the tilt is removed from the scanned data. The original vertical direction of the human body is determined only from the shape data. The trunk establishes the vertical direction of the human body, and this is not heavily affected by their tilt. The results of calculating the center of gravity

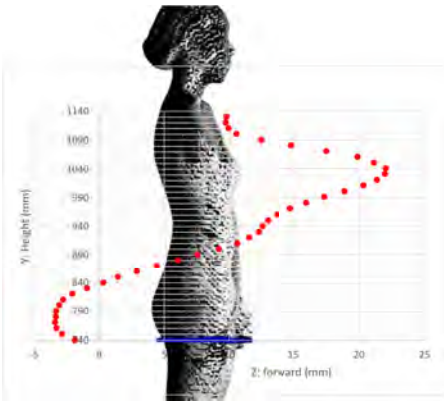


Fig. 2 Gravity centers on the horizontal section position

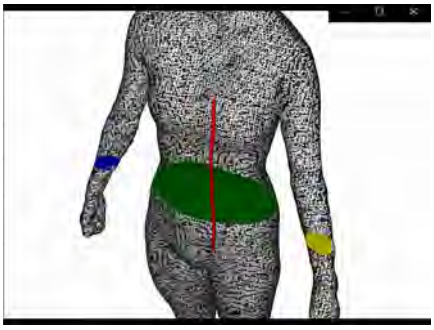


Fig. 3 Detecting the central axis of the trunk

of the cross-sectional shape at various vertical positions of the torso in **Fig. 2** show that the range from the buttocks to the bottom of the bust reflects the vertical axis of the trunk relatively well. Therefore, we extracted cross-sections from two locations at the upper and lower ends of the limited trunk range instead of extracting the tilt from the entire body; we then defined the vertical axis by connecting the plane centers of gravity, as shown in **Fig. 3**.

The tilt of the human body model can be corrected if the vertical axis of the human body is obtained. However, the tilt in the rotational direction around the vertical axis would still remain. This is not a problem when printing with a 3D printer or conducting dressing simulations, but extraction with complete symmetry is essential for generating a torso with an aesthetic sense we pursue. Therefore, we used the symmetry of the trunk, here centering on the vertical axis obtained from the trunk, to determine the angle of rotation needed to correct the frontal direction. Specifically, the horizontal cross-section of the central part of the trunk is extracted, after which the left and right sides are reversed, here assuming a temporary plane of symmetry. The plane of symmetry can be said to be optimal if the overlapping area is the largest when the reversed cross-section and original cross-section overlap with each other, as shown in **Fig. 4**.

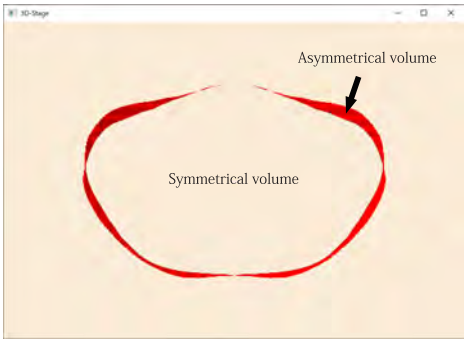
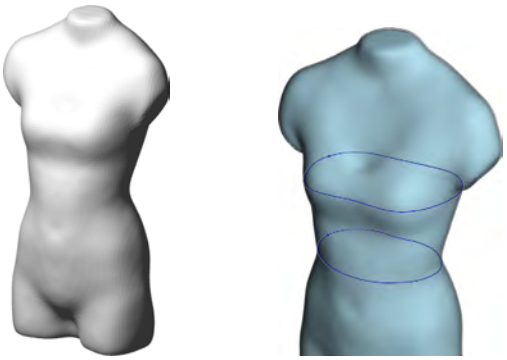


Fig. 4 Searching for symmetry planes on the human body



(a) Smoothing (b) Perimeter preserving

Fig. 5 Denoising the body model

Existing methods that correct the tilt in 3D models generally take the approach of aiming for the model's overall shape (i.e., global optimization). However, a globally optimal solution often cannot be obtained when a prerequisite is the deformation of limbs. Therefore, we conducted local optimization on the trunk of the human body by focusing on the production of an ideal torso model rather than on an entire mannequin. This approach allows for the stable extraction of the torso without relying on the position of the subject's limbs.

The polygon model can be voxelized after correcting the vertical axis tilt and left–right rotation misalignment. In our implementation, the voxel values inside the polygon model are set to 1 and those outside to 0. In addition, for the points on the boundary surface, the intermediate values are used. In other words, these points near the body define the distance field from the surface of the body. The initial polygon model is approximated by an isosurface extracted with a threshold value of 0.5.

The distance from the surface of the body should be found for all voxels when defining a distance field in a general sense. However, finding the shortest distances for positions far away from the surface of the body is a highly computationally expensive process. Therefore, we decided to form a distance field only in the near field close to the surface of the body, here based on the premise of designing an allowance with a reasonable range. The surface can be expanded or contracted within a reasonable range for a human body by changing the threshold value.

However, when voxelization is carried out without adjustment, the actual polygon data include scanning errors and noise because of the unevenness of the model's underwear, which will deteriorate the usability as a torso. To address this issue, a smoothing filter is used for the 3D voxel space in the same way as the denoising filter is used in image processing, as shown in **Fig. 5**(a).

The smoothing process controls fluctuations in the distance field by averaging the voxel values between adjacent voxels. The surface of the body, which is defined as the isosurface, also changes when the voxel value changes. Areas with a more significant local curvature (e.g., minute unevenness) are removed more often, resulting in a human body model with a smooth surface that still retains its basic shape.

Because the smoothing process decreases the spatial frequency of the voxel model, very complicated models topologically come closer to a sphere, and protrusions such as fingers and arms are shaved off. As a result, the body model deforms and deviates from the actually measured shape. However, by adjusting the threshold value for isosurfacing, a body model that retains the surface quality can be generated while maintaining the representative parameter values, such as the boundary length, as shown in Fig. 5(b).

3.2 Torso model generation

Because the human body has distortions in its posture and symmetry, even if the body model closely reproduces the actual body shape, it would lack the aesthetics of a torso for the garment design. To deal with this problem, the sagittal plane of the body is automatically detected from the body model, and the postural sway and symmetry differences of the body shape are removed.

Figure 6 shows the symmetrization process. The left–right symmetry is obtained based on the left–right symmetrical plane of the trunk, which is detected when creating the body model. Adding the symmetrical voxel values across the sagittal plane and updating the voxels on the left and right sides so that they have the same values result in the body model being inverted and superimposed on itself. The tilt of the body model is removed before voxelization; consequently, the shared volume is maximized. The completely overlapping sections have a voxel

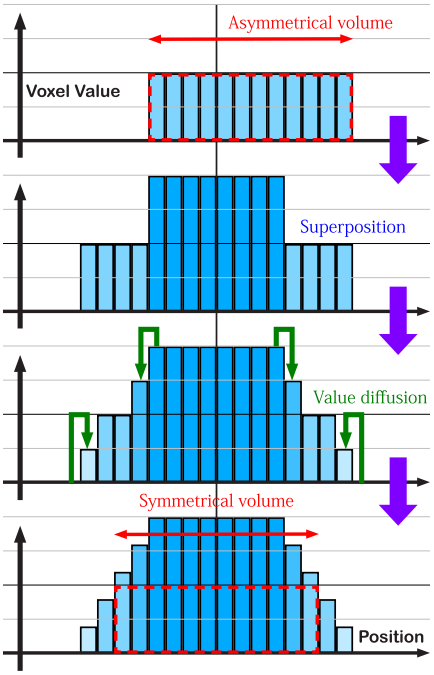
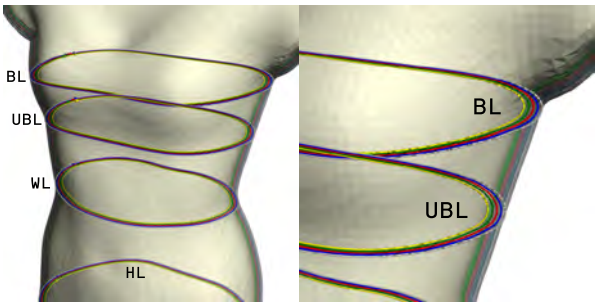


Fig. 6 Symmetrization process

value that is almost double the original, but the nonoverlapping and nonsymmetrical parts will have less values than this. Thus, the entire body shape could be made symmetrical while retaining a complete intersection (i.e., the parts that are originally symmetrical) by smoothing only the areas with low voxel values while maintaining the intersections with high voxel values. Areas outside the body where the voxel values are zero are also excluded from smoothing to prevent the base of the voxel values from widening because of smoothing. In other words, smoothing is conducted only in areas on the human body where the shape is fixed and in the regions between the spaces outside the body. The voxel values are then relaxed and symmetry is established while maintaining the basic shape as much as possible.

However, as with surface smoothing in Fig. 5(a), there will be distortions in the bust shape and other features. Therefore, by adjusting the threshold value, the isosurface is symmetrized while preserving the chest circumference length, as shown in Fig. 5(b).

Furthermore, to create a foundation suitable for designing a basic pattern for the upper body, the threshold value is intentionally shifted, and a slightly inflated torso model is generated. By definition of the distance field, at a position at any height, the closer to 1 the threshold becomes, the smaller the perimeter becomes. In other words, allowance is included in the shape in advance so that the pattern can be directly cut on the torso. **Figure 7** locates four representative cross-sectional feature

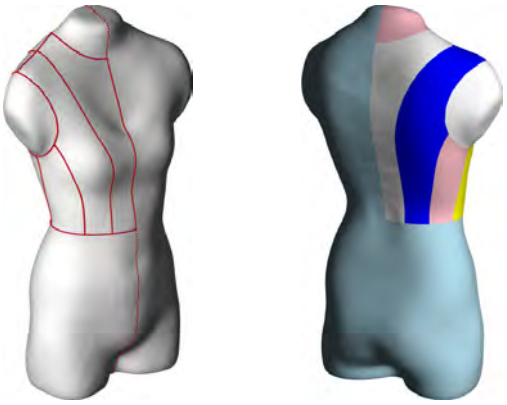


(a) Cross-sectional lines

Feature	Height	Threshold				
		0.3	0.4	0.5	0.6	0.7
Bust Line (BL)	1160	844	829	813	798	780
Under Bust Line (UBL)	1100	738	722	707	691	674
Waist Line (WL)	1010	685	669	654	639	622
Hip Line (HP)	800	928	913	898	883	866

(b) Cross-sectional perimeters (mm)

Fig. 7 Torso with a controllable threshold



(a) Design lines drawn on the body surface (b) Trimmed isosurfaces

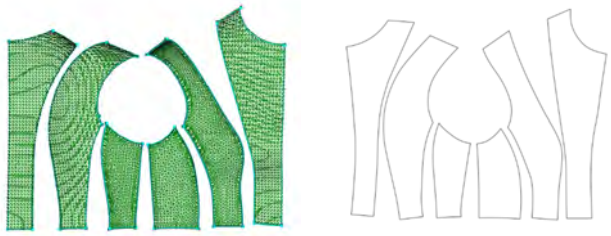
Fig. 8 Dividing by design lines

lines in (a) and lists their perimeter allowances according to different thresholds in (b).

4. 3D Patterning

To obtain the pattern needed to sew a garment, design lines (cut open lines) are drawn on the surface of the torso model, as shown in **Fig. 8**(a). A princess line (a silhouette with a narrow waist and flared hem because of the vertical switching lines) is inserted in the center of the front and back, and other design lines are inserted between the princess line and side. Once the neck and sleeve lines are inserted, the upper body pattern enveloping the upper body is complete. Fig. 8(b) shows the corresponding trimmed isosurface.

The design lines are empirically arranged to decrease the distortions when each of the areas divided by the lines is eventually converted into a flat surface pattern[6].



(a) Trimmed isosurface components (b) Flattened patterns

Fig. 9 Pattern generating process

(a) Before (b) After

Fig. 10 Virtual fitting of the basic pattern

Color	Stretch ratio
None	< 0.01
White	0.01
Yellow	0.05
Red	0.10 <

Fig. 11 Visualization of stretch vectors

Convex areas such as the bust and shoulder blade are considered cone shaped, and the dividing boundary is arranged on the top.

The curved surfaces of the patterns placed on the torso are converted into a cuttable flat surface pattern, as shown in **Fig. 9**. The polygon model that explicitly defines the surface is suitable for deforming the surface, so the pattern drawn on the isosurface is converted into a polygon mesh. Specifically, after the entire isosurface is converted into a polygon mesh by the Marching Cubes method[7], it is trimmed according to the pattern region. The polygons across the pattern boundary are divided along the design lines, and a polygon mesh, which is completely separated for each pattern, is obtained in Fig. 9(a). Although the original polygon mesh has 3D curvatures, a pattern for a flat surface can be developed by applying a simple cloth simulation in Fig. 9(b). Because deformation occurs when the curved surfaces are pressed down into a flat surface to remove the curvatures, stretch distortion of the fabric is unavoidable. However, the stretch distortion is minimized by imposing a mass-spring model with ridge lines and the polygon's vertices. The visible outline of the obtained polygon mesh becomes the basic pattern for the upper body. The geometric expandability is not strictly guaranteed with the mass-spring simulation. However, it is possible to minimize distortions through the design of the pattern line so that some expandability can be ensured.

5. Virtual Fitting

The polygon mesh of the flat surface pattern is always distorted when converting it into a flat surface, so even if the flat surface pattern is resewn into a 3D shape, it will not match the original torso shape. Therefore, it is necessary to perform virtual fitting to quantitatively evaluate the amount of silhouette distortion. **Figure 10** shows the patterns before and after the virtual fitting.

To reproduce the realistic behavior of fabrics, it is necessary to make the density of polygon meshes uniform by remeshing. After obtaining a flat surface pattern, which becomes a visible outline of the polygon mesh, the old polygon mesh is temporarily disposed, and then, the new polygon mesh is repacked. The vertices of the new polygon mesh are filled in using the Poisson-disc distribution[8], and a uniform polygon mesh that connects the vertices with a minimum distance is generated by Delaunay triangulation[9].

A cloth simulator was implemented based on the DressMakingCAD[10]freeware by Tomoyuki Ito, who is a research collaborator of this work[11]. If the pattern is designed on a flat plane, fabric distortion in the wearing simulation is large, and the pattern must often be corrected. However, because this pattern was originally designed in 3D space, fabric distortion in the wearing simulation is minimal, so no correction is needed. In **Fig. 11**, the elongation of the fabric is calculated along the polygon mesh edges and visualized as three-dimensional bars; only bars with elongation greater than 0.05 are shown, and the colors follow the legend. Incidentally, apart from the elastic distortion of the pattern, the distortion of the outline affects the garment's silhouette, so the cross-sectional shapes of the collar, sleeves, and waist are corrected, as necessary.

Table 1 Working times (minutes)

Process	Operation time	Generation time
Posture correction	10	20
Smoothing	10	10
Symmetrization	10	10
3D Patterning	30	5
2D Flattening	10	40
Virtual fitting	30	5

6. Results

Using this system, we created prototypes of apparel items for seven women (F1,...,F7) and one man (M1). The eight individuals were selected for their gender differences and body shape and posture characteristics, even if the differences in BMI were small.

All experiments were conducted on a computer with an Intel Core i7-7700 3.60 GHz CPU with 32 GB of RAM and a Windows 10 x64 operating system. The codes were implemented using Java. **Table 1** shows the operation and generation times. The generation time is longer than expected, but we believe that it can be sped up by making the program multithreaded.

A basic pattern for an upper body of female research participant F1, which was slightly adjusted using the virtual fitting apparel CAD, is shown in **Fig. 12**(a). Generally, a right pattern is prepared for women while a left pattern is prepared for men. Cutting and sewing were performed using thick sheeting. Sewing was performed with a sewing machine, and an open fastener was attached because the centers of the back formed a curved line.

The results of F1 wearing the body prototype pattern are shown in Fig. 12(c). An evaluation questionnaire with four response items regarding wearing the clothing item was carried out: "good", "somewhat good", "somewhat poor", and "poor". The silhouette was evaluated as "good" for all sides: front, sides, and back. Regarding the fit and wearing comfort, the evaluation was "good". In the free description section, responses such as "The silhouette of the chest curvature from the side is pretty. The fit of the sleeve and shoulder lines is appropriate" were given. In addition, the comment "There is a sufficient allowance, and no discomfort when moving" was received.

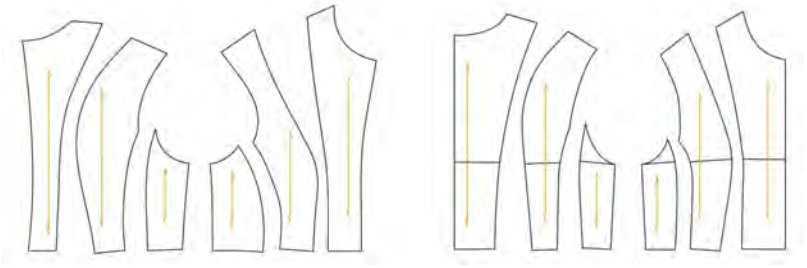
Figure 12(b) shows the body prototype pattern created by draping with the torso we developed for F1. A target line was placed on the torso using a 2 mm IC tape. The draping was conducted by roughly cutting the right front body, right front side, right back body, and right back side using thick sheeting and placing the target lines along the body. In doing so, we did not reserve any allowance. Figure 12(d) shows the results of F1 trying on the garment. The fitting evaluation questionnaire showed the following results: regarding the silhouette, an evaluation of "good" was obtained for the front, sides, and back; regarding the fit and wearing comfort, an evaluation of "good/somewhat good" was obtained; and the free responses included the following: "The silhouette around the chest is very pretty. The garment fit very well along the unevenness from the centers of the back to the sides". "The garment fits perfectly on the body, and there is no wasted space".

The free responses of F1 comparing the our new body prototype in Fig. 12(c) and the traditional torso-based body prototype in Fig. 12(d) included the following: "The new body prototype seemed to fit thinner bodies better. The garment fit the chest curvature very well, and the silhouette was pretty". "There is little difference in the frontal silhouettes of the new and traditional body prototypes, but the side silhouette of the new body prototype is more beautiful".

Comparing the body prototypes in Fig. 12(c) and Fig. 12(d), ours better fits the roundness of the chest in the front. For the back side, ours better fits the body, even though the body shape is prominent around the scapula. We were able to simulate this trend in virtual fitting as well in **Fig. 13**. Furthermore, in the comparison of the sagittal plane area, our prototype has a smaller area of allowance, as shown in **Fig. 14**. In other words, we quantitatively demonstrated that we could use our method to make aesthetically pleasing clothes with a minimal allowance when compared with the conventional 2D patterning.

Therefore, we created body prototypes for two more female research participants (F2, F3) to try on, as shown in **Fig. 15**. As with research participant F1, we carried out a four-step questionnaire. Both of them answered that the silhouette was "good" for all sides: front, sides, and back. The fit was rated as "good" by both, but the wearing comfort was divided between "good" and "somewhat good". The free description responses included the following: "The fit was strong due to the small allowance, but it fit the human body in an upright position very well".

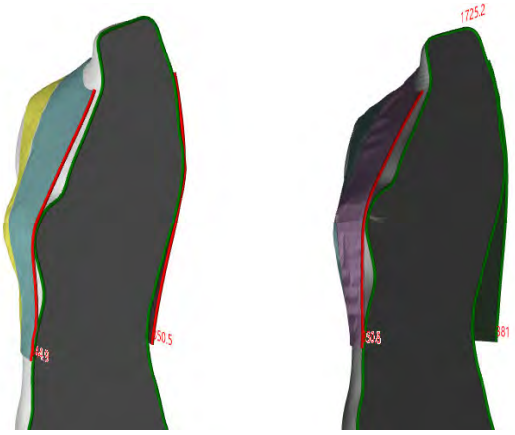
Figure 16(a) shows the men's prototype pattern (left side), which was slightly modified by virtual fitting and



(a) Basic pattern (Ours) (b) Basic pattern (Draping)

(c) Wearing the basic pattern (Ours) (d) Wearing the basic pattern (Draping)

Fig. 12 Evaluations by women's try-on

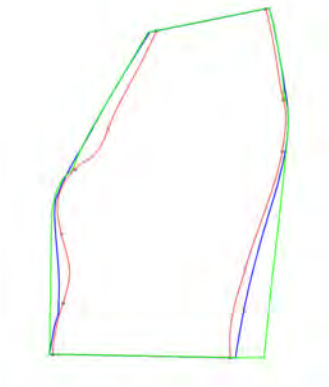


(a) Ours (b) Draping

Fig. 13 Simulation of the fitting

Basic pattern	Cross-section	Gap
Bunka-style	60804.69	7968.20
Ours	56967.68	4131.19
Original body	52836.49	

(a) Cross-sectional areas (mm^2)

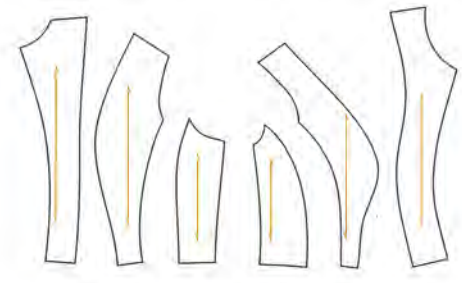


(b) Cross-sectional lines

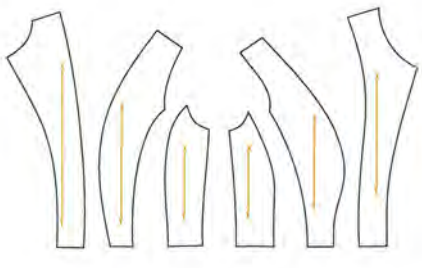
Fig. 14 Evaluation of the fitting

apparel CAD. Figure 16(c) shows the results of participant M1's (first author of this paper) fitting. The fitting evaluation questionnaire showed the following results: regarding the silhouette, an evaluation of "good" was obtained for the front, sides, and back; regarding the fit and wearing comfort, an evaluation of "good" was obtained. The free responses included the following: "The area around the waist was well done". "The overall silhouette was clean".

A men's vest[12]pattern is shown in Fig. 16(b), where drawing diagrams of the vest based on the men's basic pattern are shown on the left, and the development diagram shown on the right is a completed vest pattern. The results of M1 wearing the vest are shown in Fig. 16(d). In the evaluation questionnaire regarding wearing, the silhouette was evaluated as "good/somewhat good" for all sides: the front, sides, and back. Regarding the fit and wearing comfort, the evaluation was also "good/somewhat good". In the free description section, the responses were as follows: "Overall, the allowance is just right; the height seems slightly longer", "It fits my sloped shoulders", and "There is no unnecessary tight feeling".

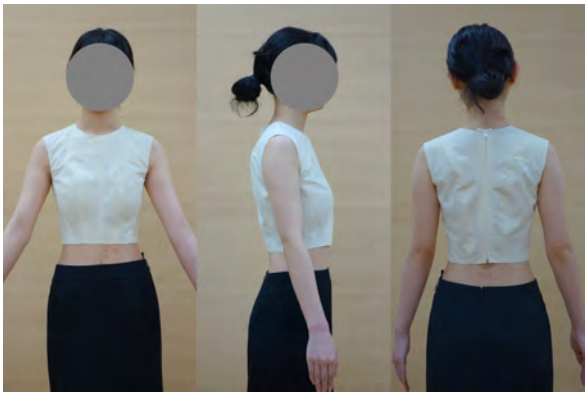


(a) Basic pattern for F2



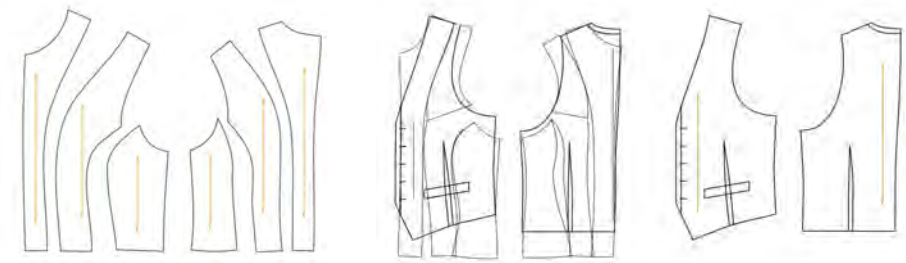
(b) Basic pattern for F3

(c) Wearing the basic pattern for F2



(d) Wearing the basic pattern for F3

Fig. 15 Evaluations by women's try-on (Ours)



(a) Basic pattern

(b) Vest pattern

(c) Wearing the basic pattern

(d) Wearing the vest

Fig. 16 Evaluation by men's try-on (Ours)

The men's vest shown in Figure 16(d) had just the right amount of allowance and resulted in a beautiful silhouette. The vest fits so well without trying it on because of the appropriateness of the original prototype data for the vest.

In addition, we made prototypes for more research participants (F4,...,F7) and performed virtual fittings. **Figure 17** shows five prototypes with distinctive body shapes. The rows show the simulation results for the same research participant, while the columns show the

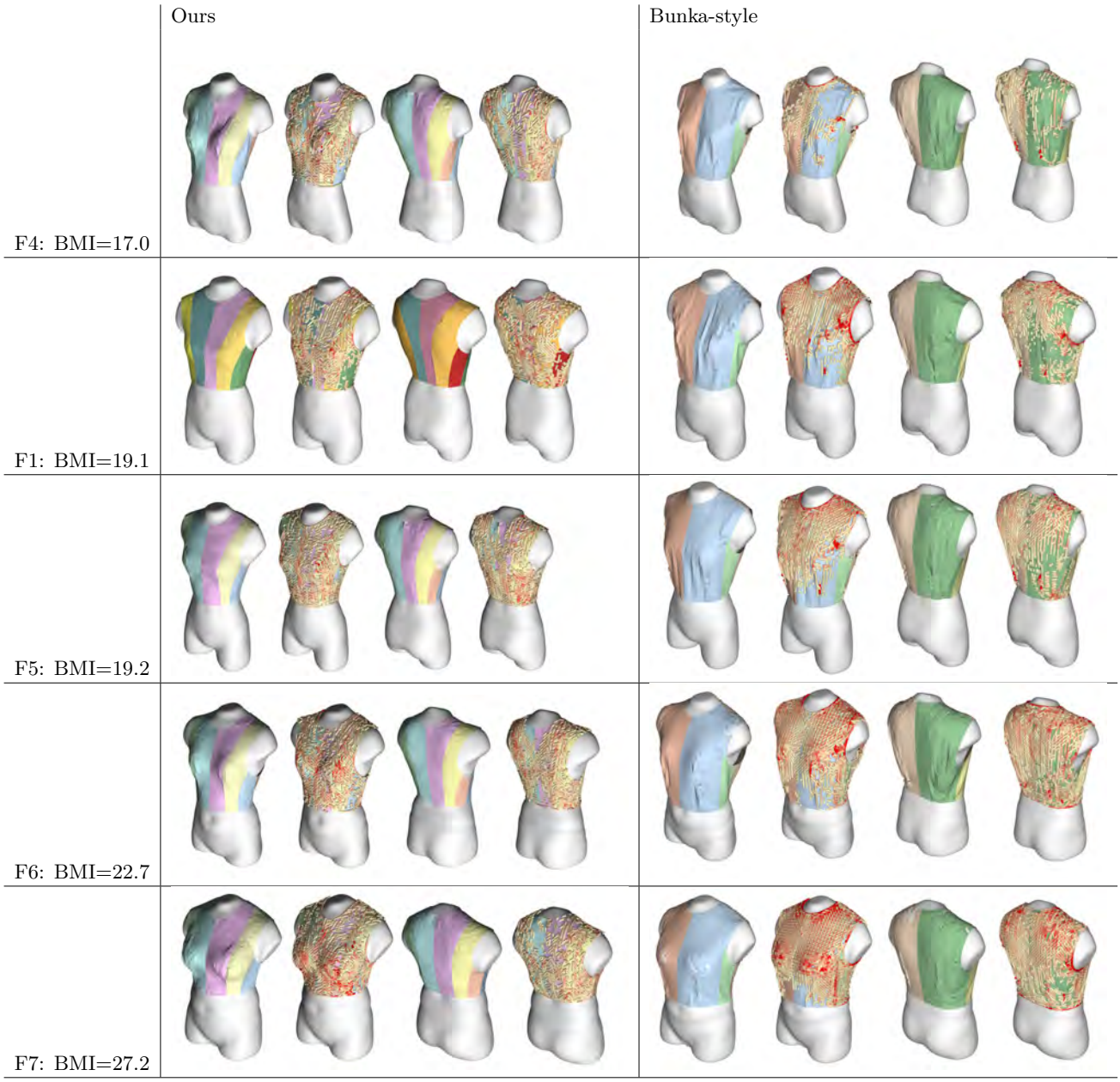


Fig. 17 Virtual fitting of various body types

different patterning methods. The four images for one garment show the wrinkles and the degree of stretching in the cloth as seen from the front and back of the human body. Many wires drawn on the surface indicate the expansion and contraction of the fabric in terms of color and direction. The colors used here follow the definition in Fig. 11.

Note that BMI is used to show the degree of obesity. The simulation shows that the clothes designed with our method fit the body better and wrinkled less than those with the traditional Bunka-style[13)]2D drawing. In addition, the conventional draping was less likely to cause stretching of the fabric because of the large overall allowance, but in the case of characteristic body shapes, it tended to stretch a great deal locally.

7. Conclusion

In the current paper, a high-quality volumetric body model for which noise was eliminated was created. Then, a highly aesthetic volumetric torso model with little symmetrical differences of the body and with a strictly designated allowance by thresholding the 3D distance field was generated. After designing a pattern for an ideal 3D form on the adaptive torso model, a flat surface pattern with minimal distortion was expanded. A basic pattern for the upper body and a men's vest were prototyped, and try-on evaluations were carried out to make sure there was enough room for each of the body types. In future research, we would like to improve the volume of the torso further and produce various apparel items.

Acknowledgments

We would like to sincerely thank Tomoyuki Ito for his collaboration in programming and the staff at BMD System Corporation for their cooperation in the 3D body measurements. This work has been partially supported by JSPS KAKENHI Grant Number JP19K02337.

References

1) DressingSim LSX (LookStailorX), http://www.digitalfashion.jp/new/product/dressingsim_lsx/ (2005).
2) Z. Stjepanovič, P. Tanja, A. Rudolf, J. Simona: "3D Virtual Prototyping of Clothing Products", Innovations in Clothing Technology and Measurement Techniques, Technical University of Lodz, pp.28–41 (2012).
3) CREACOMPO2 Torso, https://www.toray-acs.co.jp/products/creacompo2/patternmagic2-3d/ (2014).
4) N. Umetani, D. Kaufman, T. Igarashi, E. Grinspun: "Sensitive Couture for Interactive Garment Modeling and Editing", ACM Trans. on Graphics, Vol.30, No.4, pp.1–12 (2011).
5) T. Yamamoto, M. Nakayama, I. Fujishiro: "Development of Individual Torso Based on 3D Distance field", Abstracts of the 72nd Annual Meeting of Home Economics of Japan, p.7 (2020) (in Japanese).
6) T. Masuda, H. Imaoka: "Estimation of Body Form Factor for Developing Trunk Surface (Part 2), Comparison of Tight-Fitting Pattern of Front Bodice with Development Pattent of Front Trunk", Journal of Home Economics of Japan, Vol.47, No.4, pp.343–355 (1996).
7) W. Lorensen, H. Cline: "Marching Cubes: A High Resolution 3D Surface Construction Algorithm", Proc. of ACM SIGGRAPH Computer Graphics, Vol.21, No.4, pp.163–169 (1987).
8) A. Ahmed, H. Perrier, D. Coeurjolly, V. Ostromoukhov, J. Guo, D. Yan, H. Huang, O. Deussen: "Low-discrepancy Blue Noise Sampling", ACM Trans. on Graphics (Proc. of SIGGRAPH Asia), Vol.35, No.6, Article No.247 (2016).
9) J. Guo, D. Yan, G. Bao, W. Dong, X. Zhang, P. Wonka: "Efficient Triangulation of Poisson-Disk Sampled Point Sets", The Visual Computer, Vol.30, No.6–8, pp.773–785 (2014).
10) DressMakingCAD, https://xn--6xw240d.net (2015) (in Japanese).
11) K. Choi, H. Ko: "Stable but Responsive Cloth", ACM Trans. on Graphics, Vol.21, No.3, pp.604–611 (2002).
12) T. Kaneko: MEN'S Clothes for All Seasons, NIHON VOGUE Corp. (2015) (in Japanese).
13) Bunka Fashion College (Ed.): Garment Design Textbook (1) Fundamentals of Garment Design, Bunka Publishing Bureau, (2009) (in Japanese).

(Received May 31, 2021)
(Revised December 14, 2021)

Masanori NAKAYAMA (*Member*)
He is currently working as the chief priest of temple "Ankokuin" in Chiba Prefecture and as the CEO/artist of the company "LUXOPHIA", which produces and promotes computational art. He received his B.E. and M.E. at Keio University in 2002 and 2004. His current research interests include photoreal rendering, free-form surface modeling, data processing based on spherical geometry, panoramic display, stereogram, apparel CAD education, and 3D human body measurement. He is a member of IPSJ, IIEEJ, and SAS.

Takami YAMAMOTO (*Member*)
She is an Associate Professor in the Department of Fashion and Art at Wayo Women's University, Chiba, Japan. She completed a doctoral program at the graduate school of Ochanomizu University and received a Ph. D. in Integrated Science in 2006. Her current research interests include clothing construction education, apparel CAD education, and 3D human body measurement. She is a member of IPSJ, IIEEJ, JES, JSET, SAS, The Japan Society of Home Economics, and The Japan Research Association for Textile End-Uses.

Issei FUJISHIRO
(*Honorary Member/Fellow*)
He is currently a professor of information and computer science at Keio University, Yokohama, Japan. He received his Doctor of Science in information sciences from the University of Tokyo in 1988. His research interests include modeling paradigms and shape representations, applied visualization design and lifecycle management, and smart ambient media with multi-modal displays. He is a member of Science Council of Japan, a fellow of JFES, an honorary member of IIEEJ, a fellow of IPSJ, and a senior member of IEEE and ACM. He is a 2021 inductee to IEEE Visualization Academy.

IIEEJ **Paper** *-- Special Issue on Journal Track Papers in IEVC2021Part II --*

Invertible Fingerprint Replacement for Image Privacy Protection

Kazuya NAKAMURA†, Shugo YAMAGUCHI†, Hideki TSUNASHIMA†, Shigeo MORISHIMA††(*Member*)

† Waseda University, †† Waseda Research Institute for Science and Engineering

<**Summary**> The demand for privacy protection has been increasing with the widespread use of devices that can easily take high-resolution images, such as digital cameras and smartphones. In particular, fingerprint information is one of the targets for privacy protection, but there is no research which specifically deals with fingerprint information removal to our best knowledge. In this paper, we propose a method for reversibly replacing fingerprints in an image with another fake fingerprint. This method makes it possible to automatically remove the original fingerprint information in the input image and generate a natural image. Moreover, the input image ' s fingerprint information can be easily restored from the output image only by specific persons who know the key used in the image generation process. In conclusion, we confirmed that the fingerprint information removal and restoration methods are effective by using the fingerprint authentication model.

Keywords: image processing, privacy protection, fingerprint, texture synthesis

1. Introduction

As digital cameras capable of taking high-resolution images become widespread, there is an increasing risk that fingerprint information is included in images. When you take images that include your hands, your fingerprint information can be unintentionally included in the images. Ogane *et al.*[1]pointed that fingerprint information can be obtained from a finger image taken with a digital single-lens reflex camera from a distance of 3 m. There is a risk that your fingerprint can be maliciously used when you share images containing fingerprints on SNS. To avoid such a risk, there is a demand for a method that can easily remove fingerprints in images. Orekondy *et al.*[2] proposed the method to redact privacy content including fingerprints, but generated images are unnatural because privacy regions are blacked out. Sharing unnatural images is not desirable for SNS users, so the required naturalness is that the appearance of images is not changed as much as possible. Applying a Gaussian filtering method is not secure because the original fingerprint can be restored by a deblurring method such as deconvolution[3] . Therefore, a method for completely removing personal information and generating natural images is needed. Yu *et al.*[4]proposed a method which detects the privacy content in images and replaces such content with privacy-protected content generated by Generative Adversarial Networks (GAN)[5]. This method mainly deals with faces and cars as privacy content. However, there is no method that replaces fingerprint information in color images and generates natural images for privacy protection.

In this paper, we propose an automatic removal of fingerprint information in images and a natural fake fingerprint image generation method. It is possible to remove the fingerprint information in an image by replacing the original fingerprint with a generated fake fingerprint. Our method can generate a natural fingerprint image that has fingerprints' fine structure and retains the color appearance by considering the global color distribution of the fingertips in the original image. Furthermore, our method can easily restore the fingerprint information of the input image from the output image by using the key information used in the fingerprint replacement process. Only persons who know the key information can restore it by referring the least significant bit (LSB) substitution, which is the basic digital watermarking method. Our method can be applied to criminal investigations if persons with specific authority manage this key. The police can use the fingerprint information in the image on SNS for the investigation when they access the key information. In this paper,

(i) We propose a novel framework to remove original fingerprint information in images for privacy protection.

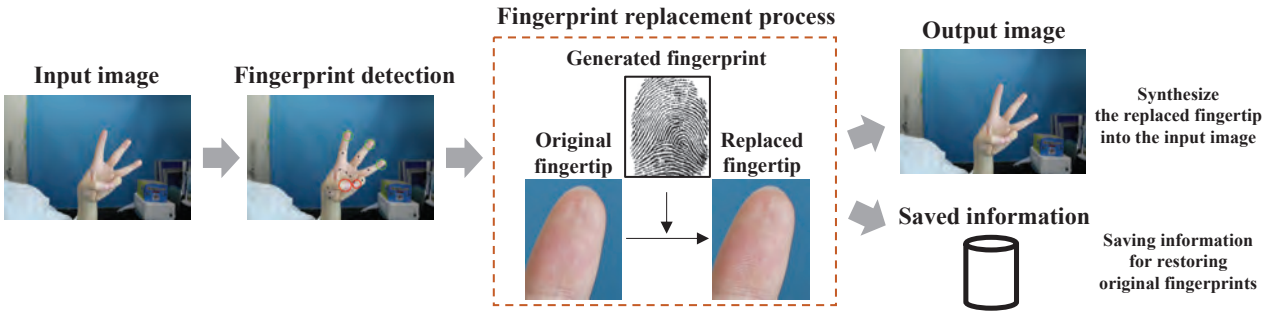


Fig. 1 The overview of our fingerprint replacement procedure

(ii) We propose a novel algorithm that replaces fingerprint information in color images with that of fake generated fingerprints and synthesizes natural fingerprint images retaining the color appearance.

(iii) We propose a novel method that allows only persons who have the key information to easily restore original fingerprint information from synthesized fingerprint images using a binary random number table.

2. Related Work

(a) Fingerprint Removal

Basic methods for removing general privacy content are blurring, pixelating, and masking[6]. Ribaric *et al.*[7]stated that similar methods could be applied to fingerprint de-identification. Recently, inpainting or replacement with privacy-protected content is adopted for removing general privacy objects[4)8)]. However, a recent study about removing privacy content, including fingerprints in images, adopts the masking method[2)], so output images are unnatural. We are reluctant to use the inpainting method for removing fingerprints because it may be difficult to represent natural fingerprint images due to the fingerprints' fine structure. Therefore, generating fingerprint images is necessary for a replacement method to prepare natural fingerprints.

(b) Fingerprint Generation

Study about fingerprint generation[9)], which aims mainly at fingerprint dataset augmentation, deals with generating monochrome fingerprint images. We hesitate to directly replace fingerprints in images with these monochrome images for privacy protection because the output images will be unnatural. There are several methods using deep learning frameworks[10)11)]such as GAN[5)] or Variational Auto-Encoder (VAE)[12)]. We can generate color fingerprint images and remove the original fingerprints in images if we use a color fingertip image dataset with the above deep learning framework. However, to our best knowledge, there is no color fingertip image dataset and it takes a high cost to make such a dataset. Therefore, it is difficult to directly generate color fingerprint images by GAN[5)]or VAE[12)].

We propose a unique method that is focused on removing fingerprint information in color images for privacy protection. Our method replaces the fingerprints' information in the images with that of fake generated fingerprint images, and this enables us to remove the original fingerprints' information, retaining the color appearance.

3. Method

We input an image including fingertips and generate a natural image by replacing the original fingerprint with a generated fake fingerprint. **Figure 1** shows an overview of our fingerprint replacement procedure. Firstly, it is determined whether the fingerprint exists in the input image for each finger. Next, for each finger of the removal target, a natural image is generated in which the fingerprint information is replaced with generated fake fingerprint information. Then, we saved the information necessary for restoring the original fingerprint information.

3.1 Fingerprint detection

We localize the detection area by the bounding box of the hands in the input image, and detect the feature points around the bounding box by OpenPose[13)]. Our experiment is based on the method by Ortega[14)]. Next, we determine whether the fingerprint exists in the image. Then, we judge the front or back of the hand by calculating the cross product of the vector from the palm to the index finger $\boldsymbol{V_1}$ and the vector from the palm to the little finger $\boldsymbol{V_2}$. The front or back of the hand can be identified by the positive or negative of the z component of this cross product. Next, we determine whether the fingerprint exists in the image if the input image includes the front of the hand. For each finger, we calculate the inner product of the vector from the palm to the base of the finger $\boldsymbol{v_1}$ and the vector from the first joint of the fin-

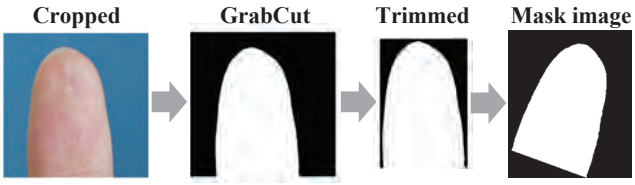


Fig. 2 The process of making a mask image

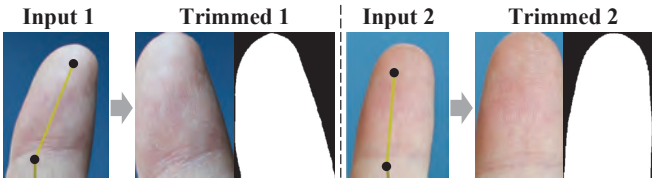


Fig. 3 The result of determining fingertip areas when the keypoints cannot be detected accurately

ger to the fingertip v_1'. If this inner product is positive, the finger is detected as a target for fingerprint removal.

3.2 Preparation for fingerprint replacement

The generated fake fingerprint used for the replacement should be different from the input fingerprint. To prepare such a fingerprint image, we use Finger-GAN[10], which can generate various grayscale fingerprint images. We use the binarized image of the generated fake fingerprint image in the fingerprint replacement process. We prepare several generated fake fingerprints and select one whose fingerprint information is different from that in the input image because the selected fake fingerprint should not be identified as the same fingerprint in the input image. If a generated fingerprint matches another fingerprint that exists in the real world, it does not matter as long as the generated fingerprint does not match the original fingerprint in the input image.

