IIEEJ Transactions on

Image Electronics and Visual Computing

Special Issue on Journal Track Papers in IEVC2019

Vol. 7, No. 2 2019

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.7 No.2 December 2019

CONTENTS

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 IEVC2019

Editor: Prof. Yuriko TAKESHIMA
Tokyo University of Technology

The 6th International Conference on Image Electronics and Visual Computing (IEVC2019) was held in Bali, Indonesia on August 21-24, 2019 as the international academic event of Image Electronics Engineers of Japan (IIEEJ). It was based on the great success of previous five workshops in 2007 (Cairns, Australia), 2010 (Nice, France), 2012 (Kuching, Malaysia), 2014 (Koh Samui, Thailand), and 2017 (Da Nang, Vietnam). The aim of the conference is to bring together researchers, engineers, developers, and students from various fields in both academia and industry for discussing the latest researches, standards, developments, implementations and application systems in all areas of image electronics and visual computing.

There were two paper categories in IEVC2019: general paper and late breaking paper (LBP), and in general paper, there were two tracks: Journal track (JT) and Conference track (CT). In IEVC2019, 33 JT papers, 52 CT papers and 24 LBP were submitted.

Journal track is a newly introduced one and has the advantage to be able to publish the paper on the journal (IIEEJ Trans. on IEVC) in the "Special Issue on Journal Track in IEVC2019" planned on December 2019 issue. JT papers were submitted as full paper version (8 pages) by the conference paper submission deadline (March 2019), to be peer-reviewed in advance. At the timing of review result notification for conference paper, initial review result for journal paper will be also sent to authors. JT papers passed the conference paper level judgement, shorter version (2-4 pages) will be asked to submit for the inclusion into the proceedings. After the conference, authors were given a period of three weeks to revise the paper, to reflect the initial review result and also comments received at the conference presentation.

This special issue is limited to JT papers of which revised version were submitted by the announced deadline, and includes seven papers which passed the review process to be in time for the publication schedule. We also plan the special issue on "Extended Papers Presented in IEVC2019" for all presenters in IEVC2019 in the next transaction.

Finally, I would like to thank all the reviewers and editors for their time and efforts towards 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 IIEEJ office for various kinds of support.

IIEEJ **Paper** -- *Special Issue on Journal Track Papers in IEVC2019* --

Binary Malignancy Classification of Skin Tissue Using Reflectance and Texture Features from Macropathology Multi-Spectral Images

Eleni ALOUPOGIANNI†, Hiroyuki SUZUKI††, Takaya ICHIMURA†††,
Atsushi SASAKI†††, Hiroto YANAGISAWA†††, Tetsuya TSUCHIDA†††,
Masahiro ISHIKAWA††††(*Member*), Naoki KOBAYASHI††††(*Fellow*), Takashi OBI††

† Tokyo Institute Of Technology, Department of Information and Communications Engineering,
†† Tokyo Institute of Technology, Research Institute for Innovation in Science and Technology,
††† Saitama Medical University, Faculty of Medicine,
†††† Saitama Medical University, Faculty of Health and Medical Care

<**Summary**> This study suggests an analysis procedure on macropathology multi-spectral images (macroMSI), for visual representation of grossly malignant regions of skin samples during excision margin pathological diagnosis. We implemented binary malignancy classification on a database of ten high-resolution 7-channel macroMSI tissue samples, captured before and after formalin fixing. We reconstructed spectral reflectance by Wiener estimation and described texture using local binary patterns (LBP). Highlighted malignancy regions were derived from an optimal classifier selected by cross-validated performance. The results show that malignant regions are highlighted fairly accurately and indicate the importance of analyzing unfixed tissue in conjunction with fixed tissue.

Keywords: macropathology, multispectral imaging, spectral reflectance, texture features, skin cancer classification

1. Introduction

Conditions related to skin cancer are a prevalent health concern in Japan. In year 2013 alone, 19.706 skin cancer incidents were recorded across the country[1]and skin cancer crude mortality rates almost doubled from 1999 in 2014[2]. Skin lesion treatment begins with macropathology, which refers to the initial examination of excised tissue specimens, prior microscopic evaluation[3]. Usual protocol requires compilation of a report describing gross features and photographs of specimens to map dissection sites after formalin fixing and bread-loafing. Relevant gross features are upheld by changes in shape, size, color or texture[6], while melanoma detection employs the commonly referred as ABCDE macroscopic descriptors (asymmetry, border, color, diameter, evolution)[7].The pathologist's goal during initial biopsy is to accurately identify critical tissue areas on the specimen and assess the condition of their margins, in order to determine whether further ablation is required. A major concern in macropathology is protocol and equipment discrepancies among different pathology laboratories. Specifications of the camera system, image acquisition algorithm, scene illumination and display device are factors which cause color variation in the resulting image[4]. Imaging quality is essential for effective pathology, taking into account that in many instances a second evaluation is necessary. Furthermore, macropathology is dependent on the age, training and experience of the physician[5], while lacking standardization and automation as a procedure. Consequently, it is time-consuming and produces high workload for the pathology laboratory. In the advent of digital pathology, macropathology remains an impractical task.

Multi-spectral images (MSI), an enhancement of the RGB format, can alleviate obstacles in image quality and color reproduction through the use of narrow band filters, expressing a spectral dimension. MSIs exhibit higher sensitivity to image features that are masked in RGB images, while MSI cameras become more affordable due to technological advancement. Additionally, MSI is preferred for retrieving spectral surface reflectance of objects compared to RGB. Machine learning classification at dermatologist level based on conventional or multispectral images has been already investigated with favorable accuracy[8),9]. However, macropathology images are generally neglected, with current focus being on multispectral digital slide images for histological analysis[10], improvement of color validity[11]and cancer tissue classification[12].

By applying traditional machine learning techniques on macropathology MSI (macroMSI) of excised skin tissue, this study aims to investigate automatic classification of tissue malignancy and provide a visual tool for diagnosis that performs in a consistent manner independently of capturing conditions.

We propose a novel quasi-automated framework based on color and texture analysis of macroMSI, capable of classifying malignancy of critical regions on excised tissue in order to assist margin identification for skin cancer in current macropathology practice. Our approach attempts to mimic the pathologists' assessment of color and texture, by combining hand-crafted features from reconstructed spectral reflectance and local binary patterns, respectively. Our aim is two-fold. First, we investigate the effectiveness of features derived from a new macroMSI dataset, composed of both fixed and unfixed skin specimens during macropathology. Secondly, we examine a visual representation of probability malignancy on critical areas based on binary classification score from traditional machine learning classifiers.

2. Materials and Methods

2.1 Dataset and capture

All skin specimens comprising our dataset were obtained from the Central Pathology Department of Saitama Medical University Hospital in Saitama, Japan. The study subjects were patients for which clinical examination indicated need for further treatment. The hospital council approved data collection and all participants gave informed consent for the scientific use of their data. A total of 10 specimens were excised from an equal number of patients after clinical examination. A trained pathologist examined each excised specimen, identified on average 5 points of interest (POI) on every specimen and labelled each POI as malignant or not. The same pathologist later performed formalin fixing and cross sectioned the specimens. The effective dataset totalled 115 POIs, of which 35 were captured before fixing (21 malignant, 14 benign), 41 after fixing (23 malignant, 18 benign) and 39 after cross sectioning (24 malignant, 16 benign).

The core element of the capturing system was the open-platform camera OLYMPUS AIR A01 (4/3" Live MOS sensor, high resolution, 1736x2320 pixels, lens M.ZUIKO DIGITAL ED 30mm F3.5 Macro) located at 20cm above the capture stage and contained in a dark box. The spectral sensitivities of the RGB camera sensor and the spectral illumination characteristics are shown in **Fig. 1**. A timer controlled sequential capture under 7 narrow-band LED illuminations in the visible spectrum, resulting in 7 raw RGB images. Based on the illumination and sensitivity overlap at each spectral frequency, either the R, G or B channel of the raw image was selected to serve as a channel subimage, resulting in a 9-channel macroMSI as shown in **Fig. 2**. For modeling and evaluation purposes, we also captured a conventional RGB image and the reflectance spectrum of each POI with a spectrometer (TOPCON SR-3AR, small area measurements). All the above steps were repeated 3 times for each specimen: immediately after excision (unfixed case), after formalin fixing (fixed case) and sectioning (cross sectioned case).