In addition, determining fingertip areas is required in the fingerprint replacement. **Figure 2** shows the process of making a mask image for determining fingertip areas. At first, we rotate the input image so that the direction of the finger is vertical, and crop the image. Next, we make a mask image by GrabCut[15]and trim the mask image, referencing the white area. Finally, we rerotate the mask image and we can determine the fingertip area in the input image. That is, we use a coarse-to-fine approach that fingertip images are roughly cropped and fingertip areas are accurately determined. This coarse-to-fine method enables us to determine the fingertip areas accurately, regardless of the accuracy of OpenPose[13]. **Figure 3** shows that our method can determine fingertip areas even if the keypoints of hands cannot be detected accurately.

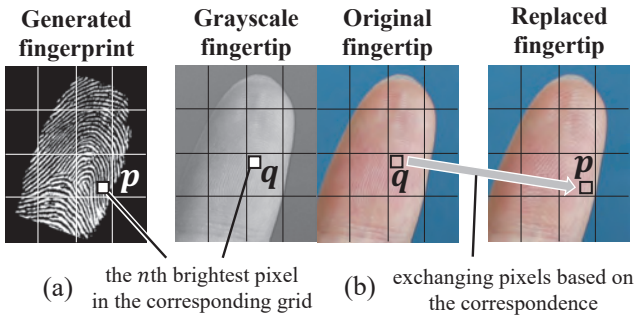


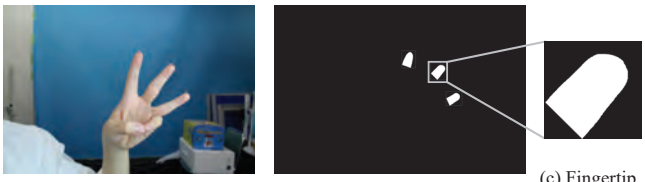
Fig. 4 The fingerprint replacement overview

3.3 Fingerprint replacement process

The fingerprint replacement process is applied to each target finger for fingerprint removal. For generating a natural image, we replace only the input fingerprint information with the generated fake fingerprint information, preserving the color appearance. **Figure 4** shows an overview of the fingerprint replacement process. In this process, we use the fingertip image, its grayscale image, and the generated fake fingerprint whose luminance is inverted from the Finger-GAN[10]output. Before the fingerprint replacement process, we divide these three images into small grids so that we can execute the replacement process locally and generate a natural image that retains the color appearance. Moreover, we smooth the generated fake fingerprint image with a Gaussian filter because the correspondence in the following step (a) of Figure 4 will be more effective. In addition, the target of the fingerprint replacement process is only the area in the fingertip.

The fingerprint replacement process is as follows. Firstly, the pixels correspond according to the order of the luminance of the generated fake fingerprint image and the grayscale image inside the fingertip area of the corresponding grid area (Figure 4 (a)). The pixel p having the nth highest luminance in a specific grid area in the generated fake fingerprint image corresponds to the pixel q having the nth highest luminance in the corresponding grid area in the grayscale image. Secondly, based on the correspondence between the pixels, the pixel p of the replaced image is replaced by the pixel q of the original image (Figure 4 (b)).

These two steps are performed for all pixels in the fingertip area estimated by GrabCut, not for pixels in the background area. This process means that the white pixels of the generated fake fingerprint representing the ridge of the fingerprint correspond to the pixels having a high luminance in the grayscale image. On the other hand, the



(a) Input image (b) Mask image (c) Fingertip mask

Fig. 5 Fingertip mask image of the input image

black pixels of the generated fake fingerprint representing the groove of the fingerprint correspond to the pixels having a low luminance in the grayscale image. After that, the fingerprint information is replaced by exchanging the pixels based on the correspondence. Furthermore, by locally performing this pixel exchanging process, it is possible to generate a natural image in which the global color distribution of the fingertips is maintained. Finally, by synthesizing the replaced fingertip image with the input image, it is possible to generate an image that the fingerprint information of the input image is replaced with that of the generated fake fingerprint image.

3.4 Saving information for restoring the original fingerprint

Since the output image is generated by shuffling the input image pixels, the fingerprint information of the input image can be restored by the reversed pixel shuffling operation on the output image. The reversed pixel shuffling operation requires both the target area of pixel shuffling and the pixel movement information. An example of the pixel shuffling area in the fingertip image is shown in **Figure 5**. White pixels in the fingertip binary mask image (Figure 5(b)) are the target of pixel shuffling in the fingerprint replacement process. By saving two types of information, the fingertip area and the pixel movement in the fingerprint-replaced image, it is possible to restore the original fingerprint information.

In this paper, in order to save the information necessary for restoring the input image, we referred to the LSB substitution method, which is the basic digital watermarking method. The LSB substitution method replaces the least significant bits with the saved information. That is, the luminance in an image is converted to an odd number where the binary saved information is 1, while it is converted to an even number where the binary saved information is 0. However, when considering the case where the binary mask of pixel shuffle area like Figure 5 (c) is saved by using the LSB substitution-like method, there is a problem that anyone can specify the pixel shuffle area(**Figure 6** (a)). This is because the parity of the luminance in the output data represents the saved binary mask of pixel shuffle area.

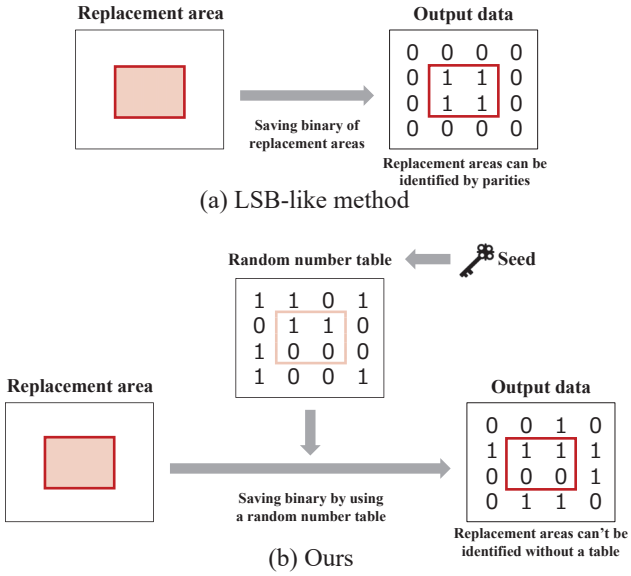


Fig. 6 Comparison of the simple LSB substitution-like method and our saving method

To solve this problem, we propose a novel LSB substitution-like method using a random binary number table. Instead of saving the binary representation information to be saved as it is, when the information to be saved is 1, the value of the random binary number table is saved as it is. On the other hand, when the information to be saved is 0, the value obtained by inverting 0 and 1 of the random binary number table is saved (Figure 6 (b)). The random number table allows us to restore the saved information in the proposed method. The saved information is proved to be 1 when the parity of the luminance in the output data and the parity of binary numbers in the random number table are matched. Otherwise, it means that the saved information is 0.

We consider that this proposed method is far more secure than the LSB substitution method because the saved information cannot be restored without the random number table. Furthermore, when a random number table is generated using a pseudo-random seed, only a specific person who knows this seed can easily restore the information. The pseudo-random seed plays a role in the fingerprint restoring process as the key information.

In this saving process, we save both the fingertip binary mask image like Figure 5 (c) and its position in the whole mask image like Figure 5 (b). Moreover, we also save the pixel movement information, which includes the grid size and the change of pixel positions in each grid.

Figure 7 shows examples of pixel shuffling in a grid

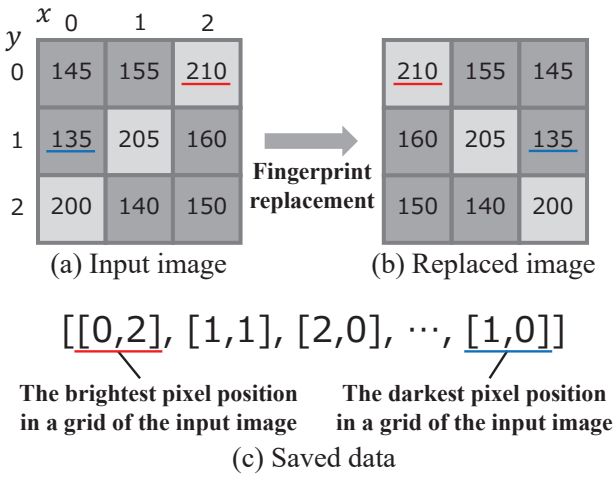


Fig. 7 Method of saving the pixel movement information; (a)(b) a grid in the fingerprint replacement process before and after this process, respectively, and (c) the saved data of the grid

and saved data. Figure 7(a) and Figure 7(b) represent pixel shuffling in a grid and the grid size is 3. Figure 7(a) means a grid in the input image and Figure 7(b) means the corresponding grid in the replaced image. For restoring the original information, we need where each pixel in the replaced image existed in the input image. Therefore, we save the pixel coordinate information in the input image like the saved data in Figure 7(c). The number of elements in the saved data of each grid is the square of the grid size (for example, the number of elements is 9 in Figure 7). The nth element in the saved data is the position (y, x) of the nth brightest pixel in the grid. By using the fact that the order of the luminances in each grid does not change, the amount of saved information is reduced, and efficient encoding is performed. Finally, we save the binarized data of both the grid size and the saved data of each grid as the pixel movement information.

4. Experiment Details

We used a Canon EOS Kiss X5 (the image size is 5184 × 3456) to capture input images and the input images' format is JPEG. For the learning of Finger-GAN[10], we used 1680 grayscale fingerprint images obtained by cropping the center of the DB2_A in FVC2006[16].

The grid in the fingerprint replacement process is required to be a minimum area that contains both ridges and grooves of the fingerprint. Therefore, we set the grid size as the distance between ridges estimated from the input image. In this estimation, we used the ratio of the fingertip size to the distance between ridges, where the fingertip size is defined as the distance between the fingertip and the first joint of the finger. This ratio is 51.2, which is calculated by a preliminary experiment. The estimated distance between ridges is calculated as follows,

$$D_r^{'} = \frac{D_f}{\alpha} \tag{1}$$

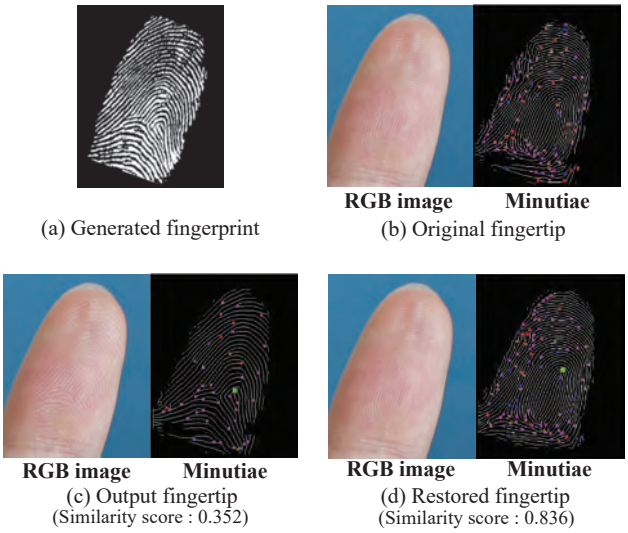
where the fingertip size is D_f, the ratio calculated in a preliminary experiment is α, and the estimated distance between ridges is $D_r^{'}$. For example, when D_f is 256 pixels, $D_r^{'}$ is 5 pixels, so we set the grid size to 5 pixels.

In the fingerprint replacement process, we select a generated fake fingerprint from several prepared fake fingerprint images. Fake fingerprints should have a different distribution of minutiae (fingerprint feature points such as ridge endings and bifurcations) from that of the original fingerprints. Therefore, we use the fingerprint authentication model to select a fake fingerprint. We construct the fingerprint authentication model using 80 fingerprint images in DB1_B in FVC2002[17], referring to FingerPrint Matching[18]. The inputs of the model are two fingerprint images, and the model calculates the similarity score between the input images[19]. It determines whether a pair of fingerprint images are identical based on the similarity score. The definition of the similarity score is as follows,

$$\text{Similarity score} = \sqrt{\frac{m^2}{n_1 n_2}} \tag{2}$$

where the number of minutiae in one input fingerprint image is n_1, that in the other input is n_2, and the number of matched minutiae in the two input fingerprint images is m. This fingerprint authentication model identifies whether the two fingerprints are matched by comparing the similarity score with the predetermined threshold. To determine the threshold, we use FMR (False Matching Rate) and FNMR (False Non-Matching Rate) curves. FMR is the rate at which two different fingerprints are recognized as the same, while FNMR is the rate at which the same two fingerprints are recognized as different fingerprints. We set the threshold to 0.38, where FMR and FNMR are equal[19].

In the fingerprint replacement process, we select one generated fake fingerprint from the prepared fake fingerprint images so that the similarity score between the original fingerprint and the replaced fingerprint is lower than the threshold of the fingerprint authentication model. This means that the replaced fingerprint is not identified as the same original fingerprint in the input image.



(a) Generated fingerprint
(b) Original fingertip
(c) Output fingertip (Similarity score : 0.352)
(d) Restored fingertip (Similarity score : 0.836)

Fig. 8 A comparison of the replaced and restored fingerprints

5. Results and Discussion

5.1 Automatic fingerprint information removal and fingerprint image generation

The output fingertip image in **Figure 8** is the result of the fingerprint replacement process when the proposed method is applied to the input image in Figure 1. We found that the distribution of the minutiae in the original image (Figure 8(b)) is entirely different from that in the output image(Figure 8(c)), as shown in the comparison of the minutiae.

Next, we use the fingerprint authentication model to confirm that the fingerprints of the output image are not identified as the original fingerprints. The fingerprint authentication model is the same as the model described in the preceding section. We confirm the existence of the proper fake fingerprint images so that the replaced fingerprint is not identified as the same as the original fingerprint. We prepare 80 fake fingerprint images generated by Finger-GAN[10]and apply the fingerprint replacement process to five finger images using all the 80 fake fingerprint images. And then, we calculate the similarity score between each replaced fingertip image and the original fingerprint in the input image. The blue histogram in **Figure 9** shows the distribution of the similarity score. Figure 9 indicates that we can get a fake fingerprint which is not recognized as the original person, as we select a generated fingerprint whose score is lower than the threshold of the fingerprint authentication model. However, if we do not select the proper fake fingerprint images regardless of the similarity score, many samples' similarity scores are higher than the threshold. The number of minutiae extracted from fingertip images captured by cameras tends to be larger than that of real fingerprints captured by fingerprint sensors because the minutiae extracted from images captured by cameras include much noise. That is why there are many fingerprint-replaced samples identified as the same as the original fingerprint. Therefore, it is important to select the proper fake fingerprint images used in the fingerprint replacement process. As long as we select a fake fingerprint image properly by using the fingerprint authentication model, the fingerprint information in the input image can be removed automatically, and we conclude that our method is effective in protecting the fingerprint information.

5.2 Restoring fingerprint information in the input image

Compared the original fingertip in Figure 8(b) with the restored fingertip in Figure 8(d), the restored image does not seem to be different from the original image. This is because the pixels inside the fingertips in the output fingertip image can be automatically returned to the original position of the input image by the saved information. The total saved information size is about 0.27MB when we use the input image (4.4MB) in Figure 1. Next, we calculate the similarity score between each fingerprint-replaced image in the preceding subsection and the original fingerprint in the input image. Figure 9 shows that the similarity score of all 400 samples is higher than the threshold. This means that the fingerprint of the restored image is recognized as the same fingerprint in the input image. Note that the similarity scores are not equal to 1 because the luminance of the restored image is changed in the JPEG compression. The result indicates that the fingerprint information in the input image can be restored from the output image.

5.3 Impact of image processing on the proposed method

In a real situation, we sometimes upload images on SNS after image processing such as filtering. Our method is applied to input images after such image processing, so we should investigate the impact of image processing on our method. We apply three types of image processing to the input images and conduct the same experiment in Sec. 5.1 and Sec. 5.2. The first method is Twitter's image filtering named Kilda, which changes the color tone more vivid. We use JPEG images which are uploaded to Twitter after filtering and downloaded as inputs. The second method is the noise reduction of the default Photo application in

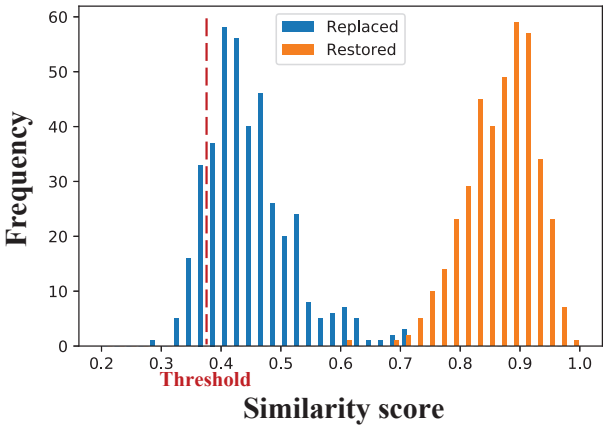


Fig. 9 The distribution of the similarity score in our experiments to check whether fingerprint replacement and restoration are conducted properly

iPhone. The third method is a sharpening filter, which enhances the edges of objects in images.

Figure 10 shows the qualitative and quantitative results of this experiment. In every type of image processing, the replaced fingertip images maintain the color appearance of the input image, and the restored fingertip images do not look so different from the inputs. In addition, these histograms show that our method can protect the original fingerprint information when we choose the proper generated fake fingerprint image whose similarity score is lower than the threshold. Moreover, all samples of the restored fingertip images are recognized as the same fingerprint in the input image. We demonstrate the practicality of the proposed method by using input images operated by several image filters that SNS users are familiar with.

5.4 The naturalness of the output image

Figure 8 shows that we can generate a natural image that retains the global color distribution of the input fingertips because the fingerprint replacement process is locally performed. In addition, there are few artifacts on the boundaries of the grid area because the grid size is very small. However, comparing the original fingertip image in **Figure 11**(a) with the output image in Figure 11(b), we found that the fingerprint of the output image is thicker than the original. This is because the fingerprint images used in the training of Finger-GAN[10] include only the central part of the fingertip, and the generated fake fingerprint is stretched over the entire fingertip in the fingerprint replacement process. That is, the problem of thicker fingerprints stems not from the fingerprint replacement algorithm, but from the fingerprint image dataset.

To confirm this, we conducted an additional experiment. Figure 11(c) shows the output fingertip image replaced with a real fingerprint whose minutiae are different from those of the input image. We found that a more natural image can be generated with a real fingerprint than a generated fingerprint because the fingerprint density of the output image is almost the same as that of the input image. Thus, our method can generate more natural fingerprint images by fake fingerprints with high fingerprint density.

6. An Example of Application

Our method can automatically remove fingerprints in images, so it can be applied to images on SNS for privacy protection. **Figure 12** shows the overview of a realistic scenario for applying our method. At first, when SNS users upload JPEG images, the SNS uploader applies our method to remove fingerprints in the images. Next, the fingerprint-replaced images are converted to the JPEG format and saved to the database of the SNS administrator because they can be displayed on SNS and it takes less time for page rendering as compared with only uncompressed images being saved. Then, SNS users can access the shared images which do not include the original fingerprints. For restoring the original fingerprints, we require where each pixel in the fingerprint-replaced image existed in the input image. Thus, the SNS administrator should also save this information, which is described in Sec.3.4. When the police need the original fingerprints in images for investigations, they can only access the restored images with the permission of the administrator. A limitation of our study is that saving information about pixel shuffling may be a burden to SNS administrators.

In a real situation, there is a potential risk that personal information unintentionally included in shared images is maliciously used. To give an example, a stalker was able to locate an idol's home through shared selfies on social media which include eye reflection[20]. Our study is important to avoid such a potential risk in shared images.

7. Conclusion

In this paper, we proposed an automatic fingerprint information replacement method in color images for privacy protection. We confirmed that the output image obtained by applying the proposed method did not include the original fingerprint information in the input image by using the fingerprint authentication model. Therefore, we conclude that the proposed method is effective in privacy

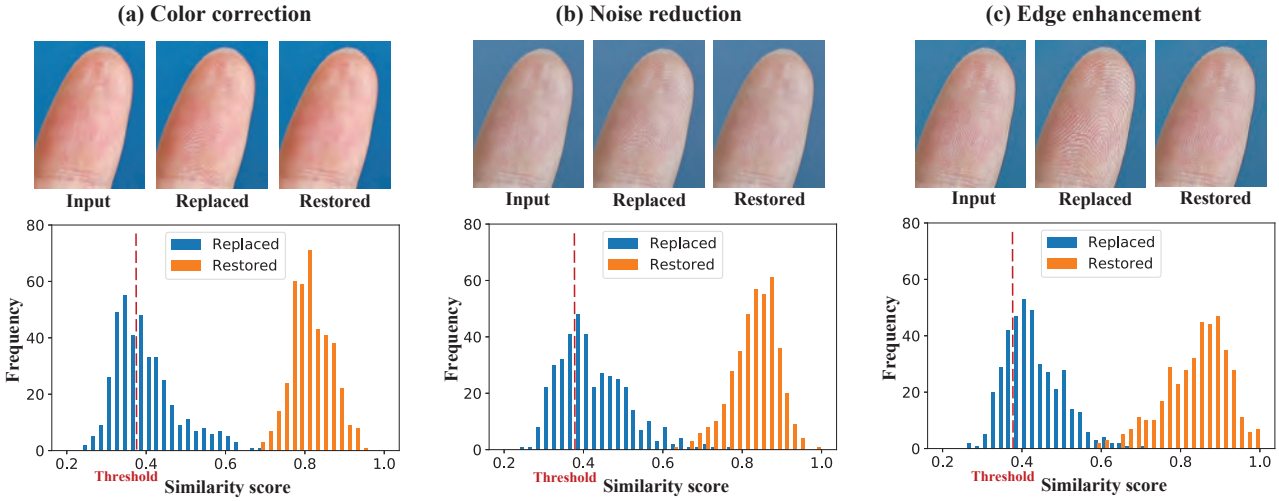


Fig. 10 The results of the same experiment in Section 5.1 and Section 5.2 with processed images as inputs of the proposed method; (a) color correction image filtering in Twitter (b) noise reduction filtering in iPhone, and (c) edge enhancement with image sharpening convolution filter

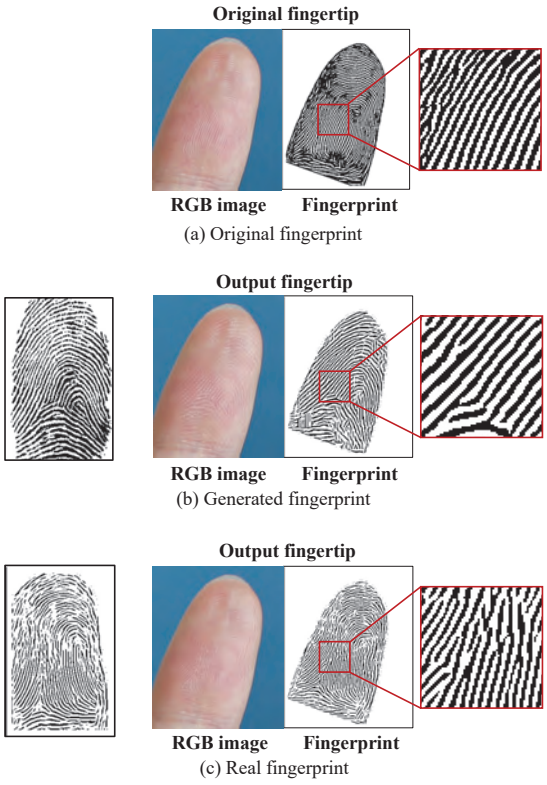


Fig. 11 A comparison of an original image and fingerprint replacement results with fake and real fingerprint

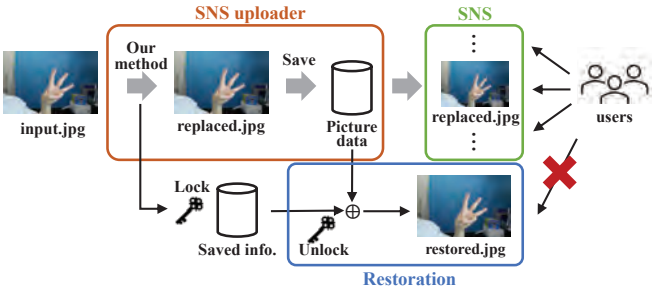


Fig. 12 Application of the proposed method for a realistic scenario

protection. Furthermore, the proposed method allows us to easily restore the original fingerprint information from the output image with the seed value of the random number table. We verified that the restored fingerprint information matched the original fingerprint information by the fingerprint authentication model and showed that it is possible to actually restore the fingerprint information by applying the improved LSB substitution-like method using a binary random number table. Our method is a big step in the research direction of fingerprint replacement because it allows only persons who have the key information to restore the original fingerprint information, unlike simple Gaussian filtering.

Acknowledgement

This research is supported by JST-Mirai Program Grant Number JPMJMI19B2, JSPS KAKENHI Grant Number JP21H05054 and JP19H01129.

References

1) T. Ogane, I. Echizen: "BiometricJammer: Method to Prevent Unauthorized Capturing of Fingerprint in Consideration of Use-friendliness", Proc. of Computer Security Symposium, Vol. 2016, No. 2, pp. 355–362 (2016).
2) T. Orekondy, M. Fritz, B. Schiele: "Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images", Proc. of Conference on Computer Vision and Pattern Recognition, pp. 8466–8475 (2018).
3) R.A. Hummel, B. Kimia, S.W. Zucker: "Deblurring Gaussian Blur", Computer Vision, Graphics, and Image Processing, Vol. 38, No. 1, pp. 66–80 (1987).

4) J. Yu, H. Xue, B. Liu, Y. Wang, S. Zhu, M. Ding: "GAN-Based Differential Private Image Privacy Protection Framework for the Internet of Multimedia Things", Sensors, Vol. 21, No. 1, p. 58 (2020).
5) I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio: "Generative Adversarial Nets", Proc. of the 27th International Conference on Neural Information Processing Systems, Vol. 2, pp. 2672–2680 (2014)
6) J. R. Padilla-López, A. A. Chaaraoui, F. Gu, F. Flórez-Revuelta: "Visual Privacy by Context: Proposal and Evaluation of a Level-Based Visualisation Scheme", Sensors, Vol. 15, No. 6, pp. 12959–12982 (2015).
7) S. Ribaric, A. Ariyaeeinia, N. Pavesic: "De-identification for Privacy Protection in Multimedia Content: A Survey", Signal Processing: Image Communication, Vol. 47, pp. 131–151 (2016).
8) R. Uittenbogaard, C. Sebastian, J. Vijverberg, B. Boom, D. M. Gavrila, P. H. N. With: "Privacy Protection in Street-View Panoramas Using Depth and Multi-View Imagery", Proc. of Conference on Computer Vision and Pattern Recognition, pp. 10581-10590 (2019).
9) R. Cappelli: "SFinGe: an Approach to Synthetic Fingerprint Generation", Proc. of International Workshop on Biometric Technologies, pp. 147-154 (2004).
10) S. Minaee, A. Abdolrashidi: "Finger-GAN: Generating Realistic Fingerprint Images Using Connectivity Imposed GAN", arXiv preprint arXiv:1812.10482 (2018).
11) M. Attia, M. H. Attia, J. Iskander, K. Saleh, D. Nahavandi, A. Abobakr, M. Hossny, S. Nahavandi: "Fingerprint Synthesis Via Latent Space Representation", Proc. of International Conference on Systems, Man and Cybernetics, pp. 1855-1861 (2019).
12) D.P. Kingma, M. Welling: "Auto-Encoding Variational Bayes", Proc. of the 2nd International Conference on Learning Representations (2014).
13) Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh: "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 43, No. 1, pp. 172–186 (2019).
14) M. Ortega, Standalone Hand Keypoint Detector, https://github.com/ortegatron/hand_standalone (2019).
15) C. Rother, V. Kolmogorov, A. Blake: ""GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts", ACM Trans. on Graphics, Vol. 23, No. 3, pp. 309–314 (2004).
16) R. Cappelli, M. Ferrara, A. Franco, D. Maltoni: "Fingerprint Verification Competition 2006", Biometric Technology Today, Vol. 15, No. 7, pp. 7–9 (2007).
17) D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman, A.K. Jain: "FVC2002: Second Fingerprint Verification Competition", Proc. of the International Conference on Pattern Recognition, Vol. 3, pp. 811–814 (2002).
18) V. K. Alilou, FingerPrint Matching: A Simple Approach, https://github.com/alilou63/fingerprint (2021).
19) T. Li: "Fingerprint Identification by Improved Method of Minutiae Matching", Electronic Thesis or Dissertation, Miami University (2017).
20) Newsweek, Stalker Finds Japanese Idol's Home from Reflections in Her Pupils, https://www.newsweek.com/stalker-finds-idol-reflection-pupils-1464373 (2019).

(Received May 28, 2021)
(Revised Feb. 1, 2022)

Kazuya NAKAMURA
He received a B.E. degree in the Department of Pure Physics from Waseda University, Japan. He is currently a student in the Graduate School of Advanced Science and Engineering at Waseda University. His research interests are image processing, computer vision, and privacy protection.

Shugo YAMAGUCHI
He received a B.E. degree in the Department of Applied Physics and an M.E. degree in the Department of Pure and Applied Physics from Waseda University, Japan. He is currently a student in the Graduate School of Advanced Science and Engineering at Waseda University. His primary research interests are in texture synthesis, digital human.

Hideki TSUNASHIMA
He received an M.E. degree in the Dept. of Electrical Engineering and Electronics from Kogakuin University, Japan. He is currently a student in the Graduate School of Advanced Science and Engineering at Waseda University. His primary research interests are in deep generative models, virtual try-on, embodied AI.

Shigeo MORISHIMA (*Member*)
He was born on August 1959. He received his B.S., M.S. and Ph.D. degrees, all in Electrical Engineering from the University of Tokyo, in 1982, 1984, and 1987, respectively. He was a visiting professor of University of Toronto from 1994 to 1995 and an invited researcher of Advanced Telecommunication Research institute from 1999 to 2011. Currently, he is a professor of School of Advanced Science and Engineering, Waseda University. He was a General Chair of ACM VRST2018 and VR/AR adviser of SIGGRAPH ASIA 2018. He is a General Chair of Pacific Graphics 2022. He received many awards and takes an administration board member of several societies.

IIEEJ Paper -- Special Issue on Journal Track Papers in IEVC2021Part II --

Game Development Using the Discrimination System to Improve Typing Skills

Surena KAWAHARA†, Teruaki HIRANO†, Noriyoshi OKAMOTO† (*Member*), Daisuke TAKAHASHI††

† Kanto Gakuin University, ††Bunka Gakuen University

<Summary> Accurate typing in the home position is recommended to type efficiently on a PC. For this reason, a lot of software has been developed to learn typing in the home position. However, few of them can determine the actual fingering simply by presenting the location of the fingers. Therefore, the system to automatically determine have been developed whether the fingering is in the correct home position using a Leap Motion. Using this discrimination system, the Typing Learning Game is constructed that can correct subjects to type in the home position. Thus, by handling the fingering data of these subjects, efficient learning can be achieved. In the experiment, 15 subjects typed 1800 characters each, and the discrimination accuracy of the Fingering Discrimination System was 98.8%, and the fingering was improved by 25.6% from the results of the Typing Learning Game. The results show that the proposed automatic Fingering Discrimination System can accurately determine the fingering of the subject.

Keywords: typing fingering, leap motion, typing learning system

1. Introduction

The rate of personal computer (referred to as PC) use among youth is low level, partly due to the widespread use of smartphones. Few situations occur in society where PCs are not used. For this reason, it is useful for youths to become accustomed to operating PCs before going out into society.

The results of a survey of PC usage in Japan is compared for youth only and for all ages, and the results are shown in **Fig.1**[1), 2)]. The survey for all ages does not distinguish between desktop PCs and laptops. This graph represents the low rate of PC use among youth, despite the comparable rate of smartphone use.

In this paper, fingering is the key word. The meaning of fingering is to specify the finger and hand positions that should be used when pressing a key. Choosing the suitable fingering will improve the speed and accuracy of typing. The home position is one of the chosen fingering where the hand is placed in the home row of the middle of the keyboard and the nearby keys are assigned to each finger placed. In this paper, the fingering as per the home position is considered as correct fingering and the other self-taught fingering is considered as incorrect fingering.

One of the reasons why youths avoid PCs is the difficulty of keyboard input. Unlike smartphones, keyboards require a large number of keys, as well as the use of all fingers. For beginners to learn the keyboard efficiently, correct fingering in the home position is recommended in most case. By entering the same key with the same finger each time, the user can memorize the key position by finger acquisition. Therefore, it is important to learn the home position. However, the problem is people who have become accustomed to inputting with their self-taught fingering. Some people type with just their index finger. Correcting a habit by yourself is not an easy task because it requires a strong consciousness to correct it.

The purpose of this paper is to correct the fingering according to the home position. Thereby, fingering learning environment is proposed. The fingering learning environment consists of two proposed methods, the Typing Learning Game and the Fingering Discrimination System. The Typing Learning Game has the effect of correcting fingering by using a unique control method that utilizes the shooting game. The Fingering Discrimination System automatically evaluates whether the typing is a correct fingering or not. Conventional typing learning software used only two pieces of information: typing speed and typing accuracy. The proposed method improves the learning effect by newly handling fingering information that has not been used in conventional software. Correcting fingering and measuring its effectiveness as an advanced typing learning environment.

The structure of this paper is described as follows. In Chapter 2, related research is referred to, respectively, Section 2.1 for typing learning and Section 2.2 for fingering

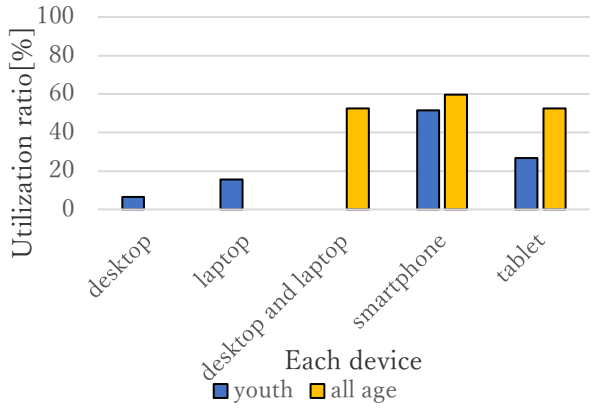


Fig.1 PC usage rates for youths and all ages

discrimination. Chapter 3 describes the proposed method in detail. Section 3.1 details the Typing Learning Game, a method for correcting fingering. Section 3.2 details the Fingering Discrimination System, a method for automatically discriminating fingering. Section 2.1 corresponds to Section 3.1, and Section 2.2 corresponds to Section 3.2 in terms of the relationship between the related research and the proposed method. In Chapter 4, the procedure and results of an experiment is described to measure the effect of the proposed Typing Learning Game on fingering correction and to determine the accuracy of fingering discrimination of the Fingering Discrimination System. In Chapter 5, the results of the experiment are discussed. Chapter 6 gives a conclusion.

2. Related Research

2.1 Related research on typing learning

First, the studies that investigated the significance of the home position are described below. Then, related studies that devise ways to make the home position learnable are described. In the last paragraph, related research is compared with this paper, and the significance of the proposed method is clarified.

Feit et al.'s research demonstrated that home position typing is not necessarily superior[3]. That research was found that pressing the same key with the same finger and using more fingers are more important factors that determine typing speed and accuracy than whether typing in the home position. Subjects who met these two conditions had outperformance typing ability regardless of their home position. However, all the subjects who excel in their own fingering had the correct fingering for F and J, which are the home position reference positions. Also, when a beginner is typing for the first time, the home position is the best way to keep these two conditions. Therefore, even with these results, the home position is still important for beginners.

Takaoka et al. developed a typing learning system for the purpose of acquiring touch typing ability[4]. To make the subject conscious of the home position, the keyboard on the screen is presented with which finger to press. In addition, the test is performed using the typing learning system with the hands hidden by a handkerchief. By hiding their hands, the subject is forced into a home position.

Takakura et al. proposed a typing learning support method using passive tactile sensation[5]. By applying vibratory stimuli to the finger that has the correct fingering according to the character to be input, the subject is encouraged to correct the fingering. The system was able to memorize the subject's key layout.

Although the above related research was effective in learning fingering, they used indirect correction methods such as displaying a guide on the screen, and relied on the subject's own consciousness to correct the fingering. To correct the subject's habit of typing with their own fingering, it is necessary to correct them directly without relying on the subject. The method of giving stimuli to the input finger can correct the habit directly, but it requires wearing special gloves on the hand. Therefore, in this paper, the Typing Learning Game and Fingering Discrimination System is proposed that can directly correct fingering without the subject being conscious of it.

2.2 Related research on fingering discrimination

As a previous research of automatic fingering discrimination, Feit et al. performed fingering discrimination by motion tracking 52 markers attached to the hand with 12 high-speed infrared cameras (already referred to in Section 2.1). In addition, a spectacle-type eye-tracking device was used to discriminate touch typing. In this research, fingering was regarded as a strategy for how to efficiently input the keyboard. The acquired fingering data was used to find out what kind of strategy was superior.

Chigira et al. proposed a typing practice system equipped with a fingering discrimination function by using Leap Motion[6]. The acquired fingering data gives feedback by detecting the weak key and asking a word containing the weak key. The conditions that are judged to be weak keys include not only typos, but also incorrect fingering that cannot be obtained with ordinary typing software.

The related research using motion markers described above is capable of discriminating with very good accuracy. However, it is not easy because of its high cost and complexity, such as the use of multiple expensive high-speed cameras. In this paper, simple and inexpensive system for learning is proposed. In addition, similar related research using Leap Motion differ from this paper in that they do not have a calibration function to obtain the key positions. By performing the calibration by the subject himself, it is possible to take into account the accuracy difference derived from the shape of individual fingers.

3. Proposed Method

The proposed Typing Learning Game is a learning software that has a home position correction effect. Next, the proposed Fingering Discrimination System can automatically discriminate whether the subject's fingering is correct or incorrect. The correlation diagram between the subjects and the proposed method is shown in **Fig.2**. This environment corrects the fingering for the subject and evaluates the fingering. The subject will receive feedback on the evaluation results. The details of the proposed Typing Learning Game are described in Section 3.1, and the details of the proposed Fingering Discrimination System are described in Section 3.2.

3.1 Typing Learning Game with fingering correction

In the Typing Learning Game proposed, the home position is corrected naturally without the subject being conscious of it. For that purpose, the control method of the shooting game is utilized. Correct the fingering by typing the player's movement control.

A captured image of the actual play screen is shown in **Fig.3**. The filled triangle at the bottom of the screen is the player. Assign the player's movement control to the F key for the left direction and the J key for the right direction. These keys are the reference positions for the home position. By using these key assignments, the left and right index fingers are naturally placed on the F and J keys when controlling. Also shoot a bullet with the space key.

Nevertheless, with only the above-mentioned movement control, only 3 keys can be corrected. The highlight function corrects the fingering of other keys. The inverted triangle with the character at the top of the screen drawn in the center is the enemy. The inverted triangle with the letters drawn in the center at the top of the screen is the enemy. As an example, let us assume that Enemy *L* and *K* are in the screen. These enemy with character will not be destroyed when hit. However, when the L key is pressed, a circle is drawn around the enemy *L* and highlighted. If the enemy is highlighted, it can be destroyed by hitting it with a bullet. On the other hand, when Enemy *L* is highlighted, Enemy *K* cannot be destroyed. Press the K key to switch the highlight display from Enemy *L* to Enemy *K*. The last pressed key is retained as the highlight display key. As a supplement, characters other than the movement keys F and J appear as enemies.

Consider the case where the subject operates this game. As the basis of a shooting game, it is necessary to always

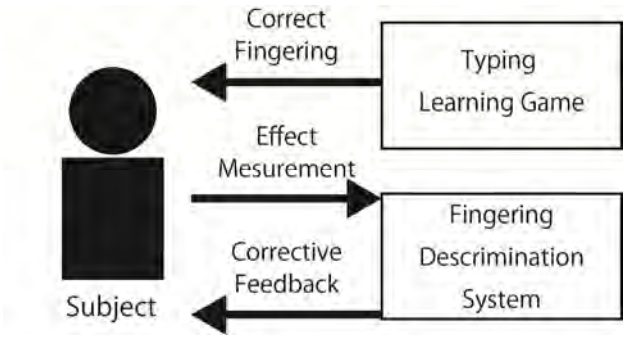


Fig.2 Fingering learning environment

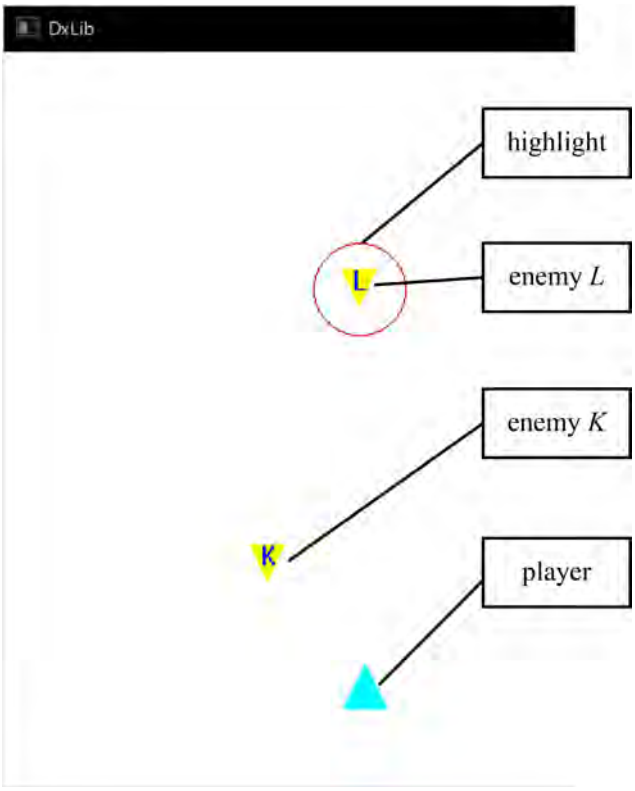


Fig.3 Play screen

move laterally with the F and J keys to align the vertical axis with the enemy and avoid the bullets of the enemy. On top of that, in this game, highlighting must be done. The index finger needs to be placed on the F and J keys as much as possible, which causes constraints on pressing the highlighting keys with other fingers. Due to this restriction, the highlight display key is also pressed with the correct fingering. Therefore, it is possible for the subject to naturally correct the home position by these control methods.

3.2 Fingering Discrimination System

The Fingering Discrimination System proposed in this paper can discriminate fingering in a simpler way. The system discriminates between correct and incorrect fingering as per the subject's home position. Improve learning is aimed by making it possible to newly collect fingering data.

In typing learning, feedback such as the amount of learning and the score to provide subjects is important[7]. The system proposed in this paper shows the subject which key fingering is most likely to be mistaken.

Use Leap Motion is a USB device that can acquire fingertip position information and finger type. This sensor is a high-performance motion sensor specialized for hand tracking. A schematic diagram of this system is shown in **Fig.4**. Using the stand in this way, the sensor is hung from above. The sensor gives a bird's-eye view of the keyboard and the hands that cover it.

To use this system, calibration to obtain the key position coordinates is necessary in the preliminary stage. Have the subject input one character at a time and obtain the fingertip coordinates as the key position. The fingertip coordinates are obtained from the sensor. Calibration allows us to tolerate differences in keyboard dimensions such as key pitch to some extent.

There is a verification to find the actual spatial resolution of Leap Motion[8]. As a result of verification, a spatial resolution of 1.2 mm on average was obtained. This is sufficient accuracy when the key pitch of the keyboard is considered.

The fingering discrimination method is described below. First, the system recognizes that a key has been pressed. The fingertip closest to the pressed key is the finger that pressed the key. Then, the system obtains the finger from the sensor. If the relationship between the key and the finger is the same as the home position, the fingering is correct; otherwise, the fingering is incorrect. For example, if the A key is pressed by the little finger of the left hand, discrimination by the system is a correct fingering, and if it is pressed by any other finger, discrimination is an incorrect fingering.

4. Experiment of Proposed Methods

4.1 Experimental procedure

The accuracy experiment of the Fingering Discrimination System and the effect measurement of the Typing Learning Game were carried out at the same time in one flow. The number of subjects is 15 (A to O).

As a beginning, subjects were asked to enter 300 Roman characters from *A* to *Z*, one by one, not as words. At that time, the Fingering Discrimination System measures the subject's fingering while typing. Next, a Typing Learning Game is played to correct the subject's fingering. After playing, the 300-character test is performed again. 3 sets of this procedure will be performed at weekly intervals, and the patient will be monitored. The 15 subjects inputting 1800 characters ((300 + 300)×3). The flowchart for one set is shown in **Fig.5**. Each set took about 20 minutes, and there were no restrictions on the subjects.

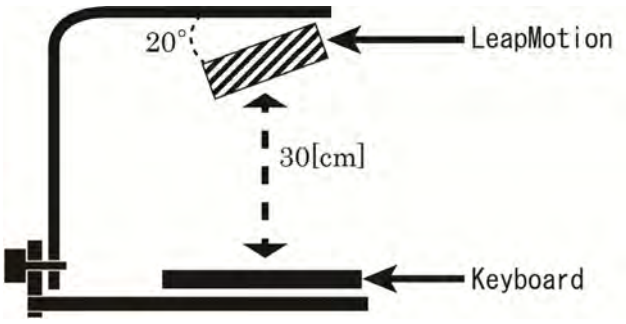


Fig.4 Outline of Fingering Discrimination System

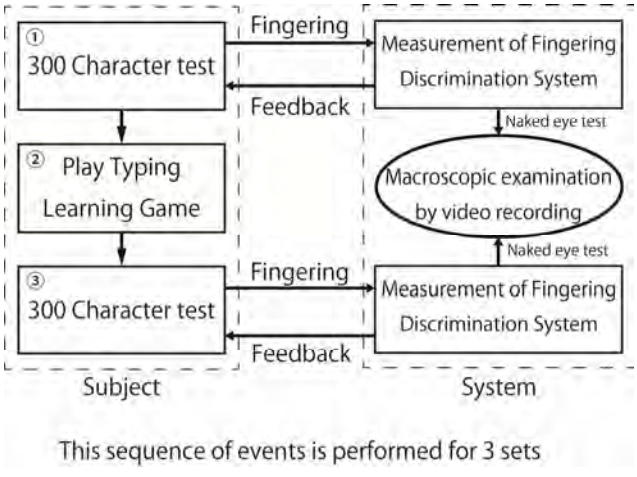


Fig.5 Flow of a procedure

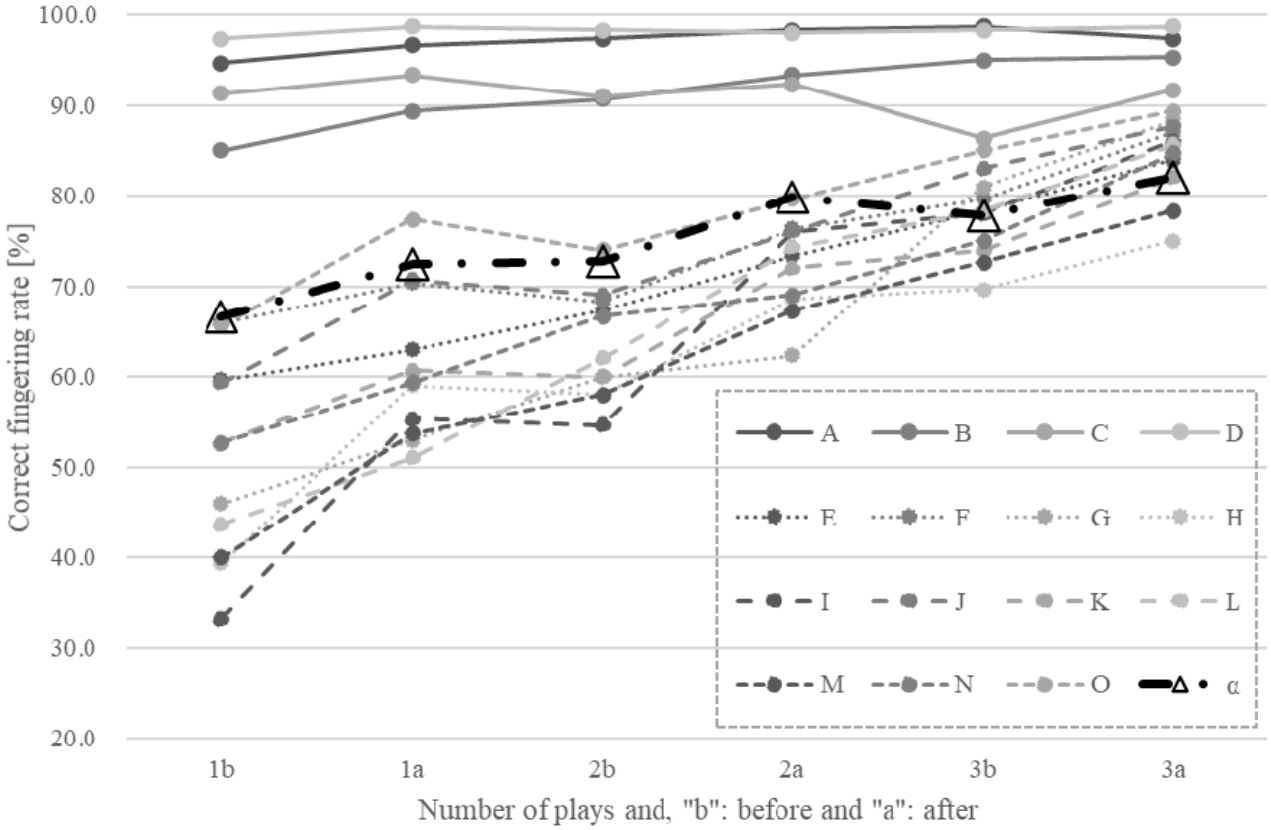


Fig.6 Transition of correct fingering rate for each set (legend *α* is conventional method)

The result of fingering discrimination by the system in the 300-character test is recorded and the actual fingering obtained with the naked eye is compared. The discrimination accuracy of the Fingering Discrimination System will be determined and discussed.

To measure the effect of a Typing Learning Game using the result of automatic fingering discrimination by a Fingering Discrimination System. By comparing the fingering before and after playing the game, the effect of this game on correcting fingering will be determined. Although typing speed is an important factor in evaluating typing ability, this paper does not consider speed because accuracy is focused on of input.

Madison et al. demanded that the larger the key pitch of the keyboard, the higher the input speed and accuracy[9]. Our proposed method use a 109-key keyboard with a key pitch of 19 mm.

The attributes of the subjects are described below. 4 subjects, A to D, are working adults who are skilled typists. The skilled subjects typed efficiently, with both speed and accuracy of input. On the other hand, the subjects from E to O either do not use the PC very often or have a habit of typing with their own fingering. The mean age of subjects E-O is 35.0 ± 30.0. These subjects tend to input slower than the others. However, subject J has a reasonably fast input speed due to his unique fingering.

4.2 Experiments to compare conventional methods

The proposed typing learning game will improve learning by correcting fingering. However, it is expected that conventional typing learning systems also have a corrective effect by the subjects themselves. The experiment was conducted by replacing the Typing Learning Game of the proposed method with a conventional one in the flow shown in Fig.5. As a representative, use the “Home Position Basic” of “Free Typing Practice Software” by P-Ken[10]. This is a web browser application that can be played on the Internet without installation. The flow of the game is the same as that of normal typing software, where the user enters the characters displayed on the screen one by one. The correct fingering for each letter is displayed on the screen to encourage learning of the home position. The number of subjects is two, different from the subjects playing proposed method, and they are beginners in typing.

4.3 Experimental results

The accuracy of the Fingering Discrimination System as a result of the experiment, the discrimination accuracy per character was as high as 98.8% out of the total number of inputs of 27,000 characters. See reference for details of the data[11]. The effect measurement results of the Typing Learning Game for each subject shown in **Fig.6** are based on the fingering data collected by this system. In addition, the legend α of the triangular markers described below is the data of the conventional method. The legend α is the average of two subjects who played conventional typing learning software, treated as one person.

5. Discussion

The correct fingering rate on the vertical axis in Fig.6 is the probability of typing in the home position per character, the higher the rate, the better. The numbers on the horizontal axis represent the number of sets, and the before and after representing whether the test results are before or after gameplay. The capital letters in the legend represent each subject.

In “1b”, the result of the test without playing the game even once, there is a large difference between those who are good at typing and those who are not. However, in "3a", the result after playing the game three times, the difference is smaller, indicating that the less skilled players have improved. When the target correct fingering rate was set at 79.3%, which was the average correct fingering rate of the advanced players in the preliminary experiment, 13 out of 15 players exceeded the target rate in "3b".

Subjects A to D have a high percentage of correct fingerings at “1b”, but the overall results show a slight blurring. From this, that even advanced users do not always type in the perfect home position can be seen.

Subjects A to D have a high correct fingering rate from "1a", and are therefore not affected by the fingering correction effect. On the other hand, subjects E to J are strongly affected by the fingering correction effect. In particular, the percentage of correct fingering tends to increase from “1b” to “1a”, the first play. However, the gradient became more gradual with each play. Therefore, the correct fingering rate is expected to become saturated after a certain level of increase.

Similar experiments were conducted with the conventional typing learning software described in Section 4.2, and the results showed that P-Ken also measured an improvement in fingering. However, the means difference between values of correct fingering rate in “1b” and that in “3a” was 15.3%±1.7, which was 12.3% less than the proposed method's 25.6%±15.5. The seven subjects who showed a positive fingering rate lower than α in the "1b" phase were preceded in the "3b" phase. This difference can be attributed to the direct fingering correction. In conventional typing learning software, a guide to correct fingering is displayed on the screen, but subjects who are accustomed to their own fingering ignore it and type. Such subjects to correct their habits spontaneously is difficult, and therefore external correction is necessary.

The discrimination accuracy of the Fingering Discrimination System was 98.8% of one character. It is thought that the calibration of acquiring the key position coordinates one character at a time enables highly accurate discrimination. Since it is necessary to acquire all the character keys, the system cannot be used immediately, but in this paper, accuracy is emphasized. During calibration, the position coordinates acquired on the sensor's camera image can be mapped on the keyboard. Therefore, if an abnormal value is found, it is corrected on the spot. On the other hand, a failure in discrimination occurs when a correct fingering is identified as an error, or vice versa. This is due to the overlapping of fingers. There were also cases where Leap Motion lost tracking of the hand.

6. Conclusion

In this paper, the fingering is focused on, and aimed to correct to the correct home position. For this purpose, the Typing Learning Game was constructed that can correct fingering. In the Typing Learning Game, the fingering correction can be done by typing the controls of a shooting game. However, only this typing learning game cannot evaluate whether the fingering is correct. Therefore, the Fingering Discrimination System is developed to measure whether the fingering is correct or not. The Fingering Discrimination System can automatically determine the fingertip position coordinates, the position of each key, and the type of finger while typing. In the experiment, the effectiveness of the Typing Learning Game and the accuracy of the Fingering Discrimination System was measured in a series of steps.

The effectiveness of the game was measured by the automatic discrimination of the Fingering Discrimination System, and the accuracy of the discrimination was determined by comparing the fingering with the actual fingering with the naked eye. As for the measurement of the effect of the Typing Learning Game, the results of the

experiment showed that the average fingering of all 15 subjects improved by 25.6% after playing 3 sets. The accuracy of the T was as high as 98.8% per character out of a total of 27,000 (300 characters ×6 sets ×15 subjects) characters. From these experiments, it was found that the beginners of typing receive a correction effect. As an immediate task, the correction effect of this system on typing beginners will be quantitatively evaluated by a t-test to see if significant differences are found.

As future works, learning environment is aimed that can correct fingering more firmly by combining these two methods. In addition, to discuss the meaningfulness of correcting the home position, the system will be constructed to automatically measure whether the subject is now touch typing and aim to create an environment that can evaluate overall typing ability.

References

1) Cabinet Office, Government of Japan, https://www8.cao.go.jp/youth/youth-harm/chousa/h29/net-jittai/pdf/sokuhou.pdf (2021).
2) Ministry of Internal Affairs and Communications, https://www.soumu.go.jp/johotsusintokei/statistics/data/180525_1.pdf (2021).
3) A. Feit, D. Weir, A. Oulasvirta: “How We Type Movement Strategies and Performance in Everyday Typing,” Proc. of The 2016 CHI Conference on Human Factors in Computing Systems (CHI’16), pp. 4262-4273 (2016).
4) E. Takaoka, T. Hashimoto: “Development of Web-Based Touch-Typing Education System and Analysis of Students’ Learning Activities,” IPSJ SIG Technical Report, Vol.2010-CE-106, No.2, pp.1-10(in Japanese) (2010).
5) R. Takakura, K. Hakka, B. Shizuki: “Exploration of Passive Haptics Based Learning Support Method for Touch Typing,” Proc. of the 31st Australian Conference on Human-Computer-Interaction (OZCHI’19), pp. 529-533 (2019).
6) Y. Chigira, M. Nishi: “Development of Typing Practice System (Fingers) by Identifying a Finger That Presses a Key with a Leap Motion Controller as a Function of Focused Typing Practice with AI,” JSTE Journal, Vol.61, No.3, pp. 223-230, (in Japanese) (2019).
7) E. Minoura, S. Takeoka, C. Liao: “Effects of Presentation of Learning History Information on Sustainable Learning,” JSET Journal, Vol.43, Suppl, pp. 37-40, (in Japanese) (2019).
8) F. Weichert, D. Bachmann, B. Rudak, D. Fisseler: “Analysis of the Accuracy and Robustness of the Leap Motion Controller,” Sensors (Basel) 2013, Vol.13, No.5, pp. 6380-6393(2013).
9) H. Madison, A. Pereira, M. Korshoj, L. Taylor, A. Barr, D. Rampel: “Mind the Gap: The Effect of Keyboard Key Gap and Pitch on Typing Speed, Accuracy, and Usability, Part3,” Human Factors The Journal of the Human Factors and Ergonomics Society, Vol.57, No.7, pp. 1188-1194 (2015).
10) ICT Proficiency Assessment, the P-Ken Assessment of ICT, https://manabi-gakushu.benesse.ne.jp/gakushu/typing/homeposition.html (2021).
11) S. Kawahara, T. Hirano, D. Takahashi, K. Nakayashiki, N. Okamoto, R. Tachino: “Development of Typing Learning Game That Can Correct Fingering by Finger Position Descrimination,” IIEEJ Reserch Proceedings, Vol.295, pp. 52-57, (in Japanese) (2021).

(Received May 31, 2021)
(Revised Jan. 29, 2022)

Surena KAWAHARA

He received B.Eng. and M.Eng. degree in 2019 and 2021, respectively, from Kanto Gakuin University. His research interests include pattern recognition, signal and image processing.

Teruaki HIRANO

He received Ph.D. degree in Informatics from the Graduate school of Kogakuin University in 2005. He became an Assistant Professor of Networking and Multimedia Course College of Science and Engineering Kanto Gakuin University in 2013. His research interests include Image processing, Augmented reality and Virtual reality.

Noriyoshi OKAMOTO (Member)

He received a Doctor of Engineering degree in 1985 from Kogakuin University. He became a Professor of Networking and Multimedia Course College of Science and Engineering Kanto Gakuin University in 1999. His research interests include biometrics, signal and image processing and pattern recognition.

Daisuke TAKAHASHI

He received a Doctor of Engineering degree from Kanto Gakuin University. He became an Assistant Professor of Bunka Gakuen University in 2021. His research interests include biometrics, Image processing and pattern recognition.

IIEEJ Short Paper -- Special Issue on Journal Track Papers in IEVC2021Part II --

High-Speed and Accurate Authenticity Judgment Using Physically Unclonable Function of Inkjet Printed Code

Kazuaki SUGAI†, Kohei SHIRAI†, Kitahiro KANEDA† (*Member*), Keiichi IWAMURA† (*Member*)

†Dept. of Electrical Engineering, Tokyo University of Science,

<Summary> The distribution of easily imitated counterfeit products, such as food packaging, brand tags, and pharmaceutical labels, has become a serious economic and safety concern. To address this issue, we propose a system for authenticity judgment of genuine and counterfeit products with high speed and accuracy, focusing on the physically unclonable function of an inkjet-printed code and a locally likely arrangement hashing (LLAH) system that performs high-speed image retrieval. In this study, we verified that the proposed system has high discriminability and stability, based on highly accurate results obtained from a dataset of up to 4,000 sheets. In addition, the effectiveness of the system was also confirmed by validating it on multiple printers and comparing it with Oriented FAST and Rotated BRIEF (ORB), a typical feature matching method, in terms of discriminability and speed.

Keywords: inkjet print, physically unclonable function, authenticity judgment, image processing

1. Introduction

In recent years, manufacturing technology for various products has advanced at a considerably fast pace, and the trade has expanded along with it. While this has increased convenience in daily life, determining the authenticity of information and distinguishing between genuine and counterfeit products has become increasingly difficult for the average consumer. According to an OECD/EIPO survey, the amount of counterfeit and pirated goods traded worldwide is approaching $500 billion, which is equivalent to approximately 2.5% of the world's total trade[1]. Counterfeit goods range from branded goods and airline boarding passes to pharmaceuticals and food products. According to a WHO report, a wide range of counterfeit products, with an increased Internet penetration rate, provide a dizzying array of both branded and generic drugs[2]. In addition, counterfeit medicines have a tremendous impact on society, not only in terms of economic damage but also as a threat to the health and safety of consumers, infringement of intellectual property rights and trademark rights, and as a source of income for illegal organizations.

In the field of security printing, holograms and security inks are used, and artifact metrics using transmitted light images have been used in Japan for some time; systems that use light scattering to determine authenticity have also been studied [3)–8)].

This study aims to construct a system that enables producers and consumers to judge product authenticity by cross-verifying printed codes on labels and tags at the manufacturing, production, and delivery stages.

At the microscale, ink grains can be observed in inkjet printouts. The shapes and positions of these grains are slightly different for each print, and the combination of these grains can be regarded as unique features. Herein, the features (shape and position of the ink grains) to determine authenticity.

As reported in Reference[7)-8)], this system uses the properties of physical unclonable function (PUF), but instead of using optical properties such as the material properties of the ink, it uses individual differences that exist during printing. The advantage of this system is that it does not require special optical analysis for verification or electrical analysis, such as the widely introduced RFID, and the equipment required for the determination is easy to use.

Then, we confirmed the accuracy of the proposed system by comparing it with a printed image that was physically reproduced by the same printer (not a malicious clone) in an ideal environment.

2. Physically Unclonable Function of Inkjet Printed Code

In this section, we introduce the physically unclonable function of inkjet-printed matter, which is used for authenticity determination.

The shape and positional relationship of ink on paper can represent a unique feature of materials printed by an inkjet printer. The following factors can cause differences in the shapes and positions of the ink,

Fig.1 Three codes and an enlarged photo of the upper left corners

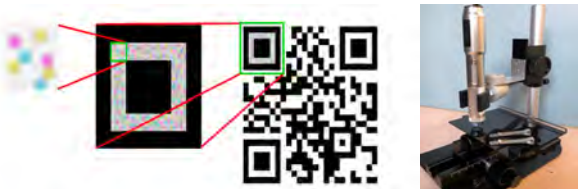


Fig. 3 Part of the QR code to be extracted and digital microscope used for capturing

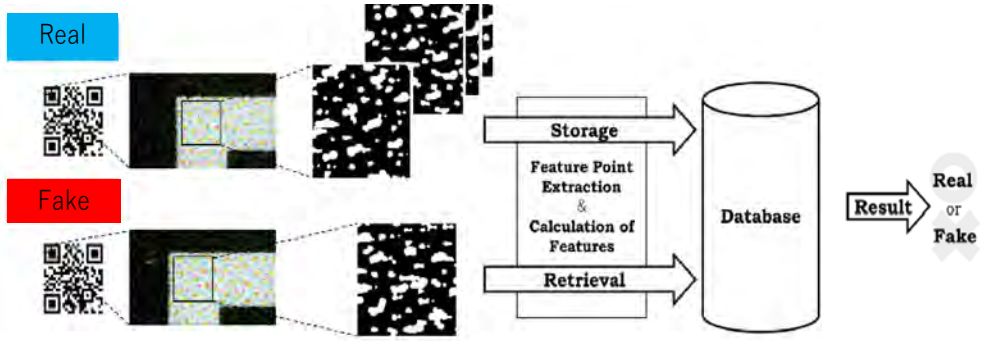


Fig. 2 Overview of authenticity judgment system

1. Slight airflow on the paper surface.
2. The direction in which the header moves when the ink is fired.
3. Air resistance of the ink in flight.
4. Differences in the fiber quality of the adherend.

These properties can be used for authenticity determination, as the uniqueness of these object fingerprints (physically unclonable function properties) for each print hinders malicious reproduction. These properties are independent of the printer type and apply to all printers.

In this study, QR codes, typically used for airline tickets and pharmaceutical labels, were used as an embedding medium to obtain characteristics unique to inkjet printing. Specifically, a single color (grayscale information) with a density of 200 (Max 255) is embedded in the white area in the upper left corner of the QR code and printed.

As an example, **Fig. 1** shows an enlarged image of the area where a single color is embedded in the position detection part of the upper left corner of three different printed QR codes. Each code has a single color embedded in the white area in the upper-left corner. The same URL can then be read from each QR code. Fig.1 shows that the shapes and positions of printed inks differ greatly among the three types of QR code in which grayscale information is embedded, making it difficult to accurately reproduce the same ink arrangement on a microscopic level. Therefore, the features of printed materials can be used for determining authenticity.

3. Authenticity Judgment System

3.1 Overview of authenticity judgment system

An overview of this system is presented in **Fig. 2**, which can be divided into the following three main stages:

1. Image capture by a microscope.
2. Image pre-processing.
3. High speed authenticity judgment using LLAH.

3.2 Image capture by microscope

First, a magnified capture of the upper left corner of the printed QR code, in which grayscale information was embedded, is captured using a digital microscope. **Figure 3** shows detailed capturing of an actual digital microscope usedure

3.3 Image preprocessing

Using the captured images, we preprocessed the images to extract the feature points of the LLAH system to be used in the authenticity assessment. Image preprocessing is performed as follows. In all the steps, OpenCV 3.30 was used.

To extract appropriate feature points from microscopic images, an appropriate binary image was created by passing the image through multiple filters and extracting the hue of the ink by dividing the grayscale portion into three colors: cyan, magenta, and yellow. **Figure 4** shows the extracted ink colors before and after image preprocessing (cyan image is shown as an example).

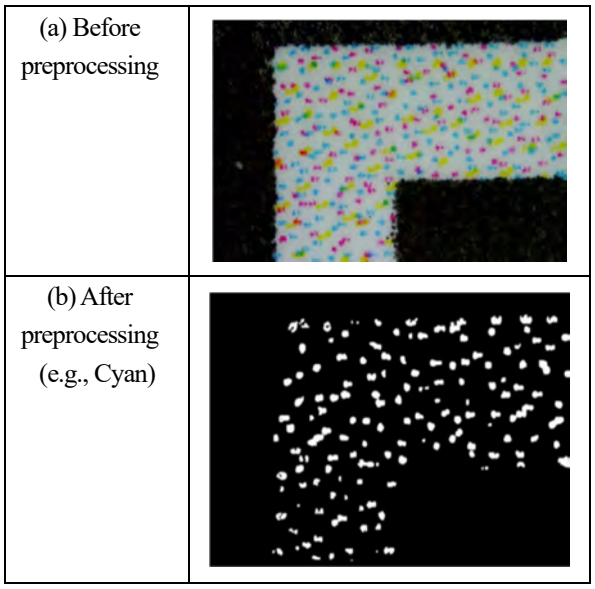


Fig. 4 Comparison before and after preprocessing

1. Smoothing by Gaussian filter.
2. Adjust contrast.
3. Extract the three ink hues by setting threshold values. Hue: Cyan: 90 to 150, Magenta: 150 to 180, Yellow: 20 to 40, Saturation All: 90 to 255, Value All: 120 to 255
4. Grayscale conversion.
5. Adaptive binarization.
6. Smoothing by Gaussian filter.
7. Re-binarization.
8. Invert bits to create an image for input into LLAH.
9. Shrink the image again to remove the noise.

3.4 LLAH extraction of feature points and calculation of feature values

For the features, we use affine invariants[9)-11)] as geometric invariants. Because we are computing features for very fine parts as geometric invariants that are robust to changes in the position of feature points, as opposed to changes in noise and camera shake, for use in smartphones in future systems. When there are four points A, B, C, and D on the same plane as shown in **Fig. 5**, the affine invariant *X* for point A is calculated using Eq. (1).

$$X = \frac{P(A,C,D)}{P(A,B,C)} \quad (1)$$

where P (A, B, C) is the area of the triangle formed by the three points A, B, and C. This feature is computed from the point closest to the point and evaluated discretely. To improve discrimination, for each point of one image, the combination of six adjacent points is calculated based on the x-coordinate and y-coordinate, and the combination of four points is then calculated from the six selected points. The area ratio of the resulting 15

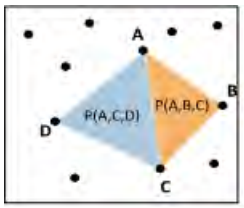


Fig.5 Calculation of affine invariants

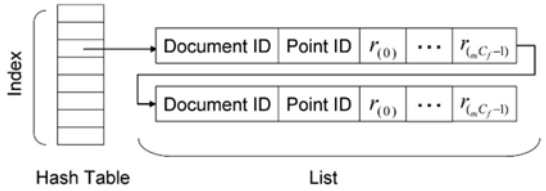


Fig.6 Configuration of the database[9)]

combinations (the triangle combination in this case) was calculated as one feature. When registering the extracted features in the database, for fast retrieval, the index (H_{index}) and quotient (Q) were calculated from the extracted features so that they could be retrieved using Eq. (2).

$$\left(\sum_{i=0}^{{}_mC_4-1} r_{(i)}\, d^i \right) = QH_{size} + H_{index}, \quad (2)$$

where r is the discrete value of the i-th affine invariant, d^i is the number of discrete value levels, and H_{size} is the size of the hash table. In this study, we set d = 10, m = 6 and H_{size} = 1086. This means that the same index, H_{index}, and quotient, Q, can be obtained from the same feature vector in the two matching images. This value is a discretization of the ratio of 15 triangles, which is very discriminative and almost unique. Therefore, during retrieval, for features with the same H_{index} (when collisions occur with the same index), quotient Q is used for comparison instead of the feature value. Therefore, as shown in **Fig. 6**, the image ID, feature point ID, and quotient Q are registered in the database. The image ID and feature point ID are the identification numbers of the image and feature point, respectively.

3.5 Outputting the result of the authenticity assessment

Matching is performed by voting on registered images using a voting table. First, as in registration, the index of the database is calculated for each feature point, and the obtained index is used to refer to the list shown in Fig.6, For each item in the list, we check if the quotient matches, and if it does, we increase the image ID item in the voting table.

Finally, the image ID with the highest number of increments is output as the retrieval result and the number of increments as the number of matches. The retrieval result is used for authenticity judgment as follows:

1. Positive judgment: If output image ID matches the database and the correct result is output, the number of matches exceeds the threshold value.
2. False judgment: If output image ID does not match the database and the wrong result is output
3. Undecided: If there is no corresponding image in the database

4. Verification

4.1 Overview

For verification, QR codes were printed by Canon IP8730 as captured in the environment described in Section 3.2. The captured images were categorized as follows:

1. Registration data: 1000 images were taken for registration.
2. Collation data A: The same 1000 QR codes were re-captured and used to match the registration images.
3. Collation data B: Dataset taken using the same procedure as dataset A to improve reliability.
4. Fake data A: Data taken from the second 1000 QR codes printed by the same inkjet printer without registration.
5. Fake data B: Data taken from the second 1000 QR codes printed by the same inkjet printer without registration to improve reliability.

The resolution of the image was 1920 pixels (width) and 2160 pixels (height). The magnification of the microscope was fixed at 40×, and the position of the microscope was adjusted using the XY stage described in section 3.2. A total of 5000 images were collected. To show that the images will not be affected by their environment they were taken on different days and under the same scale, lighting, and brightness conditions.

To verify the accuracy of system, we changed the size of the datasets registered in the database. Specifically, verifications were conducted for four dataset sizes: 100, 250, 500, and 1000 registrations, collation A B, and fake A B data.

In addition to prove that there is no printer dependency, we performed the same verification using the data of 100 prints each for both collation and fake on a home printer (Canon MP3530) and an industrial printer (Ricoh Pro C7100).

In addition, to demonstrate the superiority of LLAH in terms of accuracy and speed, it was tested on the matching and fake data of each of the 100 sheets to compare it with Oriented FAST and Rotated BRIEF (ORB)[12),13)], which are scale- and rotation-invariant like LLAH and faster than other SIFT and SURF image-matching algorithms.

For the ORB, the original images were matched with each other using Hamming distance and cross-checking, and the features with a distance of 300 or less were counted as the matched features; the total number was used as the number of matches. Subsequently, as in the LLAH system, a threshold of 5 was set to determine the authenticity for comparison.

4.2 Equipment

The following instruments were used for the verification

1. Printer:
   IP8730 (print resolution 9,600 dpi) by Canon Japan
   MG3530 (print resolution 4,800 dpi) by Canon Japan
   Pro C7100 (print resolution 4,800 dpi) by RICOH Japan
2. Microscope:
   Auto focus microscope VIEWTY (Maximum magnification 250×) by 3R SOLUTION Japan
3. XY Stage: Manual XY stage by CHUO PRECISION INDUSTRIAL Japan
4. Computer:
   OptiPlex 7050 (Intel Core i7-7700 @ 60 GHz, 64 GB RAM) by Dell America

5. Results

Table 1 shows the results of the authenticity evaluation for each registration dataset. As can be seen from Table 1, the results are highly accurate, with 100% positive results for both collation and fake data, **Table 2** shows the average number of matches, Minimum number of matches for collation data and maximum number of matches for fake data and average processing time for each registration dataset size separately for collation and fake data. Furthermore, **Fig. 7** shows a box-and-whisker diagram of the number of matches for the collation data (Fake data is omitted because the maximum value is 4). As can be seen from Table 2, the number of matches of fake data increases slightly as the registration dataset size increases: however, even with 1000 sheets, the number of matches between the collation data and fake data is average approximately 34.5 (minimum 3.25) times farther apart from Table 2, confirming that the feature points and features have very high discriminability.

In addition, the LLAH system can perform authenticity judgment at a very high speed of 13.69 ms even if the registration dataset size increases to 1000, without any change in the average processing time of the index-value-based matching system. The required storage capacity is 48.0 MB for a dataset size of 1000, which is approximately 53.3 times less than the 2.56 GB (2,560 MB) required when the original images are stored.

Table 1 Judgment results for each dataset size of the authenticity judgment system

	Dataset size 100			Dataset size 250		
	Positive judgment	False judgment	Undecided	Positive judgment	False judgment	Undecided
Collation A	100	0	0	250	0	0
Collation B	100	0	0	250	0	0
Fake B	0	0	100	0	0	250
Fake A	0	0	100	0	0	250
	Dataset size 500			Dataset size 1000		
	Positive judgment	False judgment	Undecided	Positive judgment	False judgment	Undecided
Collation A	500	0	0	1000	0	0
Collation B	500	0	0	1000	0	0
Fake A	0	0	500	0	0	1000
Fake B	0	0	500	0	0	1000

Table 2 Comparison of number of matches, and processing time for each dataset size

Dataset	Average number of matches (Collation)	Average number of matches (Fake)	Minimum number of matches (Collation)	Maximum number of matches (Fake)	Average processing time [ms]
100	44.38	1.03	14	3	13.24
250	44.28	1.21	14	3	13.05
500	44.01	1.28	13	3	13.67
1000	49.06	1.42	13	4	13.69

Table 3 Comparison of accuracy, number of matches, and processing time of averages for each dataset size

	Canon PIXUS iP8730 Data A			Canon PIXUS iP8730 Data B		
	Positive judgment	False judgment	Undecided	Positive judgment	False judgment	Undecided
Collation	100	0	0	100	0	0
Fake	0	0	100	0	0	100
	Canon PIXUS MG353			RICOH Pro C7100		
	Positive judgment	False judgment	Undecided	Positive judgment	False judgment	Undecided
Collation	100	0	0	100	0	0
Fake	0	0	100	0	0	100

Table 4 Comparison of LLAH and ORB

	Verification with LLAH			Verification with ORB		
	Positive judgment	False judgment	Undecided	Positive judgment	False judgment	Undecided
Collation	100	0	0	94	0	6
Fake	0	0	100	0	0	100

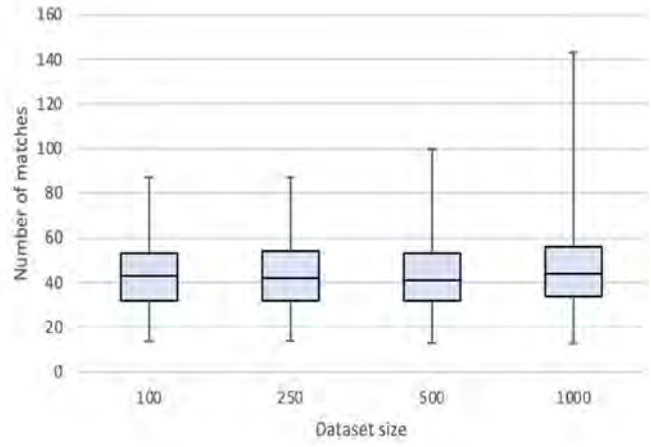


Fig. 7 Comparison of a number of matches for each dataset size

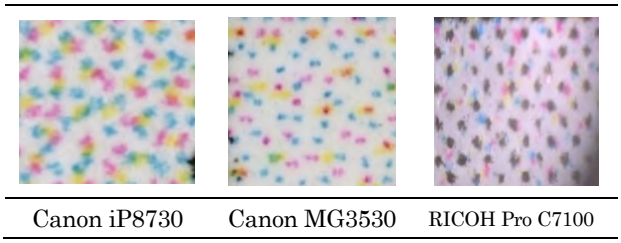


Fig. 8 Comparison of three types of inkjet printers

Fig. 9 Example of printing with a laser printer

Next, for printer dependency, results of verification using the images are shown in **Fig. 8** and **Table 3**. The figure shows that irregular scattering of ink grains, which is necessary for this verification, can be observed in all inkjet printers. The table also shows that all the printers were able to determine the authenticity of the collated and fake data at 100%. This confirms that this system can determine the authenticity independent of the printer.

Furthermore, the results of the comparison between LLAH and ORB are shown in **Table 4**. This table shows that LLAH was able to determine the authenticity of 100% of the images, while ORB was not able to match some of the images with 94% of the matching images. We can see that the average number of matching points for LLAH is 44, while the maximum number of matching points for ORB is approximately 10. In addition, the average speed of LLAH was 13.24 ms, while that of ORB was 89.30 ms. These results confirm the superiority of the LLAH in terms of both accuracy and speed.

6. Conclusion

In this study, we proposed a system that can perform authenticity judgment for inkjet-printed code with high speed and accuracy. We obtained sufficient results a 100% positive judgment rate with an average processing time of 13.69 ms with up to 4000 datasets (total of collation and fake data).

For future work, it will be desirable to move from a microscopic environment to a more practical environment to a more practical environment that combines a standard camera and the microlens of a smartphone. In addition, to maintain the current high level of discrimination even as the size of the registration dataset increases, it is necessary to consider improvements such as setting limits on the sampling and listing of feature points and features. Furthermore, although an inkjet printer was used in this verification, it is necessary to verify whether the system can determine the authenticity of images captured by laser printers (**Fig. 9**), widely used in offices.

In addition, as a long-term future work, this system could be adapted for food and pharmaceuticals that are stored in a good environment with a short cycle time but would also need to be tested for longer storage periods and in poor environments.

Acknowledgments

The authors would like to thank the reviewers for their constructive comments to improve our paper. This work was supported by the JSPS KAKENHI Grant Numbers 21K11931.

References

1) OECD and European Union Intellectual Property Offices: “Trends in Trade in Counterfeit and Pirated Goods, Illicit Trade”, OECD Publishing Paris [PDF file]. Retrieved from https://www.oecd-ilibrary.org/trade/trends-in-trade-in-counterfeit-and-pirated-goods_g2g9f533-en (2019).
2) The World Health Organization (WHO): “Growing Threat from Counterfeit Medicines. Bulletin of the World Health Organization”, Vol. 88, pp. 247–248 (2010).
3) K. -T. P. Lim, H. Liu,Y. Liu, J. -K.W. Yang; “Holographic Colour Prints for Enhanced Optical Security by Combined Phase and Amplitude Control”, Nature Communications, Vol. 10, p. 25 (2019).
4) Z. Song, T. Lin, L. Lin, S. Lin, F. Fu, X. Wang, L. Guo: “Invisible Security Ink Based on Water-Soluble Graphitic Carbon Nitride Quantum Dots”, Angewandte Chemie, Vol. 55, pp. 2773–2777 (2016).
5) H. Matsumoto, T. Matsumoto: “Artifact-Metric Systems. Technical Report of IEICE”, ISEC 2000-59, pp. 7–14 (2000).
6) T. Matsumoto, N. Iwashita: “Financial Operations and Artifact Metrics”, Financial Studies, Vol. 23, pp. 153–168 (2004). (In Japanese).

7) R. Arppe-Tabbara, M. Tabbara, T. J. Sorensen: “Versatile and Validated Optical Authentication System Based on Physical Unclonable Functions”, ACS Applied Materials & Interfaces, Vol. 11, pp. 6475-6482 (2019).
8) Y. Liu, F Han, F. Li, Y. Zhao, M. Chen, Z. Xu, X. Zheng, H. Hu, J. Yao, T. Guo, W. Lin, Y. Zheng, B. You, P. Liu, Y. Li, L. Qian: “Inkjet-printed Unclonable Quantum Dot Fluorescent Anti-Counterfeiting Labels with Artificial Intelligence Authentication”, Nature Communications, pp.1-7 (2019).
9) T. Nakai, K. Kise, M. Iwamura: “Use of Affine Invariants in Locally Likely Arrangement Hashing for Camera-Based Document Image Retrieval”, Document Analysis Systems, Vol. 7, pp. 541–552 (2006).
10) M. Iwamura, T. Nakai, M. Kise: “Improvement of Retrieval Speed and Required Amount of Memory for Geometric Hashing by Combining Local Invariants”, Proc. BMVC 2007 (2007).
11) M. Iwamura, T. Nakai, K. Kise: “Real-Time Document Image Retrieval for a 10 Million Pages Database with a Memory Efficient and Stability Improved LLAH”, Proc. International Conference on Document Analysis and Recognition, pp. 1054–1058 (2011).
12) S. A. K.Tareen, Z. Saleem: “A Comparative Analysis of SIFT, SURF, KAZE, ORB, and BRISK”, Proc. of International Conference on Computing, Mathematics and Engineering Technologies (2018).
13) H.-J. Chien, C.-C. Chuang, C.-Y. Chen, R. Klette: “When to Use What Feature SIFT, SURF, ORB, or A-KAZE Features for Monocular Visual Odometry”, Proc. of International Conference on Image and Vision Computing New Zealand (2016)

(Received May 24, 2021)
(Revised January 21, 2022)

Kazuaki SUGAI
He received the B.S. degrees in Electronical Engineering from Tokyo University of Science in 2021. Since 2021, he has been a Master course student in Electronical Engineering of Tokyo University of Science. His research interests are document understanding, information security and cyber real media processing.

Kohei SHIRAI
He received the B.S. and M.S. degrees in Electronical Engineering from Tokyo University of Science in 2019, 2021 respectively. His research interests are document understanding, information security and cyber real media processing.

Kitahiro KANEDA (*Member*)
He received the B.S. and M.S. degrees in Mechanical Engineering from Waseda University in 1984 and 1989, respectively. He also received the M.S. degree in Electrical Engineering from Duke University in 1995 and the Ph.D. degree in Electrical Engineering from Tokyo University of Science in 2010. He is now with NAGASE & CO., LTD., and visiting professor of Osaka Prefecture University. His research interests are document understanding, information security and cyber real media processing.

Keiichi IWAMURA (*Member*)
He received the B.S. and M.S. degrees in Information Engineering from Kyushu University in 1980 and 1982, respectively. During 1982-2006, he belonged to Canon Inc. He received the Ph.D in Tokyo University. He is now Professor of Tokyo University of Science. His research interests include Coding theory, Information security, Digital watermarking. He is Fellow of Information Processing Society of Japan, Chair of Technical Committee on Enriched Multimedia of the Institute of Electronics, Information and Communication Engineers in Japan.

Upon the Special Issue on CG & Image Processing Technologies Supporting and Expanding Human Creativities

Editor: Shinya KITAOKA
(DWANGO Co., Ltd.)

The AI technology is expected to become a key technology for solving social problems and SDGs (Sustainable Development Goals), such as declining birthrate, aging population, shortage of labor, depopulation area, and so on, to which mature society, especially Japan is facing. The application scope of the technology in the field of image processing is expanding beyond the image generation and object recognition to include areas related to creativity, such as attribute conditioned image generation, super-resolution, image colorization, and line art coloring.

On the other hand, deep learning techniques don't always have good nature at interpretability and explainability of results required for creativity processing. Also, there are still issues to be addressed in AI technology, such as the ineffectiveness to the human-computer interface field by a simple application. So, technologies to support and expand the human creativity are gathering much concern, and quite promising in wide variety of applications.

Based on the above observation, Trans. on IE and VC has planned the special issue on "CG & Image Processing Technologies Supporting and Expanding Human Creativities" targeting the June 2022 issue.