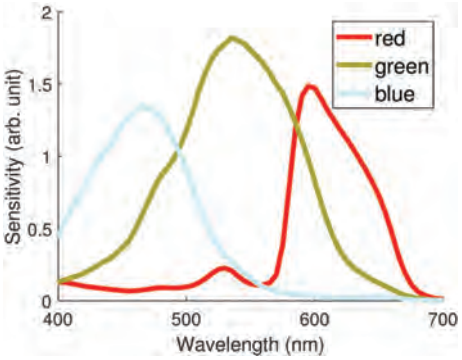


Fig. 1 Spectral characteristics (in arbitrary units) of the camera (a) sensitivity and (b) luminous intensity

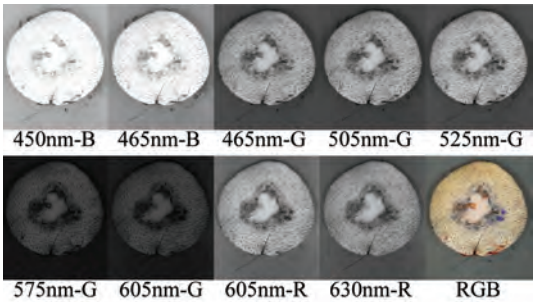


Fig. 2 Example of the 9 channels of the macroMSI and its respective conventional RGB image

2.2 Suggested analysis framework

In order to visualise malignancy probability from macroMSI, we propose a unified work flow as described in **Fig. 3**. After obtaining the 9-channel macroMSI, we apply white balance correction, where scene illumination is estimated with the Gray World assumption on a calibration image of a Macbeth color chart. We identify the label POIs and segment a respective region of interest (ROI). Afterwards, from each ROI we reconstruct its reflectance spectrum and LBP-based texture. The feature vector is produced by concatenating the two components and served as classification input. For the sake of selecting the optimal classifier, various classifier configurations are trained and validated using 80% and 10% of the dataset as the respective train and validation sets. Validation is achieved by 5-fold Stratified Cross Validation (CV). At each CV iteration, features of the train set are used for training dimension reduction, scaling and training the classifier on the reduced transformation of the train features. Subsequently, performance is validated by classifying the test set features, after transforming them using previously trained scaling and dimension reduction. Through validation stage we retain only trained classifiers with cross-validated Area Under the ROC curve(AUC)[17] larger than 0.75. Optimal classifiers are trained again on previously used 90% of the dataset and tested on the remaining 10%. During testing, for each POI we obtain a binary class label as malignant/benign together with a classification prediction score, whose range and value differ depending on the classifier. We normalize prediction scores on a range from 0 to 100, in order to express the percentage of malignancy probability. Higher normalized prediction score as belonging to the malignant class was interpreted as higher malignancy probability. Therefore, we can visualize the classification results on an sRGB image obtained from the macroMSI, as a mapping of predicted class together with malignancy percentage using an intensity color map. The trained optimal classifier can then be incorporated in diagnosis. The proposed processing steps are described in detail in the following sections.

2.3 ROI segmentation

Each single pixel POI identified by the pathologist, needs to be expanded to a wider ROI with same intensity characteristics. This step reduces the influence of noisy pixels in the analysis and enables malignancy visualisation. We achieve ROI segmentation through region growing, with the pathologist-selected pixel as the seed. Starting from the seed, the ROI grows by including neighbouring pixels with intensity value within 8% difference from the current ROI pixels. Region growing is applied on every channel of the macroMSI in order to produce binary masks for the ROI. The final mask for the ROI consists of pixels enclosed in at least 6 of the channel binary masks. In order to speed up feature extraction without loss of generality, we use a ROI where we limit region growing at a 30 pixel radius from the seed. Additionally, we use the entire grown ROI without radius limit during the visualization stage.

2.4 Reflectance reconstruction

Reconstruction of the reflectance spectrum $\hat{r}$ from MSI pixel intensity values g is possible, if the camera spectral sensitivity and illumination are known. We can regard it as a linear inverse problem solved by Wiener estimation. In a matrix form in the discrete space, the generalized solution for estimated reflectance becomes:

$$\hat{r} = MH^T(HM^{-1}H^T + K_n)^{-1}g \tag{1}$$

with the help of a smoothing matrix M, camera system matrix H and noise matrix K_n. We use a smooth-

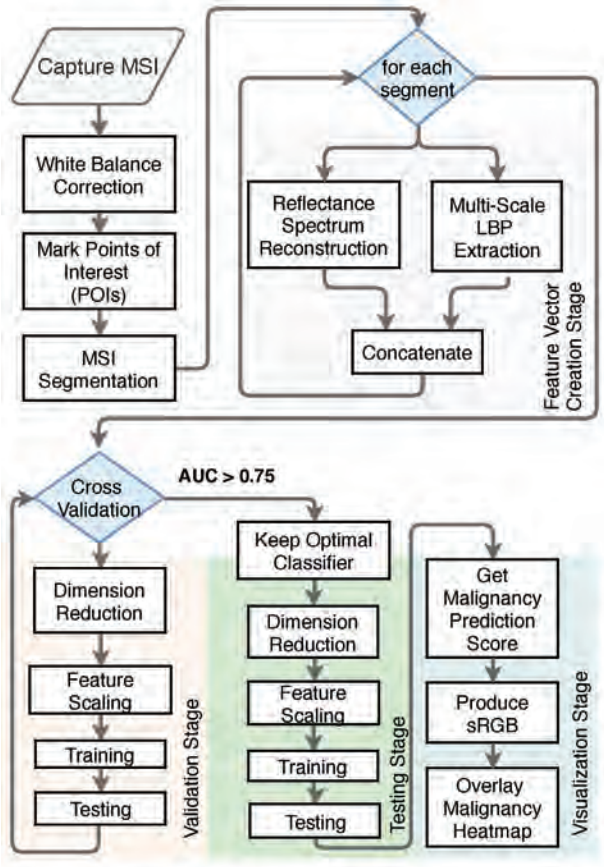


Fig. 3 Flowchart of the proposed analysis framework for macroMSI

ing matrix modelled by the autocorrelation of the measured reflectance spectra from skin specimens grouped by anatomical location. The noise matrix is modelled with diagonal covariance σ^2 in the order of 0.1, different for every channel. In order to further improve reconstruction, we considered an additional spatial noise model[13], which denoises pixels via Bayesian Interference from the statistical information of a pixel's neighbourhood. Reconstructed reflectance was computed as the average of reconstructed spectra from every pixel inside the ROI. The results of reconstruction were evaluated by average Normalized Root Mean Square Error (NRMSE). Additionally, reconstructed reflectances of the whole macroMSI were used to generate a standard RGB (sRGB) image for visualisation.

2.5 Local binary patterns

Local Binary Patterns (LBP) are an efficient and widely used texture desciptor. Conventional LBP functions as a spectral histogram of intensity differences in a neighbourhood of P pixels across 2^P bins, computed in scale R from a central pixel. Due to the frequent appearance of uniform patterns[14], a uniform LBP descriptor can decrease feature size without substantial performance loss. In our case pattern rotation is also irrelevant, as homogeneous skin tissue surrounds lesions, allowing the use of Rotation Invariant Uniform LBP (RIU-LBP) variation. However, application of LBP on macroMSI shows intricacies because texture does not progress across the spectral dimension, thus we investigate three variations of 3D LBP. For the first two variations, RIU-LBP histogram is calculated independently for each channel. Afterwards, all band results are concatenated (CatLBP, N histograms of $(P+2)$ bins) or summed across bins (SumLBP, $(P+2)$ bins). The third variation is implemented as the Multispectral Multiscale LBP (MMLBP)[15], which expands RIU-LBP by adding a multispectral component computed across P_λ adjoining spectral channels ($(P+2)2^{P_\lambda}$ bins). We calculated the various LBP descriptors over 8 spatial neighbours, 2 spectral neighbours, at scales 1 and 2, and concatenated the scale features.

2.6 Dimension reduction

An overtly large feature vector not only increases required time for training, but also causes computational instability and overshadows important components, along with introducing problems deriving from the curse of dimensionality. In our case, inherently, both neighbouring spectral frequencies and neighbouring spatial brightnesses display correlation, which can be resolved by dimension reduction. For this purpose we applied Principal Component Analysis (PCA) and Independent Component Analysis (ICA). PCA projects data on an axis that maximizes variance, whereas ICA identifies orthogonal independent components of a non-gaussian linear mixture model for the data[16]. In order to preserve the physical meaning of reflectance and textural features, we applied dimension reduction independently on each of them. Furthermore, in order to avoid regressing reduced reflectance components back to the channel number, we train dimension reduction on the respective measured spectra. The number of components kept after reduction was tuned among values $[10, 20, 30]$.

2.7 Classification

Three traditional classifiers were compared for the task of malignancy classification; Support Vector Machine (SVM), Random Forest (RF) and k-Nearest Neighbour (KNN). SVM is a supervised machine learning algorithm that tries to project data onto a hyperplane which maximizes classification margin and has been used extensively for medical classification. A sigmoid kernel accomplished projection and parameter C was tuned in range [0, 1.5]. On the other hand, RF is a collection of decision trees searching for the best feature among a subset of features. Entropy was used as a decision criterion for 50 components. KNN is a simple classifier, which implements non-parametric, lazy classification using a similarity distance metric in the feature space. We tuned neighbour number among values {1, 3, 5} with correlation as distance metric. AUC and overall accuracy of malignancy predictions compared to ground truth labels of the pathologist are used to select optimal classifiers. By design, SVM and KNN provide values of the discrimination function in order to compute the necessary thresholds for class separation during AUC calculation. On the other hand, RF provides prediction probability values which can substitute discrimination function values in AUC calculation. The feature set was chosen among combinations of reflectance spectrum (Spect) plus LBP variations. During classification, as fixed dataset we considered images of both fixed and sectioned tissue, because sectioning is performed after formalin fixing.

3. Results

3.1 Performance of feature extraction

The feature vector contained two types of features: re-

flectance features and texture features. Employing Wiener estimation, we reconstructed the reflectance spectrum of two individual sources: a conventional RGB image and an MSI captured from the same camera system. **Table 1** shows the average NRMSE of reconstruction for different tissue datasets. MSI-based reflectance reconstruction outperformed RGB-based reconstruction, achieving an average NRMSE of 0.0543 in the complete dataset, almost half the error of the RGB-based case. Observing the reconstructed spectra, it is evident that although reconstruction fails to reproduce non-smooth parts of the reflectance curve, the relative height of the curve is reconstructed accurately, as exemplified in **Fig. 4**. Light-green solid line denotes the target measured reflectance, while x markers indicate center wavelengths of the MSI camera channels. Dashed blue and solid magenta lines denote an example of MSI-based and RGB-based reconstructed reflectance, respectively. Regarding image sets for the estimation, MSI-based reconstruction from imaging of unfixed tissue with NRMSE 0.0405 was more successful compared to both fixed and cross sectioned tissue. The situation was reversed for RBG-based reconstruction, with the fixed dataset being superior to unfixed. Although cross-section ROIs are important for understanding the interior structure of the tissue, the small size of sectioned pieces which contained thin, noisy ROIs at the edge of the specimen, hindered reconstruction.

We described texture in each ROI of the specimen using variations of the LBP operator. For the MSI case, three variations of LBP were extracted; CatLBP, SumLBP and MMLBP. For the RGB case, conventional LBP was applied on the grayscale-transformed image. Although, the larger number of channels of the MSI resulted in larger feature vectors for MSI-based texture, it also allowed for greater detail in texture description. For example, even though the LBP histograms for RGB-LBP and SumLBP had the same number of bins, different values were assigned to each bin, due to the small variance of LBP values at each image pixel obtained from the former. The remaining two descriptors, namely CatLBP and MMLBP, can describe texture in greater detail due to the larger amount of bins assigned to different spectral channels (CatLBP) or different relations among adjoining channels (MMLBP). The effect of applying LBP operator in the spectral dimension is more evident in the classification task and will be discussed in section 3.3.

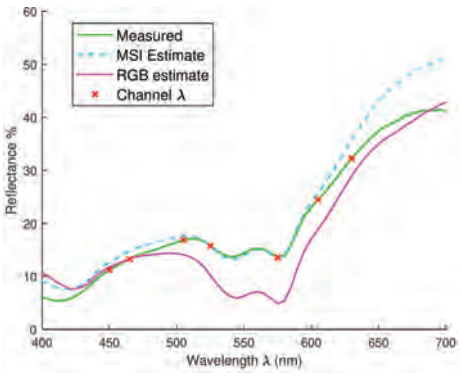


Fig. 4 Comparative example of reconstructed reflectances

Table 1 Average NRMSE of reflectance reconstruction from RGB and MSI image source

Source	Fixed	Unfixed	Cross Sectioned	Overall
RGB	0.0955	0.1043	0.1257	0.1085
MSI	0.0643	0.0405	0.0559	0.0543

3.2 Dimension reduction performance

After concatenating reflectance and LBP values, the feature vector ended up having up to 351 values (in the case of CatLBP). Dimension reduction proved especially useful on texture features. The optimal selection was PCA on reflectance and ICA on texture. PCA had larger effect on AUC value when applied on the reflectance spectrum, whereas ICA improved AUC when applied on all 3 texture descriptors. Judging by the definitions of the two dimension reduction methods, it is evident that the difference in performance can be attributed to the different underlying models for reflectance and texture features. Indeed, the reflectance spectrum is strongly correlated in neighbouring wavelengths, while texture is a mixture of materials and surfaces. Moreover, PCA improved AUC for RF classifiers, but didn't have any noticeable effect on KNN or SVM. The AUC results for the case of RF classifier are shown in **Table 2**. While the first 3 components of PCA in most cases could explain above 96% of the total variance, use of only 3 components did not achieve sufficient performance.

3.3 Validation performance

During validation, the classification task was investigated using a variety of configurations based on SVM, RF or KNN classifiers. Performance comparison of tuned

Table 2 Cross-validated AUC per dimension reduction scheme for RF classifier and fixed tissue input

DimRed	Spect	+CatLBP	+MMLBP	+SumLBP
None	0.708	0.749	0.749	0.749
PCA	0.774	0.728	0.745	0.728
ICA	0.733	0.794	0.794	0.794

classifiers during validation showed that cross-validated AUC of all classifiers improved with the inclusion of texture descriptors in the feature vector, rather than using only spectral reflectance, as shown in **Fig. 5**. RF (50 components, entropy criterion) achieved highest AUC of 0.896 using an input set comprising of reflectance spectrum and CatLBP texture features with PCA and ICA dimension reduction respectively. SVM had the second best AUC overall again with CatLBP used as texture descriptor. KNN generally had the worst performance among the three classifiers. The choice of a feature set using Spect+MMLBP with PCA and ICA dimension reduction respectively was the most effective choice in terms of AUC. However, comparison with cross-validated accuracy of label prediction against ground truth labels of the pathologist during validation phase in **Table 3** shows that the remaining LBP descriptors have better malignancy prediction properties. In this regard the choice of Spect+MMLBP or SumLBP. The rationale behind this compromise is that, while AUC expresses separability of the binary malignancy class, prediction score actually expresses binary malignancy probability, thus feature vector choice should account for prediction accuracy. Additionally, images of unfixed tissue are proven more successful in terms of AUC as input of the classifier compared to the fixed tissue (Fig. 5). The mixed dataset had an intermediate performance. This result corroborates the conclusion of reflectance reconstruction, which indicates unfixed tissue as a better source for reconstruction. When the feature set was calculated from an MSI rather than an RGB image source, classification improved. As shown in Table 3, classification from MSI using reflectance spectrum and LBP achieved 80.33% accuracy and 0.82 AUC during validation, whereas the best RGB-based classification scheme achieved 69.42% with 0.50 AUC. The latter result is unacceptable under the AUC criterion that we set to select optimal classifiers through validation.

3.4 Testing performance

After selecting a handful of optimal classifiers through validation stage, we tested them on the remaining testing dataset using mixed tissue. Ground truth of the test specimen together with the POI locations that were pinpointed by the pathologist are presented in **Fig. 6**. ROI labels are marked with a triangle for malignancy and a circle for benignity. The combination of RF with either Spect+CatLBP or Spect+MMLBP features was the best classifier after testing, with achieved malignancy prediction accuracy at 83.33% on the test specimen as shown in **Table 4**. On the other hand, SVM achieved less than 70% prediction accuracy. Contrary to expectations raised from the performance of SVM during validation, SVM had low performance during testing, which may indicate over-fitting to the training data.

For classification with RF, the mispredicted cases referred to a false positive at one of the two POIs on tissue covered with black hair, as indicated in Fig. 6. Although RF with Spect+SumLBP achieved perfect prediction on the test set, the prediction scores were all close to 50%, thus it was discarded. The SVM classifier predicted both false negatives and false positives. Visualization of class prediction scores during testing as malignancy probabilities of the test ROIs resulted in the heat values shown in Fig. 6 for unfixed tissue. We observe that the segmented regions are wide and generally coincide with visually identifiable regions. Comparing the malignancy probabilities produced from the two RF classification schemes, the case of Spect+MMLBP shows more distributed malignancy probability. Such probability distribution is more accurate and representative of the pathologist ground truth, considering that the non-malignant area should have a much lower malignancy probability than the ma-

Table 3 CV malignancy prediction accuracy (%) per classification scheme during Validation Stage (top to bottom: SVM, KNN, RF)

RGB		MSI			
Spect	+LBP	Spect	+CatLBP	+MMLBP	+SumLBP
62.36	63.99	60.83	63.33	65.14	68.29
69.42	61.48	65.14	63.33	65.14	68.29
52.50	54.16	70.75	73.33	74.97	80.33

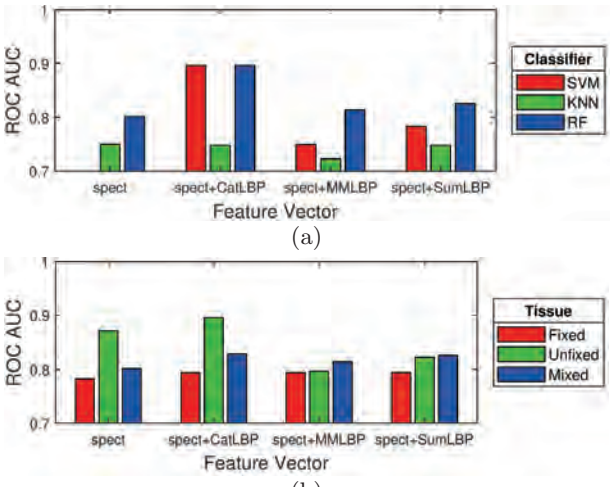
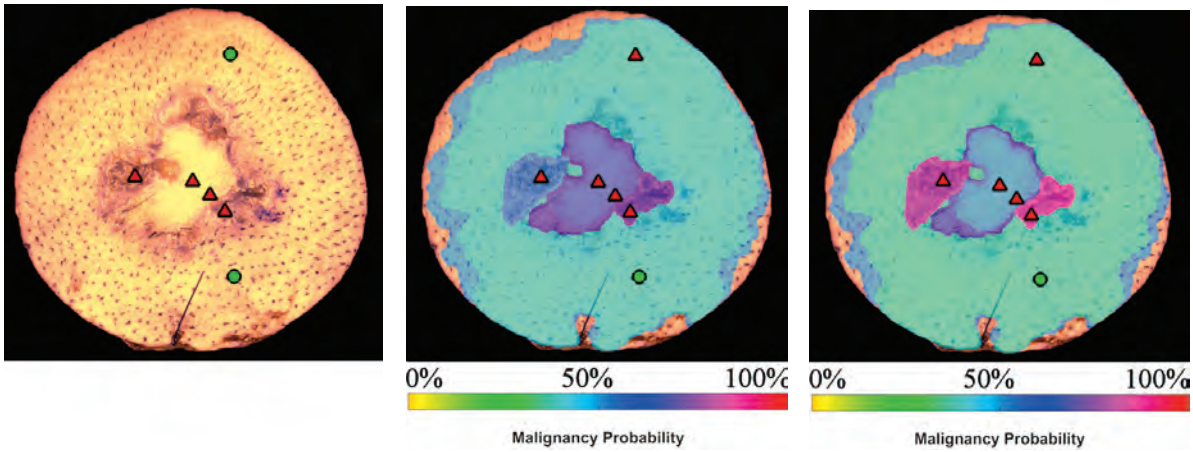


Fig. 5 AUC comparison of optimally tuned classifiers after cross validation for all combinations of classifier by feature set (a) and fixing condition of the input data (b)



(a) Ground truth (b) RF with Spect+CatLBP (c) RF with Spect+MMLBP

Fig. 6 Malignancy classification results visualised on an sRGB image of an unfixed test specimen

Table 4 Malignancy prediction accuracy (%) per optimal classifier during Testing Stage

Classifier	Features	Parameters	Accuracy
RF	+CatLBP	50comp, entropy	83.33
RF	+MMLBP	50comp, entropy	83.33
SVM	+CatLBP	sigmoid, C=1	66.67

lignant area at the center of the specimen. While the entire core of the test specimen contained malignancy areas, the dark coloured side ROIs are expected to have higher malignancy probability, compared to the white center, which contains mostly atypical cells, as achieved by the Spect+MMLBP feature vector. In the case of Spect+CatLBP, malignancy probability for both malignant and non-malignant areas is almost the same around 50% regardless of the correct binary prediction. This fact highlights the importance of taking into account not only accuracy, but also validity of the malignancy probability in the context of macropathology. Even though MMLBP features during validation were overshadowed by CatLBP, final testing proved MMLBP as a more robust feature vector.