In this special issue, 2 papers that were accepted within the publication schedule of June issue, are contained. The topics are "Face Reconstruction Algorithm based on Lightweight Convolutional Neural Networks and Channel-wise Attention" and "Investigation of New Design Method Rooted in Local History and Culture through Application of Photogrammetry: A Case Study on Application of Maruko-bune, a Traditional Boat Unique to Lake Biwa in Japan, to Architectural Design". We believe this special issue theme will be common to various study items and the results of these papers will stimulate the eagerness of the related activities.

Last but not least, I would like to thank all the reviewers and editors for their contribution to improving the quality of papers. I would also like to express my deepest gratitude to the members of the editorial committee of IIEEJ and the staff at the IIEEJ office for various kinds of support.

IIEEJ **Paper** -- *Special Issue on CG & Image Processing Technologies Supporting and Expanding Human Creativities* --

Face Reconstruction Algorithm based on Lightweight Convolutional Neural Networks and Channel-wise Attention

Haoqi GAO† (*Student Member*) , Koichi OGAWARA†

† The Faculty of System Engineering, Wakayama University

<**Summary**> 3D face reconstruction and face alignment are two highly relevant topics in face research. However, for these tasks, computational complexity is another consideration besides the training accuracy of the model. Our goal is to regress the 3D facial geometry and dense correspondence information from the given 2D image. Thus, in this paper, we fit the 3D morphable model based on a lightweight convolution neural network of the ShuffleNetV2 Plus series network and channel-wise attention model, which can improve the representation ability of the network and the performance of the 3D face reconstruction task without increasing the number of network parameters. Evaluations on test datasets show that our approach achieves significant performance improvements on both 3D face reconstruction and dense face alignment tasks. Alignment performances evaluate on AFLW2000-3D, and our method obtains a lower mean Normalized Mean Error (NME(%)) of 3.694.

Keywords: 3D face reconstruction, Attention model, ShuffleNet

1. Introduction

Since 2D images do not provide any depth information, 3D face reconstruction from 2D images has been a fundamental task and challenge in computer vision and graphics in recent years. Recovering 3D face geometry can address numerous challenges (e.g., large poses and occlusions). Traditional approaches employ a set of 3D base shapes to capture morphological models of face shape variations[1]. Blanz et al.[2),3)] suggested that the geometric structure and texture of a face can be approximate by a linear combination of orthogonal basis vectors obtained by Principal Components Analysis (PCA) and proposed a typical statistical 3D face model called 3D Morphable Model (3DMM). Based on 3DMM, most of the previous methods are mainly based on optimization algorithms to obtain the coefficients of 3DMM. But such approaches are usually time-consuming due to the high optimality complexity and suffer from local optimal solutions and poor initialization. With the development of deep learning, using convolutional neural networks (CNNs) can solve numerous computer vision problems. There have been several works to use CNNs for estimating 3DMM coefficients[4)–8)] to recover the corresponding 3D information from 2D face images, which significantly improves the reconstruction quality and efficiency. Although CNN methods are applied to solve 3D face reconstruction problems, most previous 3DMM-based networks have complex structures and large model parameter spaces, making it difficult to achieve convergence in network training.

Our goal is to regress the 3D geometry of the face and its dense counterpart information.

Inspired by the effectiveness of the Lightweight network, which exploits pointwise group convolution and channel shuffle. Our network structure combines three major achievements: 3DMM[2),3)], ShuffleNetV2 Plus series of units[9),10)] and Squeeze-and-excitation (SE) attention mechanism[11)] to improve the representation ability of the network. For 3D face reconstruction, the proposed network significantly reduces the computational cost while maintaining the model accuracy.

Additionally, we report the experimental result of ours with Selective kernel network (SK) attention[12)], Coordinated attention (CA)[13)], and no attention mechanisms respectively. It demonstrates the effectiveness of the SE attention mechanism embedded in our network.

Our method performs better than 3DDFA[5)] and DAMDNet[1)] with a reduction in Normalized Mean Error (NME(%)) of 1.725 and 0.202 on the AFLW2000-3D dataset respectively. Experimental results show that our algorithm improves the performance of 3D reconstruction.

2. Related Work

We reviewed several previous works on the area of face reconstruction in this section.

2.1 3D face reconstruction

Traditionally, a sparse set of 2D facial fiducial points represents face shape[5]. Cootes et al.[14),15)] suggested that subspace analysis can be used to model shape variations. Unlike 2D Shape Models, a 3DMM model separates rigid (pose) and nonrigid (shape and expression) transformations, which allows the model to cover a range of shape variations and preserve shape simultaneously[5]. 3DMM based methods are popular for 3D face reconstruction tasks, and a large amount of work proposes to improve the performance of 3DMM based modeling. Most previous approaches regress the 3DMM coefficients by solving a nonlinear optimization problem to establish the point correspondence between a 2D facial image and a canonical 3D face model, including facial landmarks[16)–21)] and local features[22)–24)]. However, such methods are usually time-consuming and rely heavily on the accuracy of landmarks or other feature points[25)].

CNN-based approaches have achieved remarkable success in many areas. In contrast to nonlinear optimization, CNNs can be used as regressors to estimate 3DMM coefficients directly, which can significantly improve the reconstruction quality and efficiency. Alp. Guler et al.[26)] designed a fully convolutional network to estimate the dense correspondence between a given 3D template and an input image. Jourabloo et al.[4)] introduced cascaded CNNs to regress the 3DMM parameters, which takes lots of time due to the multiscale and iterations. More researchers have proposed to obtain the reconstructed 3D face directly bypassing 3DMM coefficient regression. For example, Jackson et al.[27)] devised a 3D binary volumetric as a new representation of the 3D structure. Deng et al.[28)] proposed a multi-task that the reconstructed branch of the 3D vertex representation incorporated with the existing box and 2D landmark regression branches during joint training. Feng et al.[29)] suggested UV position map representation, which records the 3D shape of a face in UV space. Although their approach is no longer limited to the space of 3DMMs, it needs complex network structures and lots of time to predict voxel or mesh information[25)]. Some approaches proposed to reconstruct multiple faces simultaneously. Like Zhang et al.[30)] suggested a single-shot multi-face reconstruction framework in a fully weakly-supervised fashion. Furthermore, unsupervised approaches explored, Tewari et al.[31)] proposed an encoder network with an expert-designed generative model that can be trained end-to-end in an unsupervised manner. Nevertheless, unsupervised methods do not perform well in large poses and heavily occluded faces.

3. Method Overview

Real-world tasks have motivated much work to design lightweight architectures to achieve better accuracy, which includes Xception[32)], MobileNet[33),34)], ShuffleNetV2[9),10)], and GhostNet[35)]. Group-wise Convolution and Depthwise Convolutions (DWConv) are both crucial in these works[10)].

Model's architecture contains convolutional layers with H-swish activation[34)], a stack of reconstructed Shufflenet block units that are structured in four stages, finally, with FC layers. The pipeline of our method is shown in **Fig. 1**, where blocks in gray are 3×3 ShuffleNet unit, larger green blocks are ShuffleNet Xception unit, which is deeper than 3×3 ShuffleNet unit, yellow and orange are also ShuffleNet unit with the sizes of kernels as 5 and 7, respectively. H-swish[x]$=x\frac{\mathrm{ReLU6(x+3)}}{6}$ was utilized as a drop-in replacement for ReLU. ReLU6 is a modification of ReLU where we constrain the activation to a maximum size of 6. Attention mechanisms can automatically learn the importance of each feature channel and enhance useful features and suppress non-informative features by the learned importance metric, which uses widely in many applications[11)–13)]. The addition of SE attentional mechanisms can enhance the representation power of our network structure. The computation distribution in the ShufflenetV2 network is small on DWConv, and the main computation is on 1×1 convolution. Therefore, extending the convolution kernel of DWConv can improve the effect without increasing the computational weights. So we reconstructed the Shuffle Xception unit with a size is 5.

The issue with achieving high model capacity and efficiency is how to maintain a large number and equally wide channels with neither dense convolution nor too many groups[9),10)]. Following the insight from some literature[32)–34)], an efficient ShuffleNet unit included DWConv with Batch Normalization layer. **Figure 2** shows the operation of the reconstructed ShuffleNet unit in our network structure. The high efficiency in the shuffle of each building block applied 'channel split', 'concat', 'channel shuffle' operations. This structure enables using more

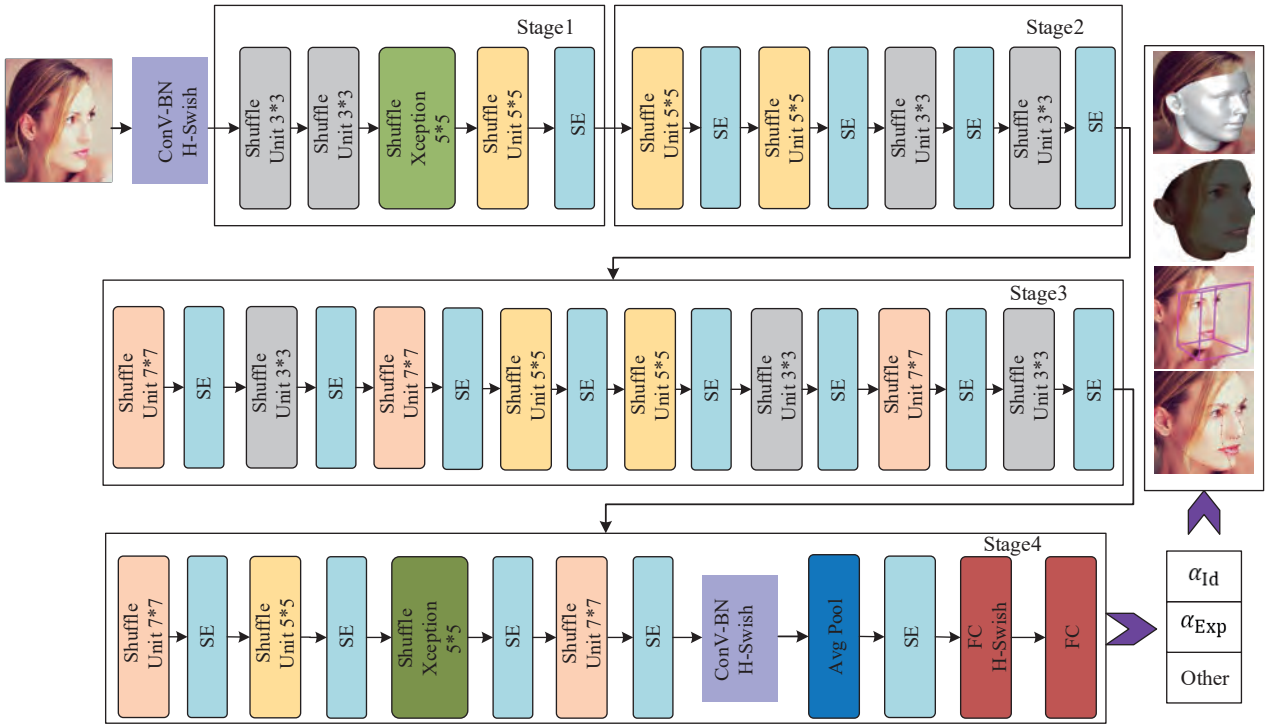


Fig. 1 The flowchart of our network based on ShuffleNet units[9),10)]

feature channels for information communication, which is regarded as a kind of feature reuse. It also improves network efficiency by reducing the degree of network fragmentation and element-wise operations. SE layer induces the model to pay more attention to the contribution of feature areas of the human face and reduces the influence of other unrelated features.

3.1 3D Morphable Model

Blanz et al.[2),3)] design 3D Morphable Model (3DMM) to recover the 3D facial geometry. It is composed of a parameterized generative 3D shape, a parameterized albedo model, together with an associated probability density on the model coefficients[4)]. The 3DMM renders a 3D face shape of S with a linear combination over a set of Principal Opponent Analysis (PCA) basis functions. It is one of the most widely used methods for describing the space of 3D faces nowadays. The 3DMM model is expressed as follows:

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \ , \qquad (1)$$

where $\bar{S}$ represents the mean shape. The identity basis A_{id} and the expression basis A_{exp} come from the Basel Face Model (BFM)[36)] and the Face Warehouse model[37)] respectively. Zhu et al.[5)] used 40 bases from BFM to generate the face shape component and 10 bases from Face Warehouse to generate the face expression component. α_{id} and α_{exp} are the corresponding coefficient of identity and expression.

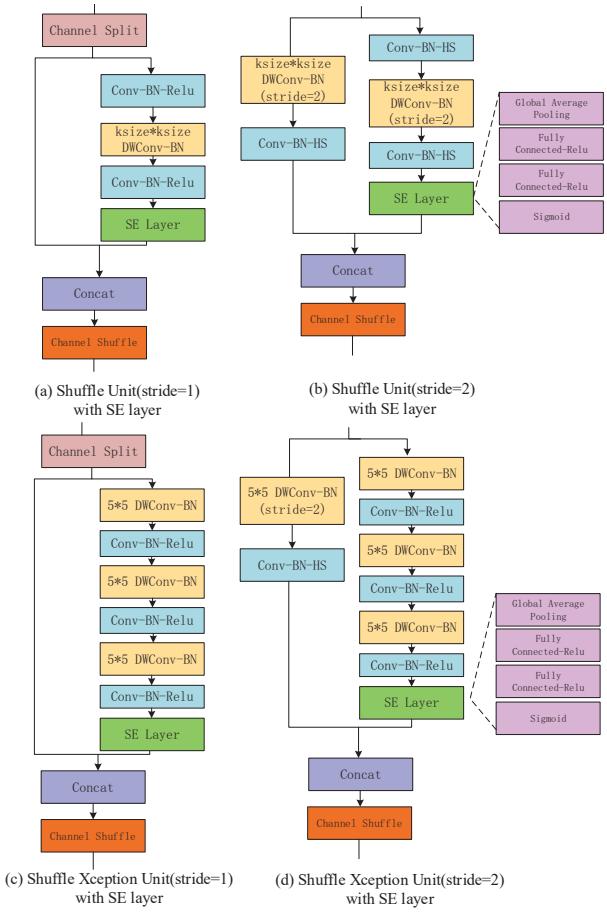


Fig. 2 Shuffle unit with SE layer and reconstructed Shuffle Xception unit with SE layer[14),15)]

In the 3DMM fitting process, we use Weak Perspective

Projection to project 3DMM onto the 2D face plane[1]. It can be expressed as follows:

$$V_p = f * P_r * R * S + t \ , \tag{2}$$

where V_p stores the coordinates of the 3D vertices projected onto the 2D plane, f is the scale factor, P_r is the orthographic projection matrix $P_r = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, R is a rotation matrix consisting of 9 parameters and t is the translation vector. Our goal is to predict the coefficients $p = [f, R, t, \alpha_{id}, \alpha_{exp}]^T$ for rendering a 3D face, which includes a 40d identity parameter vector, a 10d expression parameter vector, and other 12d vector.

3.2 Loss function

Previous literatures on 3D reconstruction[1),5),26)] have described some widespread losses.

Weighted Parameter Distance Cost (WPDC)[5]: minimize the distance between the ground truth p^g and the model predicted parameter p^0. The researchers claimed that the parameters in the 3DMM contribute to the accuracy of fitting with different impacts.

$$\begin{aligned} E_{wpdc} &= (p^g - p^0)^T W (p^g - p^0) \\ W &= diag(w_1, ..., w_{62}) \ , \end{aligned} \tag{3}$$

$$w_i = ||V(p^0(i)) - V(p^g)|| / \sum w_i \ , \tag{4}$$

where the diagonal matrix W contains the weights of each parameter, w_i indicates the importance of the i-th coefficients, computed from how much error it introduces to the locations of 2D landmarks after projection.

Vertex Distance Cost (VDC)[5]: minimize the vertex distances between the fitted and the ground truth 3D face.

$$E_{vdc} = ||V(p^g) - V(p^0)||^2 \ , \tag{5}$$

where $V(\cdot)$ is the face construction and weak perspective projection as Eq.(2).

For 3D face vertices reconstructed with the estimated 3D parameters, Wing Loss[1),38)]is another loss function E_{wing} which define as:

$$E_{wing} = \begin{cases} \omega \, ln(1 + |\Delta T(P)|)/\epsilon & if \ \ |\Delta T(P)| < \epsilon \\ |\Delta T(P)| - C & otherwise \ , \end{cases} \tag{6}$$

$$\Delta T(P) = V(p^g) - V(p^0); \ \ C = \omega - \omega \, ln(1 + \omega/\epsilon) \ , \tag{7}$$

where $\omega = 10$ sets the range of the Non-linear part to ($-\omega, \omega$), $\epsilon = 2$ constrains the curvature of the Non-linear region, and C is a constant that smoothly links piecewise-defined linear and Non-linear parts[1),37)].

Here, we compare with networks proposed earlier (e.g., 3DDFA[48] and DAMDNet[1]), using their same loss experimental function L_{wpdc_vdc} and L_{wpdc_wing} setting during the training, respectively. By comparing the experimental results in **Table 1** and **Fig. 3**, it shows the accuracy of the results.

$$\begin{aligned} L_{wpdc\_vdc} &= \lambda_1 * E_{vdc} + E_{wpdc} \\ L_{wpdc\_wing} &= \lambda_2 * E_{wing} + E_{wpdc} \end{aligned} \tag{8}$$

In **Fig. 3**, we give the results of different loss function with different λ_1 and λ_2 settings. As shown in the figure, our network optimizes with the L_{wpdc_wing} loss function and sets parameter $\lambda_2 = 0.7$.

4. Experiments

4.1 Datasets

(a) 300W-LP

The 300W-LP dataset[5] contains 61,225 synthetic face images across large poses (1,786 from IBUG[40], 5,207 from AFW[41], 16,556 from LFPW[42], and 37,676 from HELEN[43]) along with their corresponding 3DMM annotation coefficient values. These images are synthesized from 300W[40] through a morphable model-based 3D profiling algorithm proposed in the paper[5]and are of coverage across large pose ranges from -90 to 90 degrees.

(b) AFLW

The ALFLW dataset[39] contains 21,080 faces in the wild, which is a large-scale face database including multi-poses and multi-views, and each face annotates with 21 feature points. At test time, we divide it into 3 subsets based on absolute yaw angle: $[0°, 30°]$, $[30°, 60°]$ and $[60°, 90°]$ with 11,596, 5457, and 4027 samples, respectively.

(c) AFLW2000-3D

The AFLW2000-3D dataset is a sample of 2,000 faces selected from the AFLW dataset[39]. Zhu et al.[5] introduced this dataset and annotated its corresponding 3DMM coefficients and the corresponding 68 3D facial landmarks. For evaluation, we also split it into 3 subsets according to their absolute yaw angles: $[0°, 30°]$, $[30°, 60°]$ and $[60°, 90°]$ with 1,312, 383 and 365 samples, respectively.

(d) Menpo-3D

The Menpo-3D[44] dataset contains 8,955 challenging frames varying in illuminations, poses, and occlusions.

4.2 Evaluation metric

A commonly used metric for 3D face alignment tasks is the Normalized Mean Error (NME), which define as

Table 1 NME errors by diagonal length of the bounding box on the AFLW and AFLW2000-3D datasets

	AFLW DataSet(21 pts)					AFLW 2000-3D Dataset(68 pts)				
Method	[0°,30°]	[30°,60°]	[60°,90°]	Mean	Std	[0°,30°]	[30°,60°]	[60°,90°]	Mean	Std
RCPR[45]	5.43	6.58	11.530	7.85	3.24	4.26	5.96	13.18	7.8	4.74
ESR[46]	5.66	7.12	11.94	8.24	3.29	4.6	6.7	12.67	7.99	4.19
SDM[47]	4.75	5.55	9.34	6.55	2.45	3.67	4.94	9.76	6.12	3.21
3DDFA[48]	5.00	5.06	6.74	5.60	0.99	3.78	4.54	7.93	5.42	2.21
3DDFA+SDM[48]	4.75	4.83	6.38	5.32	0.92	3.43	4.24	7.17	4.94	1.97
3DFAN[49]	-	-	-	-	-	3.38	4.46	5.59	4.48	1.11
SRN[50]	-	-	-	-	-	2.97	3.85	5.09	3.97	1.07
DAMDNet[1]	4.359	5.209	6.028	5.199	0.835	2.907	3.830	4.953	3.897	1.02
Weakly-Supervised[30]	-	-	-	-	-	2.60	3.48	4.78	3.62	1.10
RetinaFace[28]	-	-	-	-	-	**2.57**	**3.32**	**4.56**	**3.48**	1.01
Our(($wpdc\_vdc\_1$)	4.871	6.019	6.857	5.916	0.997	3.302	4.333	5.662	4.432	1.183
Our($wpdc\_wing\_0.7$)	**4.217**	**5.004**	**5.77**	**4.997**	**0.777**	2.795	3.593	4.695	3.694	**0.954**

the average of visible landmark error normalized by the bounding box size[1),5)] instead of the common inter-pupil. The formula for NME can be written as Eq.(9).

$$NME(k_i, k_i^*) = \frac{1}{M}\sum_{i=1}^{M}\frac{||k_i - k_i^*||_2}{\sqrt{wbbox \times hbbox}}, \qquad (9)$$

Herein, M is the number of facial landmarks. k_i is the estimated landmarks points, k_i^* is their corresponding ground truth. *wbbox* and *hbbox* define as the width and height of the ground-truth bounding box.

4.3 Training details

Our proposed method is implemented on an Nvidia GTX 2080 Ti GPU using Pytorch. By applying random perturbations to the pitch, yaw, and roll angles of the 300W-LP dataset to obtain 687,854 training images[48)]. The datasets were cropped around the facial region and resized to 120 × 120. We train the network using SGD optimizer with a learning rate of 0.02, a momentum of 0.9, and a weight decay of 5e-4. For a total of 60 training epochs with a batch size of 256, we adjust the learning rate to 0.004, 0.0008, and 0.00016 after 30, 40, and 50 epochs, respectively.

4.4 Experimental results

Table 1 lists NME value by diagonal length of the bounding box with the best results highlighted. We evaluate our model for the 3D dense face alignment task on the AFLW2000-3D and AFLW datasets, which are divided into three groups by comparing it with several baseline methods(e.g., 3DDFA[48)], 3DFAN[49)], DAMDNet[1)], SRN[50)], Weakly-Supervised[30)] and RetinaFace[28)]). The results in **Table 1** show that in comparison to these baseline methods, our method can achieve a comparable result on facial landmark localization. Weakly-Supervised[30)], and RetinaFace[28)], which are used 256d 3DMM coefficients and 1k 3D vertices regression for 3D face reconstruction, respectively. In contrast, our method only used public 62d 3DMM coefficients. Compared with Weakly-Supervised[30)], the NME values of our algorithm decrease under large yaw angles. Our method is not designed for such tasks specifically, it still outperforms the most similar approach DAMDNet[1)] and 3DDFA[48)], which are jointly detecting and reconstructing 3D face shape[30)].

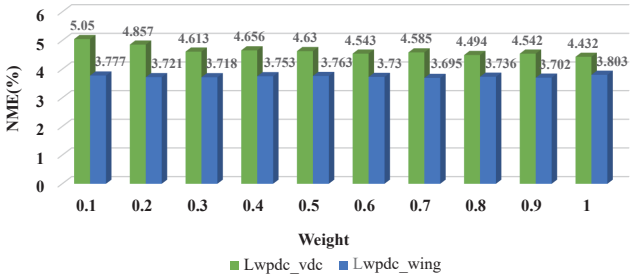


Fig. 3 Different loss optimization results for different weights

Figure 3 shows the mean NME(%) values of optimization with different loss functions L_{wpdc_vdc} and L_{wpdc_wing} in different λ_1, λ_2 for our model.

Lower NME is much better. The results show that optimizing the model with L_{wpdc_wing} (blue) will be better than L_{wpdc_vdc} (green). Also, we can see that for the L_{wpdc_vdc} loss function, the minimum value of NME(%) obtain when the λ_1 is equal to 1, and for the L_{wpdc_wing} loss function, the model performs best when the λ_2 is equal to 0.7, which can significantly improve the face alignment accuracy in a full range of poses.

In real-life scenarios, faces often have multiple challenging elements, such as makeup faces occluded simultaneously (Makeup_occ) and large-scales expressions face due to blurring (Exp_blur).

Table 2 NME errors and comparisons of model size and complexity on AFLW and AFLW2000-3D dataset

	GFLOPs	Params	AFLW DataSet(21 pts)					AFLW 2000-3D Dataset(68 pts)				
Method			[0°,30°]	[30°,60°]	[60°,90°]	Mean	Std	[0°,30°]	[30°,60°]	[60°,90°]	Mean	Std
MobileNet[33)]	0.70	12.82M	4.31	5.044	5.874	5.076	0.782	2.853	3.757	4.921	3.844	1.037
GhostNet[35)]	**0.05**	3.98M	4.472	5.285	6.149	5.302	0.839	3.039	3.9	5.2	4.047	1.088
DAMDNet[1)]	0.12	**2.76M**	4.359	5.209	6.028	5.199	0.835	2.907	3.830	4.953	3.897	1.02
ShuffleNet[10)]	0.12	4.41M	**4.204**	5.058	6.127	5.13	0.963	2.797	3.629	4.834	3.753	1.025
With SK[12)]	0.17	7.29M	4.413	5.187	6.17	5.256	0.881	2.995	3.9	5.26	4.052	1.14
With CA[13)]	0.12	4.65M	4.229	5.054	5.919	5.067	0.845	2.862	3.693	4.812	3.789	0.979
Our	0.12	5.58M	4.217	**5.004**	**5.77**	**4.997**	**0.777**	**2.795**	**3.593**	**4.695**	**3.694**	**0.954**

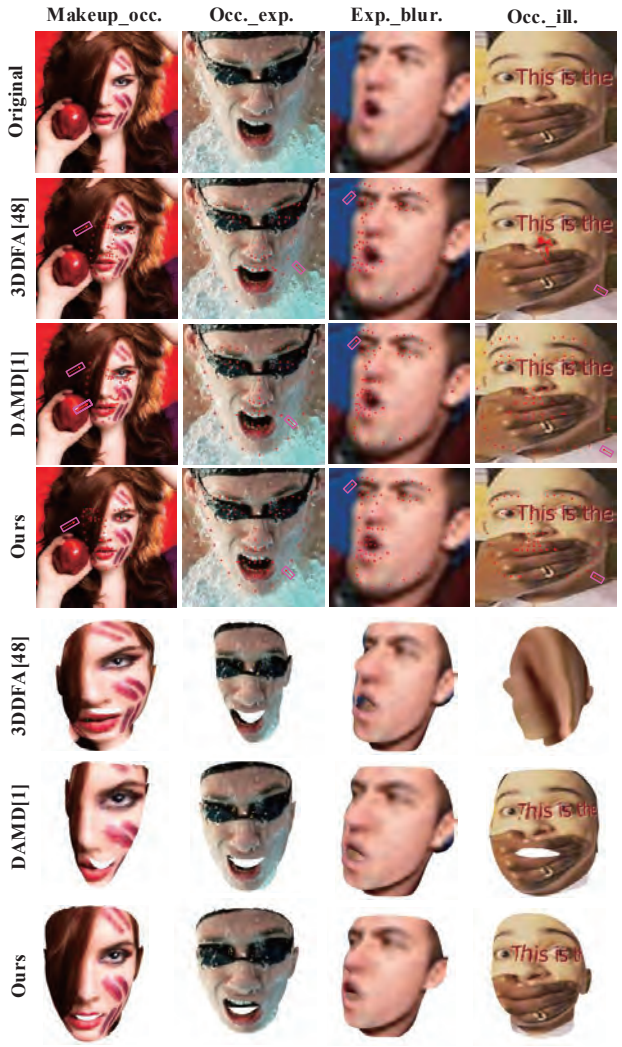


Fig. 4 Comparison results with 3DDFA[48)] and DAMDNet[1)] methods

Figure 4 shows some comparative visualization results with latest method. When the test image under multiple challenging conditions (e.g., Makeup_occ), our method can accurately get the location points of occlusion or makeup parts and provide complete 3D face structure when compared to 3DDFA[48)], and DAMDNet[1)].

The core of lightweight networking is to lighten the network in size and speed while maintaining as much accuracy as possible. Several lightweight network architectures proposed in the last two years mainly include MobileNet[33),34)], ShuffleNet[9),10)], GhostNet[35)], Xception[32)] and so on. For the operational efficiency of networks, the commonly used evaluation factors are GFLOP and parameters. These two factors are used to measure the complexity of the model to judge the algorithm performance. The parameter value is related to the model size, GFLOP value is related to the model speed.

In **Table 2**, we learn with L_{wpdc_wing} loss function and compare each of these models with our model. In terms of accuracy, the result of our experiments performs better. Also, we list the results of different attention mechanisms embedded in our network model. Here, we compare the original ShuffleNet network without attention mechanism, with SK attention mechanism[12)], and with CA attention mechanism[13)]. Although the GFLOPs and parameters of ShuffleNet are much higher than GhostNet's GFLOPs and DAMDNet's parameters. However, the increase in model size and speed is acceptable compared to the increased accuracy. In comparison with CA and SK attention mechanisms, the SE module can improve the precision of our network. In summary, our framework can significantly improve the accuracy of the network without adding too many network parameters and GFLOPs.

Figure 5 shows more examples of random selection from different challenging situations. Top row are the input images which come from Menpo-3D[44)], AFLW2000-3D[5)] and 300-test-3D[40)]. The second and third rows are the 3D landmarks (68 keypoints) plotted for different display views. Our algorithm not only predicts the key points but also estimates the 3D face structure. The fourth and fifth rows represent the images of the predicted 3D pose and depth estimates, respectively. The last two rows are the 3D face model with the texture image of the face and the reconstructed 3D face projection on the input images. As can be seen, our method is robust to occlusions, illumination, and large pose and expression variations.

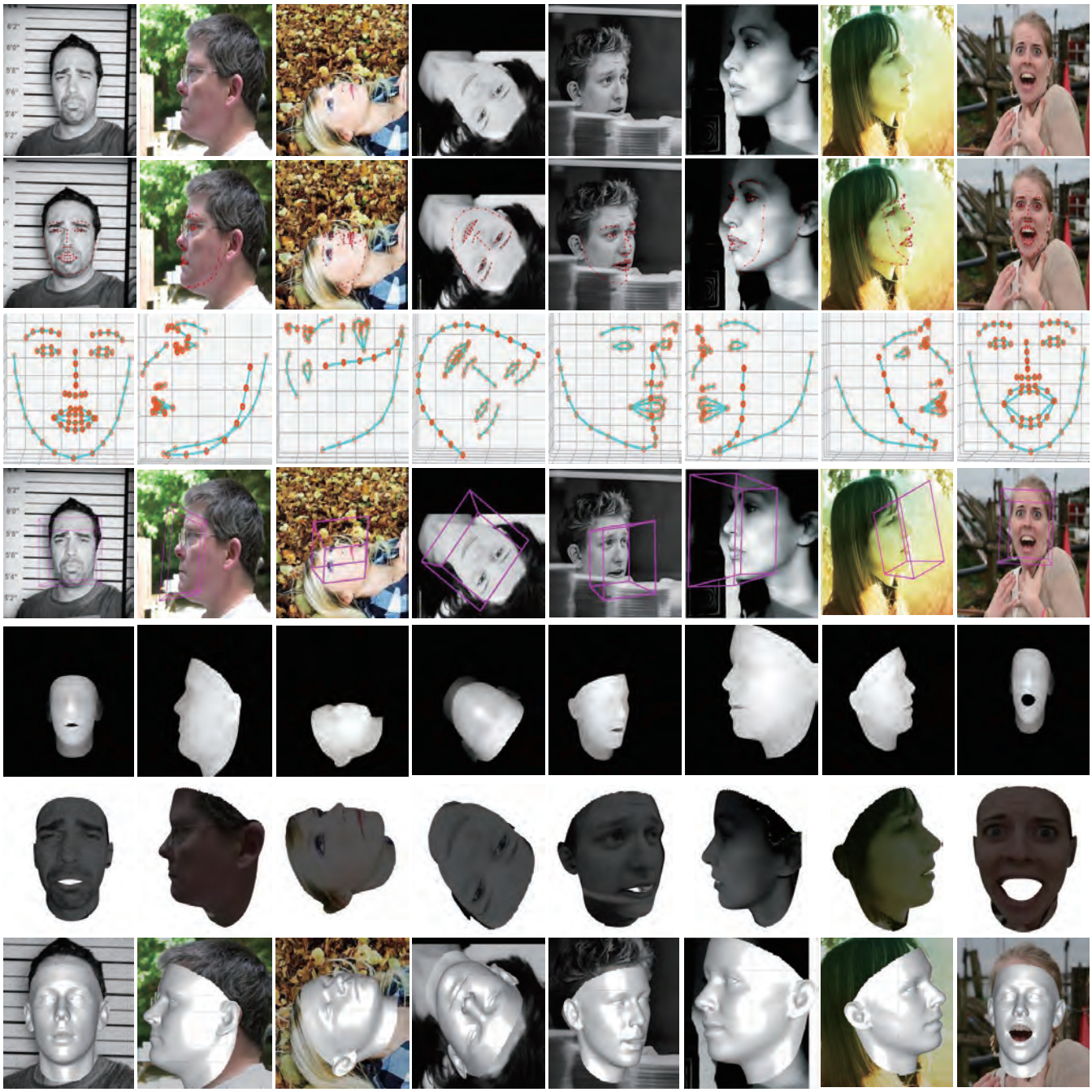


Fig. 5 Face alignment and 3D face reconstruction visualization results

5. Conclusions

In this paper, we focus on solving the 3D facial reconstruction problem. Briefly, we took both the face image and 3DMM coefficients as inputs and learned the model to evaluate the intrinsic consistency between the predicted 3DMM coefficients and the corresponding face image, offering supervision for 3D face model learning. Our method is based on lightweight convolutional neural networks to address relatively complex structures of previous networks. In addition, our network incorporates a channeled attention mechanism that can improve the representational power of the network. Extensive experiments on two challenging datasets (AFLW and AFLW2000-3D) demonstrate that our method obtains the mean NME(%) of 4.997 and 3.694, respectively, which is a significant improvement over existing similar methods. In addition, we visualize the 3D face reconstruction results of our model on Menpo-3D datasets. Our algorithm could substantially improve the quality of the 3D face reconstruction.

References

1) L. Jiang, X. J. Wu, J. Kittler: "Dual Attention MobDenseNet(DAMDNet) for Robust 3D Face Alignment", Proc. of IEEE/CVF International Conference on Computer Vision

Workshop (ICCVW), pp.504–513 (2019).
2) V. Blanz, T. Vetter: "A Morphable Model for the Synthesis of 3D Faces", Proc. of 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194 (1999).
3) V. Blanz, A. Mehl, T. Vetter, H. P. Seidel: "A Statistical Method for Robust 3D Surface Reconstruction from Sparse Data", Proc. of 2nd International Symposium on 3D Data Processing, Visualization and Transmission, pp. 293–300 (2004).
4) A. Jourabloo, X. Liu: "Large-pose Face Alignment via CNN-based Dense 3D Model Fitting", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4188–4196 (2016).
5) X. Zhu, X. Liu, Z. Lei, S. Z. Li: "Face Alignment in Full Pose Range: A 3D Total Solution", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.41, No.1, pp. 78–92 (2017).
6) E. Richardson, M. Sela, R. Kimmel: "3D Face Reconstruction by Learning from Synthetic Data", Proc. of Fourth International Conference on 3D Vision (3DV), pp. 460–469 (2016).
7) E. Richardson, M. Sela, R. Or-El, R. Kimmel: "Learning Detailed Face Reconstruction from a Single Image", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1259–1268 (2017).
8) F. Liu, D. Zeng, Q. Zhao, X. Liu: "Joint Face Alignment and 3D Face Reconstruction", Proc. of European Conference on Computer Vision(ECCV), pp. 545–560 (2016).
9) X. Zhang, X. Zhou, M. Lin, J. Sun: "Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6848–6856 (2018).
10) N. Ma, X. Zhang, H. T. Zheng, J. Sun: "Shufflenet v2: Practical Guidelines for Efficient Cnn Architecture Design", Proc. of European Conference on Computer Vision(ECCV), pp. 116–131 (2018).
11) J. Hu, L. Shen, G. Sun: "Squeeze-and-Excitation Networks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018).
12) X. Li, W. Wang, X. Hu, J. Yang: "Selective Kernel Networks", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 510–519 (2019).
13) Q. Hou, D. Zhou, J. Feng: "Coordinate Attention for Efficient Mobile Network Design", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13713–13722 (2021).
14) T. F. Cootes, G. J. Edwards, C. J. Taylor: "Active Appearance Models", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.23, No.6, pp. 681–685 (2001).
15) T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham: "Active Shape Models-Their Training and Application", Computer Vision and Image Understanding, Vol.61, No.1, pp. 38-59 (1995).
16) Y. J. Lee, S. J. Lee, K. R. Park, J. Jo, J. Kim: "Single View-based 3D Face Reconstruction Robust to Self-occlusion", EURASIP Journal on Advances in Signal Processing, Vol.2012, No.1, pp. 1–20 (2012).
17) X. Zhu, Z. Lei, J. Yan, D. Yi, S. Z. Li: "High-fidelity Pose and Expression Normalization for Face Recognition in the Wild", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 787–796 (2015).
18) J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Nießner: Face2face: "Face2face: Real-time Face Capture and Reenactment of RGB Videos", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387–2395 (2016).
19) P. Huber, G. Hu, R. Tena, P. Mortazavian, W. P. Koppen, W. Christmas, M. Rätsch, J. Kittler: "A Multiresolution 3D Morphable Face Model and Fitting Framework", Proc. of 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 79–86 (2016).
20) C. Cao, Q. Hou, K. Zhou: "Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation", ACM Transactions on Graphics (TOG), Vol.33, No.4, pp. 1–10 (2014).
21) L. A. Jeni, J. F. Cohn, T. Kanade: "Dense 3D Face Alignment from 2D Videos in Real-time", Proc. of 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 1, pp. 1–8 (2015).
22) C. M. Grewe, S. Zachow: "Fully Automated and Highly Accurate Dense Correspondence for Facial Surfaces", Proc. of European Conference on Computer Vision (ECCV), pp. 552–568 (2016).
23) P. Huber, Z. H. Feng, W. Christmas, J. Kittler, M. Rtsch: "Fitting 3D Morphable Face Models using Local Features", Proc. of IEEE International Conference on Image Processing (ICIP), pp. 1195–1199 (2015).
24) S. Romdhani, T. Vetter: "Estimating 3D Shape and Texture using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior", Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 986–993 (2005).
25) X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo ... J. Feng: "3D Face Reconstruction from a Single Image Assisted by 2D Face Images in the Wild", IEEE Trans. on Multimedia, Vol. 23, pp. 1160–1172 (2020).
26) R. Alp Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, I. Kokkinos: "Densereg: Fully Convolutional Dense Shape Regression in the Wild", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6799–6808 (2017).
27) A. S. Jackson, A. Bulat, V. Argyriou, G. Tzimiropoulos: "Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression", Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 1031–1039 (2017).
28) J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou: "Retinaface: Single-shot Multi-level Face Localisation in the Wild", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5203–5212 (2020).
29) Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou: "Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network", Proc. of European Conference on Computer Vision (ECCV), pp. 534–551 (2018).
30) J. Zhang, L. Lin, J. Zhu, S. C. Hoi: "Weakly-Supervised Multi-Face 3D Reconstruction", In: Computing Research Repository (CoRR), Vol. abs/2101.02000 (2021).
31) A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, C. Theobalt: "Mofa: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction", Proc. of IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 1274–1283 (2017).
32) F. Chollet: "Xception: Deep Learning with Depthwise Separable Convolutions", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258 (2017).
33) M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen: "Mobilenetv2: Inverted Residuals and Linear Bottlenecks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520 (2018).
34) A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M.

Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, H. Adam: "Searching for Mobilenetv3", Proc. of IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324 (2019).

35) K. Han, Y. Wang, Q.Tian, J. Guo, C. Xu, C. Xu: "Ghostnet: More Features from Cheap Operations", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1580–1589 (2020).

36) P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter: "A 3D Face Model for Pose and Illumination Invariant Face Recognition", Proc. of sixth IEEE International Conference on Advanced Video and Signal based Surveillance, pp. 296–301 (2009).

37) C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou: "Facewarehouse: A 3D Facial Expression Database for Visual Computing", IEEE Trans. on Visualization and Computer Graphics, Vol.20, No.3, pp. 413–425 (2013).

38) Z. H. Feng, J. Kittler, M. Awais, P. Huber, X. J. Wu: "Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2235–2245 (2018).

39) M. Koestinger, P. Wohlhart, P. M. Roth, H. Bischof: "Annotated facial landmarks in the wild: A Large-scale, Real-world Database for Facial Landmark Localization", Proc. of IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 2144-2151 (2011).

40) C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic: "300 Faces in the Wild Challenge: The First Facial Landmark Localization Challenge", Proc. of IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 397–403 (2013).

41) X. Zhu, D. Ramanan: "Face Detection, Pose Estimation, and Landmark Localization in the Wild", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886 (2012).

42) P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, N. Kumar: "Localizing Parts of Faces using a Consensus of Exemplars", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.35, No.12, pp. 2930–2940 (2013).

43) E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin: "Extensive Facial Landmark localization with Coarse to Fine Convolutional Network Cascade", Proc. of IEEE international conference on computer vision workshops (ICCVW), pp. 386–391 (2013).

44) S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, J. Shen: "The Menpo Facial Landmark Localisation Challenge: A Step towards the Solution", Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 170–179 (2017).

45) X. P. Burgos-Artizzu, P. Perona, P. Dollr: "Robust Face Landmark Estimation under Occlusion", Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 1513–1520 (2013).

46) X. Cao, Y. Wei, F. Wen, J. Sun: "Face Alignment by Explicit Shape Regression", International Journal of Computer Vision, Vol.107, No.2, pp.177–190 (2014).

47) X. Xiong, F. De la Torre: "Supervised Descent Method and Its Applications to Face Alignment", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (2013).

48) X. Zhu, Z. Lei, X. Liu, H. Shi, S. Z. Li: "Face Alignment across Large Poses: A 3D Solution", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.146–155 (2016).

49) A. Bulat, G. Tzimiropoulos: "How Far are We from Solving the 2D & 3D Face Alignment Problem?(and a Dataset of 230,000 3D Facial Landmarks)", Proc. of IEEE International Conference on Computer Vision(ICCV), pp.1021–1030 (2017).

50) X. Luo, P. Li, F. Chen, Q. Zhao: "Improving Large Pose Face Alignment by Regressing 2D and 3D Landmarks Simultaneously and Visibility Refinement", Chinese Conference on Biometric Recognition, pp.349–357 (2018).

(Received September 30, 2021)
(Revised February 1, 2022)

Haoqi GAO (*Student Member*)
She received the M.E. degree from Beihang University of Software in 2016. From 2019 she stared a Ph.D. in system engineering at Wakayama University. Her research interests include face detection, 3D face reconstruction, medical image processing, and deep learning.