4. Discussion

Multispectral imaging of unfixed skin tissue proved to be an effective input set for the binary classification task for malignancy detection, with over 80% accuracy during testing. Use of MSIs outperformed the use of conventional RGB images at every stage of the analysis. Furthermore, the inclusion of both reflectance and texture features, which mimic the pathologist's approach during diagnosis, improved classification accuracy as well as AUC value. Unfixed and fixed tissue samples were evaluated, revealing the superiority of inclusion of unfixed tissue in the analysis. The logic behind this fact may lie on the improved reflectance reconstruction accuracy that the unfixed dataset provides. Additionally, considering that formalin fixing is a chemical process, which affects the condition of the tissue, it modifies color and reflectance of the unprocessed specimen. Moreover, the unfixed dataset contained skin samples from the interior of the specimen, which may have different constitution compared to exterior tissue and in turn affected classification performance. Nonetheless, cross sectioned tissue should be analysed, due the deep tissue information it provides, therefore inclusion of unfixed tissue in the analysis is recommended.

Regarding potent texture descriptors, MMLBP was revealed to be a compact and practical choice for the analysis of skin macropathology images. Validation results proved RF classifier as a suitable classification model, which is consistent with its value as a relatively unbiased predicting model. However, it should be emphasized that obtaining a large and complete dataset for training is necessary in order to avoid over-fitting. Dimension reduction seemed more appropriate for textural features rather than for reflectance features, due to the fact that reflectance was already reconstructed from a limited amount of values equal to the number of multispectral channels. The choice of PCA for reconstructed reflectance spectrum and ICA for texture features is recommended. The justification of the latter may lie in the mathematical definition of LBP as a descriptor of frequency of brightness difference. Although the size of the feature vector could be reduced from a few hundreds to just a few tens of values, training dimension reduction on a larger dataset can further limit the important components in the feature vector and accelerate the classification output.

Although the dataset of the present study was enough

for investigative purposes, extensive experimentation on a wider dataset of macroMSIs is required. Analysis of macropathology images poses difficulties due to the high variance of the human skin (skin complex, anatomical location, age, etc.). Moreover, skin cancer related diseases can have more stages than the dipole malignancy versus benignity. Such difficulties were evident during our analysis and our limited dataset inhibited classification performance. Moreover, captured macroMSIs displayed patches with high saturation and bright reflections. Therefore, image quality should be increased by thoroughly wiping liquids from the skin tissue and by incorporating a polarizing filter in the camera system. A notable limitation of the proposed framework is the need for the pathologist to pinpoint areas respective to the different tissue areas on the specimen. In order for the procedure to be fully automated it is imperative that the centers of different tissue ROIs are automatically identified with adaptive region segmentation on the MSI, so that the rest of the steps can be applied. In order to achieve more accurate visual results, malignancy scores could be adjusted together with clinical information such as sex, age or medical history, information that is taken into account during macropathology. Due to the nature of skin lesions, that exist on the surface of the human body, adaptations of the proposed multispectral analysis framework can be employed during skin cancer screening and dermoscopy for early detection purposes.

5. Conclusion

We proposed a framework using spectral reflectance and texture features for visualising cancer malignancy probability on macroMSI of skin specimens to assist margin identification during macropathology. Final visual results of the proposed procedure depicted fairly accurately both regions with similar tissue characteristics as well as their malignancy probabilities. Experiment results suggest RF classifier with a feature vector consisting of reconstructed spectral reflectance and MMLBP texture features for the proposed framework. We highlighted the importance of including unfixed tissue during analysis and training, as well as the superiority of MSI images to conventional RGB images. Although further investigation on a larger dataset is necessary, the suggested analysis framework of macroMSI was proven to be suitable for enhancing the present pathology practice.

References

1) Cancer Information Service, National Cancer Center, Japan: "Cancer Incidence and Incidence Rates in Japan in 2009: A Study of 32 Population-Based Cancer Registries for the Monitoring of Cancer Incidence in Japan (MCIJ) Project", Japanese Journal of Clinical Oncology, Vol. 45, pp. 884–891 (2015).
2) M. Nishi: "Epidemiology of Skin Cancer in Japan", Journal of Tumor, Vol. 4, No. 2, pp. 369–373 (2016).
3) R. Romaguear, M. Nassiri, A. R. Morales: "Tools to Facilitate and Standardize Grossing", Histologic, No. 1, pp. 17–21 (2003).
4) Y. Yagi: "Color Standardization and Optimization in Whole Slide Imaging", Diagnostic Pathology, BioMed Central, Vol. 6, No. 1, p. S15 (2011).
5) A. F. Jerant, J. T. Johnson, C. D. Sheridan, T. J. Caffrey: "Early Detection and Treatment of Skin Cancer", American Family Physician, Vol. 62, pp. 357–368 (2000).
6) R. L. Siegel, K. D. Miller, A. Jemal:: "Cancer Statistics, 2019", CA: A Cancer Journal for Clinicians, Vol. 69, No. 1, pp. 7–34 (2019).
7) N. R. Abbasi, H. M. Shaw, D. S. Rigel, R. J. Friedman, W. H. McCarthy, I. Osman, A. W. Kopf, D. Polsky: "Early Diagnosis of Cutaneous Melanoma: Revisiting the ABCD Criteria", Jama, Vol. 292, No. 22, pp. 2771–2776 (2004).
8) J. J. Squiers, W. Li, D. R. King, W. Mo, X. Zhang, Y. Lu, E. W. Sellke, W. Fan, J. M. Dimaio, J. E. Thatcher: "Multispectral Imaging Burn Wound Tissue Classification System: a Comparison of Test Accuracies Between Several Common Machine Learning Algorithms", Proc. of Medical Imaging 2016: Computer-Aided Diagnosis, Vol. 9785, p.97853L (2016).
9) L. Rey-Barroso, F. Burgos-Fernandez, X. Delpueyo, M. Ares, S. Royo, J. Malvehy, S. Puig, M. Vilaseca: "Visible and Extended Near-Infrared Multispectral Imaging for Skin Cancer Diagnosis", Sensors, Vol. 18, No. 5, pp. 1441 (2018).
10) A. Vasaturo, S. Diblasio, D. Verweij, W. Blokx, A. M. Willeke, J. H. Vankrieken, I. J. Devries, C. G. Fidgor: "Multispectral Imaging for Highly Accurate Analysis of Tumour-Infiltrating Lymphocytes in Primary Melanoma", Histopathology, Vol. 70, No. 4, pp. 643–649 (2017).
11) F. Saleheen, A. Badano, W.-C. Cheng: "Evaluating Color Performance of Whole-Slide Imaging Devices by Multispectral-Imaging of Biological Tissues", Proc. of Medical Imaging 2017: Digital Pathology, Vol. 10140, p.101400R (2017).
12) H. Xu, C. Lu, R. Berendt, N. Jha, M. Mandal: "Automated Analysis and Classification of Melanocytic Tumor on Skin Whole Slide Images", Computerized Medical Imaging and Graphics, Vol. 66, pp. 124–134 (2018).
13) P. Urban, M. R. Rosen, R. S. Berns: "A Spatially Adaptive Wiener Filter for Reflectance Estimation", Color and Imaging Conference, Society for Imaging Science and Technology, Vol. 2008, No. 1, pp. 279–284 (2008).
14) T. Ojala, M. Pietikainen, T. Maenpaa: "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, pp. 971–987 (2002).
15) R. Peyret, A. Bouridane, F. Khelifi, T. Muhammad, S. Al-Maadeed: "Automatic Classification of Colorectal and Prostatic Histologic Tumor Images Using Multiscale Multispectral Local Binary Pattern Texture Features and Stacked Generalization", Neurocomputing, Vol. 275, pp. 83–93 (2018).
16) A. Hyvarinen, E. Oja: "Independent Component Analysis: Algorithms and Applications.",Neural networks: the official journal of the International Neural Network Society, Vol. 13, pp.

411–430 (2000).
17) M. S. Pepe, T. Cai, G. Longton:"Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve", Biometrics, Vol. 62, No. 1, pp. 221–229 (2005).

(Received September 12, 2019)
(Revised November 10, 2019)

Eleni ALOUPOGIANNI
She received a B.Sc&M.Sc. diploma from National Technical University of Athens, Greece, in 2016. She earned a M.Eng. degree and is currently enrolled in a Ph.D. course on Information and Communications Engineering in Tokyo Institute of Technology, Japan. She is currently working on medical image processing and multi-spectral images.

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 Frontier Collaborative Research Center, Tokyo Institute of Technology. From 2004 to 2016, he was a Researcher with Frontier Collaborative Research Center, Tokyo Institute of Technology. Since 2016, he has been an Assistant Professor with Institute of Innovative Research, Tokyo Institute of Technology. His research interests include optical information security, holograms, biometric authentication, and medical and healthcare information systems.

Takaya ICHIMURA
He earned M.D., Ph.D. degrees and is an assistant professor of the Saitama Medical University. He received the Ph.D. degree from Kumamoto University, Kumamoto, Japan, in 2005. His current research interests include nuclear atypia and the molecular nature of chromatin.

Atsushi SASAKI
He received M.D. and Ph.D. degrees from Gunma University School of Medicine in 1980 and 1984, respectively. He received the Neuropathology Best Paper Award in 2001, and the Brain Tumor Pathology Best Paper Award in 2002. He is currently a professor at the Department of Pathology, Saitama Medical University. His research interests include brain tumor pathology and microglia. Dr. Sasaki is a member of the International Society of Neuropathology (member of the council) and the American Association of Neuropathologists (active member).

Hiroto YANAGISAWA
He graduated from the Saitama Medical University, Japan. He is currently a medical staff at the Department of Dermatology, Saitama Medical University. His clinical and research interests include the areas of skin cancer and surgery. He is a member of Japanese Dermatological Association.

Tetsuya TSUCHIDA
He received M.D. and Ph.D. degrees from the University of Tokyo, Japan in 1978 and 1986, respectively. He is currently a professor at the Department of Dermatology, Saitama Medical University. His clinical and research interests include the areas of collagen disease, skin cancer and dermoscopy. He is a member of Japanese Dermatological Association.

Masahiro ISHIKAWA (*Member*)
He received a Ph.D degree from Niigata University, Niigata, Japan, in 2006. He is currently an Assistant Professor at the Saitama Medical University. His current research interests include image processing 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 2008. 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 IEVC2019* --

Robust, Efficient and Deterministic Planes Detection in Unorganized Point Clouds Based on Sliding Voxels

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

† Shinshu University, †† National Defense Academy of Japan

<**Summary**> Planes detection in unorganized point clouds is a fundamental and essential task in 3D computer vision. It is a prerequisite in a wide variety of tasks such as object recognition, registration, and reconstruction. Conventional plane detection methods are remarkably slow because they require the computation of point-wise normal vectors and are non-deterministic due to their dependency on random sampling. Therefore, we propose a drastically more efficient and deterministic approach based on sliding-voxels. A sliding voxel is an overlapping grid structure in which we analyze the planarity of the points distributions to extract hypothetical planes efficiently. Each possible plane is validated globally by weighing and comparing its co-planarity with other sliding voxels' planes. Experimental results with simulated and realistic point clouds confirmed that the proposed method is several times faster, more accurate, and more robust to noise than conventional methods.

Keywords: planes detection, point clouds, pca, ransac, hough transform

1. Introduction

Point clouds are sets of points in a 3D space that represent the surface of objects and scenes. There are many point cloud sources such as LiDAR, structured light sensors, stereo vision cameras, and so on. However, high variations of sensors noise patterns[1),2)] are challenging problems when dealing with point cloud data.

In intelligent robotics, a core task is to recognize environment patterns, especially those from human crafted objects and scenes which are describable by a set of planar structures. Therefore, the detection of planar surfaces such as obstacles, floor, stairs, and tables is crucial to several applications; i.e., virtual keypoints detection[3)], object detection[4)], reconstruction[5)], localization and mapping[6)–9)], roofs detection[10)] and augmented reality[11)]. Besides, planes detection is involved in essential robotics tasks such as placing objects onto planar surfaces[12)] or climbing stairs[13)].

Conventional planes detection algorithms[3),14)] depend on inefficient point-wise normal vectors computations. When using 2.5D data (organized point clouds or depth images), there exist fast algorithms to compute them [15),16)] at the cost of losing accuracy. In addition, their efficient and robust computation in unorganized point clouds is still an open topic.

Therefore, in this work we contribute with a drastically faster approach: a sliding voxel–based algorithm. It does not need pre-computed normal vectors; instead, it directly works with the points distribution of overlapping 3D voxels to acquire surface information. We analyze the scattering of each sliding voxel to locate co-planar regions and detect hypothetical planes. Then, planes are extracted from the validation of hypothetical planes against an enhanced subset of the point cloud.

Since ground truth planes data can be obtained from 3D models, we provided experiments with realistic point cloud simulations generated from these models as well as ground truth data. To verify the performance of the proposed method, we measured its efficiency and error against the most popular algorithms: two RANSAC variations[3),14)] and the Randomized Hough Transform[17)] for planes detection.

The structure of this paper is as follows. In section 2, we briefly make definitions before explaining how the conventional methods work and a critical analysis of these. Section 3 describes in detail the proposed method, while section 4 defines the datasets, their acquisition, as well as the results of quantitative and qualitative experiments. Lastly, in section 5 we summarize the results and provide an insight about future works.

2. Planes Detection in Unorganized Point Clouds and Related Works

2.1 Planes detection definition

Let $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3 : 1 \leq i \leq n_p\}$ where n_p is the number of points in a point cloud. These points are unorganized, hence, algorithms commonly build structures from them, namely octrees. They are sampled from the surface of objects, and may not be complete due to occlusion and the range of sensors. Furthermore, sampling density, noise patterns, and artifacts can vary in large amounts even in adjacent regions, due to how the points are sampled by a variety of sensing techniques.

Now, let $\mathbf{P}^{det} = \{\mathbf{h}_i \in \mathbb{R}^4 : 1 \leq i \leq n_h \text{ and } h_i = \{n_i^x, n_i^y, n_i^z, d_i\}\}$ where n_h is the number of planes detected from a point cloud. The first 3 components correspond to the ith plane normal vector $\hat{\mathbf{n}}_i$, and the last d_i component to the plane normal distance to the origin. Then, we define a plane detector as a function that inputs a point cloud $\mathcal{P}$ and outputs a set of planes coefficients $\mathbf{P}^{det}$.

Figure 1 shows a diagram of a generic planes detector. It inputs a set of 3D points $\mathcal{P}$ with noise and sensor artifacts and outputs a set of detected planes coefficients. $\mathcal{P}$ may consist of several planar and curved surfaces, and its density is not uniform.

A planes detector should consider the presence of the aforementioned nuisances. Therefore, it selects planar points within a distance threshold ϵ, and the planes of $\mathbf{P}^{det}$ are estimated from these using PCA or least squares.

If the ground truth planes of a point cloud $\mathcal{P}$ is $\mathbf{P}^{gt}$, ideally, the output of a planes detector should satisfy $\mathbf{P}^{det} = \mathbf{P}^{gt}$. Because $\mathbf{P}^{gt}$ is unknown by the detector, it has to figure out how to generate hypothetical planes. In the remaining parts of this section, we briefly explain the conventional methods algorithms.

2.2 RANSAC

RANSAC[18] is a simple but robust model fitting algorithm, in computer vision is widely used for feature matching but also can be used for lines and planes fitting.

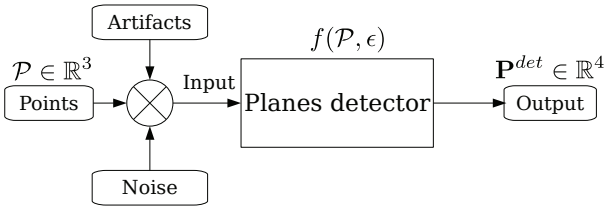


Fig. 1 Diagram of a generic plane detector

It is an iterative approach consisting mainly in hypothesis and verification steps. A subset of n random samples are drawn from the point cloud ($|\mathcal{P}| > n$) and a hypothetical model $\mathcal{M}$ is generated. For each $\mathcal{M}$, an error metric between the points and $\mathcal{M}$ is computed. The model is verified by counting the number of inliers τ within an error threshold ϵ. If τ does not reach a threshold τ', the process is repeated until k iterations. Once a proper $\mathcal{M}$ is found, then its coefficients should be refined by other methods such as least mean squares.

k is calculated from the probability of drawing good samples from the data, i.e., samples that generate accurate models. Let w be the probability of drawing an inlier from the data, then $b = w^n$ will be the probability of finding a good set of samples from $\mathcal{P}$, then to ensure a probability z of RANSAC to withdraw a valid model it defines k as

$$k = \frac{\log(1-z)}{\log(1-b)}. \tag{1}$$

For instance, suppose we have a noiseless point cloud with 3 planes. If we want RANSAC to find one plane model with high probability, let's say $z = 0.99$, $w = 1/3$ and $n = 3$, then we have to iterate up to 122 times per plane. However, if we have noise and other non-planar points (outliers) then we have to modify w and set a lower value, then the maximum iterations per plane increase drastically. When we test for $w \in \{1/4, 1/5, 1/6\}$ the maximum iterations per-plane increases from 122 to 292, 573, and 992 respectively.

This method is non–deterministic and it has two main disadvantages when applying it to point cloud data. First, to find a model we have to assume that the inliers/outliers ratio w is relatively low, since even in noiseless point clouds it should approximate the ratio between the number of points of the smallest plane over the whole point cloud, thus increasing the maximum iterations per plane. Second, since point clouds are multi-model, we have to execute RANSAC several times. Clearly, this adds finishing conditions difficult to configure.