Koichi OGAWARA
He received his Ph.D. degree in Information and Communication Engineering from the University of Tokyo, Japan, in 2002. From 2006 to 2011, he was a guest Associate Professor at Kyusyu University. Currently, he is an Associate Professor at Wakayama University. His research interests include robotics, computer vision, and machine learning. He won the Best Vision Paper Award at the IEEE International Conference on Robotics and Automation 2007.

IIEEJ **Short Paper** -- *Special Issue on CG & Image Processing Technologies Supporting and Expanding Human Creativities* --

Investigation of New Design Method Rooted in Local History and Culture through Application of Photogrammetry: A Case Study on Application of Maruko-bune, a Traditional Boat Unique to Lake Biwa in Japan, to Architectural Design

Akari YOSHIDA†, Toshitomo SUZUKI†† (*Member*), Hiroyuki TAGAWA††

†Daiwa House Industry, ††Department of Architecture, Mukogawa Women's University

<Summary> This study aims to investigate new design method rooted in the history and culture of a region through photogrammetry and the reconstruction of shapes as an expansion of the possibilities of using photogrammetric technology, which has advanced rapidly in recent years. Specifically, as an example of something rooted in local history and culture, we focused on Maruko-bune, which were developed as the boats unique to Lake Biwa in Japan but are no longer built or used. We measured the Maruko-bune using photogrammetry and constructed its 3D model. We then attempted to design architecture by decomposing and reconstructing the shape of the 3D model. As a result, we found that the photogrammetry and 3D modeling technologies have reached the point where the shape of the Maruko-bune can be measured and reproduced by photogrammetry and 3D modeling using software and a smartphone camera that can easily be used without special calibration by architects and designers who are not experts in photogrammetry, and its curves and surfaces can be used for architectural design.

Keywords: photogrammetry, Maruko-bune, mesh data, 3D model, architectural design

1. Introduction

The purpose of this research is to investigate a new design method rooted in the history and culture of a region by photogrammetry and the reconstruction of shapes as an expansion of the possibilities of using photogrammetric technology, which has advanced rapidly in recent years. As an example, we will attempt to design new architecture by using photogrammetry and constructing the shape of a traditional Japanese boat.

Boats have been used as a means of transportation since ancient times and have been built with the best technology and in accordance with the culture of each era and region. Looking at these boats, we can see that they are not only practical but also beautiful, rooted in the history and culture of their regions.

Today's ships and boats are built more for practicality and construction efficiency and less in relation to local materials and technology. Along with this, as with much contemporary architecture, their relevance to local history and culture has diminished.

In this study, the Maruko-bune, which was developed as a boat peculiar to Lake Biwa in Japan but is no longer built and used, and whose exact 3D shape has never been surveyed, is measured by photogrammetry, and a 3D model that captures its characteristics is constructed. Then, we try to design a building by disassembling and reconstructing the reproduced shape. The architectural design is based on the assumption that a theater will be built on the shores of Lake Biwa in Shiga Prefecture.

2. Related Works

We have been studying the use of 3D laser scanners to measure and reveal the shape of handmade pantiles used on a house built around 1900. However, photogrammetry is more suitable for surveying larger scale objects such as ships and boats, which can be directly used to shape the appearance of buildings.

The technology of photogrammetry has improved in recent years and is used in a variety of fields, including not only engineering such as architecture and naval engineering but also geography and archaeology. There have been many studies on photogrammetry in the field of architecture, such as those by Kersten, Acevedo Pardo, and Lindstaedt[1], Shults[2], and Pepe and Costantino[3]. The application of photogrammetry to ships has been studied such as by Ahmed, Jamail, and Yaakob[4] and by Burdziakowski and Pawel[5]. The previous studies of photogrammetry in engineering have focused on accurately measuring and reproducing the shape of the object.

However, although data that is too precise or too complex can be useful when it is used to recreate a shape as it is, it is more of a hindrance when it is used to design a new piece of

architecture or something else. Alby, Grussenmeyer, and Perrin[6] pointed out through their research on photogrammetry and 3D modeling of architectural works that compromises are necessary between photogrammetric measurements and architectural representations, which seems to overlap with what we want to argue. In any case, we do not know of any studies like ours where the shapes of artifacts rooted in local history and culture are not directly reproduced and utilized as 3D models, but rather the shapes are decomposed and reconstructed to be applied to new designs that are different from the originals. Also, we cannot find any studies that apply the result of photogrammetry of a ship or boat rooted in local history and culture to the design of architecture on the land, which is different from that of a ship or boat.

Most architects and industrial designers are not experts in photogrammetry. If the goal is to make the photogrammetric survey as accurate as possible and faithfully reproduce the shape, it would be better to ask a photogrammetric expert. However, an enormous cost is likely to be incurred by conducting accurate photogrammetry on large artifacts that cannot be easily moved from their locations, such as architecture and the Maruko-bune displayed indoors in this study. Nevertheless, if the goal is to restore or maintain an existing historical building, it is possible to adopt the idea that the top priority is to grasp its shape as accurately as possible. However, when deconstructing and reconstructing the shape to create a new design, photogrammetry is one piece of information that architects and designers cite to create a new design. Overly complex data makes it difficult to process and in fact becomes an obstacle to the use of data in design. Therefore, rather than incurring costs by improving the accuracy of the photogrammetry itself, it is more critical for architects and designers to be able to do the photogrammetry themselves as cheaply and easily as possible.

Architects and designers use a camera frequently in their daily design work. On the other hand, they can hardly be assumed to have a dedicated camera for photogrammetry. For better photogrammetric accuracy, it is certainly desirable to calibrate the camera to reduce the effects of lens deformation, as already pointed out in several studies[2)3)]. Requiring a special, complex camera calibration for photogrammetry would mean that the architects and designers would have to revert to the settings for other works. This could easily become an obstacle to widespread adoption of photogrammetry among architects and designers. The setting process also needs to be simplified to promote the use of photogrammetry.

In this study, we will explore the extent to which the results of photogrammetry can be used in architectural design when photogrammetry is performed using Agisoft Metashape, which is software that can be easily used by architects and designers who are not experts in photogrammetry, and a smartphone camera, which is a device that they use on a daily basis, without any special calibration.

3. Features of Maruko-bune

Maruko-bune were wooden sailing boats with unique structures that were the mainstay of Lake Biwa transportation from the beginning of the modern era to before World War II at the latest. They were used for not only transporting goods and people but also fishing. A small number of engine-equipped Maruko-bune was still built after World War II.

Maruko-bune are unique to Lake Biwa and have completely different structures from typical Japanese boats used on the open seas during the same period, such as Kitamae-bune (**Fig. 1**).

A Maruko-bune has four main characteristics.

(1) The hull is longer and deeper than that of a Kitamae-bune.

(2) The bottom of the boat is a half arc while a Kitamae-bune had a three-story shelf structure and an inverted triangle shape.

(3) Half logs called "*omogi*" are used for the port side panels.

(4) Many strips of steel plates are used in the bow section.

The main characteristic of the Maruko-bune is expressed in its cross section. The roundness is created by three parts: the *shiki* at the bottom, the *frikake* at the lower side, and the *omogi* at the upper side. For the *shiki*, lumbered cedar boards are glued together using special boat's nails called sewing nails to form a gentle curve. The *omogi* on the port side is a cedar log cut in half lengthwise, with the outside face of the log facing outward. Between the *shiki* and the *omogi*, a horizontal piece of *frikake* is attached diagonally.

The angle at which the *frikake* is attached is an important value that determines the overall balance of the boat. Mr.

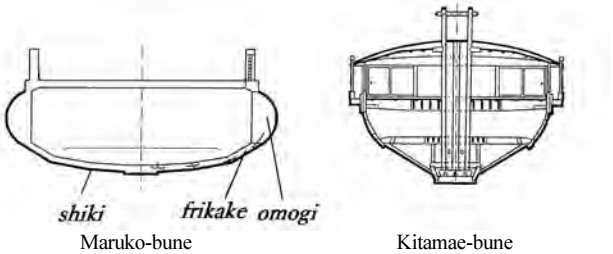


Fig. 1 Central sections of a Maruko-bune and Kitamae-bune drawn by the authors on the basis of photographs taken in the Kita-Omi Museum of Maruko-bune

Sanshiro Matsui, a Maruko-bune carpenter, kept this value as a board chart for over half a century[7]. The name "Maruko-bune" is said to come from the *maru* (roundness) of its shape, and this roundness is best shown in the cross section. Maruko-bune is also said to have been developed as a special type of Japanese boat because the closed nature of Lake Biwa prevented the exchange of shipbuilding technology with other regions. Because the *maruta* (logs) are split in two and attached to the sides, the logs appear to stick out when viewed from the outside (**Fig. 2**).

Due to this external characteristic, Maruko-bune was also called "Maruta-bune" or "Maru-bune." Maruko-bune played an important role in the transportation of goods, numbered more than 1,000 at any one time until the end of the 18th century, and ranged from large ones carrying 400 *koku* to small ones carrying 6 *koku* (1 *koku* = about 180 liters). Of these, the Maruko-bune with a loading capacity of about *hyaku* (100) *koku* were the most popular, and thus were sometimes called "Hyakkoku-bune."

4. Photogrammetry

In the past, boat carpenters did not draw any designs when building Maruko-bune; they made the boats by measuring the dimensions of already existing Maruko-bune[8]. Although there are drawings (**Fig. 3**) that the boat carpenters were forced to draw, these drawings alone do not provide enough information to specify the shape of the Maruko-bune and create its 3D model.

Therefore, we attempted to reconstruct the shape of the existing Maruko-bune by photogrammetry. The preserved Maruko-bune is located at the Lake Biwa Museum (Kusatsu City, **Fig. 4**) and the Kita-Omi Museum of Maruko-bune (Nagahama City, **Fig. 5**). However, it was difficult to measure the shape of the Maruko-bune in the Lake Biwa Museum by photogrammetry due to its installation condition. Therefore, we decided to conduct a photogrammetric survey of the Maruko-bune in the Kita-Omi Museum of Maruko-bune.

The Maruko-bune (Fig. 5) in the Kita-Omi Museum of Maruko-bune was built in 1931 and was actually used until around 1965. It was brought to its current location in conjunction with the construction of the museum. It has an engine room because it was equipped with an engine (**Figs. 6 and 7)** instead of sails.

The smartphone used for photography was Apple's iPhone XS. As already mentioned, which is important in this research is not so much the accuracy of the photogrammetry, but how easily architects and designers, who are not experts in photogrammetry, can use the photogrammetry technology on

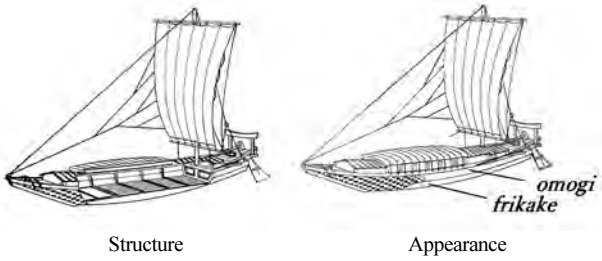


Fig. 2 Structure and appearance of a Maruko-bune drawn by the authors on the basis of photographs taken in the Kita-Omi Museum of Maruko-bune

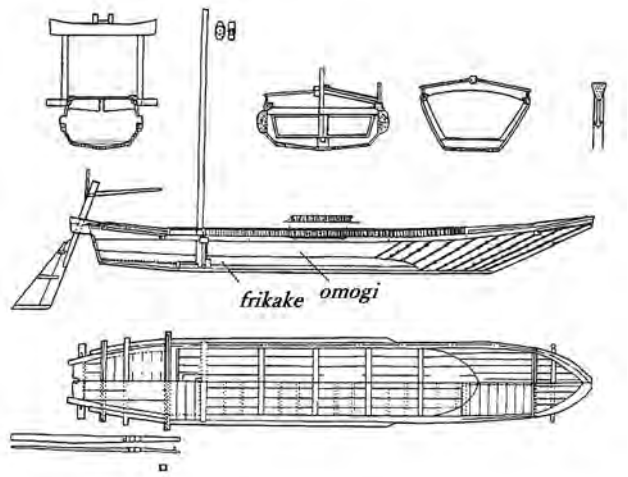


Fig. 3 General structural drawings of a Maruko-bune before World War II [8]

Fig. 4 Maruko-bune in the Lake Biwa Museum (Photographed in November 2019)

Fig. 5 Maruko-bune in the Kita-Omi Museum of Maruko-bune (Photographed in November 2019)

Fig. 6 The reconstruction result of most of the port side of the Maruko-bune in the Kita-Omi Museum of Maruko-bune

Fig. 7 The reconstruction result of the part from the rudder to the starboard side in the rearmost part of the Maruko-bune in the Kita-Omi Museum of Maruko-bune

a widely available device. Therefore, no special calibration was done for the photography.

Instead, we tried to take photos from as many angles as possible in a short period of time, within the constraints of the size of the exhibition room. Architects and designers can easily take a lot of photos in this way. As shown in the left side of Fig. 5, there are stairs and other obstacles on the starboard side. The photos on the port side (opposite side of Fig.5) were used for photogrammetry, except for the rearmost part, because the shape of the Maruko-bune is considered to be symmetrical. The rearmost part was photographed from both the port and starboard sides because of the thin rudder protruding and the lack of obstructions on both sides. A summary of the EXIF data of the 105 photos taken during a 6-minute period and used for the photogrammetry is given in **Table 1**.

We then imported the photos into Agisoft Metashape Professional and tried to reconstruct the boat's shape. The camera positions obtained as a result of the software calculations are shown in **Fig. 8**, and the photogrammetric results are shown in Figs. 6 and 7. Since there were not many obstacles around the port side and rearmost part of the Maruko-bune and the space was bright, the overall shape and surface irregularities were clearly visible. However, we could not automatically connect the part from the rudder to the starboard side in the rearmost part (Fig. 7 and lower in Fig. 8) with most of the port side (Fig. 6 and upper in Fig. 8) of the Maruko-bune. This may have been influenced by the protruding rudder.

Table 1 Summary of EXIF data for 105 photos used for photogrammetry

Dimensions		4032 x 3024
Device make		Apple
Device model		iPhone XS
F number		f/1.8
Exposure time	mean	1/64.4
	maximum	1/40
	minimum	1/122
ISO speed	mean	125
	maximum	320
	minimum	40
Focal length		4 mm
Focal length in 35mm		26 mm

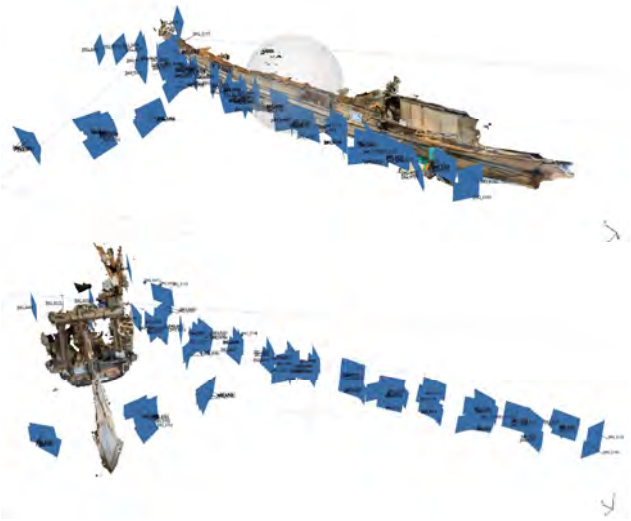


Fig. 8 Camera positions obtained from photogrammetry of the Maruko-bune in the Kita-Omi Museum of Maruko-bune

Fig. 9 Mesh data of pieced-together the Maruko-bune

5. Construction of 3D Model

The mesh data obtained in Figs. 6 and 7 were imported in VectorWorks, a computer-aided design (CAD) software, for use in architectural design. Next, the port side data shown in Fig. 6 was inverted to create the starboard side data, and the port side data, starboard side data, and stern data shown in Fig. 7 were integrated (**Fig. 9**). We then compared this mesh data with the drawing on display at the Kita-Omi Museum of Maruko-bune (**Fig. 10**).

From Fig. 10, we can see that the front part of the mesh data is a little bulkier than the drawing, and the rear part of the mesh data is slightly missing from the side view. As already

mentioned, the Maruko-bune were not built in accordance with the drawings, and the shapes of the boats and the drawings may not perfectly match. Therefore, this degree of difference may occur, and reproducibility is impossible to evaluate. The overall shape of the mesh data and drawing are similar. The effect of the deformation of the camera lens can hardly be observed in Figs. 8 and 9. Therefore, it is considered that a certain level of accuracy can be obtained without calibration. Although the position of the rudder at the stern of the ship is shifted significantly, this is not a problem because the rudder is not fixed.

The mesh data created in Fig. 9 is a polyhedron, not a curved surface. Also, the amount of data is too large to be processed and used for architectural design by CAD software. Therefore, the mesh data cannot be used for design as it is. Therefore, we decided to create a 3D model of the curved surface of the Maruko-bune on the basis of this mesh data. First, we created multiple longitudinal cross sections by referring to the mesh data and attempted to create a 3D model using the Loft Surface tool of VectorWorks (**Fig. 11**). However, the longitudinal cross-section could not accurately represent the *omogi*, a feature of the Maruko-bune.

Therefore, we then attempted to create a 3D model by creating shortitudinal cross sections. Because the total length of this boat is 17.2 m, we decided to divide the length into 16 equal parts and create cross-sections every 1,075 mm. The cross-sections were created separately for *omogi*, which characterizes the shape of the Maruko-bune, and for the rest of the main body (**Fig. 12**). In addition, cross sections of *omogi* were added where the shape of the section changed abruptly.

Then we created the 3D model by using the cross-section divided into *omogi* and the main body. The 3D model could not be created smoothly with the Loft Surface tool of VectorWorks because the curved surface was twisted and unevenness was created due to the error of each cross section. Therefore, we used the Loft tool on Rhinoceros to generate smooth curved surfaces without torsion. Because the architectural design in our study was carried out using VectorWorks, the data of the curved surface generated by Rhinoceros was imported back into VectorWorks (**Figs. 13** and **14**).

Then, the *omogi* data and the data of the main body were combined to complete the 3D model of the Maruko-bune (**Fig. 15)**. In this way, we were able to create a 3D model of the Maruko-bune on CAD. We compared this 3D model with the drawing in Fig. 10 (**Fig. 16**).

In the 3D model in Fig. 16, the rear part of the boat is slightly missing from the plan and side drawings, but the shape of the model and the drawing are closer than in Fig. 10, except for the rudder and engine room, which were omitted in creating the 3D model. As already mentioned, this Maruko-bune was not built on the basis of drawings. However, even taking this into account, the mesh data obtained from the photogrammetry was replaced by a 3D model, which is considered to more faithfully reproduce the boat's shape.

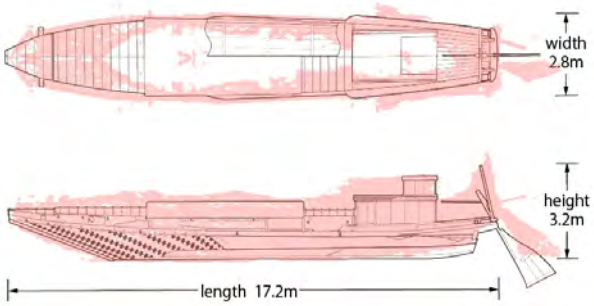


Fig. 10 Comparison of mesh data and drawings (Red: mesh data, black line: drawing)

Fig. 11 3D model created with longitudinal cross-sectional view

Fig. 12 Cross-sectional lines of the main body every 1,075 mm

Fig. 13 3D model of *omogi* created with Rhinoceros

Fig. 14 3D model of the main body created with Rhinoceros

Fig. 15 3D model of a Maruko-bune merged in VectorWorks

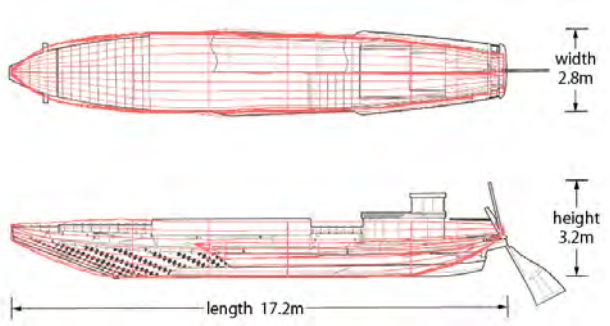


Fig. 16 Comparison of 3D model and drawings
(Red: 3D model, black: drawing).

6. Application to Architectural Design

As already mentioned in the introduction, this section assumes the design of a theater to be built on the shore of Lake Biwa in Shiga Prefecture and examines how the shape of the Maruko-bune obtained by photogrammetry and its 3D modeling can be used for its architectural design.

6.1 Investigation using study models

Prior to using the 3D model shown in Figs. 15 and 16 for the design of the theater, study models were first created with paper clay from the image of a Maruko-bune and waves (**Fig. 17**).

On the basis of this created paper clay study model (Fig. 17), we made study models with aluminum foil to express their curves and curved surfaces (**Fig. 18**).

Study models 1 and 2 are composites of the decomposed shapes of the boat. Comparing study models 1 and 2, we can feel the shape of the boat in study model 1, but we can feel both the image of the waves and the shape of the boat more in study model 2. Therefore, we decided to adopt the idea of study model 2 and proceed with the design. Study model 3 (**Fig. 19**) is based on study model 2, in which more details were made clear and the image was given shape. On the basis of study model 3, the 3D model was created in VectorWorks. **Figure 20** shows the perspective view of this 3D model from the eye level.

On the basis of the shape of study model 3 and its 3D model, architectural design will be done by reconstructing the 3D model of the shape of the Maruko-bune (Figs. 15 and 16). The design will make use of the curves and curved surfaces of the Maruko-bune (**Fig. 21**).

6.2 Use of curves and curved surface of the Maruko-bune

Since roundness is a characteristic of Maruko-bune, its curves were incorporated. The curves of the plan as seen from above of the 3D models of *omogi* and main body were extracted. The curves were then incorporated into the design by enlarging and reducing them to fit the shape of 3D model of study model 3 (**Fig. 22**).

Also, the Maruko-bune were the boats that had been widened by increasing the number of *shiki*, thus increasing their capacity. Taking this into account, the curved surface of

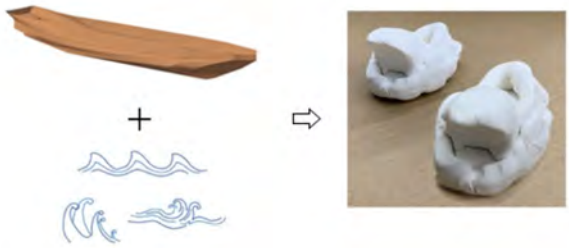


Fig.17 Diagram of the study with images of a round boat and waves

Study model 1 Study model 2

Fig. 18 Study models made by using aluminum foil

Fig. 19 Study model 3 with aluminum foil and wire mesh

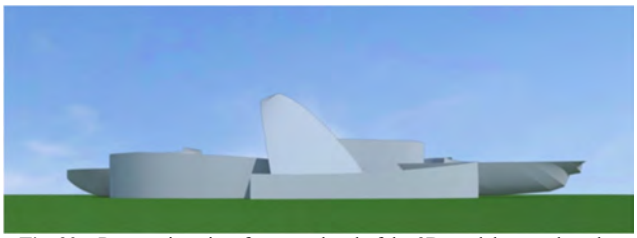


Fig. 20 Perspective view from eye level of the 3D model created on the basis of study model 3.

Fig. 21 Diagram of the reconstruction of the shape of a Maruko-bune

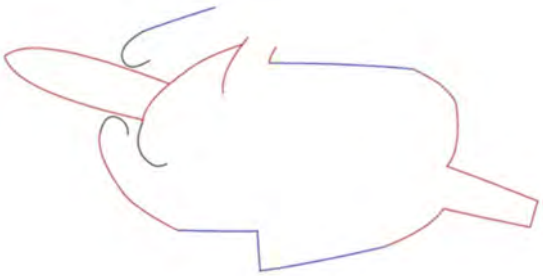


Fig. 22 Plan incorporating the curve of the boat
(Blue line: *omogi*, red line: main body)

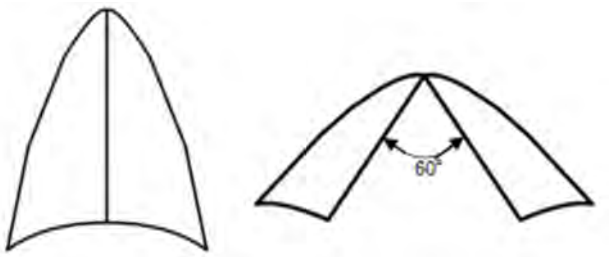


Fig. 23 Curved surface of the forward part of the boat cut and opened 60° at the top

the forward part of the boat, which had been widened by the *shiki*, was incorporated into the shape of the fly tower, which has a larger volume. In fact, we first cut the curved surface of the main body of the Maruko-bune in half. After that, the tip of the boat was opened 60° at the top, and the two halves were connected by another curved surface (**Fig. 23**).

Figures 24 through **26** shows the floor plan, elevation, section, and hand-drawn perspectives of the architectural design whose preliminary design was completed utilizing the explained design method. We were able to design an architecture with an impressive curved surface that inherited the shape of the Maruko-bune.

7. Conclusion

In this study, we examined architectural design methods by photogrammetry, 3D model construction, and reconstruction of the shape of a Maruko-bune, a boat that was developed unique to Lake Biwa in Japan but is no longer built or used today, to investigate a new design method rooted in the history and culture of the region by the photogrammetry. As a result, we found that the photogrammetry and 3D modeling technologies have reached the point where the shape of the Maruko-bune can be measured and reproduced by photogrammetry and 3D modeling using software and a smartphone camera that can be easily used without special calibration by architects and designers who are not experts in photogrammetry, and its curves and surfaces can be used for architectural design.

This design method may require some adjustments

Fig. 24 First floor plan of the theater

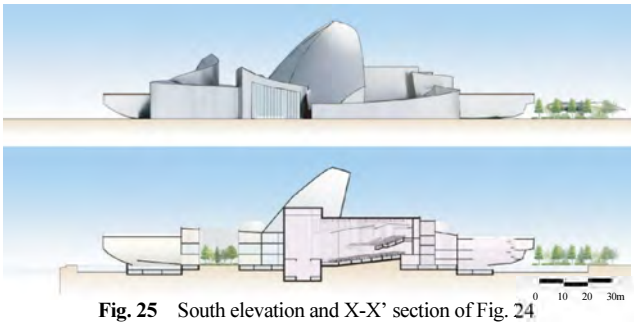


Fig. 25 South elevation and X-X' section of Fig. 24

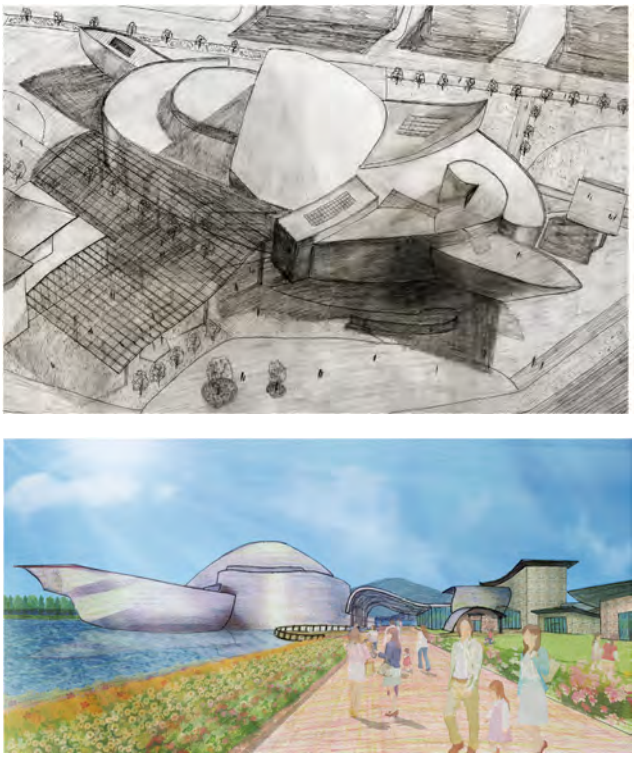


Fig. 26 Hand-drawn perspectives

depending on the photogrammetric and design objects, and the design concept. In any case, we have shown that the history and culture of a region like the Maruko-bune can be passed on to the future in a new form by applying this method.

However, we could not conduct a detailed study on the appropriateness of changing the scale of curves and surfaces obtained by photogrammetry in this architectural design. This is an issue to be addressed in the future.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP17K00741, JP20K12520.

References

1) T. Kersten, C. Acevedo Pardo, M. Lindstaedt: “3D Acquisition, Modelling and Visualization of North German Castles by Digital Architectural Photogrammetry,” Proc. of ISPRS XXth Congress. Vol. XXXV, Commission V, Part B2, pp.126–132 (2004).
2) R. Shults: “New opportunities of low-cost photogrammetry for Culture Heritage preservation,” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLII-5/W1, No.42, pp.481–486 (2017).
3) M. Pepe, D. Costantino: “UAV Photogrammetry and 3D Modelling of Complex Architecture for Maintenance Purposes: the Case Study of the Masonry Bridge on the Sele River, Italy”, Periodica Polytechnica Civil Engineering, (2020).
4) Y. M. Ahmed, A. B. Jamail, O. Yaakob: “Boat Survey Using Photogrammetry Method,” International Review of Mechanical Engineering (IREME), Vol.6, No.7, pp. 1643–1647 (2012).
5) P. Burdziakowski, P. Tysiac: “Combined Close Range Photogrammetry and Terrestrial Laser Scanning for Ship Hull Modelling,” Geosciences, Vol.9, No.5, p.242 (2019).
6) E. Alby, P. Grussenmeyer, J. Perrin: “Visualization of Architectural Works by Photogrammetry: a Compromise Between Measurement and Representation,” Workshop on Vision Techniques Applied to the Rehabilitation of City Centres, 2004, Lisbonne, Portugal. pp.1–11 (2004).
7) K. Makino: “Marukobune,” the traditional sailing boats in the Lake Biwa region, Japan., p.77, Yuzankaku (2013).
8) M. Yoda: “Maruko-bune,” Shiga Prefectural Azuchi Castle Archaeological Museum, Nagahama Castle Historical Museum, eds.: *Biwako no Fune ga Musubu Kizuna —Maruki-bune, Maruko-bune kara “Uminoko” made—* [Bonds Connected by Lake Biwa's Ships and Boats: From Maruki-bune and Maruko-bune to “Uminoko”], pp.47–49, Sunrise Publishing (2012). (in Japanese)

(Received Nov. 26, 2021)
(Revised Jan. 31, 2022)

Akari YOSHIDA

She received her B.Arch. and M. Arch degrees from Mukogawa Women’s University, Hyogo, Japan, in 2020 and 2022, respectively. Since April 2022, she has worked at Daiwa House Industry Co., Ltd.

Toshitomo SUZUKI (*Member*)

He received his B.Eng., M.Eng. and Dr. Eng. degrees from Kyoto University, Japan, in 1998, 2000 and 2003, respectively. He had worked at the Architectural Research Association, Kyoto, Japan, in 2003. He had been a Lecturer from 2004 to 2011, an Associate Professor from 2012 to 2017, and has been a Professor since 2018, of the Department of Architecture in Mukogawa Women’s University, Japan. His research interests focus on architectural planning and design, and design science. He is a member of AIJ, IEICE, IIEEJ, IPSJ and JSPE.

Hiroyuki TAGAWA

He received his M.S. degree and Ph.D. degree in Structural Engineering & Mechanics from University of Washington, Seattle, USA, in 2000 and 2005, respectively. He had been an associate professor from 2014 to 2020, and has been a professor since 2021, of the Department of Architecture in Mukogawa Women’s University, Japan. His research interests include architectural and structural design, structural analysis including finite element analysis, structural morphogenesis. He is a member of AIJ, IASS and JSCES.

IIEEJ **Paper**

Generative Image Quality Improvement in Omnidirectional Free-Viewpoint Images and Assessments

Qiaoge LI†, Oto TAKEUCHI†, Hidehiko SHISHIDO†, Yoshinari KAMEDA†(*Member*), Hansung KIM††, Itaru KITAHARA†(*Member*)

†University of Tsukuba, ††University of Southampton

<**Summary**> This paper proposes a method that improves the quality of omnidirectional free-viewpoint images by generative adversarial networks. Omnidirectional images are a popular way of obtaining three-dimensional (3D) visual information, while free-viewpoint images are essential to Virtual Reality (VR) and Mixed Reality (MR) applications. Therefore, we generated free-viewpoint images with 3D information estimated by the captured omnidirectional images. The quality of the generated images is deteriorated by the 3D reconstruction error due to occlusion and miss-correspondences. In this work, we proposed a method that uses Generative Adversarial Networks (GAN) to solve this problem. We focused on the structural information of various perspectives and applied a "divide and conquer" approach by separating the images into perspectives before training and recombining them at a later stage. At the same time, we conducted a comprehensive, multi-faceted evaluation of the proposed method to verify its effectiveness in improving image quality. Based on the actual information distribution in the equirectangular images, we analyze the adaptability of different image quality evaluation methods. After careful assessment, we consider that the proposed method can generate highly accurate, omnidirectional free-viewpoint images.

Keywords: visual reconstruction, free-viewpoint image, omnidirectional image, image quality improvement, generative adversarial networks

1. Introduction

Photography with omnidirectional (360 degrees) cameras is an effective technique for observing the surrounding environment. This technique has received more attention in recent years due to its ability to achieve immersive viewing combined with head-mounted displays. Google Street View achieves multiple-directional observations by switching the viewpoints as instructed and appropriately selecting the omnidirectional images taken from multiple perspectives.

By applying a three-dimensional (3D) estimation process such as Structure from Motion (SfM) to omnidirectional multi-view images, the position and rotation of the omnidirectional camera and the 3D shape of the target space can be estimated. We previously proposed a generation method of bullet-time video using omnidirectional cameras, which used estimated 3D information to switch viewpoints freely while focusing on a point in the captured space[1]. In this method, omnidirectional observation is only available at the captured viewpoints, not at the non-captured positions. The smoothness of the viewpoint motion decreases when the intervals widen between multi-viewpoint cameras. Another significant issue is that the viewpoints are entirely stationary and cannot be moved from the capturing positions.

Free-viewpoint image generation for reproducing the appearance of any perspective has been greatly researched in computer vision over the past twenty years[2)–14)]. Nonetheless, this field still has many unsolved or partially solved issues. For example, image quality is reduced by 3D reconstruction errors caused by correspondence search errors and occlusion artifacts. Such quality reduction remains an important research issue. Even though the accuracy of 3D reconstruction can be improved using devices that obtain depth information, such as RGB-D cameras[4),10),11)], the simplicity of the capturing system is reduced, causing complications for practical applications. We proposed a solution to this problem using an omnidirectional camera. In multiple omnidirectional images, there are many overlapping regions due to the wide field of view. Hence, the same areas of the 3D

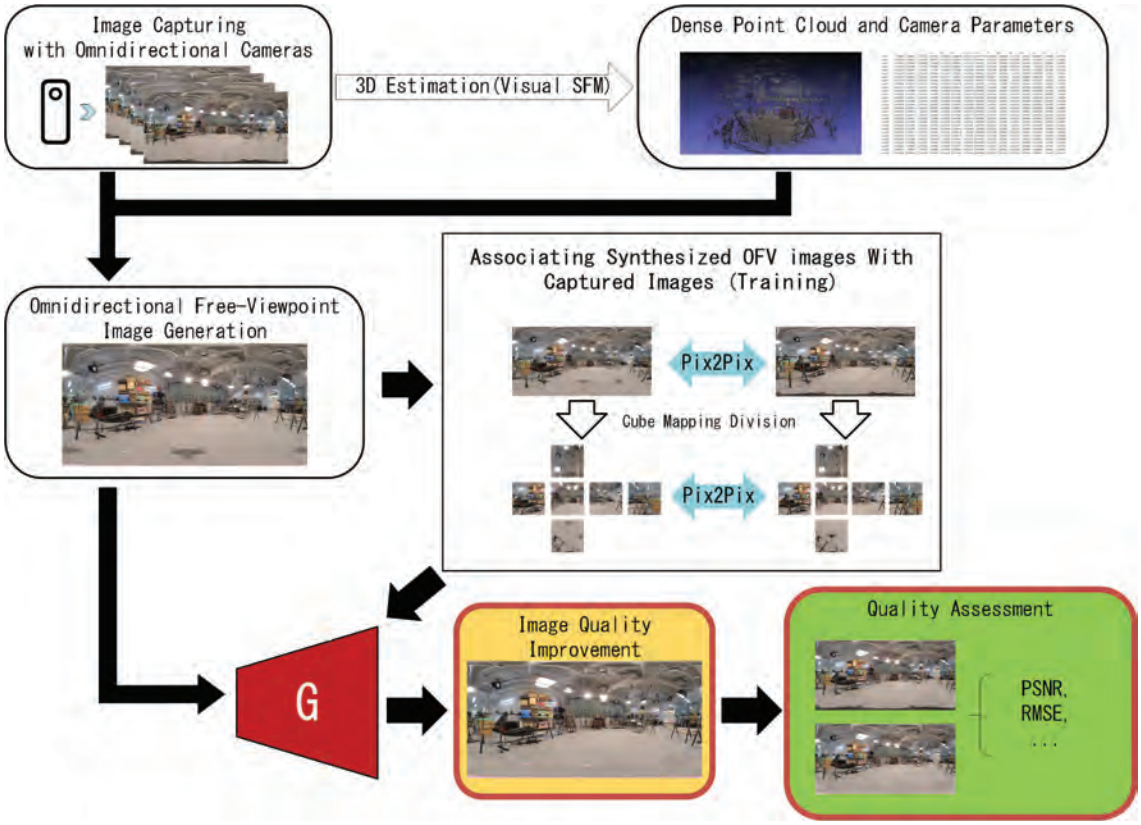


Fig. 1 Overview of image-quality improvement in OFV images

space are viewed from different viewpoints, thus improving the accuracy of the correspondence search. Furthermore, the omnidirectional image also has high adaptivity to the Virtual Reality (VR) content creation for the immersive experience.

Barnes et al.[15]proposed a method to restore the deteriorated image quality by using image reconstruction techniques in 2009. In recent years, deep learning approaches have also been presented[16),17)]for more natural image quality improvements. However, these algorithms assume that the shape, size, and location of the area that needs to be restored are known. In our case, it is quite difficult to recognize regions with low quality for free-viewpoint image generation, which depends on the capturing conditions. This situation complicates applying conventional image reconstruction techniques to solve the problem of image quality deterioration for generated free-viewpoint images.

Figure 1 shows the overview of image-quality improvement in OFV images. This paper makes two main contributions. The first proposes a generation method of high-quality Omnidirectional Free-Viewpoint (OFV) image with deep learning. OFV images are generated from 3D information and camera parameters obtained by SfM and the captured omnidirectional images. Subsequently, Generative Adversarial Networks (GAN) are adopted to eliminate the artifacts and the missing regions within the OFV images. Using this approach, the image quality of OFV images is improved. The GAN performance is significantly improved by image division due to the reduced appearance diversity and the greatly reduced number of structural features to be learned per model. The second contribution is image quality verification by several assessment metrics; conventional metrics which are commonly used for such task, and novel metrics, which have a high correspondence to human perception. Based on our previous work[18)], we conducted additional experiments in different scenes using more suitable evaluation metrics to validate our method comprehensively.

2. Related Works

2.1 Representation of omnidirectional multi-viewpoint images

In Google Street View (Google, 2007), omnidirectional images enable observation of surrounding scenery from an arbitrary perspective. By switching the images of adjacent viewpoints specified by viewers, one can get a more detailed overview of the situation while observing the surrounding scenery. An image blending process and image shape transformation enable viewers to experience a feeling of moving through the scene. Many previous types

of research also used omnidirectional images for the localization of cameras and representation of the captured scene[19]. On this basis, we estimated the 3D shape of the captured space, the position, and the rotation of omnidirectional cameras by processing the 3D reconstruction of the captured scene using the multi-viewpoint omnidirectional images. We developed the bullet-time video generation method using the estimated 3D information to allow both the viewpoint and the perspective to be switched while focusing on a point in the captured space[1]. Unfortunately, the omnidirectional image switching method limits the viewer to the capturing positions but not beyond them. Bertel et. al.[20]brought out a technique to acquire omnidirectional free-viewpoint images with 6-degree-of-freedom(6-DoF). However, the images generated by their approach do not move effectively in the vertical direction. Further, their shooting technique severely limits the range of free-moving viewpoints that the images can present, resulting in a relatively small range that can be observed.

2.2 Free-viewpoint images

There has been a vast amount of research on free-viewpoint images. Model-Based Rendering (MBR)[2),4),6)–8),11)] employed 3D Computer Graphics (CG) models, which are reconstructed from multi-view images of the captured scenes to reproduce views from arbitrary perspectives. Image-Based Rendering (IBR)[3),5),9),10)] directly synthesizes appearances from captured multi-view images. These methods have also been applied in the most recent researches that proposed methods for synthesizing multi-viewpoint 360 images[21)–26)].

In MBR, the quality of the generated free-viewpoint images depends on the accuracy of the reconstructed 3D CG model. For this reason, errors can be significant when performing a 3D reconstruction for a complex scene, and artifacts may appear in the generated view. In addition, when using multiple cameras for observation, the inherent occlusion makes it challenging to reconstruct accurate 3D shapes and reduces the quality of the generated images[8)].

Since IBR does not explicitly reconstruct a 3D shape but instead applies a simple shape, free-viewpoint images can be generated without considering the complexity of the captured space. However, when the captured space's shape significantly differs from the actual shape, the appearance of the generated view can be distorted considerably by excessive image fitting errors. To minimize this distortion and generate an acceptable result, the number of cameras must be increased.

2.3 Image quality improvement

Research on image quality improvement has been actively conducted for about ten years[15)–17),27)]. One way to improve the appearance of an image is by finding corresponding image information using the continuity of the image itself[15)]. Even though this technique has also complemented free-viewpoint videos[27)], it cannot reconstruct unobserved information in images. Various methods have been proposed to reconstruct information not contained in images using convolutional neural networks and GANs. However, all of these methods assume that no information is contained within the missing regions and the shapes, sizes, and positions of these regions are known[16),17)]. We proposed a method that reproduces an equivalent appearance to the captured image by compensating for the deterioration of the image quality due to viewpoint movement by applying reconstruction using GAN[28)]

2.4 Omnidirectional image quality improvement

The combination of omnidirectional photography and image quality improvement is still a new emerging region in computer vision. To the best of our knowledge, there is currently no direct application of image quality improvement methods for panoramas. However, methods that have a similar effect, such as super-resolution and image inpainting, have been applied to equirectangular omnidirectional images with satisfying results[29)–32)]. However, the super-resolution method cannot solve the problem of missing information, while image inpainting requires prior knowledge for the missing regions. Since the missing information of free-viewpoint images is not regularly distributed. The existing method cannot increase the quality of OFV images. We proposed a method that reproduces an equivalent appearance to the captured images by compensating for the deterioration of the image quality due to the viewpoint movement by applying reconstruction using GAN. Our approach fills the gaps in addressing this particular task.