Furthermore, from a geometric perspective, RANSAC is only testing planes models against point-model distance inliers, i.e., it is not aware of points curvatures, hence generating spurious models across noisy regions in the case of realistic point clouds. There has been application-driven RANSAC improvements; therefore, in this work, we evaluate two geometry-aware RANSAC variations: Coarse-

to-Fine RANSAC[3] and Efficient RANSAC[14].

2.3 Coarse-to-Fine RANSAC

An iterative Coarse-to-Fine RANSAC, hereinafter referred to as CFRANSAC, was developed to be used for virtual keypoints detection[3]. This method employs RANSAC iteratively to detect each model in a coarse-to-fine approach, with Euclidean clustering spatially separating plane inliers in segments before refining its coefficients.

Geometry-aware RANSAC methods like this filter distance inliers by using point normals: a local feature of point clouds that is not sensed and has to be estimated. Apart from the distance threshold ϵ, it introduces an angular threshold θ between the hypothetical plane normal vector and each point normal vector.

First the coarse step detects surfaces which are roughly planar with big threshold values. However, when these surfaces are neighbors of similar planar surfaces RANSAC tends to report inaccurate results. Therefore, the coarse inliers are refined by using RANSAC again but with stricter thresholds.

Since nearly parallel planes far from each other can be erroneously detected as one, it can produce an error when refining the model coefficients. Thus, planes inliers are spatially clustered, where each cluster is treated as a different model if they meet a specified points number threshold τ'. Lastly, plane candidates are refitted against their inliers in a least squares sense.

2.4 Efficient RANSAC

Efficient RANSAC[14] (EFRANSAC) executes iteratively RANSAC on disjoint random subsets under the assumption that valid planes models will be detected in most of the subsets. Therefore, it only accepts models that are prominent in the number of inliers, and were found in most of the disjoint subsets. Every accepted model remove its inliers from the point cloud and the process iterates again until the finishing criteria is met.

Unlike CFRANSAC, EFRANSAC filters out points with large deviations among their respective normal vectors during random sampling. It also filters points with normals deviating more than a defined angle from the plane normal vector. Moreover, it filters from the inliers only the largest connected component on the plane model by discretizing the inlier points translated to the plane coordinates. Finally, the candidate shapes are refitted using their inliers and removed from $\mathcal{P}$.

Even though its core functionality is faster than conventional RANSAC methods, it requires point-wise normals estimation, increasing its computational cost drastically.

2.5 Randomized Hough Transform

While the Standard Hough Transform performs numerous point-wise calculations, the Randomized Hough Transform[19], RHT, exploits random sampling to accelerate the voting process.

Similar to RANSAC, in each iteration, RHT selects a sample of 3 points within a given distance, then the spanning plane for those points is calculated using their cross product. Finally, the plane is voted once in an accumulator. Then, a plane model is detected by the algorithm when a cell in the accumulator reaches a given amount of votes.

The RHT is not aware of the points geometry, but of the parameter space during the voting process. Thus, it does not need normal vectors to compute planes coefficients with reasonable accuracy and speed. A drawback of this approach is that it is not aware of the points distributions of the detected plane models.

Even though, the ball accumulator solved some accumulator problems[17], it is non–trivial to adjust the discretization parameters, among others who affect directly the precision of the algorithm, including measurements to control the minimum and maximum distance between the random samples and a restriction on the smallest eigenvalue size of the spanned planes.

The latter restrictions are not possible only with the plane voting process. Therefore, the RHT needs to be aware of the inliers points distribution by recurring to a sorted clustering of the most prominent planes.

Since it is non deterministic, the finishing conditions have to be chosen carefully. The algorithm will stop when the remaining points go lower a given threshold, or if the algorithm fails to build a plane model in a given amount of iterations. If these conditions are not properly configured, the algorithm may not find all the most prominent planes, or it may take longer to stop. Additionally, the performance of the RHT decrease drastically when there is a large number of planes to be detected, or if the number of non-planar points is increased as shown in the hall and arena model experiments of Borrmann et. al[17].

3. Proposed Method

We focused on two main problems of conventional planes detection methods: speed and robustness. The major speed problem of conventional methods is that

they need to compute point-wise normals on overlapping neighborhoods[20].

Even though planar surfaces can be described locally, it is more precise to describe them globally. Nonetheless, algorithms that work on the whole set of points are very slow, e.g., the Standard Hough Transform.

Quantized noise from RGB-D and structured light sensors makes it even harder to describe locally planar surfaces, forcing algorithms to increase their search radius, threshold or voxel size; thus preventing to detect smaller objects.

Speed achieved by downsampling can also vanish small surfaces and extremely deviate their normal vectors, as the neighborhood has to be expanded way farther the downsampling region to compute them.

Therefore, we propose a method that efficiently detects planes via sliding voxels. Based on the Sliding Window in images, the proposed method uses a 3D Sliding Window implemented with octree voxels. It travels through occupied voxels of a point cloud and calculates geometric information about the points distribution using neighbor voxels.

Figure 2 outlines the proposed method algorithm. First, we build an octree with voxel size V_s and calculate tangent planes. This local plane fitting provides us with curvature information; therefore, we use sliding voxels to estimate the degree of coplanarity of each voxel.

Since coplanar voxels are more likely to be part of a prominent plane, we sort and mark them as hypothetical planes. Finally, planes are extracted from the validation of hypothetical planes against a geometrically enriched subset of the whole point cloud.

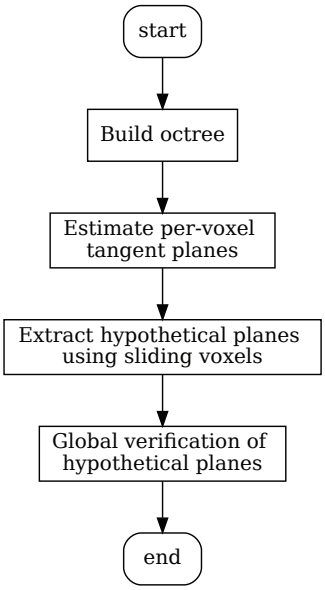


Fig. 2 Flowchart of the proposed method

3.1 Hypothetical planes extraction from coplanar voxels

For each voxel, its centroid $\mathbf{c}$, unit normal vector $\hat{\mathbf{n}}$ and a planarity value P are calculated using the eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3$ and their corresponding eigenvectors $\mathbf{v}_i$ of the covariance matrix, where:

$$\hat{\mathbf{n}} = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} \tag{2}$$

and

$$P = \frac{\lambda_1}{\lambda_2}. \tag{3}$$

After the voxel information is processed, the Sliding Voxel walks through the occupied voxels and calculates an overall score $\mathcal{S}_r$ for the whole neighborhood, i.e. 26 neighbors plus the current voxel:

$$\mathcal{S}_r = \sum_{i=1}^{n} \mathcal{S}_i : i \in \mathbb{Z}^+_{\leq 27}, \tag{4}$$

where $\mathcal{S}_i$ is the planarity P of the i-th neighbor voxel. As this score gets bigger, the less planar is the neighborhood. Therefore, a low score means the current voxel is coplanar. For a more straightforward parametrization, this planarity measure is normalized using the maximum and minimum observed values:

$$\mathcal{S}^*_r = \frac{\mathcal{S}_r - \min\limits_{i=1}^{n} \mathcal{S}_{r_i}}{\max\limits_{i=1}^{n} \mathcal{S}_{r_i}}, \tag{5}$$

where n is the number of occupied voxels.

Figure 3 illustrates in warmer colors how this score can approximate regions that have low probability of being coplanar.

Coplanar voxels are further filtered and sorted in ascending order by their score $\mathcal{S}^*_r$. The resulting coplanar subset $\mathcal{V}_p$ is used to find planes in a centroids point cloud $\mathcal{C}$, that was enriched with their normal vector to form an approximate representation of non-overlapping tangential

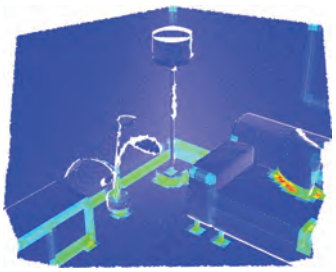


Fig. 3 Point cloud of a room model, color represents a heatmap of the scores $\mathcal{S}_r$ of each voxel

planes.

3.2 Global verification

At this point, we can map a voxel centroid to their voxel information such as its score $\mathcal{S}_r^*$, and the total number of points inside the voxel. Simultaneously, its normal vector $\hat{\mathbf{n}}$ and its centroid $\mathbf{c}$ are used to approximate the tangential plane $\mathbf{p}_v$ at $\mathbf{c}$, i.e., a geometrically enriched version of the original point cloud: $\mathcal{C}$.

The purpose of this point cloud is to provide an efficient representation to validate hypothetical planes in a global sense. For all coplanar voxels $\mathcal{V}_p$, their tangential plane $\mathbf{p}_v$ is validated against $\mathbf{c} \in \mathcal{C}$ using a decomposed plane-to-plane distance: the Euclidean distance to each $\mathbf{c}$ and the angular deviation between their normal vectors.

Ideally, the magnitude of the inner product between n normal vectors of a planar surface tend to be 1, i.e. $\frac{1}{n}\sum_{i=0,j=0}^{n} |\hat{\mathbf{n}}_i \cdot \hat{\mathbf{n}}_j| \approx 1$ where $i \neq j$. Therefore, inliers are selected by thresholding their Euclidean distance and the normalized angular distance between the hypothetical plane normal vector $\hat{\mathbf{n}}_v^h$, and the corresponding tangent plane normal vector $\hat{\mathbf{n}}_v^t$. This metric is defined by the positive cosine distance $\cos_d^+$ calculated from its similarity $\cos_s$ as follows

$$\cos_s = \frac{\hat{\mathbf{n}}_v^h \cdot \hat{\mathbf{n}}_v^t}{\|\hat{\mathbf{n}}_v^h\| \, \|\hat{\mathbf{n}}_v^t\|} \mapsto [-1, 1], \tag{6}$$

$$\cos_d^+ = \frac{2}{\pi} \cos^{-1}(|\cos_s|) \mapsto [0, 1]. \tag{7}$$

When $\cos_d^+$ takes the value of 1, the planes are orthogonal, and if it is 0, they are parallel. In the proposed method, it is used as a parameter to decide whether an inlier will be rejected or not.

3.3 Planes extraction

Inliers of the global point cloud are sometimes too disperse and that can lead to false positives. To avoid this issue, a fast 1-cluster euclidean clustering is performed from the refined inliers using a distance threshold defined by the octree voxel size $Th = 4V_s$, which is twice the maximum possible distance between centroids.

From the resulting inliers, a cluster is constructed and its plane is calculated from these via Principal Component Analysis. Since it is impossible to calculate a plane if the number of inliers is less than 3, the plane coefficients are copied from the tangent plane of the voxel with the lowest score in the cluster.

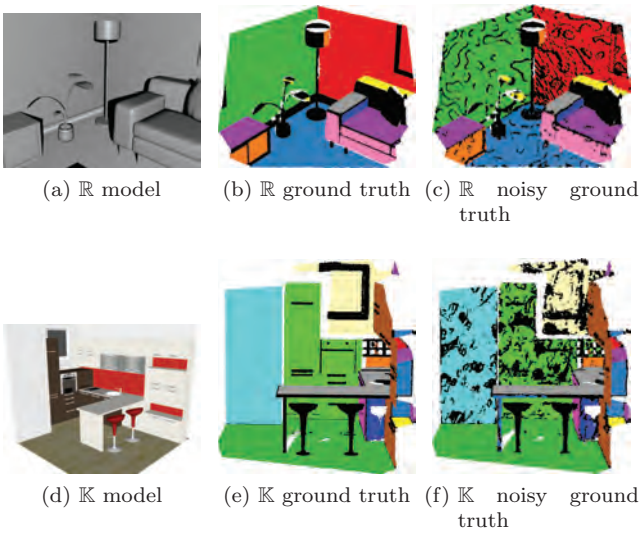
If the cluster does not map enough points on the original point cloud, or if its planarity P (see Eq. (3)) is not low enough, then the cluster is rejected. Otherwise, it is added to a list of detected clusters which includes its plane coefficients and inliers.

After all hypothetical planes are processed, they are sorted by their number of inliers. This allows us to reuse $\mathcal{C}$ and remove inliers progressively from the most to the least prominent plane.

Once the algorithm removes all the centroids inliers of the hypothetical planes, the process finishes and the list of clustered hypothetical planes becomes the detected planes list. A summary of the proposed method parameters is shown in **Table 1**.

Table 1 Proposed method parameters

Parameter	Description
Voxel size[m]	Size of the octree leaves
Planarity threshold	Maximum value for $\mathcal{S}_r^*$
Inliers threshold[m]	Max. euclidean distance to plane
Max cosine distance	Max. $\cos_d^+$ of plane and inliers
Min plane size[#]	Min. support of planes



(a) $\mathbb{R}$ model (b) $\mathbb{R}$ ground truth (c) $\mathbb{R}$ noisy ground truth

(d) $\mathbb{K}$ model (e) $\mathbb{K}$ ground truth (f) $\mathbb{K}$ noisy ground truth

Fig. 4 Datasets models

4. Experimental Results

4.1 Datasets and evaluation method

The datasets and ground truth planes used in the experiments can be seen in **Fig. 4**. Point clouds from Kinect V1 sensor simulation were built using Blensor 1.0.18[21].

A room model[22] $\mathbb{R}$ was used to generate noisy and noiseless simultaneous scans. To create a more complicated scenario, a kitchen model[23] $\mathbb{K}$ was scanned by translating and rotating 10 times the sensor in the simulation software. All scans were performed in world coordinates to correctly register the points via concatenation. The resulting point cloud has a wide variety of noise patterns which makes difficult even for humans to detect

Table 2 Dataset information

Name	BBDD[m]	Points[#]	Planes[#]
$\mathbb{R}$	6.09	295,144	9
$\mathbb{R}$ noisy	6.09	295,144	
$\mathbb{K}$	12.19	249,348	14
$\mathbb{K}$ noisy	12.19	503,398	

small planar surfaces in some locations. In addition to the registered noisy scan, we also simulated a clean registration which does not have noise or quantization artifacts. Due to registration, both $\mathbb{K}$ point clouds are incredibly dense. Hence, voxel grid filter of leaf size 0.01[m] was applied to them.

Numerical information of the datasets can be seen in **Table 2**, where BBDD stands for Bounding Box Diagonal Distance, i.e. the length of a diagonal line that crosses the bounding box of the point cloud.

For both $\mathbb{K}$ and $\mathbb{R}$, plane models were extracted directly from the polygons inside Blensor. For each planar surface, plane coefficients were generated from a polygon over each plane by using its normal vector and barycenter. This ground truth data allows us to measure numerically and precisely the accuracy of each plane detection method.

Regarding the processing time T_w, we used wall time since its clock has more resolution. We note that we excluded the time it takes to load a point cloud from disk, also, we included the time for the required preprocessing of each algorithm. For CFRANSAC and EFRANSAC we included the calculation of normal vectors inside T_w since they rely heavily on them. For the proposed method, we also included the neighborhood-aware octree preparation time.

For each run, a list of ground truth planes $\mathbf{P}^{gt}$ and detected planes $\mathbf{P}^{det}$ coefficients is prepared. Then, we defined 2 error metrics. First, the angular error ω (in degrees) between the normal vector of an i-th ground truth plane $\hat{\mathbf{n}}_i^{gt} \subset \mathbf{P}_i^{gt}$ and the normal vector of a j-th detected plane $\hat{\mathbf{n}}_j^{det} \subset \mathbf{P}_j^{det}$, is defined as

$$\omega = 180 \cos_d^+(\hat{\mathbf{n}}_i^{gt}, \hat{\mathbf{n}}_j^{det}) \mapsto [0, 180]. \tag{8}$$

Second, the offset difference δ between the i-th ground truth plane $d_i^{gt} \subset \mathbf{P}_i^{gt}$ and the j-th detected plane $d_j^{det} \subset \mathbf{P}_j^{det}$ is

$$\delta = \left| |d_i^{gt}| - |d_j^{det}| \right|. \tag{9}$$

For a ground truth plane $\mathbf{p}_i^{gt} \in \mathbf{P}^{gt}$, the best detection match is a plane from $\mathbf{p}_j^{det} \in \mathbf{P}^{det}$ that minimizes their positive cosine distance $\cos_d^+$ as vectors in $\mathbb{R}^4$, then similar to Eq. (7)

$$\cos_d^+ = \frac{2}{\pi} \cos^{-1} \left(\left| \frac{\mathbf{p}_i^{gt} \cdot \mathbf{p}_j^{det}}{\left\|\mathbf{p}_i^{gt}\right\| \left\|\mathbf{p}_j^{det}\right\|} \right| \right). \tag{10}$$

Also, a match is rejected if their $\omega > 15$[deg] and $\delta > 20$[cm]. Thus, its result will be a list of matches $\mathbf{K}$ of ground truth planes associated with their best detected plane match satisfying the above conditions.

Let $\mathbf{K}_i = \{\mathbf{P}_i^{gt}, \mathbf{P}_i^{det}\}$, where $i = \{1, 2, \ldots, M\}$ and M is the total number of matched planes (true positives), then we can define the precision γ as

$$\gamma = \frac{M}{|\mathbf{P}^{det}|}, \tag{11}$$

and recall ζ of the detection as

$$\zeta = \frac{M}{|\mathbf{P}^{gt}|}, \tag{12}$$

where $|\mathbf{P}^{det}|$ is the number of detected planes and $|\mathbf{P}^{gt}|$ is the number of ground truth planes. Therefore, the harmonic mean between precision and recall, i.e. the F_1 score[24] is defined as

$$F_1 = \frac{2}{\gamma^{-1} + \zeta^{-1}}. \tag{13}$$

The range of the above metrics is $[0, 1]$ where higher values mean better results.

On the other hand, the (true) efficiency E_{ff} measures how fast the evaluated methods detected correct planes, i.e., the number of true positives M, over the processing time T_w

$$E_{ff} = \frac{M}{T_w}. \tag{14}$$

Note that since the conventional methods are non–deterministic, we used the average and standard deviation of 50 runs.

Qualitative results are evaluated by segmenting the output planes list $\mathbf{P}^{det}$ of each method. Although inlier points can be obtained from every method natively, their patterns can confuse the reader and generate an unfair comparison. For instance, RHT results may look cleaner while it mistakenly selects protruding noise patterns as part of a planar surface, producing more significant errors in the coefficients of the resulting planes.

A native output of the proposed method is shown in **Fig. 5**, for every plane in the detection result, the centroids inliers are shown, then for each plane, a different color was chosen according to the Glasbey lookup table[25], otherwise, points are kept black. Therefore, we

Fig. 5 Detection example of the proposed method

Table 3 Parameters of the proposed method in the evaluation experiments

Parameter	$\mathbb{K}$	$\mathbb{K}$ noisy	$\mathbb{R}$	$\mathbb{R}$ noisy
Voxel size[m]	0.14	0.16	0.12	0.12
Planarity threshold	0.001	0.1	0.02	0.02
Inliers threshold[m]	0.04	0.07	0.07	0.07
Max cosine distance	0.04	0.2	0.05	0.05
Min plane size[#]	500	2500	1300	1000

Table 4 Parameters values of the segmentation algorithm

Parameter	$\mathbb{K}$	$\mathbb{K}$ noisy	$\mathbb{R}$	$\mathbb{R}$ noisy
k[#]	50	100	50	100
d_ϵ[m]	0.03	0.05	0.025	0.05
θ_ϵ[deg.]	45	45	45	45

segmented the results only by looking at the standard output of the evaluated algorithms: the list of planes coefficients $\mathbf{P}^{det}$.