2.5 Image quality assessments

When assessing the quality of images, Peak Signal-to-Noise Ratio (PSNR), Root Mean Squared Error(RMSE), and Structural Similarity Index Measure (SSIM) are the most common evaluation metrics[33)]. However, the evaluation results obtained by these approaches do not fully match our subjective visual perception. PSNR relies on

(a)before interpolation (b) after interpolation

Fig. 2 Omnidirectional depth images

comparing the L1 distance of each pixel of two images to count the overall mean square error, which is insensitive to the overall structural differences between the two images. RMSE has a similar disadvantage to PSNR. SSIM incorporates the statistics of geometric differences, but its results are still affected by the local perceptual field size, and the overall quality assessment is not stable. In addition, geometric differences, luminance, and contrast are given the same weight in the default algorithm, which does not necessarily match the human perceptual quality. To solve this problem, researchers have proposed a deep feature-based picture quality evaluation criterion. One of the most commonly used is Learned Perceptual Image Patch Similarity (LPIPS)[34]and Fréchet Inception Distance (FID)[35]. FID scores can represent the performance of GAN to a certain extent. LPIPS focuses on measuring the similarity of images using deep-level features that assess their quality similarly to human visual perception. Both metrics have been well tested and compared and significantly correlate with the similarity of pictures that can be perceived by humans.

3. Generation Method for Omnidirectional Free-Viewpoint Image

3.1 Image capturing and 3D estimation

Multiple omnidirectional images are acquired at different viewpoints around a target space. As a result of continuous research of 3D information estimation for multi-viewpoint images, some excellent SfM libraries[36)–39)]have been developed. These libraries, however, are usually based on perspective projection images for incremental SfM, which is different from the projection geometry of the omnidirectional images. Therefore, in our approach, each omnidirectional image is divided into multiple perspective images. We used perspective geometry to virtually set up cameras with partial overlap between different perspective images at identical positions and applied the SfM library to each perspective projection image generated from omnidirectional images. Thus, the camera parameters of the image and the sparse 3D point cloud are estimated. The position and orientation of each omnidirectional camera can be calculated from the estimated camera parameters of the corresponding virtual cameras[1)] . Based on the parameters and the sparse 3D point cloud, multi-view stereo processing[40)]obtained a dense 3D point cloud.

3.2 Generation of omnidirectional depth image

We obtained sparse omnidirectional depth images (**Fig. 2**(a)) by calculating the distance from each viewpoint of the omnidirectional camera to the 3D point cloud estimated in section 4.1. We calculated the color differences in the CIELAB color space between the projected 3D point cloud and the pixels of the captured image at the viewpoint where the depth information is generated. This color difference increases when the 3D information of the point cloud is estimated incorrectly. To reduce the amount of incorrectly estimated 3D information, a threshold of 20 was set on the color differences. The depth value is not added to the calculation when the color difference exceeds the threshold.

Due to our inability to estimate the depth values of the unprojected pixels of the 3D point cloud (Fig. 2(a)), many regions are missing in the depth image. These regions were interpolated using a cross-bilateral filter[41)], which refines one of the images based on another image that has less observation noise. In our example, the captured color image is used to filter the depth image. The filtering equation is as follows:

$$D_p = \frac{\Sigma_{r\in N} d(p,r)c(I_p,I_r)D_r}{\Sigma_{r\in N} d(p,r)c(I_p,I_r)}, \tag{1}$$

$$d(p,r) = exp[-\frac{(p-r)^2}{2\sigma_1^2}], \tag{2}$$

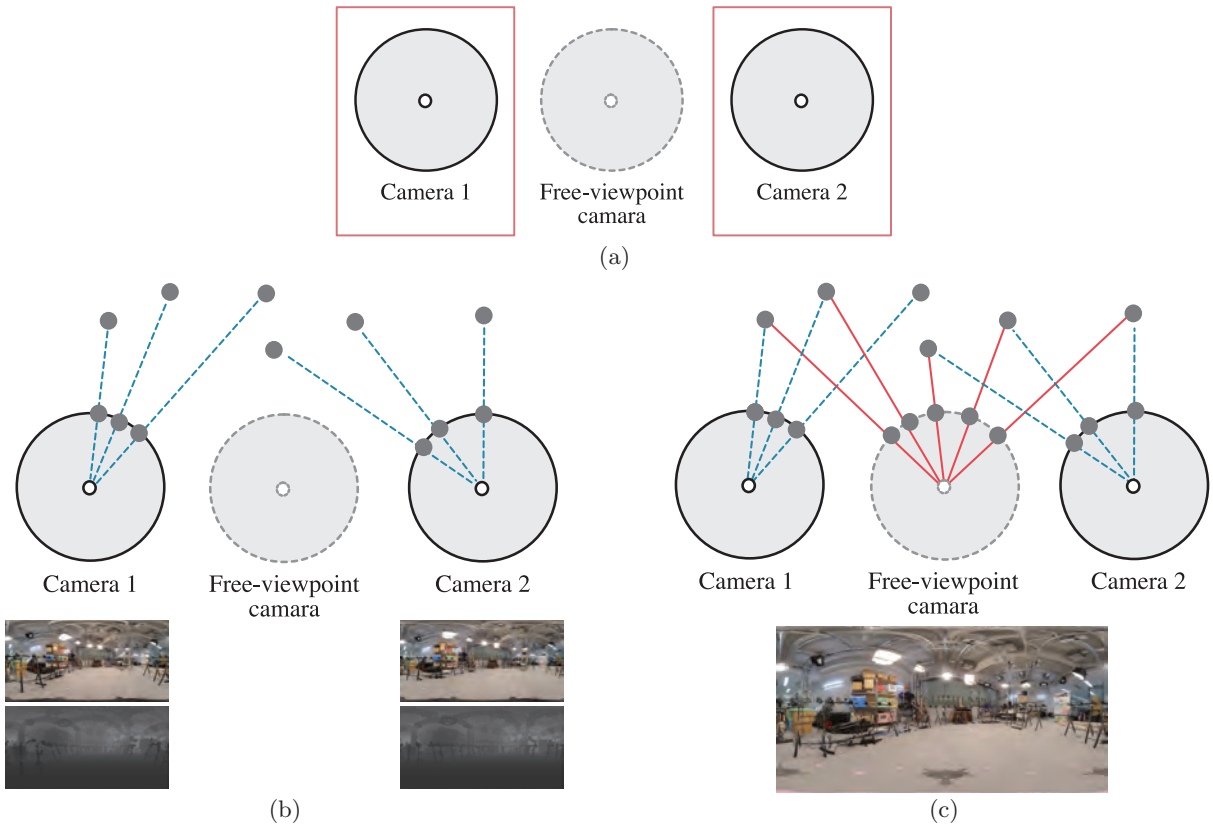


Fig. 3 Generation method of OFV image; (a) virtual viewpoint and two nearest omnidirectional cameras for generating OFV images, (b) projection of each pixel value (color information) of acquired omnidirectional camera with reference to depth information, and (c) process of reprojecting color information in space onto omnidirectional image sphere to generate omnidirectional OFV image

$$c(I_p, I_r) = exp[-\frac{(I_p - I_r)^2}{2\sigma_2^2}], \tag{3}$$

where p represents the pixel on the RGB equirectangular image at the same location as the invalid pixel from the point cloud, r is the adjacent reference pixels of p. D is the depth value, N is the set of reference pixel coordinates, I is the luminance value, σ_1 and σ_2 are constants. $d(p, r)$ is the spatial similarity, which represents the spatial weight, and $c(I_p, I_r)$ represents the weight of color similarity. The invalid depth values are calculated by weighting the spatial distance and the color similarity between the reprojected point cloud and the captured images. Thus, as shown in Fig. 2(b), the depth map can be interpolated while maintaining the captured image's contours.

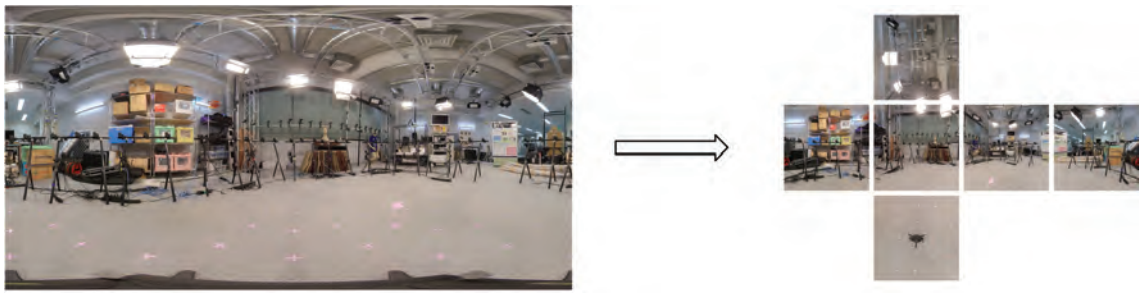
3.3 Generation of omnidirectional free-viewpoint image

We can determine the relative positions of the cameras used for shooting in the scene and define them as actual viewpoints. Next, we specify the location of the viewpoint for which we want to generate the OFV and calculate the distance from this novel viewpoint to all the actual viewpoints, selecting the spots representing the two closest cameras for the next step of the synthesis (**Fig. 3**(a)). The color information contained in each pixel of the omnidirectional images captured at the two points mentioned above is projected into 3D space to generate a dense 3D point cloud model (Fig. 3(b)). The OFV image is generated by projecting these 3D point clouds back onto the omnidirectional image sphere of the new viewpoint. When different points are projected on the same pixel of the free-viewpoint image, the point closer to that viewpoint is adopted to eliminate the hidden surface (Fig. 3(c)). With this method, we can generate free-viewpoint images at any point around where multiple omnidirectional images were captured.

However, significant artifacts and blurring can be observed in the generated OFV images. Causes of these artifacts include errors in 3D estimation and missing 3D information due to occlusions in the captured space. The quality of generated OVF needs to be improved.

4. Image Quality Improvement

This section describes how to reduce artifacts and blurring within the OVF images using GAN. In this re-



(a) (b)

Fig. 4 Example of cube mapping division on omnidirectional image; (a) one of the captured omnidirectional equirectangular images and (b) cube mapping images after image division

search, we adopted Pix2Pix[28] as a way to implement a GAN. Pix2Pix is a conditional GAN that learns the correspondence between two different styles of images, such as sketches and photographs or aerial photographs and maps, and then converts one to the other. In this research, Pix2Pix is applied to image conversion between free-viewpoint images and captured images to improve the quality of the former.

Pix2Pix consists of two networks: an image generator and a discriminator. A prepared pair of pre-conversion and post-conversion images is used as training data. The pre-conversion image is input to the image generator. The generated images or the prepared post-conversion images are input to the discriminator, which decides which image is input. In the learning procedure, the image generator tries to deceive the discriminator, while the discriminator tries to make accurate judgments. In addition to the conditional GAN (cGAN) and the Least Absolute Deviations (L1) loss functions[28], that are used in Pix2Pix, the perceptual loss function is used to calculate the total loss function[42] to learn the features of perception and to make the result closer to the ground truth on the structural similarity. Thus, the following functions represent the objective generator:

$$G = arg \min_{G} \max_{D} (L_{cGAN} + \lambda L_1 + \eta L_{percep}), \quad (4)$$

$$L_{cGAN} = E_{x_i}[logD(x_i, y_i)] + E_{x_i,z}[log(1 - D(x_i, G(x_i, z))], \quad (5)$$

$$L_1 = E_{x_i} \| G(x_i, z) - y_i \|, \quad (6)$$

$$L_{percep} = E_{x_i} \| VGG19_{relu22}(G(x_i, z)) - VGG19_{relu22}(y_i) \|, \quad (7)$$

where L_{cGAN} represents cGAN loss function, L_1 is the L1 loss function, L_{percep} is the perceptual loss function, D is the discriminator, G is the generator, x_i and y_i are the input image and the corresponding ground truth, respectively. z is the arbitrary noise, and $VGG19_{relu22}$ is the activation map of our model that is used to extract the deep structural features and calculate the perceptual loss function, here we used a pre-trained 19-layer VGG network[43], where $relu22$ represents the features obtained by the 2nd convolution before the 2nd max-pooling layer. λ and η are coefficients that balance the different loss terms. After the entire learning process is complete, the OFV images are processed using the image generator.

As training data, the OFV image with a synthetic size at the viewpoint captured in section 4.3 is used as the pre-transformation image, while the omnidirectional image captured in section 4.1 is used as the post-transformation image. After training the image generator with the training data, the OFV image at the virtual viewpoint is input to the learned image generator to generate highly realistic images with less degradation of image quality.

We focus on learning the projection geometry of the images to achieve learning efficiency. However, omnidirectional images based on equirectangular projections can lead to significant changes in appearance based on their projection characteristics leading to shifts in viewpoints, which in turn leads to increased diversity in appearance among the learned samples and more complicated learning. Therefore, by dividing the omnidirectional equirectangular image into multi-view projection images, the effect of projection geometry on appearance diversity can be reduced, and then GAN learning for image contents can be well implemented. In this paper, as shown in **Fig. 4**, the cube mapping method is used to divide the omnidirectional equirectangular image into six perspectives. The image generators are constructed using the perspective projection images on each plane separately. The OFV images generated in section 4.3 are divided by

using cube mapping, sequentially processed by the generators that are trained in all six directions and recombined to improve the quality.

5. Omnidirectional Image Quality Assessment

When assessing the quality of images, Peak Signal-to-Noise Ratio (PSNR), Root Mean Squared Error (RMSE) and Structural Similarity Index Measure (SSIM) are the most common evaluation metrics. Previous work also pointed out that these metrics allow objective evaluation of image quality for omnidirectional images[33]. However, assessments on omnidirectional images is quite different from perspective ones. One of main reasons is the differences of information distribution between the image formats. In equirectangular image, information are projected unevenly along the vertical axis, since its original source of information is a sphere, the projected information distributed more densly in the equator than polar areas. This unevenness makes the direct application of traditional evaluation metrics (PSNR, RMSE, SSIM) on equirectangular images less rigorous. In contrast, the use of criteria that are more relevant to human perception (FID, LPIPS) can more effectively reflect the human eye's evaluation of image quality. Therefore, in our experiments, all four evaluation criteria mentioned above were applied. Furthermore, the structure of the equirectangular image is distorted compared to the real scene it presents, and this distortion may make the quality assessment based on the panoramic image not necessarily consistent with the quality of the image observed in real applications (e.g. VR, AR). Therefore, we perform two evaluation processes for each set of images, one directly based on the equirectangular images, and the other we calculate the average of the evaluation results obtained by applying all criteria to the perspective views in 6 directions (up, down, left, right, front and back) as a supplement to provide a quality evaluation that is informative for real applications.

6. Experiments

6.1 Experiment environment

We conducted a comprehensive experiment to evaluate the impact of deep learning on the quality of OFV images and image division on learning efficiency. We took omnidirectional shots at three indoors scenes (University of Tsukuba) to validate our proposed method. Notice that the second and third scenes are captured in the same space but with different camera alignment and layouts. A flow chart of our complete experiment at the first scene is shown in **Fig. 5**. **Figure 6** shows the layout of our cameras in the experiment of the first scene, note that the cameras are placed by hand, therefore the distance between every 2 cameras may not be the same. We captured omnidirectional images at 54 viewpoints as shown and divided them into two parts: part 1 (30 images) for generating the OFV images of the capturing position of part 2 (24 images), and images of part 2 as the ground truth of the generated OFV images. Then images in part 1 were used to generate dense 3D point cloud using open-source software: VisualSFM. Simultaneously, those cameras' position and orientation were obtained from the bundle adjustment file from the SFM results. To acquire the extrinsic camera parameters of part 2, we fixed the camera position in part 1 and added images of part 2 to the incremental SFM process. Notice that images in part 2 are only added to SFM for obtaining the camera parameters but do not contribute to the 3D point cloud of the captured scene. The 3D point cloud is projected onto each omnidirectional image sphere to generate a sparse depth image based on the estimated camera parameters. A dense omnidirectional depth image is then generated at each viewpoint by interpolating the gaps between the projected points. With the captured omnidirectional image as texture and integrating it with the corresponding depth image, an omnidirectional image at any viewpoint can be synthesized. Thus, a dataset of actual captured omnidirectional images and synthesized ones at the same viewpoints is obtained. The following are the PC specifications for the data processing: CPU AMD Threadripper 3960X, RAM: 128GB, and GPU: NVIDIA RTX 2080ti.

Then we conducted the experiments on the image quality improvement on the generated OFV images. We planned to use Pix2Pix for a style transfer between the generated OFV and the captured images. However, due to the limitations of our GPU's VRAM, it was impossible to train a model with images at the original resolution. Therefore, bilinear downsampling was performed while generating the cube mapping images. The resolution of each perspective projection image was 512 × 512 pixels, and the resolution of each recombined equirectangular image was 2048 × 1024 pixels. We aligned the captured images of part 1 with the same pose as the OFV images, split them by cube mapping and considered them true images. Then we generated OFV images at the same positions of the images in part 1 and split them by cube

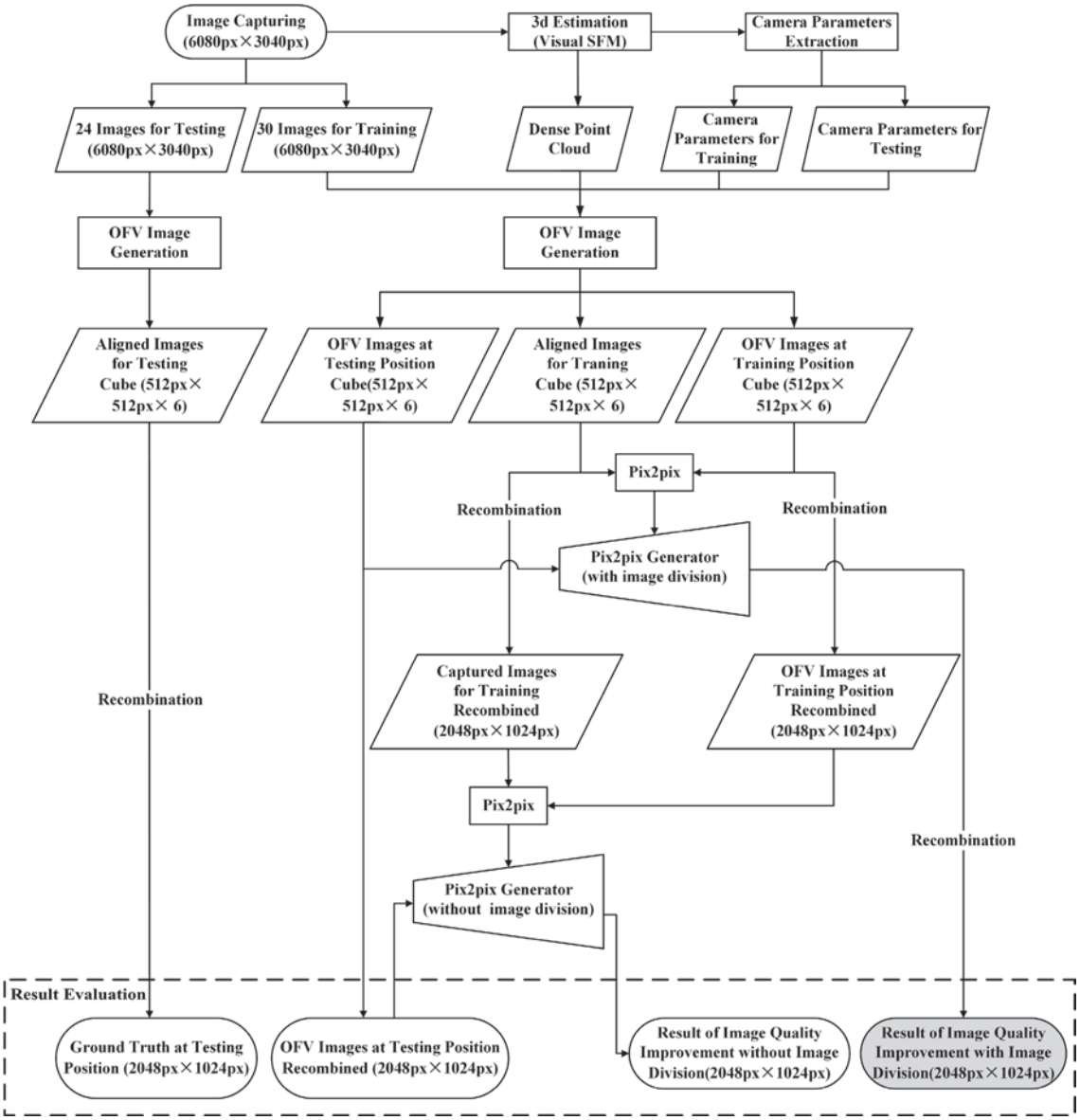


Fig. 5 Flow chart of our full experiment at Scene 1

mapping. These are the fake images for the discriminator to classify. Pix2Pix was employed to realize the style transfer between the generated OFV images and the captured ones on both the cube mapping images and the recombined equirectangular ones. Our approach for generating OFV is based on an assumption that the shooting environment is static, which means there shouldn't be anything in the scene change during the capturing time. However, the tripod is inevitable for capturing omnidirectional images. Also the segment of the cube mapped images that contain the tripod are not our main focus for generating the free-viewpoint images. Hence, we didn't improve their quality, just kept them as the originally generated. We set the learning steps at 1000 epochs, according to our experience for the satisfying results. The OFV images were subsequently processed by the generator of both approaches to obtain their results.

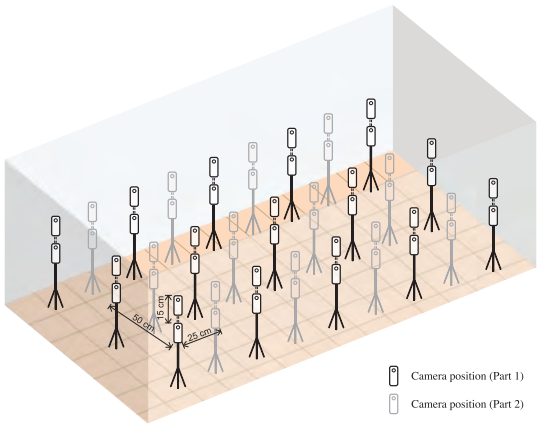
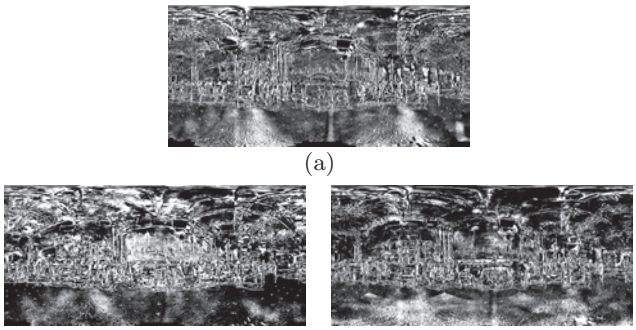


Fig. 6 Layout of the shooting points during experiments

6.2 Results and assessments

Figure 7 shows the error maps of each generated OFV image results relative to the ground truth, where the black area indicates no difference between the image to be measured and the ground truth. The larger the proportion of black areas is, the closer the image is to the ground truth. We grayscaled the error map and increased the exposure rate to facilitate observation. Figure 7(a) shows the comparison result of the original OFV image. Figure 7(b) and (c) represent the comparison result of the OFV images with image quality improvement. The training images for Fig. 7(b) are equirectangular ones, while the images used in Fig. 7(c) are segmented with cube mapping. The black area in Fig. 7(c) has the highest proportion from these three figures, indicating that the images processed by our proposed method with image division are the closest to the actual image data.

Figure 8 shows several examples of the curve of loss with increasing epochs during our picture quality improvement training using GAN. The L_{vgg} in the figures represents the L_{percep} in equation 4.7. Figure 8 (a), (b), (c) represent the training curves using cube mapping maps in three experimental scenarios, respectively.



(a) (b) (c)

Fig. 7 Error map of OFV results; (a) the original OFV image, and (b), (c): OFV images with two different quality improvement approaches

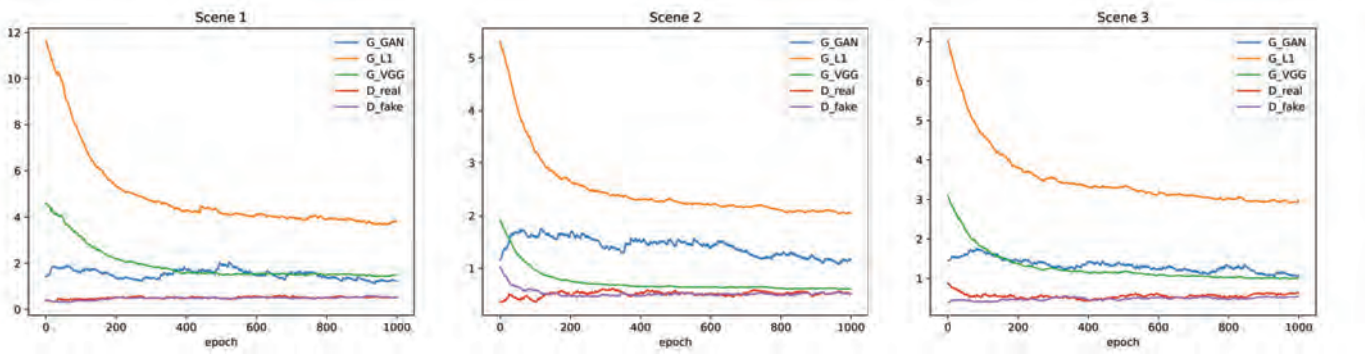
Figure 9 provides a comparison of the results holistically for all three experiments. **Figure 10** provides a detailed visual comparison of the image results. Figure 10(a) presents the generated OFV images without image quality improvement, Fig. 10(b) and (c) show the OFV images with improved image quality, and Fig. 10(d) is the ground truth. When comparing Fig. 10(a) with Fig. 10(b), the overall picture noise is reduced compared to the generated OFV images, even though the improvement of the image details is small. When we compare Fig. 10(a) with Fig. 10(c), most artifacts and overshadowing areas were corrected, and the image quality improved significantly.

We evaluated the generated results quantitatively with PSNR, SSIM, RMSE, LPIPS, and FID, respectively, for different scenes. When assessing the quality of cube mapping results, we calculated the average score of 6 perspectives (up, down, front, back, left, and right) with different evaluation metrics.

In **Table 1**, the image division method resulted in better quality than the original image in all scenes and metrics. However, the method without image division does not perform consistently. In some cases, it performs even better than the method with image division, while in some cases, the results are worse than original images. The results clarify that image division is necessary when using the GAN method to improve the quality of the results consistently.

7. Discussion and Future Work

One of the applications of our research is to generate free-viewpoint tour videos for museums or world heritage sites by walking through the scene with an omnidirectional camera in hand. Many of these environments are



(a) (b) (c)

Fig. 8 Curve of loss during training; (a), (b), (c): the loss curves when training with cube map images captured in the three experiment scenes respectively

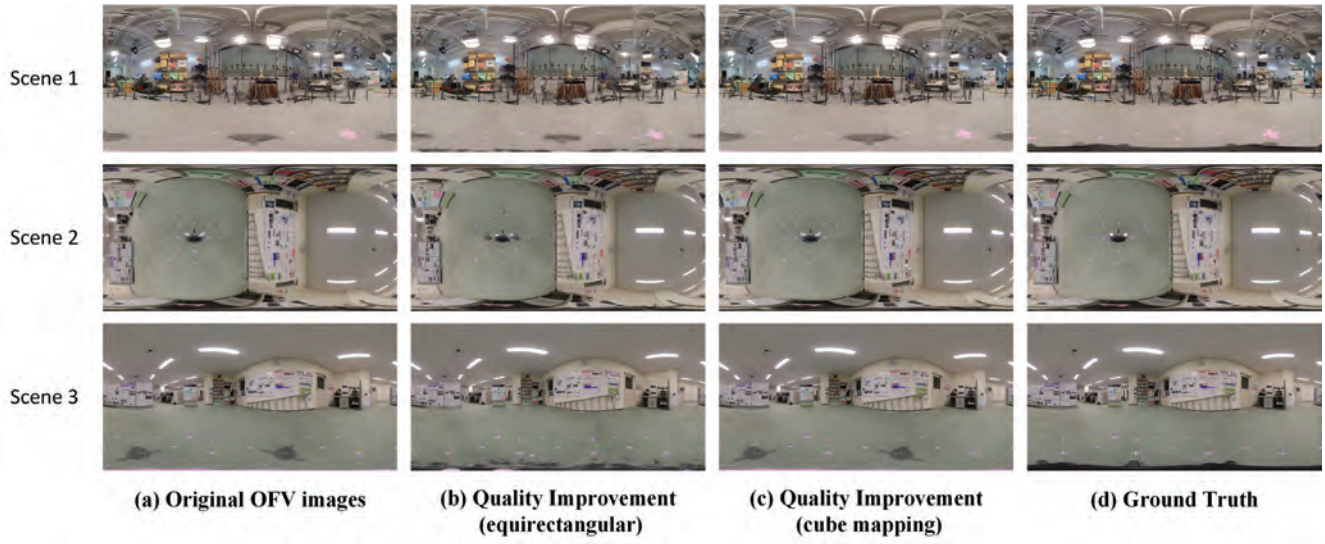
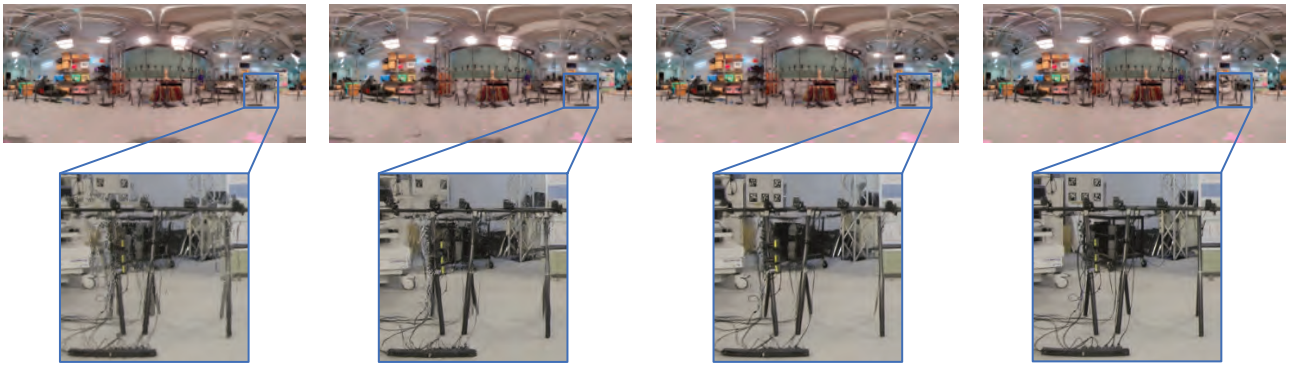


Fig. 9 Results of generated OFV images and image quality improvement with proposed method in different scenes



(a) Original OFV images (b) Quality improvement (equirectangular) (c) Quality improvemnet (cube mapping) (d) Ground truth

Fig. 10 Detail comparison for OFV images

complex and have confounding factors. Our image quality improvement approach could eliminate such kinds of artifacts. The limitations of our system are the processing time and overall complexity. For the step of 3D estimation, the processing time increases exponentially with the number of images. Also, training separate models for different perspectives also costs plenty of time. For example, in our experiment, the training process cost about 6 hours for the model of each perspective. In future work, we plan to optimize the current procedure and the tools we use and generate high-quality omnidirectional images in real-time.

8. Conclusions

We proposed a generation method of high-quality omnidirectional free-viewpoint images. We used the omnidirectional images captured by a camera to estimate the 3D information of a captured scene (i.e., dense point cloud) and camera parameters, which then generated a depth image at any given position. Subsequently, the depth information was back-projected by several captured images to obtain an OFV image at that viewpoint. We employed the deep learning method to eliminate the artifacts and the missing regions within the OFV image (i.e., GAN). The performance of GAN was significantly improved by image division due to the reduced diversity of appearance and the much-reduced number of structural features to be learned per model. We verified the improvement of our proposed method by several assessment metrics, including both conventional and novel metrics that have a high correspondence to human perception.

This work references the article "Image-quality Improvement of Omnidirectional Free-viewpoint Images by Generative Adversarial Networks" published in VISAPP2020. Additional experiments were conducted by changing the shooting scenes and devices.

Table 1 Quantitative comparison on images of different scenes ("Equirectangular" and "Cube Mapping" labels: the format of images for evaluation)

Methods	PSNR↑	RMSE↓	SSIM↑	LPIPS↓	FID↓
Scene 1					
Equirectangular					
Original	19.8702	0.1962	0.8837	0.1791	61.5231
GAN without image division	**20.7632**	**0.1670**	0.8332	0.1682	55.9696
GAN with Image division	19.8974	0.1732	**0.8894**	**0.1406**	**50.5382**
Cube Mapping					
Original	24.2378	0.1151	0.8817	0.1761	113.4358
GAN without image division	25.3958	0.1106	0.8783	0.1812	114.9271
GAN with image division	**25.5883**	**0.1104**	**0.8901**	**0.1062**	**75.0969**
Scene 2					
Equirectangular					
Original	27.3838	0.0817	0.9194	0.0877	44.4624
GAN without image division	28.0804	0.0754	**0.9429**	0.1140	42.2554
GAN with Image division	**28.6569**	**0.0707**	0.9267	**0.0779**	**38.6189**
Cube Mapping					
Original	29.3740	0.0609	0.9077	0.1058	36.511
GAN without image division	30.4743	0.5583	0.9323	0.1387	49.2194
GAN with image division	**31.192**	**0.5337**	**0.9359**	**0.0710**	**32.8505**
Scene 3					
Equirectangular					
Original	21.4703	0.1657	0.8891	0.1379	66.7209
GAN without image division	**24.9985**	**0.1103**	**0.8952**	0.1551	**37.2136**
GAN with image division	22.5691	0.1438	0.8919	**0.1207**	64.5273
Cube Mapping					
Original	28.8787	0.0673	0.8915	0.1041	51.6350
GAN without image division	**30.2735**	0.0666	**0.9233**	0.1312	82.7078
GAN with image division	29.3582	**0.0607**	0.9160	**0.0610**	**46.1974**

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 17H01772 and by JST CREST Grant Number JPMJCR14E2, Japan

QL and OT designed this system with support from IK, HS, YK, and HK. QL summarized and improved the previous OT research results, improved the loss function of GAN by adding a loss term to the perception, and performed a comprehensive validation and evaluation of the whole system. QL also added evaluation criteria and wrote the paper. KI and OT participated in its review and revision.

References

1) O. Takeuchi, H. Shishido, Y. Kameda, H. Kim, I. Kitahara: "Generation Method for Immersive Bullet-Time Video Using an Omnidirectional Camera in VR Platform", Proc. of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Co-located with MM 2018(AVSU 2018), pp. 19–26 (2018).

2) I. Kitahara, Y. Ohta: "Scalable 3D Representation for 3D Video Display in a Large-Scale Space", Proc. of IEEE Virtual Reality, Vol. 2003-Janua, No. 2, pp. 45–52 (2003).

3) W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, L. McMillan: "Image-Based Visual Hulls", Proc. of the 27th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '00, pp.369–374. (2000).

4) R. A. Newcombe, A. Fitzgibbon, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton: "KinectFusion: Real-Time Dense Surface Mapping and Tracking", Proc. of 2011 10th IEEE International Symposium on Mixed and Augmented Reality, pp.127–136 (2011).

5) S. M. Seitz, C. R. Dyer: "View Morphing", Proc. of the 23rd Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '96, pp.21–30 (1996).

6) T. Kanade, P. Rander, P. J. Narayanan: "Virtualized Reality: Constructing Virtual Worlds from Real Scenes", IEEE Multimedia, Vol. 4, No. 1, pp. 34–47 (1997)

7) S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, R. Szeliski: "Building Rome in a day", Communications of the ACM, Vol. 54, No. 10, pp. 105–112 (2011).

8) T. Shin, N. Kasuya, I. Kitahara, Y. Kameda, Y. Ohta: "A Comparison between Two 3D Free-Viewpoint Generation Methods Player-Billboard and 3D Reconstruction", Proc. of 3DTV-CON 2010: The True Vision - Capture, Transmission and Display of 3D Video, pp. 3–6 (2010).

9) M. Tanimoto: "FTV: Free-Viewpoint Television", Signal Processing: Image Communication, Vol. 27, No. 6, pp. 555–570 (2012).

10) P. Hedman, T. Ritschel, G. Drettakis, G. Brostow: "Scalable Inside-Out Image-Based Rendering", ACM Trans. on Graphics, Vol. 35, No. 6, pp. 1–11 (2016).

11) S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, S. Izadi.: "Holoportation", Proc. of the 29th Annual Symposium on User Interface Software and Technology, pp. 741–754 (2016).

12) N. Inamoto, H. Saito: "Free Viewpoint Video Synthesis and Presentation from Multiple Sporting Videos", Electronics and Communications in Japan, Part III: Fundamental Electronic Science (English translation of Denshi Tsushin Gakkai Ronbunshi), Vol. 90, No. 2, pp. 40–49 (2007).

13) O. Grau, A. Hilton, J. Kilner, G. Miller, T. Sargeant, J. Starck: "A Free-Viewpoint Video System for Visualization of Sport Scenes", SMPTE Motion Imaging Journal, Vol. 116, No. 5–6, pp. 213–219 (2007).

14) M. P. Tehrani, T. Tezuka, K. Suzuki, K. Takahashi, T. Fujii: "Free-Viewpoint Image Synthesis Using Superpixel Segmentation", APSIPA Trans. on Signal and Information Processing, Vol. 6, No. 2017, pp.1–12 (2017).

15) C. Barnes, E. Shechtman, A. Finkelstein, D. B. Goldman: "PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing", ACM Trans. on Graphics, Vol. 28, No. 3, Article No. 24, pp.1–11 (2009).

16) D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros: "Context Encoders: Feature Learning by Inpainting", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016-Decem, pp. 2536–2544 (2016).

17) S. Iizuka, E. Simo-Serra, H. Ishikawa: "Globally and Locally Consistent Image Completion", ACM Trans. on Graphics, Vol. 36, No. 4, Article No. 107, pp.1–14 (2017).

18) O. Takeuchi, H. Shishido, Y. Kameda, H. Kim, I. Kitahara: "Image-Quality Improvement of Omnidirectional Free-Viewpoint Images by Generative Adversarial Networks", Proc. of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020), Vol. 4, pp. 299–306 (2020).

19) D. Caruso, J. Engel, D. Cremers: "Large-Scale Direct SLAM for Omnidirectional Cameras", Proc. of IEEE International Conference on Intelligent Robots and Systems, Vol. 2015-Decem, pp. 141–148 (2015).

20) T. Bertel, M. Yuan, R. Lindroos, and C. Richardt: "OmniPhotos: Casual 360 VR Photography with Motion Parallax", ACM Trans. on Graphics, Vol. 39, No. 6, Article No.266, pp.1–12 (2020).

21) M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, P. Debevec: "Immersive Light Field Video with a Layered Mesh Representation", ACM Trans. on Graphics, Vol. 39, No. 4, pp. 1–15 (2020).

22) B. Attal, S. Ling, A. Gokaslan, C. Richardt, J. Tompkin: "MatryODShka: Real-Time 6DoF Video View Synthesis Using Multi-Sphere Images", Proc. of Computer Vision – 16th European Conference (ECCV 2020), pp.441–459 (2020).

23) K. E. Lin, Z. Xu, B. Mildenhall, P. P. Srinivasan, Y. Hold-Geoffroy, S. DiVerdi, Q. Sun, K. Sunkavalli, R. Ramamoorthi: "Deep Multi Depth Panoramas for View Synthesis", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 12358 LNCS, pp. 328–344 (2020).

24) A. P. Pozo, M. Toksvig, T. F. Schrager, J. Hsu, U. Mathur, A. Sorkine-Hornung, R. Szeliski, B. Cabral: "An Integrated 6DoF Video Camera and System Design", ACM Trans. on Graphics, Vol. 38, No. 6, Article No. 216, pp.1–16 (2019).

25) A. Serrano, I. Kim, Z. Chen, S. DIVerdi, D. Gutierrez, A. Hertzmann, B. Masia: " Motion parallax for 360 RGBD Video", IEEE Trans. on Visualization and Computer Graphics, Vol. 25, No. 5, pp. 1817–1827 (2019).

26) R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, P. Debevec: "A System for Acquiring, Processing, and Rendering Panoramic Light Field Stills for Virtual Reality", ACM Trans. on Graphics, Vol. 37, No. 6, pp. 1–15 (2019).

27) H. Shishido, K. Yamanaka, Y. Kameda, I. Kitahara: "Pseudo-Dolly-In Video Generation Combining 3D Modeling and Image Reconstruction", Adjunct Proc. of the 2017 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct 2017, pp. 327–333 (2017).

28) P. Isola, J. Y. Zhu, T. Zhou, A. A. Efros: "Image-to-Image Translation with Conditional Adversarial Networks", Proc. of -30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Vol. 2017-Janua, pp. 5967–5976 (2017).

29) H. Nagahara, Y. Yagi, M. Yachida: "Super-Resolution from an Omnidirectional Image Sequence", Proc. of 2000 26th Annual Conference of the IEEE Industrial Electronics Society, Vol. 4, pp. 2559–2564 (2000).

30) C. Ozcinar, A. Rana, A. Smolic: "Super-Resolution of Omnidirectional Images Using Adversarial Learning", Proc. of 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP) (2019).

31) E. Upenik, P. Akyazi, M. Tuzmen, T. Ebrahimi: "Inpainting in Omnidirectional Images for Privacy Protection", Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), pp. 2487-2491 (2019).

32) N. Kawai, K. Machikita, T. Sato, N. Yokoya: "Video Completion for Generating Omnidirectional Video without Invisible Areas", IPSJ Trans. on Computer Vision and Applications, Vol. 2, pp.200-213 (2010).

33) C. Li, M. Xu, S. Zhang, P. Le Callet: "State-of-the-Art in 360 Video/Image Processing: Perception, Assessment and Compression", IEEE Journal of Selected Topics in Signal Processing, Vol. 14, No. 1, pp. 5–26 (2019).

34) R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang: "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, No. 1, pp. 586–595 (2018).

35) M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter: "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium", Advances in Neural Information Processing Systems, Vol. 2017-Decem, No. Nips, pp. 6627–6638 (2017).

36) A. Al-Saidi, N. J. Avis, I. J. Grimstead, O. F. Rana: "Distributed Collaborative Visualization Using Light Field Rendering", Proc. of 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGRID 2009, pp. 609–614 (2009)

37) C. Wu: "Towards Linear-Time Incremental Structure from Motion", Proc. of IEEE 2013 International Conference on 3D Vision, pp.127–134. (2013).

38) C. Sweeney, T. Hollerer, M. Turk: "Theia: A Fast and Scalable Structure-from-Motion Library", Proc. of MM 2015 - the 2015 ACM Multimedia Conference, pp. 693–696 (2015).

39) J. L. Schönberger, J. M. Frahm: "Structure-from-Motion Revisited", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016 Decem, pp. 4104–4113 (2016).

40) S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski: "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 519–526 (2006).

41) L. Chen, H. Lin, S. Li: "Depth Image Enhancement for Kinect Using Region Growing and Bilateral Filter", Proc. of International Conference on Pattern Recognition, No. January 2012, pp. 3070–3073 (2012).

42) X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, X. Tang: "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks", Proc. of the European Conference on Computer Vision (ECCV) Workshops, pp. 63–79 (2018).

43) K. Simonyan, A. Zisserman: "Very Deep Convolutional Networks for Large-Scale Image Recognition", Proc. of 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–14 (2015).

(Received May 31, 2021)
(Revised December 18, 2021)

Qiaoge LI
He recieved his B.E. degree in Engineering from Dalian University of Technology in 2016, and his M.S.E. degree in Mechanical Engineering from University of Washington in 2019. Since 2020, he became a Ph.D. student in Empowerment Informatics at the University of Tsukuba. His research interests include novel view synthesis and virtual reality.

Oto TAKEUCHI
He recieved B.E. in Science Engineering from the University of Tsukuba in 2018. In 2017, he joined Computer Vision and Image Media Laboratory supervised by Itaru Kitahara. He recieved his M.E. degree in Intelligent and Mechanical Interaction Systems from University of Tsukuba. His research interest was Computer Vision.

Hidehiko SHISHIDO
He received his M.E. and Ph.D. degrees in Engineering from the University of Tsukuba, Japan, in 2013 and 2016, respectively. In 2016, he served as a researcher at the Japan Institute of Sports Sciences. Since 2017, he has been an Assistant Professor at the University of Tsukuba.

Yoshinari KAMEDA (*Member*)
He received a B. Eng., M. Eng., and Ph. D. in Information Science in 1991, 1993, and 1999 from Kyoto University. He joined University of Tsukuba in 2002 and now a full professor at University of Tsukuba. His research interests include computer vision, 3D data processing, VR, AR, Mixed reality, and 3D user interface.

Hansung KIM
He received his Ph.D. degree from Yonsei University, South Korea, in 2005. He is currently an associate professor in the School of Electronics and Computer Science at the University of Southampton, UK. His research interests include 3D computer vision, audio-visual data processing, and virtual reality.

Itaru KITAHARA (*Member*)
He received his B.E. and M.E. degrees in Science Engineering from University of Tsukuba, Japan in 1994 and 1996, respectively. In 1996, he joined Sharp Corporation. 2000-2003, he was a research associate of University of Tsukuba. He received his Ph.D. in 2003. 2003-2005, he was a researcher at ATR. 2005-2019, he was an assistant professor and associate professor at the University of Tsukuba. Since 2019, he has been a professor at the Center for Computational Sciences, University of Tsukuba. His research interests include computer vision, mixed reality, and intelligent image media.

IIEEJ Paper

Computerized Classification Method for 1p/19q Codeletion in Low Grade Gliomas from Brain MRI Images Using Three Dimensional Radiomics Features

Daiki TANAKA† (*Student Member*), Akiyoshi HIZUKURI†, Ryohei NAKAYAMA†

†Graduate School of Science and Engineering, Ritsumeikan University

<Summary> The purpose of this study was to develop a computerized classification method for 1p/19q codeletion in low grade gliomas (LGGs) from brain MRI (magnetic resonance imaging) images using three dimensional (3D) radiomics features. Our database consisted of brain T2 weighted MRI images (102 LGGs with 1p/19q codeletion and 57 LGGs without it) obtained from 159 patients. In the proposed method, 107 3D radiomics features were extracted from LGG region in T2 weighted MRI images. The feature selection was performed with a least absolute shrinkage and selection operator to reduce redundancy among the extracted 3D radiomics features. A support vector machine (SVM) with the selected 3D radiomics features evaluated the likelihood of 1p/19q codeletion in LGG. A three-fold cross validation method was employed to train and test the proposed method. The classification accuracy, the sensitivity, the specificity, and the area under the receiver operating characteristic curve with the proposed method were 80.5%, 83.3%, 75.4%, and 0.836, respectively, showing an improvement when compared with SVM using 2D radiomics features (74.2%, 78.4%, 66.7%, and 0.783; $p = 0.03$). The proposed method with 3D radiomics features achieved high classification accuracy for 1p/19q codeletion in LGG from brain MRI images and would be useful for determining the patient managements.

Keywords: low grade glioma, 1p/19q codeletion, brain MRI image, 3D radiomics features

1. Introduction

Glioma is a primary tumor in brain, accounting for about 40% of central nervous system tumors[1]. It is classified four grades (grades I-IV) according to its aggressiveness[2]. Low grade gliomas (LGGs) classifying into grades II and III includes three molecular subtypes depending on the presence of mutations in the isocitrate dehydrogenase (IDH) gene and the co-deficiency of the short arm of chromosome 1 and the long arm of chromosome 19 (1p/19q codeletion)[3]. Appropriate treatment and the prognosis depend on those molecular subtypes. Especially, LGGs with 1p/19q codeletion are longer survival and positive response to chemotherapy[4-6]. Therefore, World Health Organization recommends determining the treatment plan of patient with LGG based on the diagnostic result of 1p/19q codeletion by genetic analysis. However genetic analysis for 1p/19q codeletion, which is an invasive examination, causes a heavy burden on patient[7].

To overcome this problem, some investigators have developed computerized classification methods for 1p/19q codeletion in LGGs from brain MRI images. Hirano et al. proposed an estimation method for 1p/19q codeletion using a linear discriminant analysis with two dimensional (2D) radiomics features[8]. Zhou et al. developed a classification method for 1p/19q codeletion using a random forest with 2D radiomics features[9]. Those 2D radiomics features were determined by applying a large number of feature extractors to LGG in MRI images[10]. Some studies have applied convolutional neural networks (CNNs) to classify 1p/19q codeletion in LGGs from brain MRI images. Gonzalez et al. developed a classification method with GoogLeNet[11]. Akkus et al. also used multi-scale CNNs that integrate feature maps of different sizes[12]. The CNN approach achieved higher classification performance than the traditional approach based on a classifier with handcraft features. However, those studies only analyzed a LGG in a slice image with a maximum diameter of the LGG. Therefore, the LGG has not been analyzed in the through-plane direction, which means that only part of the LGG has been analyzed.

So, to analyze the whole LGG three-dimensionally in volume data will worth checking to improve the classification accuracy. The purpose of this study is to develop a computerized classification method for 1p/19q codeletion in LGGs from brain MRI images using three dimensional (3D) radiomics features for the whole LGG.

2. Materials and Methods

2.1 Materials

Our database is consisted of brain T2 weighted MRI images

for 159 patients (mean age 42; age range 13-84 years) obtained from The Cancer Imaging Archive[13]. It includes 102 LGGs with 1p/19q codeletion and 57 LGGs without 1p/19q codeletion. **Figure 1** shows an example of LGGs with/without 1p/19q codeletion in brain T2 weighted MRI images. The size of the T2 weighted MRI image is 256×256 pixels, whereas the number of slices is from 20 to 60. The spatial resolutions are 0.78×0.78 mm and 0.94×0.94 mm. The slice thicknesses are also 3.0 and 7.5 mm. Those images are resized using a trilinear interpolation[14] to be an isotropic voxel with a pixel size of 1 mm. The trilinear interpolation can be seen as a linear interpolation of two bilinear interpolations at through-plane direction. In the Cancer Imaging Archive, all LGG regions are provided as mask images.

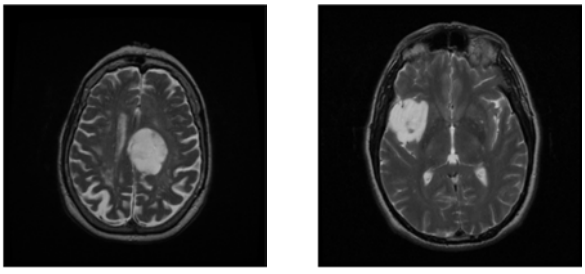
The proposed method was developed and evaluated using python3.7.0 on a workstation (CPU: Intel Core i7-9700K processor, RAM: 48 GB, and GPU: NVIDIA GeForce RTX 2070 SUPER).

2.2 Extraction of 3D radiomics features

To classify 1p/19q codeletion in LGG, 107 3D radiomics features based on seven categories were extracted from a LGG region based on the provided mask image in T2 weighted MRI images using pyradiomics[15]. The seven categories were described below:

(1) Morphological Characteristics (MC)[16] : 14 features
(2) Intensity First Order Statistics (IFOS)[17] : 18 features
(3) Gray Level Co-occurrence Matrix (GLCM)[18] : 24 features
(4) Gray Level Size Zone Matrix (GLSZM)[19), 20] : 16 features
(5) Gray Level Run Length Matrix (GLRLM)[21), 22] : 16 features
(6) Neighboring Gray Tone Difference Matrix (NGTDM)[20), 23] : 5 features
(7) Gray Level Dependence Matrix (GLDM)[24), 25] : 14 features

Category (1) evaluates the shape information of the LGG, while the six remaining categories (2) - (7) evaluates the intensity variations and the tissue textures. Pyradiomics has extended the analysis of the matrixes such as GLCM, GLSZM, and GLRLM to 3D and can determine 3D radiomics features from the whole LGG region (see pyradiomics documentation[26] for details). The IFOS is determined from the histogram based on the signal intensities in the entire LGG region. The GLCM represents the probabilities of occurrences of combination of the signal intensities at a particular distance and direction. In this study, the GLCMs for each of 13 angles in 3D (26-connectivity) were determined with the distance of 1 pixel. The GLSZM and the GLRLM represent the size and the length of connected area with the same voxel intensity, respectively. The NGTDM shows the difference between the signal intensities and the average signal intensity within a particular distance, whereas the GLDM represents the relationship between every signal intensity in the entire LGG region and all intensities at a particular distance. Those different matrices evaluate the uniformity and the regularity of the signal intensities. The GLSZMs, GLRLMs, GTDMs and GLDMs were determined from 13 angles.



(a) LGG with 1p/19q codeletion (b) LGG without 1p/19q codeletion

Fig.1 Example of brain T2 weighted MRI images

2.3 Feature selection

A least absolute shrinkage and selection operator (LASSO)[27] was employed to perform feature selection by reduce redundancy among the extracted 3D radiomics features. The LASSO which is one of the linear regression models is a regularization least-square method using the sum of absolute values of regression coefficients. The linear regression model including the LASSO was defined by

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{e}. \quad (1)$$

Here, $\boldsymbol{y}$ was the teacher signals (LGG with/without 1p/19q codeletion), and $\boldsymbol{x}$ was the extracted 3D radiomics features. $\boldsymbol{\beta}$ and $\boldsymbol{e}$ were the regression coefficients and the error term, respectively. In the LASSO, the regression coefficients were determined by

$$\boldsymbol{\beta} := arg\min_{\boldsymbol{\beta}}(\|\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1), \quad (2)$$

where $\|\boldsymbol{\beta}\|_1$ was the L_1 norm constraint of $\boldsymbol{\beta}$. λ was the tuning parameter that controls the strength of the constraint. Some of the regression coefficients that satisfy Eq. (2) were given as zero. Note that the 3D radiomics features had the coefficient of zero were assumed to be redundant. The 3D radiomics features given non-zero were determined as the features for classifying 1p/19q codeletion in LGG. The parameter λ was determined by grid search. Here, the range of λ was set to from 0.01 to 0.02 in increments of 0.001.

2.4 Classification of 1p/19q codeletion in LGG

A support vector machine (SVM)[28] was employed to distinguish LGGs with 1p/19q codeletion from LGGs without it. The SVM determines the hyperplane that maximizes the margin (distance) from the hyperplane to the closest point across both classes.

The 3D radiomics features selected with the LASSO was inputted to the SVM. Here, those features were normalized. The output of the SVM provided the likelihood of 1p/19q codeletion in LGG. Gaussian kernel[29] was empirically determined as a kernel function in the SVM. The parameters C and γ in the SVM were determined by grid search. The range of C that controls misclassification was set to from 1 to 1000, whereas that of γ that controls the complexity of the hyperplane was set to from 0.0001 to 0.1 in increments of 10 times.

2.5 Evaluation indices

The classification accuracy, the sensitivity[30], the specificity[30] were used to evaluate the classification performance of the proposed method. The area under the receiver operating characteristic curve (AUC)[31] was also used for evaluation of classification performance.

A k-fold cross-validation method[32] with $k = 3$ was used for training and testing of the proposed method. This method divided our database randomly into three groups. One group was used as a test dataset, whereas the remaining two groups were used as a training dataset. This process was repeated three times until every group was used as test dataset.

To evaluate the usefulness of 3D analysis, the classification performance for the proposed method with 3D radiomics features was compared to that for the classification method with 2D radiomics features. In the classification method with 2D radiomics features, 107 2D radiomics features, which were the same as the 3D radiomics features, were determined from a LGG region. Feature selection and classification process were the same manner as the proposed method.

3. Results

Figure 2 shows the selected 3D radiomics features by the LASSO in each training dataset using the three-fold cross validation method. With the grid search, the parameter λ of the LASSO in each dataset was determined to be 0.016 for dataset 1, 0.019 for dataset 2, and 0.019 for dataset 3, respectively. The numbers of the selected 3D radiomics features in the training datasets 1, 2, and 3 were 17, 20, and 19, respectively, whereas 11 3D radiomics features were selected in common for all training datasets. The 11 3D features were Elongation, Maximum, Minimum, Cluster Shade, Inverse Difference Moment Normalized, Maximal Correlation Coefficient, Long Run Low Gray Level Emphasis, Strength, Coarseness, Dependence Variance, and Large Dependence High Gray Level Emphasis.

Based on the highest the F1-score[33] in each training dataset, the parameter C of the SVM with the selected 3D radiomics features were determined to be 100 for dataset 1, 1,000 for dataset 2, and 1,000 for dataset 3, respectively. The parameter γ was also determined to be 0.001 in all training datasets.

Table 1 shows the classification performances for the proposed method with 3D radiomics features and the computerized method with 2D radiomics features. Seven 2D features (Maximum, Kurtosis, Cluster Shade, Long Run Low Gray Level Emphasis, Coarseness, Dependence Variance, and Large Dependence High Gray Level Emphasis) were selected in common for all 2D training datasets. Of the 11 3D features selected in all training datasets, Elongation, Minimum, Inverse Difference Moment Normalized, Maximal Correlation Coefficient, and Strength were not included in the 2D features selected in common. Those features evaluate mainly the local homogeneity in signal intensities within LGG region. The classification accuracy, the sensitivity, the specificity, and the AUC for the proposed method were 80.5% (128/159), 83.3% (85/102), 75.4% (43/57), and 0.836, respectively, which were greater than those for the computerized method with 2D radiomics features (74.2%, 78.4%, 66.7%, and 0.783; p = 0.03).

4. Discussion

In this study, 3D radiomics features for the whole LGG were used to classify 1p/19q codeletion in LGG from brain T2 weighted MRI images. The 3D radiomics features that also analyze in the through-plane direction can significantly improve the classification performance with the 2D radiomics features.

Elongation, Maximum, Minimum, Cluster Shade, Inverse Difference Moment Normalized, Maximal Correlation Coefficient, Long Run Low Gray Level Emphasis, Strength, Coarseness, Dependence Variance, and Large Dependence High Gray Level Emphasis were selected in common for all training datasets as shown in Fig.2. Most of those features are related to the complexity and the momentum in signal intensities within LGG region. Therefore, we believe the distribution of the signal intensities within LGG region contributes to classify 1p/19q codeletion. The regression coefficient for the Dependence Variance tended to be high in each training dataset. This feature is related to the homogeneity in signal intensities. Therefore, the homogeneity in signal intensities

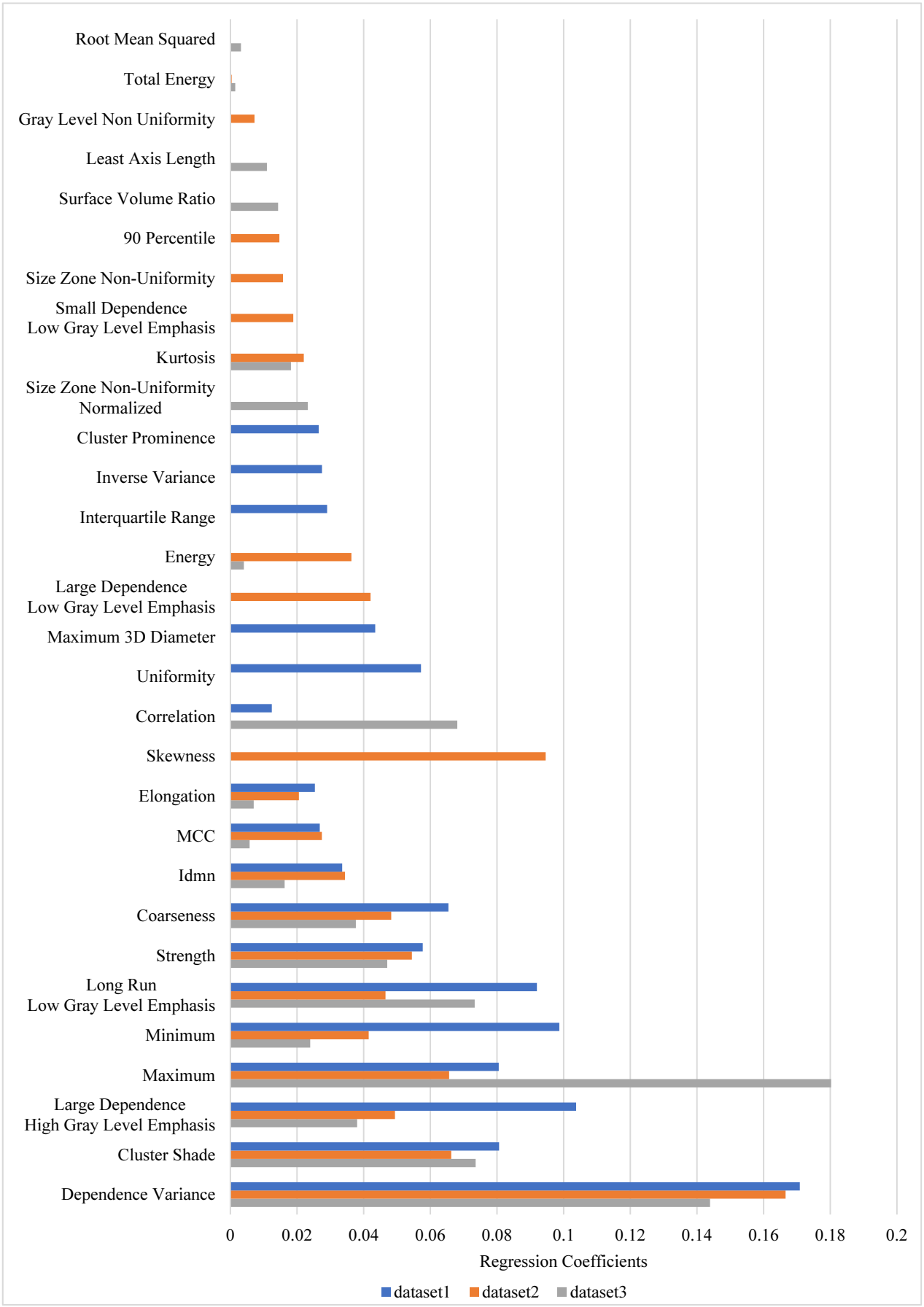


Fig.2 Regression coefficients of the LASSO for selected 3D radiomics features in each training dataset

Table 1 Comparison of classification performances for the proposed method and the computerized method with 2D radiomics features

Method	*k*	Classification accuracy [%]	Sensitivity [%]	Specificity [%]	AUC
Computerized method with 2D radiomics features	1	77.4 (41/53)	82.4 (28/34)	68.4 (13/19)	0.797
	2	67.9 (36/53)	76.5 (26/34)	52.6 (10/19)	0.714
	3	77.4 (41/53)	76.5 (26/34)	78.9 (15/19)	0.844
	Ave.	74.2 (118/159)	78.4 (80/102)	66.7 (38/57)	0.783*
Proposed method	1	77.4 (41/53)	85.3 (29/34)	63.2 (12/19)	0.807
	2	81.1 (43/53)	82.4 (28/34)	78.9 (15/19)	0.879
	3	83.0 (44/53)	82.4 (28/34)	84.2 (16/19)	0.882
	Ave.	**80.5 (128/159)**	**83.3 (85/102)**	**75.4 (43/57)**	**0.836**

*p = 0.03

might be effective for the classification of 1p/19q codeletion.

To evaluate the usefulness of the 3D radiomics features, the proposed method was compared to the computerized method with 2D radiomics features corresponding to the selected 3D radiomics features. The classification accuracy, the sensitivity, the specificity, and the AUC for the computerized method with 2D radiomics features corresponding to the selected 3D radiomics features were 67.9% (108/159), 75.5% (77/102), 54.4% (31/57), and 0.726 (p = 0.001), respectively, which were lower than those for the proposed method with the selected 3D radiomics features. 2D radiomics features would not have evaluated the image features and the texture features accurately in the whole LGG region.

In the LGGs with 1p/19q codeletion which were correctly classified by the proposed method and were incorrectly classified by the computerized method with 2D radiomics features, the signal intensities within the LGG regions tended to be uniform in the slice plane, but tended to be non-uniform in the through-plane direction. On the other hand, in the LGGs without 1p/19q codeletion which were correctly classified by the proposed method and were incorrectly classified by the computerized method with 2D radiomics features, the signal intensities within the LGG regions appeared non-uniform in the slice plane, but could be evaluated as uniform throughout the LGGs. The 2D analysis of LGG in the slice image was inaccurate, and the 3D analysis proved to be important.

In previous studies, the CNN approach achieved higher classification performance than the traditional approach based on a classifier with handcraft features[11), 12)]. We also compared the proposed method with the 3D convolutional neural networks (3D CNNs)[34)]. The architecture of the 3D CNNs was ResNet-18[35)] extended to 3D. The classification accuracy, the sensitivity, the specificity, and the AUC for the proposed method were higher than those for the 3D CNNs (68.6%, 72.5%, 61.4%, and 0.687; p =0.007). As our database was too small to provide the enough training data for the 3D CNNs it was impossible to learn the difference between LGG with 1p/19q codeletion and LGG without 1p/19q codeletion, in signal pattern. On the other hand, the SVM used in this study has only two parameters which were cost (C) and gamma (γ). Thus, it is easy to adjust those parameters compared with the 3D CNNs.

To investigate the usefulness of the feature selection based on the LASSO, we compared the proposed method with the selected 3D radiomics features to the SVM with all 3D radiomics features (107 3D radiomics features). The classification accuracy, the sensitivity, the specificity, and the AUC for the proposed method were substantially higher than those for the SVM with all 3D radiomics features (77.4%, 81.4%, 70.2%, and 0.799; p = 0.04). Therefore, the feature selection by the LASSO was shown to be useful in the classification of 1p/19q codeletion in LGG.

There are some limitations in this study. One limitation is that our database in this study was small. In the future study, we need to expand our database, and then evaluate the proposed method for the classification of 1p/19q codeletion. The second limitation is that the mask images provided from The Cancer Imaging Archive were used as the LGG regions. It would be boring for radiologists to manually trace the LGG regions at clinical practice. Therefore, we have to develop a segmentation method for LGG.

5. Conclusion

In this study, 3D radiomics features for the whole LGG were used to improve the classification accuracy for 1p/19q codeletion in LGG from brain MRI images. The proposed method exhibited a higher classification accuracy when compared to the computerized method with 2D radiomics features.

Acknowledgements

This work was supported by JSPS department expense19K20719.

References

1) K. U. Rathod, Y. D. Kapse: “Automated Brain Tumor Detection and Brain MRI Classification Using Artificial Neural Network – A Review”, International Journal of Science and Research, Vol. 5, No.7, pp. 175–179 (2016).
2) M. Nitta, T. Komori: “Outline and Problems of the WHO 2016 Classification of Tumors of the Central Nervous System”, Japanese Journal of Neurosurgery, Vol. 25, No. 11, pp. 782–791 (2017).
3) D. N. Louis, A. Perry, G. Reifenberger, A. V. Deimling, D. Figarella-Branger, W. K. Cavence, H. Ohgaki, O. D. Wiestler, P. Klcihues, D. W. Ellison: “The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary”, Acta Neuropathologica, Vol. 131, No. 6, pp. 803–820 (2016).
4) K. Ichimura, H. Arita, Y. Narita: “Genetic Basis for the Development of Adult Gliomas”, Japanese Journal of Neurosurgery, Vol. 23, No. 7, pp. 532–540 (2014).
5) G. Cairncross, M. Wang, E. Shaw, R. Jenkins, D. Brachman, J. Buckner, K. Fink, L. Souhami, N. Laperriere, W. Curran, M. Mehta: “Phase III Trial of Chemoradiotherapy for Anaplastic Oligodendroglioma: Long-term Results of RTOG 9402”, Journal of Clinical Oncology, Vol. 31, No. 3, pp. 337–343 (2013).
6) M. W. Ruff, J. Uhm: “Anaplastic Glioma: Treatment Approaches in the Era of Molecular Diagnostics”, Current Treatment Options in Oncology, Vol. 19, No. 12, p. 61 (2018).
7) S. Miki, K. Ichimura, Y. Narita: “How to Understand The Results of Basic Glioma Genome Sequence Data”, Japanese Journal of Neurosurgery, Vol. 26, No. 11, pp. 806–816 (2017).
8) N. Hirano, Y. Uchiyama: “Radiomics for Estimating 1p/19q Codeletion in Brain Tumor Using Magnetic Resonance Imaging”, International Journal of Computer Assisted Radiology and Surgery, Vol. 15, No. 1, pp. S126–S127 (2020).
9) H. Zhou, D. Chang, H. -X. Bai, B. Xiao, C. Su, W. -L. Bi, P. -J. Zhang, J. T. Senders, M. Vallieres, V. K. Kavouridis, A. Boaro, O. Arnaout, L. Yang, R. -Y. Huang: “Machine Learning Reveals Multimodal MRI Patterns Predictive of Isocitrate Dehydrogenase and 1p/19q Status in Diffuse Low and High-grade Gliomas”, Journal of Neuro-Oncology, Vol. 142, No. 2, pp. 299–307 (2019).
10) H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, P. Lambin: “Decoding Tumour Phenotype by Noninvasive Imaging Using A Quantitative Radiomics Approach”, Nature communications, Vol. 5, No. 1, pp. 1–9 (2014).
11) S. R. Gonzalez, I. Zemmoura, C. Tauber: “Deep Convolutional Neural Network to Predict 1p19q Co-deletion and IDH1 Mutation Status from MRI in Low Grade Gliomas”, Proc. of the 10th International Conference on Pattern Recognition Systems (ICPRS) (2019).
12) Z. Akkus, I. Ali, J. Sedlar, J. P. Agrawal, I. F. Parney, C. Giannini, B. J. Erickson: “Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence”, Journal of Digital Imaging, Vol. 30, pp. 469–476 (2017).
13) B. Erickson, Z. Akkus, J. Sedlar, P. Korfiatis: “Data From LGG-1p19qDeletion”, The Cancer Imaging Archive (2017).
14) P. Thévenaz, T. Blu, M. Unser: “Image Interpolation and Resampling”, In: Handbook of Medical Image Processing and Analysis, edited by Bankman IN. Second Edition. Academic Press, pp. 465–493 (2009).
15) J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, H. J. W. L. Aerts: “Computational Radiomics System to Decode The Radiographic Phenotype”, Cancer Research, Vol.77, No. 21, pp. e104–e107 (2017).
16) W. E. Lorensen, H. E. Cline: “Marching Cubes: A High Resolution 3D Surface Construction Algorithm”, Proc. of ACM SIGGRAPH Computer Graphics, Vol. 21, No. 4, pp. 163–169 (1987).
17) N. Aggarwal, R. K. Agrawal: “First and Second Order Statistics Features for Classification of Magnetic Resonance Brain Images”, Journal of Signal and Information Processing, Vol. 3, No. 2, pp. 146–153 (2012).
18) B. Song, G. Zhang, H. Lu, H. Wang, W. Zhu, P. J. Pickhardt, Z. Liang: “Volumetric Texture Features from Higher-order Images for Diagnosis of Colon Lesions via CT Colonography”, International Journal of Computer Assisted Radiology and Surgery, Vol. 9, No. 6, pp. 1021–1031 (2014).
19) G. Thibault, B. Fertil C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, J. L. Mari: “Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification”, Pattern Recognition and Information Processing (PRIP), pp. 140–145 (2009).
20) M. Valliéres, C. R. Freeman, S. R. Skamene, I. EI. Naqa: “A Radiomics Model from Joint FDG-PET and MRI Texture Features for The Prediction of Lung Metastases in Soft-tissue Sarcomas of The Extremities”, Physics in Medicine & Biology, Vol. 60, No. 14, pp. 5471–5496 (2015).
21) M. M. Galloway: “Texture Analysis Using Gray Level Run Lengths”, Computer Graphics and Image Processing, Vol. 4, No. 2, pp. 172–179 (1975).
22) T. -Y. Kim, H. -J. Choi, H. -G. Hwang, H. -K. Choi: “Three-dimensional Texture Analysis of Renal Cell Carcinoma Cell Nuclei for Computerized Automatic Grading”, Journal of Medical Systems, Vol. 34, No. 4, pp. 709–716 (2010).
23) M. Amadasun, R. King: “Textural Features Corresponding to Textural Properties”, IEEE Trans. on Systems, Man, and Cybernetics, Vol. 19, No. 5, pp. 1264–1274 (1989).
24) C. Sun, W. G. Wee: “Neighboring Gray Level Dependence Matrix for Texture Classification”, Computer Vision, Graphics, and Image Processing, Vol. 23, No. 3, pp. 341–352 (1983).

25) F. Bianconi, M. L. Fravolini, I. Palumbo, G. Pascoletti, S. Nuvoli, M. Rondini, A. Spanu, B. Palumbo: “Impact of Lesion Delineation and Intensity Quantisation on The Stability of Texture Features from Lung Nodules on CT: A Reproducible Study”, Diagnostics, Vol. 11, No. 7, p. 1224 (2021).
26) Pyradiomics, https://pyradiomics.readthedocs.io/en/latest/index.html, (2022).
27) R. Tibshirani: “Regression Shrinkage and Selection via The Lasso”, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1, pp. 267–288 (1996).
28) V. N. Vapnik: “Statistical Learning Theory”, John Wiley & Sons, (1998).
29) W. Wang, Z. Xu, W. Lu, X. Zhang: “Determination of the Spread Parameter in the Gaussian Kernel for Classification and Regression”, Neurocomputing, Vol. 55, No. 3–4, pp. 643-663 (2003).
30) A. Benjaminse, A. Gokeler, C. P. van der Schans: “Clinical Diagnosis of an Anterior Cruciate Ligament Rupture: A Meta-analysis”, Journal of Orthopaedic and Sports Physical Therapy, Vol. 36, No. 5, pp. 267–288 (2006).
31) T. Fawcett: “ROC Graphs: Notes and Practical Considerations for Data Mining Researchers”, Technical Report HPL-2003-4, HP Labs (2003).
32) R. Kohavi: “A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection”, Proc. of the 14th International Joint Conference on Artificial Intelligence, Vol. 2, No. 12, pp. 1137–1143 (1995).
33) A. P. Zijenbos, B. M. Dawant, R. A. Margolin, A. C. Palmer: “Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation”, IEEE Trans. on Medical Imaging, Vol. 13, No. 4, pp. 716–724 (1994).
34) M. Decuyper, S. Bonte, K. Deblaere, R. V. Holen: “Automated MRI Based Pipeline for Glioma Segmentation and Prediction of Grade, IDH Mutation and 1p19q Co-deletion”, Computerized Medical Imaging and Graphics, Vol. 88, No. 101831, pp. 1–-9 (2021).
35) K. He, X. Zhang, S. Ren, J. Sun: “Deep Residual Learning for Image Recognition”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).

(Received Nov. 19, 2021)
(Revised Jan. 22, 2022)

Daiki TANAKA (*Student Member*)

He is a Master course student at the Graduate School of Engineering, Ritsumeikan University. He is involved in research on machine learning.

Akiyoshi HIZUKURI

He received his Ph.D. degree from Mie University in 2014. During 2014-2018, he was working at Mizuho Research & Technologies, Ltd. Since 2018, he has been working at Ritsumeikan University. He is currently an assistant professor at the Graduate School of Engineering, Ritsumeikan University. His research interests include machine learning. He is a member of IEICE (Japan), IEEJ, JSMP, and JAMIT.

Ryohei NAKAYAMA

He received his Ph.D. degree from Mie University in 2005. During 2005-2015, he was an assistant professor at Mie University Hospital. Since 2015, he has been working at Ritsumeikan University. He is currently a professor at the Graduate School of Engineering, Ritsumeikan University. During 2008-2009, he was a visiting assistant professor at University of Chicago. His research interests include medical image recognition, and machine learning. He is a member of IEEE, SPIE, IEICE (Japan), IPSJ, JAMIT, ISJ, and JSRT.

IIEEJ **Paper**

Improvements over Coordinate Regression Approach for Large-Scale Face Alignment

Haoqi GAO† (*Student Member*), Koichi OGAWARA†

† The Faculty of System Engineering, Wakayama University

<**Summary**> The facial alignment task has been well-studied extensively and achieved significant progress in recent years. However, previous works remain challenging due to the ambiguity of invisible landmarks under extreme viewpoints (e.g., large pose and expression). This paper proposes a novel dense network on top of the coordinate regression method to improve face coordinates localization in extreme environments. The attention mechanism can better guide the model on which information to emphasize or suppress. In our network, we embed a Convolutional Block Attention Module (CBAM) in three stages of the densely connected convolutional network (DenseNet) respectively for face alignment tasks. Then we concatenate landmark location labels with bounding box location and head pose value as the guide information for our network. We demonstrate the performance of our network through the AFLW2000-3D, AFLW2000-3D-Reannotated, Menpo-3D test datasets. Comparative experiments reveal that our network with lower mean NME (3.2%) outperforms the baseline DenseNet (3.66%), ShuffleNet (4.39%), and DenseNet+SE (3.93%) on AFLW2000-3D-Reannotated. We conclude that our network obtains improved performances for face landmarks prediction even in extreme conditions.

Keywords: face alignment, coordinate regression, CBAM, DenseNet

1. Introduction

Facial alignment[1)–3)]aims to estimate the projections of facial key points (e.g., eyes-corners, nose tip, eyebrows, chin center, and mouth corners) onto the face images, which is a necessary pre-processing for many facial tasks. Face alignment is considered an essential intermediate step for face analysis. Many relevant tasks including face recognition[4),5)], 3D reconstruction[6),7)], and face attribute estimation[8),9)]need to consider it. Neural network-based algorithms have made substantial progress in face alignment and achieved high accuracy rates. Previous face alignment approaches are parametric fitting[10)–12)], regression-based methods[13)–17)]. Regression-based algorithms can further divide into coordinate regression, which takes the landmark coordinates as the regression target, and heatmap regression, which outputs the likelihood response for each landmark. An advantage of a model-fitting-based approach is that it establishes point correspondences between 2D facial images and typical 3D facial models, making it easier to handle facial alignment in complex scenes, especially for invisible points predicted in large-scale poses. However, the fitting-based method is typically time-consuming, and the network structure used is relatively complex, which makes the model's parameter space large. Several studies have shown that regression-based methods perform well and achieve high precision for frontal and near-frontal face images. However, in real-world applications, human faces are often exposed to uncontrolled and unconstrained environments, which opens up many possibilities for capturing face images and brings various challenges to the existing regression-based model training process. For instance, some landmarks become invisible under large-scale pose situations, and even expert operators find it hard to calibrate landmarks positions accurately.

In this work, our goal is to improve the performance of facial landmark estimation in extreme environments based on regression-based methods. In this paper, We recombine and integrate the features extracted from previous dense block layers[18)], comprehensively exploiting both bottom-level high-resolution features and top-level semantics powerful features. Also, inspired by the attention mechanism, we propose a novel dense network for face alignment by embedding the Convolutional Block Attention Module (CBAM)[19)]into a densely connected

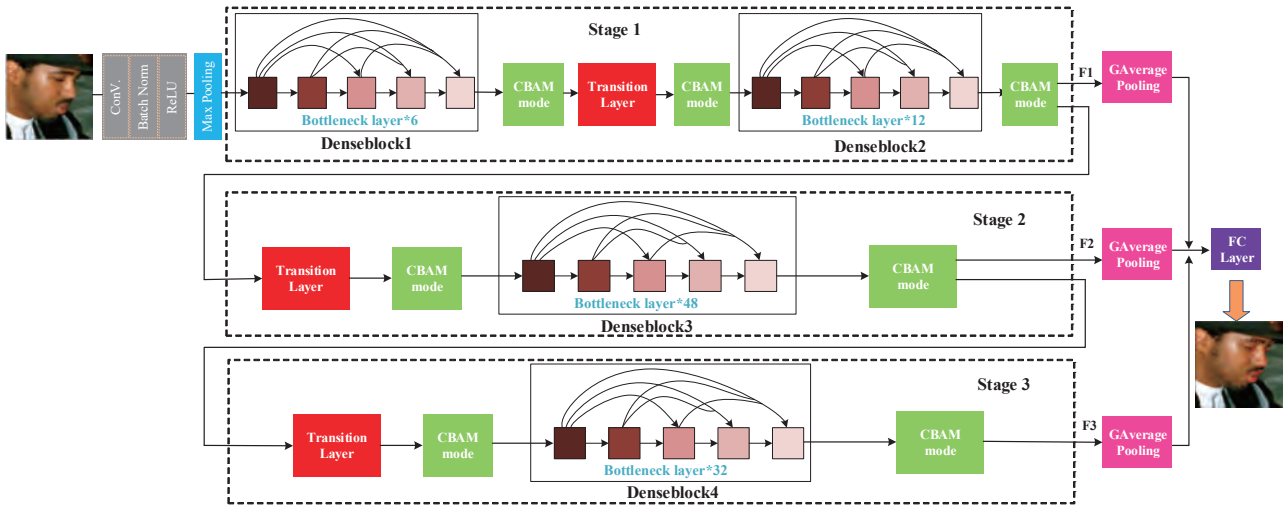


Fig. 1 The proposed network architecture

Convolutional Net. Next, We propose a new loss that incorporates landmark localization loss, GIoU loss of face bounding box, and head pose regression loss to guide our landmark localization model. Moreover, We conduct comparative experiments of Normalized Mean Error (NME) metrics based on AFLW2000-3D-Reannotated datasets, AFLW2000-3D-Test datasets, 300W-Testset-3D, and Menpo-3D datasets. Our method obtains a lower NME value in the experiment compared to other classical network structures like MobileNet[20),21)], ShuffleNet[22),23)]. Furthermore, experimental results show that our method outperforms DenseNet[18)]and DenseNet with Squeeze-and-Excitation(SE)[24)].

2. Related Work

Face alignment task is a well-studied area in the field of computer vision. Based on existing face alignment methods, we categorize methods into three subcategories: coordinate regression-based methods, heatmap regression-based methods, and 3D model fitting-based methods. In this section, we briefly review some previous works of each subcategory.

The output of coordinate-based regression methods is a vector of landmark coordinates. Sun et al.[25)]used a CNN-based coordinate regression method to regress facial landmarks location for the first time. Zhu et al.[26)]was introduced the ODN network to overcome the occlusion problem. A multi-stage model[27)], using spatial transformer networks, hourglass networks, and exemplar-based shape constraints to improve face alignment accuracy. PFLD[28)] employs a branch of the network to estimate geometric information for each face sample, subsequently regularizing the landmark localization. Retinaface[29)]jointly learning facial bounding box locations, facial landmarks, and 3D vertices force the network to learn exclusive facial features.

Heatmap-based regression methods generate heatmaps as the output for each landmark. Newell et al.[30)]were pioneers in designing a stacked hourglass (SH) network to generate heatmaps for human pose estimation. The state-of-the-art performance on facial alignment held for some time by using Hourglass models or other deformable Hourglass models[31),32)]. Deng et al.[33)]adopted a Cascaded Multi-view Hourglass scheme for face alignment. Researchers proposed SRN[2)]for landmark estimation, which can simultaneously obtain the 2D and 3D coordinates of facial landmarks. MobileFAN[34)], using lightweight architectures MobileNetV2 as the backbone to achieve better accuracy.

The method based on 3D model fitting, which offers more expressive and occluded information than its 2D counterpart, has received lots of attention from researchers in recent years. Earlier researchers[1)]explored face alignment to large-pose face images by combining cascaded CNN regressor and 3DMM model. PR-Net[6)] proposes a 2D representation called UV Positional Map that records the 3D shape of a face in UV space, then trains a simple Convolutional Neural Network to regress it from images. VRN[35)]performs a direct regression of the volumetric representation of 3D facial geometry from the 2D image. Wei et al.[36)]propose a graph convolution network to regress 3D face coordinates, which directly performs feature learning on the 3D face mesh. Recent researches for the challenge of wild face images include 2DASL[7)], and 3DDFA2[37)], which show state-of-the-arts for both 3D face reconstruction and dense face alignment

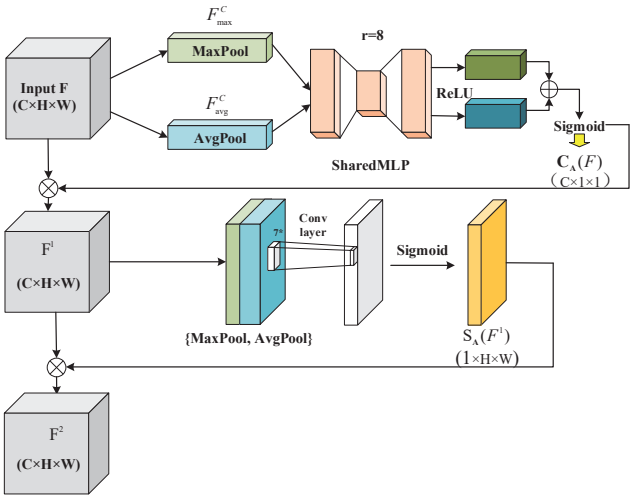


Fig. 2 The structure of CBAM model[19)]

by a large margin.

3. Method Overview

3.1 Network architecture

We employ a hierarchical architecture with several convolutional layers to extract features and make landmark predictions. Considering the satisfactory performance of DenseNet techniques[18)], which is a densely connected network that can perform feature multiplexing well. Attention mechanism improves the representation of interest[19),24),38)]. In our approach, we embed the CBAM module[19)]into DenseNet to efficiently assist the information flow within the network by learning which information to emphasize or suppress.

The pipeline of the proposed method is shown in **Fig. 1**. The detail of the architecture is depicted blockwise: the gray blocks are convolution blocks with the batch norm. The number of bottleneck layers for the dense blocks is given by [6, 12, 48, 32]. The green, blue and red blocks are CBAM models, Max pooling layers, and transition layers, respectively. At the end of each stage, we extract the features defined as F1, F2, and F3, respectively. The output channel dimensions of these features are 512, 1792, and 1920, respectively. Finally, we concatenate these features through Global Average Pooling layers and pass them to an FC layer to obtain the final output.

An illustration of the CBAM model detail provides in **Fig. 2**.

The channel attention block focuses on what is meaningful given an input image. The given feature map $F \in \mathbf{R}^{C\times H\times W}$ is passed through max pooling and average pooling and then through an MLP with a reduction ratio of $r = 8$. The purpose of r is to reduce the parameter's computational effort and effectively suppress overfitting. The SharedMLP is a two-level-fully connected layer. Because for the attention model, we want to make the weights correlate with the input, thus avoiding the issues with fixed weights in the FC layer. A sigmoid activation with smoothing advantages use to generate the final channel attention feature map $C_A(F) \in \mathbf{R}^{C\times 1\times 1}$. Different from channel attention, spatial attention focuses on where the informative part is, which is complementary to channel attention[19)]. In the spatial attention block, channel-based global max pooling and global average pooling are applied to the input F^1, then a convolutional layer with a kernel size of 7×7 is applied. Finally, the attention features $S_A(F) \in \mathbf{R}^{1\times H\times W}$ are generated by a sigmoid function.

Mathematical expressions for the process of the channel and spatial attention mechanisms are given in Eq. (1).

$$F^1 = C_A(F) \otimes F; F^2 = S_A(F^1) \otimes F^1, \tag{1}$$

where $\otimes$ denotes element-wise multiplication. F^2 is the final refined output. The transition layer consists of a batch norm layer, a ReLU layer, a 1×1 convolutional layer, and a 2×2 average pooling layer. Its role is to reduce the dimension of each output channel of the Dense block[18)].

3.2 Loss function

Based on landmark loss function (L), we also add head pose (H) (yaw, pitch, and roll) loss that can guide the distribution of facial rotation and add bounding box (B) loss to help improve the accuracy of the model further.

3.2.1 Landmark loss (L)

Wing loss[39)]for landmark regression to improve the ability to cope with minor and medium-range errors during deep network training. It is defined as follows:

$$W(x) = \begin{cases} \omega \ln\left(1 + |\frac{x}{\epsilon}|\right) & |x| < \omega \\ |x| - C & otherwise \end{cases}, \tag{2}$$

$$C = (\omega - \omega \ln\left(1 + (\frac{\omega}{\epsilon})\right)). \tag{3}$$

$l_i = \{(l_{x_1}, l_{y_1}), ..., (l_{x_{68}}, l_{y_{68}})\}_i$ represent the predicted landmarks coordinate(face bounding, eye, eyebrows, nose, and mouth) and $l_i^* = \left\{(l_{x_1}^*, l_{y_1}^*), ..., (l_{x_{68}}^*, l_{y_{68}}^*)\right\}_i$ represent the landmarks coordinate of the ground-truth, and x=$|l_i - l_i^*|$ and M is the number of images. According to Wing loss[39)], we also set $\omega = 10, \epsilon = 2$.

3.2.2 Head pose loss (H)

For head pose, $p_i = \{p_1, p_2, p_3\}_i$ represents the value of the predicted Euler angle (the three unit vectors of

the face gaze direction, that is, pitch yaw and roll). $p_i^* = \{p_1^*, p_2^*, p_3^*\}_i$ represents the value of the ground-truth Euler angle. Here, the loss we used for the pose estimator is the mean squared error (MSE) loss function.

$$L_{MSE} = \frac{1}{M}\sum_{i=1}^{M}(p_i - p_i^*)^2 \tag{4}$$

3.2.3 Bounding box loss (B)

Bounding box regression is one of the most fundamental parts of many computer vision tasks such as object localization, object detection, object tracking. One improvement is the Intersection over Union (IoU) loss. However, when two objects do not overlap, the IoU value will be zero and not reflect how far the two shapes are from each other. For non-overlapping targets, if the IoU uses as a loss, its gradient will be zero and cannot be optimized. Inspired by[40], we incorporate Generalized Intersection over Union (GIoU) loss into the bounding box loss. This loss preserves the main properties of IoU while having a gradient in all potential cases, including non-overlapping situations. The calculation process of GIOU and the difference with the IOU algorithm in the **Algorithm 1**.

$$L_{GIoU} = \sum_{i=1}^{M} 1 - GIoU(b_i, b_i^*) \tag{5}$$

$b_i = \{b_{x_1}, b_{y_1}, b_{x_2}, b_{y_2}\}_i$ represents the coordinates of the predicted bounding box and $b_i^* = \{b_{x_1}^*, b_{y_1}^*, b_{x_2}^*, b_{y_2}^*\}_i$ represents the coordinates of the ground-truth bounding box.

For each face image i, its multi-task loss defines as:

Algorithm 1 A Comparison for Iou and GIou algorithm

Input:
Predicted bounding box coordinate B^p, and groundtruth bounding box coordinate B^g;
Output:
GIoU and IoU;
1: For $B^p = (b_{x_1}, b_{y_1}, b_{x_2}, b_{y_2})$, calcuate $\bar{b}_{x_1}, \bar{b}_{y_1}, \bar{b}_{x_2}, \bar{b}_{y_2}$;
$\bar{b}_{x_1} = min(b_{x_1}, b_{x_2})$, $\bar{b}_{x_2} = max(b_{x_1}, b_{x_2})$;
$\bar{b}_{y_1} = min(b_{y_1}, b_{y_2})$, $\bar{b}_{y_2} = max(b_{y_1}, b_{y_2})$;
2: The area $A^p = (\bar{b}_{x_2} - \bar{b}_{x_1}) \times (\bar{b}_{y_2} - \bar{b}_{y_1})$;
The area $A^g = (b_{x_2}^* - b_{x_1}^*) \times (b_{y_2}^* - b_{y_1}^*)$;
3: The intersection $I = (b_{x_1}^\circ, b_{y_1}^\circ, b_{x_2}^\circ, b_{y_2}^\circ)$;
$b_{x_1}^\circ = max(\bar{b}_{x_1}, b_{x_1}^*)$, $b_{x_2}^\circ = min(\bar{b}_{x_2}, b_{x_2}^*)$;
$b_{y_1}^\circ = max(\bar{b}_{y_1}, b_{y_1}^*)$, $b_{y_2}^\circ = min(\bar{b}_{y_2}, b_{y_2}^*)$;
4: The area $A^I = (b_{x_2}^\circ - b_{x_1}^\circ) \times (b_{y_2}^\circ - b_{y_1}^\circ)$
if $b_{x_2}^\circ > b_{x_1}^\circ, b_{y_2}^\circ > b_{y_1}^\circ$; else $A^I = 0$;
5: The smallest enclosing box $B^c = (b_{x_1}^c, b_{y_1}^c, b_{x_2}^c, b_{y_2}^c)$;
$b_{x_1}^c = min(\bar{b}_{x_1}, b_{x_1}^*)$, $b_{x_2}^c = max(\bar{b}_{x_2}, b_{x_2}^*)$;
$b_{y_1}^c = min(\bar{b}_{y_1}, b_{y_1}^*)$, $b_{y_2}^c = max(\bar{b}_{y_2}, b_{y_2}^*)$;
6: The area $A^c = (b_{x_2}^c - b_{x_1}^c) \times (b_{y_2}^c - b_{y_1}^c)$;
7: IoU=$\frac{A^I}{U}$ where $U = A^p + A^g - A^I$;
8: GIoU=$IoU - \frac{A^c - U}{A^c}$;

$$Loss = \lambda_1 \sum_{i=1}^{M} W(x) + \lambda_2 L_{GIoU} + L_{MSE} \tag{6}$$

Empirically, we set $\lambda_1 = 0.1$, $\lambda_2 = 0.25$.

4. Experiments

4.1 Datasets

"300W-LP[1]" dataset contains 61,225 synthetic face images along with their corresponding 2D and 3D ground-truth landmark-based annotations. It is synthesized from 300W[41]through a morphable 3D model profiling algorithm proposed in paper[1]and coverage across large pose ranges from -90 to 90 degrees.

"300-VW[16]" is the large-scale face tracking dataset that contains 114 videos. Among these, 64 videos were used for testing and the rest 50 videos were used for training with 95,192 frames in total. For the test videos, we have three categories (CatA, CatB, and CatC). And CatC is considered the most challenging video due to its low resolution and poor facial quality.

"300W-Testset-3D[41]" contains two categories: Indoor and Outdoor. The dataset of the 300-W Challenge includes a total of 600 images from the in-the-wild.

"AFLW2000-3D[1]" contains 2,000 face samples chosen from the AFLW dataset[42], introduced by Zhu et al. along with 300W-LP datasets with varying expressions and illumination conditions.

"AFLW2000-3D-Reannotated[16]" is re-annotated to make the ground truth more accurate because for faces with difficult poses, the method from paper[1]fails to produce precise landmarks. It provides along with the LS3D-W dataset. Following the literature of[2], we also categorize the face images in AFLW2000-3D-Reannotated into three groups of views [0°, 30°], [30°, 60°], [60°, 90°]. It includes 1312, 390, and 296 images in the three groups, respectively.

"Menpo-3D[43]" dataset contains a total of 8,955 challenging frames that vary in illuminations, poses, and occlusions.

4.2 Evaluation metrics

A metric used for facial landmark localization algorithm is the point-to-point Euclidean distance, normalized by the square root of the ground-truth bounding box size instead of the common inter-pupil[16]. The formula for NME can be written as Eq.(7).

$$GTE(l_i, l_i^*) = \frac{1}{N}\sum_{i=1}^{N}\frac{||l_i - l_i^*||_2}{d_i} \tag{7}$$

Table 1 NME compared with different network model

Method	AFLW2000-3D-Reannotated NME(%)				
	[0°,30°]	[30°,60°]	[60°,90°]	Mean	Std
MobileNet[20]	5.10	6.58	7.29	6.32	1.12
ShuffleNet[22]	2.98	4.15	5.15	4.09	1.09
DenseNet+SE[24]	3.01	4.21	4.58	3.93	0.82
DenseNet[18]	2.76	3.99	4.24	3.66	0.79
DenseNet+CBAM	**2.37**	**3.33**	**3.89**	**3.20**	**0.77**

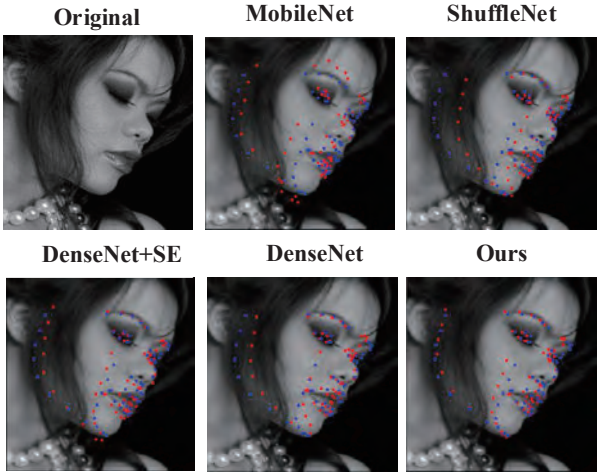


Fig. 3 Comparison results of different networks with Ours model (red–predicted,bule–ground truth)

Herein, N is the number of landmarks, l_i is the predicted landmarks points, l_i^* is their corresponding ground truth, and d_i denotes the distance for the i-th frame. It is computed as $d_i = \sqrt{w_b h_b}$. w_b and h_b are the width and height of the bounding box.

4.3 Experimental results and discussion

In **Table 1**, we compare with other existing mainstream neural network architectures (e.g.,ShuffleNet [22),23)], MobileNet[20),21)]and DenseNet[18)]). Attention mechanism SE[24)]uses global average pooled features to compute the channel-wise attention. Model DenseNet+SE means SE attention added into the DenseNet architecture.

We learn and compare each of these models with the same (Ours(L+B+H)) loss function. As the Table 1 shown, our model (DenseNet+CBAM) has better performance than other advanced networks. Compared to the original DenseNet and DenseNet+SE structures, our model (DenseNet+CBAM) performs with a reduction in NME(%) of 0.46 and 0.73, respectively.

In **Fig. 3**, we give comparative visualization results for different networks. Our model predicts the location with the smallest distance from the ground truth.

In **Table 2**, we compared Ours(L), Ours(L+B), and Ours(L+H) to verify the effectiveness of both head pose and bounding box loss functions under same (DenseNet+CBAM) model. Ours(L) only employs landmark coordinates, Ours(L+B) adds the bounding box position based on Ours(L), and Ours(L+H) adds the head pose information. Ours(L+B+H) represents an experiment concatenating landmark points, head pose, and bounding box information.

Table 2 NME compared with different loss functions

Method	AFLW2000-3D NME(%)				
	[0°,30°]	[30°,60°]	[60°,90°]	Mean	Std
RCPR[44]	4.26	5.96	13.18	7.80	4.74
ESR[45]	4.60	6.7	12.67	7.99	4.19
SDM[46]	3.67	4.94	9.76	6.12	3.21
Yu et al.[47]	3.62	6.06	9.56	6.41	2.99
3DDFA[1]	3.78	4.54	7.93	5.42	2.21
3DDFA+SDM[1]	3.43	4.24	7.17	4.94	1.97
3DSTN[48]	3.15	4.33	5.98	4.49	1.42
ExpNet[49]	4.01	5.46	6.23	5.23	1.13
FAME[50]	3.11	**3.84**	6.60	4.52	1.84
SS-SFN[51]	3.09	4.27	5.59	4.31	1.25
Ours(L)	3.18	4.78	5.59	4.52	1.23
Ours(L+B)	3.03	4.76	5.42	4.40	1.23
Ours(L+H)	2.99	4.59	5.31	4.30	1.19
Ours(L+B+H)	**2.96**	4.38	**4.95**	**4.10**	**1.02**

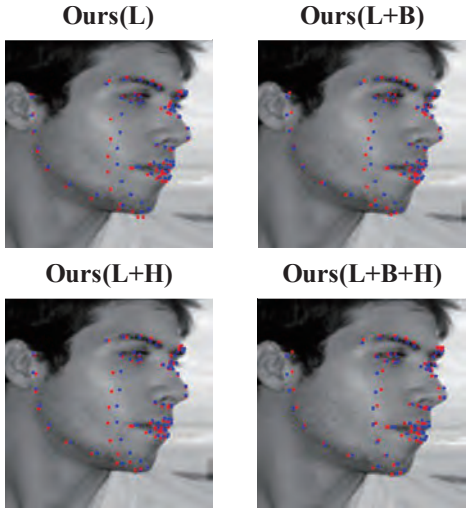


Fig. 4 Comparison results of Ours(L), Ours(L+B), Ours(L+H), and Ours(L+B+H) with the ground truth (red–predicted,bule–ground truth)

Table 2 results show our method drops 0.21 in Mean NME values compared with SS-SFN[51)]. Even though FAME[50)]performs well in [30°, 60°] range, our method achieves lowest value in both the [0°, 30°] and [60°, 90°].

In **Fig. 4**, we also give comparative visualization results for different loss functions.

Figure 4 results show that Ours(L+B+H) model predicts keypoint locations (red points) with the smallest distance compared to ground truth (blue points). It demonstrates that incorporating both bounding box and head

pose information can assist the model in improving face alignment accuracy.

In **Table 3**, we evaluated our model on 300W-Testset-3D and Menpo-3D test datasets, which collect under challenging conditions. Our proposed method achieves lower NME values when compared with other methods.

In **Fig. 5**, we compared our model with traditional CNN+3D method[52]on 300W private datasets. As shown in the figure above, the third test image from the left contains the glasses obscuration challenge and the exaggerated expression pose challenge. Traditional algorithms have large errors in estimating key points of the mouth and face contours. However, our model can accurately locate the facial key points with small errors.

Table 3 Comparison of NME on the Menpo-3D and 300W testset datasets

Method	300WTestset NME(%)	Menpo3D NME(%)
FAN[16]	2.83	3.70
SRN[2]	2.77	3.35
Ours	**2.38**	**2.84**

Fig. 5 Comparing results of our method with CNN+3D[52] on images 300W private test set

More landmark detection visualization results of our model are shown in **Fig. 6**. The main test cases focus on large-scale expressions, illumination, and 90-degree side face challenges. Figure 6 result shows that our model can locate landmarks on face images clearly and accurately in those challenging scenarios and can also accurately detect invisible points for the side-face dataset.

We also compare our method's predictions (red points) with the ground truth (blue points) in **Fig. 7**. The results clearly show that our methods can produce landmarks with minor errors compared to the ground truth.

All experiment results on various test datasets (e.g., AFLW2000-3D, 300W-Testset-3D, and Menpo-3D) demonstrate the robustness of our model even under large poses, occlusions, exaggerated expressions, low-resolution, and different light conditions.

Next, to evaluate the processing speed, we calculate our algorithm's efficiency on an Nvidia GTX 2080 Ti GPU. Our model's performance computes on the resolution of 256×256. To evaluate the model size and computation complexity, we calculate the number of network parameters (#Parameters) and Frames Per Second (FPS), a measure of computational speed.

Table 4 lists comparison reports of network complexity and runtime with other methods.

Fig. 6 Example results of our method on image from 300W-Testset-3D and Menpo-3D databases

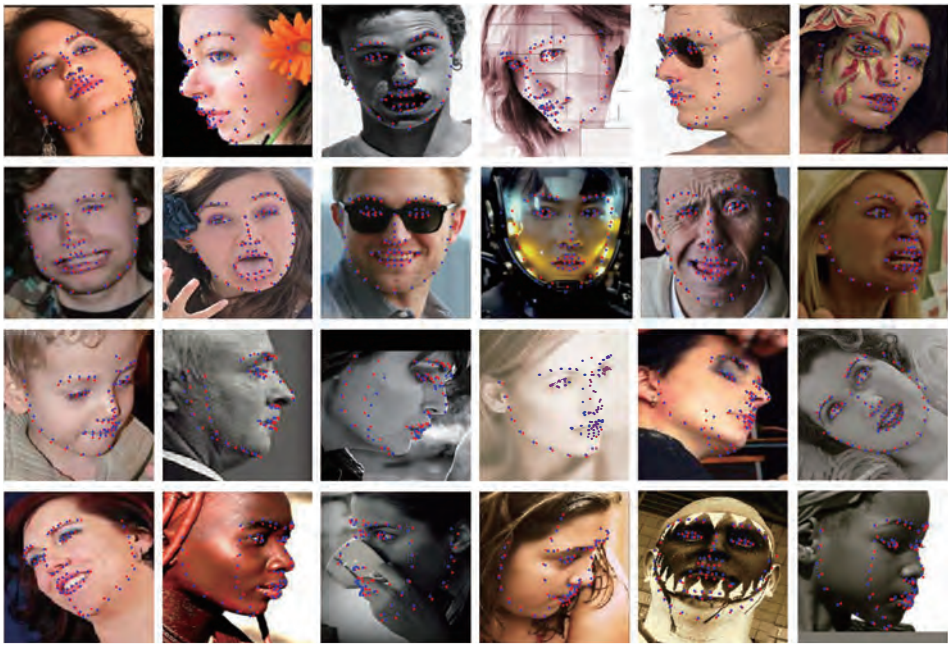


Fig. 7 Comparing results of our method with ground truth on images from the 300W-Testset-3D(first row), AFLW2000-3D(second-row), and Menpo-3D(third and fourth-row) databases respectively

Table 4 Comparison in terms of network complexity and runtime

Method	Year	Runtime		
		#Parameters	GPU	FPS
CALE[53]	2016	∼ 140M	-	3
DVLN[54]	2017	∼ 132M	-	-
DAN[55]	2017	∼ 22M	-	45
SAN[56]	2018	∼ 57.4M	343ms	50
LAB[57]	2018	∼ 50.7M	60ms	6
GEAN[58]	2020	-	58.8.0ms	-
LUVLi[59]	2020	-	17.0ms	-
Ours	-	∼ 20.4M	14ms	75

The overall number of parameters is about 20.4M in our network, and it takes 33min to train one epoch on the 300W-LP dataset (61,225 images) and 300VW-train dataset (95,192 images). The inference speed is around 75 fps. As can be seen, our model has smaller parameters and lower computational complexity than other methods (e.g., LAB[57], GEAN[58], LUVLi[59]).

5. Conclusion

In this paper, we propose an improved coordinate regression method to estimate the location of facial landmarks. We build upon the superior performance of Densenet techniques[18]by applying multi-scale feature maps and adding attention modules that can emphasize or suppress information flow within the network by learning. We further guide the detection by concatenating landmark coordinates, head pose information, and bounding box locations. Experimental results show that our approach outperforms existing comparative methods in challenging large-scale face alignment tasks. In addition, extensive experiments on challenging datasets (such as large poses, occlusions, exaggerated expressions, low-resolution, and different light conditions) also show that our training model is stable and robust.

References

1) X. Zhu, Z. Lei, X. Liu, H. Shi, S.-Z. Li: "Face Alignment Across Large Poses: A 3D Solution", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.146–155 (2016).
2) X. Luo, P. Li, F. Chen, Q. Zhao: "Improving Large Pose Face Alignment by Regressing 2D and 3D Landmarks Simultaneously and Visibility Refinement", Proc. of Chinese Conference on Biometric Recognition. Springer, Cham, pp.349–357 (2018).
3) Y. Liu, A. ourabloo, W. Ren, X. Liu: "Dense Face Alignment", Proc. of IEEE International Conference on Computer Vision Workshops (ICCVW), pp.1619–1628 (2017).
4) Y. Tai, J. Yang, Y. Zhang, L. Luo, J. Qian, Y. Chen: "Face Recognition with Pose Variations and Misalignment via Orthogonal Procrustes Regression", IEEE Trans. on Image Processing, Vol.25, No.6, pp.2673–2683 (2016).
5) S. Bhattacharya, G. S. Nainala, S. Rooj, A. Routray: "Local Force Pattern (LFP): Descriptor for Heterogeneous Face Recognition", Pattern Recognition Letters, Vol. 125, pp.63–70 (2019).
6) Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou: "Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network", Proc. of European Conference on Computer Vision (ECCV), pp. 534–551 (2018).
7) X. Tu, J. Zhao, Z. Jiang, Y. Luo, M. Xie, Y. Zhao, L.X. He, Z. Ma, J. Feng: "3D Face Reconstruction From A Single Image Assisted by 2D Face Images in the Wild", IEEE Trans. on Multimedia, Vol. 23, pp. 1160–1172 (2020).
8) T. Devries, K. Biswaranjan, G. W. Taylor: "Multi-Task Learning of Facial Landmarks and Expression", Proc. of Canadian Conference on Computer and Robot Vision, pp.98–103 (2014).

9) M. I. N. P. Munasinghe: "Facial Expression Recognition Using Facial Landmarks and Random Forest Classifier", Proc. of IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), pp. 423–427 (2018).
10) I. Matthews, S. Baker: "Active Appearance Models Revisited", International Journal of Computer Vision, Vol.60, No.2, pp.135–164 (2004).
11) T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham: "Active Shape Models-Their Training and Application", Computer Vision and Image Understanding, Vol.61, No.1, pp.38–59 (1995).
12) D. Cristinacce, T. F. Cootes: "Feature Detection and Tracking with Constrained Local Models", Proc. of British Machine Vision Virtual Conference (BMVC), Vol.1, No.2, pp.3 (2006).
13) S. Zhu, C. Li, C. Change Loy, X. Tang: "Face Alignment by Coarse-to-Fine Shape Searching", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4998–5006 (2015).
14) Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang: "Quantized Densely Connected U-Nets for Efficient Landmark Localization", Proc. of European Conference on Computer Vision (ECCV), pp.339–354 (2018).
15) J. Yang, Q. Liu, K. Zhang: "Stacked Hourglass Network for Robust Facial Landmark Localisation", Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.79–87 (2017).
16) A. Bulat, G. Tzimiropoulos: "How Far are We from Solving the 2D & 3D Face Alignment Problem?(and a Dataset of 230,000 3D Facial Landmarks)", Proc. of IEEE International Conference on Computer Vision (ICCV), pp.1021–1030 (2017).
17) S. Ren, X. Cao, Y. Wei, J. Sun: "Face Alignment at 3000 Fps via Regressing Local Binary Features", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1685–1692 (2014).
18) G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger: "Densely Connected Convolutional Networks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4700–4708 (2017).
19) S. Woo, J. Park, J. Lee, Y., I. So Kweon: "Cbam: Convolutional Block Attention Module", Proc. of European Conference on Computer Vision (ECCV), pp.3–19 (2018).
20) M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen: "Mobilenetv2: Inverted Residuals and Linear Bottlenecks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520 (2018).
21) A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. X. Tan, W. J. Wang, Y. K. Zhu, R. M. Pang, V. Vasudevan, H. Adam: "Searching for Mobilenetv3." Proc. of IEEE International Conference on Computer Vision (ICCV), pp.1314–1324 (2019).
22) N. Ma, X. Zhang, H.T. Zheng, J. Sun: "Shufflenet V2: Practical Guidelines for Efficient CNN Architecture Design", Proc. of European Conference on Computer Vision (ECCV) pp. 116–131 (2018).
23) X. Zhang, X. Zhou, M. Lin, J. Sun: "Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6848–6856 (2018).
24) J. Hu, L. Shen, G. Sun: "Squeeze-and-Excitation Networks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018).
25) Y. Sun, X. Wang, X. Tang: "Deep Convolutional Network Cascade for Facial Point Detection", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3476–3483 (2013).
26) M. Zhu, D. Shi, M. Zheng, M. Sadiq: "Robust Facial Landmark Detection via Occlusion-Adaptive Deep Networks", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.3486–3496 (2019).
27) H. Wang, R. Cheng, J. Zhou, L. Tao, H. K. Kwan: "Multistage Model for Robust Face Alignment Using Deep Neural Networks", Cognitive Computation, pp.1–17 (2021).
28) X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, H. Ling: "PFLD: A practical facial landmark detector." Computer Research Repository (CoRR), Vol.abs/1902.10859 (2019).
29) J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou: "Retinaface: Single-Shot Multi-Level Face Localisation in the Wild", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5203–5212 (2020).
30) A. Newell, K. Yang, J. Deng: "Stacked Hourglass Networks for Human Pose Estimation", Proc. of European Conference on Computer Vision (ECCV), pp.483–499 (2016).
31) J. Deng, G. Trigeorgis, Y. Zhou, S. Zafeiriou: "Joint Multi-View Face Alignment in the Wild", IEEE Trans. on Image Processing, Vol.28, No,7, pp.3636–3648 (2019).
32) X. Wang, L. Bo, L. Fuxin: "Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression", Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 6971–6981 (2019).
33) J. Deng, Y. Zhou, S. Cheng, S. Zaferiou: "Cascade Multi-View Hourglass Model for Robust 3D Face Alignment", Proc. of IEEE International Conference on Automatic Face Gesture Recognition (FG) pp.399–403 (2018).
34) Y. Zhao, Y. Liu, C. Shen, Y. Gao, S. Xiong: "MobileFAN: Transferring Deep Hidden Representation for Face Alignment", Pattern Recognition, Vol. 100, pp.107–114 (2020).
35) A. S. Jackson, A. Bulat, V. Argyriou, G. Tzimiropoulos: "Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression", Proc. of IEEE International Conference on Computer Vision (ICCV), pp.1031–1039 (2017).
36) H. Wei, S. Liang, Y. Wei: "3D Dense Face Alignment via Graph Convolution Networks", Computer Research Repository (CoRR), Vol. abs/1904.05562 (2019).
37) J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, S. Z. Li: "Towards Fast, Accurate and Stable 3D Dense Face Alignment", Proc. of Computer Vision ECCV European Conference, Glasgow, Part XIX 16. Springer International Publishing, pp. 152–168 (2020).
38) B. Chen, J. Wang, Z. Chi: "Improved DenseNet with Convolutional Attention Module for Brain Tumor Segmentation", Proc. of Third International Symposium on Image Computing and Digital Medicine, pp. 22–26 (2019).
39) Z. H. Feng, J. Kittler, M. Awais, P. Huber, X. J. Wu: "Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2235–2245 (2018).
40) H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese: "Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666 (2019).
41) C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic: "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge", Proc. of IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 397–403 (2013).
42) M. Koestinger, P. Wohlhart, P. M. Roth, H. Bischof: "An-

notated Facial Landmarks in the Wild: A Large-Scale, Real-World Database for Facial Landmark Localization", Proc. of IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 2144–2151 (2011).

43) S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, J. Shen: "The Menpo Facial Landmark Localisation Challenge: A Step towards the Solution", Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 170–179 (2017).
44) X. P. Burgos-Artizzu, P. Perona, P. Dollar: "Robust Face Landmark Estimation under Occlusion", Proc. of IEEE International Conference on Computer Vision (ICCV), pp.1513–1520 (2013).
45) X. Cao, Y. Wei, F. Wen, J. Sun: "Face Alignment by Explicit Shape Regression", International Journal of Computer Vision, Vol.107, No.2, pp.177–190 (2014).
46) X. Xiong, F. De la Torre: "Supervised Descent Method and its Applications to Face Alignment", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.532–539 (2013).
47) R. Yu, S. Saito, H. Li, D. Ceylan, H. Li: "Learning Dense Facial Correspondences in Unconstrained Images", Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 4723–4732 (2017).
48) C. Bhagavatula, C. Zhu, K. Luu, M. Savvides: "Faster than Real-Time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses", Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 3980–3989 (2017).
49) F. J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, G. Medioni: "Expnet: Landmark-Free, Deep, 3D Facial Expressions", Proc. of 13th IEEE International Conference on Automatic Face Gesture Recognition (FG), pp.122–129 (2018).
50) F. J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, G. Medioni: "Deep, Landmark-Free Fame: Face Alignment, Modeling, and Expression Estimation", International Journal of Computer Vision, Vol.127, No.6, pp. 930–956 (2019).
51) B. Chaudhuri, N. Vesdapunt, B. Wang: "Joint Face Detection and Facial Motion Retargeting for Multiple Faces", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9719–9728 (2019).
52) R. Valle, J. M. Buenaposada, A. Valdes, L. Baumela: "Face Alignment Using a 3D Deeply-Initialized Ensemble of Regression Trees", Computer Vision and Image Understanding, Vol.189, pp. 102846 (2019).
53) A. Bulat, G. Tzimiropoulos: "Convolutional Aggregation of Local Evidence for Large Pose Face Alignment", Proc. of British Machine Vision Conference (BMVC) (2016).
54) W. Wu, S. Yang: "Leveraging Intra and Inter-Dataset Variations for Robust Face Alignment", Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 150–159 (2017).
55) M. Kowalski, J. Naruniec, T. Trzcinski: "Deep Alignment Network: A Convolutional Neural Network for Robust Face Alignment", Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 88–97 (2017).
56) X. Dong, Y. Yan, W. Ouyang, Y. Yang: "Style Aggregated Network for Facial Landmark Detection", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 379–388 (2018).
57) W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, Q. Zhou: "Look at Boundary: A Boundary-Aware Face Alignment Algorithm", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2129–2138 (2018).
58) S. M. Iranmanesh, A. Dabouei, S. Soleymani, H. Kazemi, N. Nasrabadi: "Robust Facial Landmark Detection via Aggregation on Geometrically Manipulated Faces", Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 330–340 (2020).
59) A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X.M. Liu, C. Feng: "LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood", Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8236–8246 (2020).

(Received September 30, 2021)
(Revised February 21, 2022)

Haoqi GAO (*Student Member*)
She received the M.E. degree from Beihang University of Software in 2016. From 2019 she stared a Ph.D. in system engineering at Wakayama University. Her research interests include face detection, 3D face reconstruction, medical image processing, and deep learning.

Koichi OGAWARA
He received his Ph.D. degree in Information and Communication Engineering from the University of Tokyo, Japan, in 2002. From 2006 to 2011, he was a guest Associate Professor at Kyusyu University. Currently, he is an Associate Professor at Wakayama University. His research interests include robotics, computer vision, and machine learning. He won the Best Vision Paper Award at the IEEE International Conference on Robotics and Automation 2007.

Call for Papers

Special Issue on Data Interwork, Sharing, and Reusing Technologies Supporting Digital Transformation Era

IIEEJ Editorial Committee

In order to solve various social issues speedy and economically, Digital Transformation (DX) is expected to realize Fully Digitalized Society. Therefore, not only the natural language and image/video of the outer world, individual behavior and even the motivation and Kansei of the sole human are going to be processed digitally. The concept of smart home and smart city will be the example of such trend. The utilization of the data automatically acquired from the user will improve the specification of the products, and sharing service will degrade the hurdle of possession of the expensive products for the customers, in such perspective. To promote such tide, the interwork, sharing and reusing the digital data without converting into analogue data or without truncating it is very important.

Especially in image related area, the technologies to convert and process the image data and its attributes will contribute to produce the virtual outlook of the products without manufacturing and the image analysis by AI will provide further benefit in business.

This special issue of the paper targets various image-related technologies such as image processing, generation, expression, transmission and combined applications of these that support and promote data interwork, sharing, and reusing for the digital transformation era, and calls for any category (Ordinary paper, Short paper, System development paper, Data paper, Practice Oriented Paper) of papers.

1. Topics covered include but not limited to

   Image Processing, Image Recognition, Image Detection, Image/Video Communication, Image Input/Output, Computer Graphics, Visualization, Binocular Vision, 3D image processing, Computer Vision, Pattern Recognition, Big Data, Image Databases, Machine Learning, Deep Learning, Understandability, Annotation, Attribute Presumption, Web-based Technology, Security. Creativity, Usability, Interpretability, Human Interface and Interaction, User Experience, Ubiquitous, AR/VR/MR, Other related fundamental / application / systemized technologies.

2. Treatment of papers
   Submission paper style format and double-blind peer review process are the same as the regular paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as the regular contributed paper. We ask for your understanding and cooperation.

3. Publication of Special Issue:
   IIEEJ Transactions on Image Electronics and Visual Computing Vol.10, No.2 (December 2022)

4. Submission Deadline:
   Friday, July 29, 2022

5. Contact details for Inquires:
   IIEEJ Office E-mail: hensyu@iieej.org

6. Online Submission URL: http://www.editorialmanager.com/iieej/

Call for Papers

Special Issue on
Image-Related Technology for Digital Fabrication

IIEEJ Editorial Committee

In recent years, there has been a lot of research on digital fabrication, which is the process of creating something based on digital data, and various fabrication tools such as 3D printers, laser cutters, 3D scanners, CNC milling machines, etc. have become accessible to the general public, enabling individuals to create new things. As a result, it has become possible to create new products on a personal level. In addition to plastic modeling, metal and wood are also available as materials, and the use of various materials in combination, such as printing electronic circuits, has been proposed, and the development and application of MEMS is also expected. Thus, digital fabrication is not only a rapid prototyping of products, but also a technology that can be adopted in a wide range of fields such as food, clothing, architecture, medicine, education, and entertainment.

In this special issue, we invite category (Ordinary paper, Short paper, System development paper, Data paper, Practice Oriented Paper) of papers to realize digital fabrication. We would be grateful if you could submit your papers.

1. Topics covered include but not limited to

   Computer Graphics, Image Processing, Image Analysis, Image Recognition, Interaction, Computer Vision, Machine Learning/Deep Learning, User Interface, VR, AR, MR, 3D Related Technologies, Image Systems, Image Applications, General Image Basic Technologies

2. Treatment of papers

   Submission paper style format and double-blind peer review process are the same as the regular paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as the regular contributed paper. We ask for your understanding and cooperation.

3. Publication of Special Issue:

   IIEEJ Transactions on Image Electronics and Visual Computing Vol.11, No.1 (June 2023)

4. Submission Deadline:

   Monday, October 31, 2022

5. Contact details for Inquires:

   IIEEJ Office E-mail: hensyu@iieej.org

6. Online Submission URL: http://www.editorialmanager.com/iieej/

Revised: January 6, 2017

Revised: July 6, 2018

Guidance for Paper Submission

1. Submission of Papers

(1) Preparation before submission

- The authors should download "Guidance for Paper Submission" and "Style Format" from the "Academic Journals", "English Journals" section of the Society website and prepare the paper for submission.
- Two versions of "Style Format" are available, TeX and MS Word. To reduce publishing costs and effort, use of TeX version is recommended.
- There are four categories of manuscripts as follows:
  - Ordinary paper: It should be a scholarly thesis on a unique study, development or investigation concerning image electronics engineering. This is an ordinary paper to propose new ideas and will be evaluated for novelty, utility, reliability and comprehensibility. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Short paper: It is not yet a completed full paper, but instead a quick report of the partial result obtained at the preliminary stage as well as the knowledge obtained from the said result. As a general rule, the authors are requested to summarize a paper within four pages.
  - System development paper: It is a paper that is a combination of existing technology or it has its own novelty in addition to the novelty and utility of an ordinary paper, and the development results are superior to conventional methods or can be applied to other systems and demonstrates new knowledge. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Data Paper: A summary of data obtained in the process of a survey, product development, test, application, and so on, which are the beneficial information for readers even though its novelty is not high. As a general rule, the authors are requested to summarize a paper within eight pages.
- To submit the manuscript for ordinsry paper, short paper, system development paper, or data paper, at least one of the authors must be a member or a student member of the society.
- We prohibit the duplicate submission of a paper. If a full paper, short paper, system development paper, or data paper with the same content has been published or submitted to other open publishing forums by the same author, or at least one of the co-authors, it shall not be accepted as a rule. Open publishing forum implies internal or external books, magazines, bulletins and newsletters from government offices, schools, company organizations, etc. This regulation does not apply to a preliminary draft to be used at an annual meeting, seminar, symposium, conference, and lecture meeting of our society or other societies (including overseas societies). A paper that was once approved as a short paper and being submitted again as the full paper after completion is not regarded as a duplicate submission.

(2) Submission stage of a paper

- Delete all author information at the time of submission. However, deletion of reference information is the author's discretion.
- At first, please register your name on the paper submission page of the following URL, and then log in again and fill in the necessary information. Use the "Style Format" to upload your manuscript. An applicant should use PDF format (converted from dvi of TeX or MS Word

format) for the manuscript. As a rule, charts (figures and tables) shall be inserted into the manuscript to use the "Style Format". (a different type of data file, such as audio and video, can be uploaded at the same time for reference.)

http://www.editorialmanager.com/iieej/

- If you have any questions regarding the submission, please consult the editor at our office.

Contact:
Person in charge of editing
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-Ku, Tokyo 116-0002, Japan
E-mail: hensyu@iieej.org
Tel: +81-3-5615-2893, Fax: +81-3-5615-2894

2. Review of Papers and Procedures

(1) Review of a paper

- A manuscript is reviewed by professional reviewers of the relevant field. The reviewer will deem the paper "acceptance", "conditionally acceptance" or "returned". The applicant is notified of the result of the review by E-mail.
- Evaluation method

  Ordinary papers are usually evaluated on the following criteria:
  - ✓ Novelty: The contents of the paper are novel.
  - ✓ Utility: The contents are useful for academic and industrial development.
  - ✓ Reliability: The contents are considered trustworthy by the reviewer.
  - ✓ Comprehensibility: The contents of the paper are clearly described and understood by the reviewer without misunderstanding.

  Apart from the novelty and utility of an ordinary paper, a short paper can be evaluated by having a quickness on the research content and evaluated to have new knowledge with results even if that is partial or for specific use.

  System development papers are evaluated based on the following criteria, apart from the novelty and utility of an ordinary paper.
  - ✓ Novelty of system development: Even when integrated with existing technologies, the novelty of the combination, novelty of the system, novelty of knowledge obtained from the developed system, etc. are recognized as the novelty of the system.
  - ✓ Utility of system development: It is comprehensively or partially superior compared to similar systems. Demonstrates a pioneering new application concept as a system. The combination has appropriate optimality for practical use. Demonstrates performance limitations and examples of performance of the system when put to practical use.

  Apart from the novelty and utility of an ordinary paper, a data paper is considered novel if new deliverables of test, application and manufacturing, the introduction of new technology and proposals in the worksite have any priority, even though they are not necessarily original. Also, if the new deliverables are superior compared to the existing technology and are useful for academic and industrial development, they should be evaluated.

(2) Procedure after a review

- In case of acceptance, the author prepares a final manuscript (as mentioned in 3.).
- In the case of acceptance with comments by the reviewer, the author may revise the paper in consideration of the reviewer's opinion and proceed to prepare the final manuscript (as

mentioned in 3.).

- In case of conditional acceptance, the author shall modify a paper based on the reviewer's requirements by a specified date (within 60 days), and submit the modified paper for approval. The corrected parts must be colored or underlined. A reply letter must be attached that carefully explains the corrections, assertions and future issues, etc., for all of the acceptance conditions.
- In case a paper is returned, the author cannot proceed to the next step. Please look at the reasons the reviewer lists for the return. We expect an applicant to try again after reviewing the content of the paper.

(3) Review request for a revised manuscript

- If you want to submit your paper after conditional acceptance, please submit the reply letter to the comments of the reviewers, and the revised manuscript with revision history to the submission site. Please note the designated date for submission. Revised manuscripts delayed more than the designated date be treated as new applications.
- In principle, a revised manuscript willl be reviewed by the same reviewer. It is judged either accceptance or returned.
- After the judgment, please follow the same procedure as (2).

3. Submission of final manuscript for publication

(1) Submission of a final manuscript

- An author, who has received the notice of "Acceptance", will receive an email regarding the creation of the final manuscript. The author shall prepare a complete set of the final manuscript (electronic data) following the instructions given and send it to the office by the designated date.
- The final manuscript shall contain a source file (TeX edition or MS Word version) and a PDF file, eps files for all drawings (including bmp, jpg, png), an eps file for author's photograph (eps or jpg file of more than 300 dpi with length and breadth ratio 3:2, upper part of the body) for authors' introduction. Please submit these in a compressed format, such as a zip file.
- In the final manuscript, write the name of the authors, name of an organizations, introduction of authors, and if necessary, an appreciation acknowledgment. (cancel macros in the Style file)
- An author whose paper is accepted shall pay a page charge before publishing. It is the author's decision to purchase offprints. (ref. page charge and offprint price information)

(2) Galley print proof

- The author is requested to check the galley (hard copy) a couple of weeks before the paper is published in the journal. Please check the galley by the designated date (within one week). After making any corrections, scan the data and prepare a PDF file, and send it to our office by email. At that time, fill in the Offprint Purchase Slip and Copyright Form and return the scanned data to our office in PDF file form.
- In principle, the copyrights of all articles published in our journal, including electronic form, belong to our society.
- You can download the Offprint Purchase Slip and the Copyright Form from the journal on our homepage. (ref. Attachment 2: Offprint Purchase Slip, Attachment 3: Copyright Form)

(3) Publication

- After final proofreading, a paper is published in the Academic journal or English transaction (both in electronic format) and will also be posted on our homepage.

IIEEJ Trans. on Image Electronics and Visual Computing

Vol. 10, No. 1 2022

Editor in Chief: Mei Kodama
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan

Print: ISSN 2188-1898
Online: ISSN 2188-1901
CD-ROM: ISSN 2188-191x