As a prerequisite, point-wise normal vectors are calculated within a support of k-neighbors via local plane fitting. The segmentation algorithm share similarities with the last step of the proposed method. For each plane $\mathbf{p}_j^{det} \in \mathbf{P}^{det}$, plane-to-plane inliers are selected within a distance threshold of d_ϵ and an angular threshold θ_ϵ. Planes are sorted in descending order by their amount of inliers. In that order, we extract and remove the inliers of each plane from the point cloud. This ensures the most prominent planes are segmented correctly and the segmentation of false positives is minimum.

The parameters of the proposed method are shown in **Table 3** and for the segmentation are shown in **Table 4**.

4.2 Experiments results and discussion

The computer used to run the experiments has a CPU Intel Core i7-6700K with 32GB of RAM, it runs on Ubuntu 18.04.2 with PCL 1.8.1[26], CGAL 4.11[27], and clang++ 3.8.0. The proposed method was implemented using routines of the PCL library with O3 compiler optimizations.

Figure. 6 and **Fig. 7** show the visual results of executing the segmentation algorithm over the ground truth and the resulting planes of the evaluated methods.

Figure 6 shows the visual results of the noiseless point clouds. Figure 6(a) and (f) depict the segmented ground truth planes of $\mathbb{R}$ and $\mathbb{K}$ respectively. In Fig. 6(a), we can visualize the ground truth planes selection for the $\mathbb{R}$ model, avoiding slightly curved surfaces such as the backrest of the sofa and the pillow. Additionally, we avoided selecting parallel planes that are not far from each other, because it would be sporadic to detect those structures in noisy point clouds.

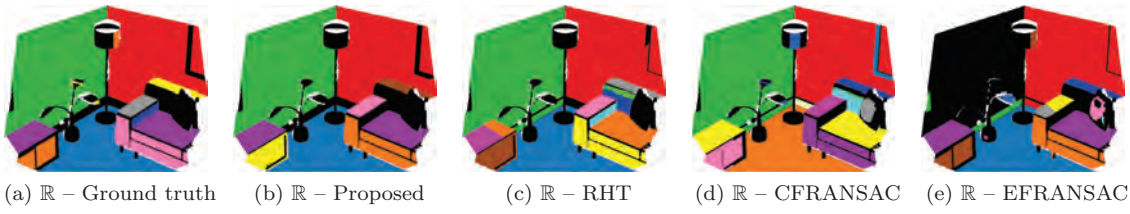
The upper row of Fig. 6 shows the results of the $\mathbb{R}$ point cloud. Noticeably, the proposed method has zero false positives in Fig. 6(b). RHT detected spurious planes in the bookshelf as well as several planes over the sofa backrest in Fig. 6(c). CFRANSAC erroneously detected the pillow and the sofa backrest as planar in Fig. 6(d). Although EFRANSAC detected some planes with good precision in Fig. 6(e), it tends to detect several spurious planes while it failed to detect even a prominent planar structure such as the wall on the left side.

The second row of Fig. 6 shows the results of the $\mathbb{K}$ point cloud. Because of its lack of noise and higher density, most methods performed accurately except for RHT, which could not detect the dining table as seen in Fig. 6(h). EFRANSAC performed fairly good in Fig. 6(j) because the higher density of $\mathbb{K}$ allows its random subsets to be more descriptive. However, its precision and recall are lower than the proposed method (as described numerically later).

Figure 7 shows the results of executing the evaluated methods over the noisy datasets. Figure 7(a) and (f) depict the segmented ground truth planes similar to Fig. 6.

The first row of Fig. 7 illustrates the segmentation results on the $\mathbb{R}$ noisy point cloud. In Fig. 7(b), the proposed method detected most of the planar structures with high accuracy and no spurious planes, showing similar segmentation patterns when comparing its results with the ground truth. Noticeably, CFRANSAC detected several spurious planes in the noisiest region of the point cloud in Fig. 7(d), whereas EFRANSAC detected less spurious planes than CFRANSAC in Fig. 7(e).

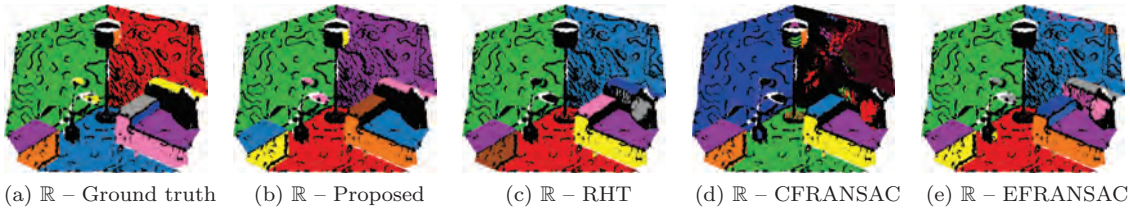
The second row of Fig. 7 shows the segmentation results on the $\mathbb{K}$ noisy point cloud. Figure 7(g) shows that the proposed method detected most of the planes while having no false positives. In Fig. 7(h), it illustrates that RHT was able to identify the most prominent planes; nonetheless, it detected fewer planes than the proposed method. CFRANSAC detected more false positives in Fig. 7(i), while it failed to detect several planar struc-



(a) $\mathbb{R}$ – Ground truth (b) $\mathbb{R}$ – Proposed (c) $\mathbb{R}$ – RHT (d) $\mathbb{R}$ – CFRANSAC (e) $\mathbb{R}$ – EFRANSAC

(f) $\mathbb{K}$ – Ground truth (g) $\mathbb{K}$ – Proposed (h) $\mathbb{K}$ – RHT (i) $\mathbb{K}$ – CFRANSAC (j) $\mathbb{K}$ – EFRANSAC

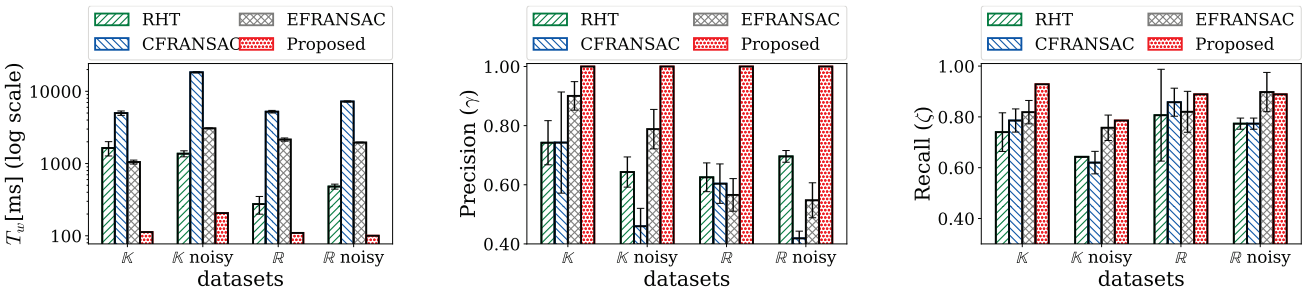
Fig. 6 Planes detection inliers using the noiseless dataset



(a) $\mathbb{R}$ – Ground truth (b) $\mathbb{R}$ – Proposed (c) $\mathbb{R}$ – RHT (d) $\mathbb{R}$ – CFRANSAC (e) $\mathbb{R}$ – EFRANSAC

(f) $\mathbb{K}$ – Ground truth (g) $\mathbb{K}$ – Proposed (h) $\mathbb{K}$ – RHT (i) $\mathbb{K}$ – CFRANSAC (j) $\mathbb{K}$ – EFRANSAC

Fig. 7 Planes detection inliers using the noisy dataset

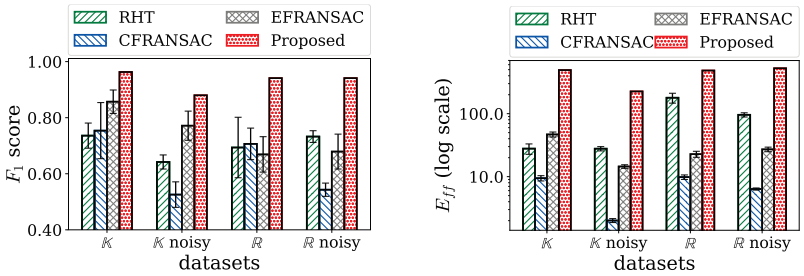


(a) Processing time T_w (lower is better) (b) Precision γ (higher is better) (c) Recall ζ (higher is better)

Fig. 8 Processing time, precision and recall of the evaluated methods

tures. In Fig. 7(j), EFRANSAC detected slightly more planes than the proposed method because it got benefited with the higher density of the $\mathbb{K}$ noisy point cloud. However, its precision is inferior as described numerically later.

While we confirmed the robustness of the proposed method visually, now we show objective assessment. Processing time, precision and recall are shown in **Fig. 8**. Each bar represents the result of executing an evaluated method over a point cloud of the dataset, where smaller bars denote better results. For the conventional methods we used the average of 50 executions and show their



(a) F_1 score (higher is better) (b) Efficiency E_{ff} (higher is better)

Fig. 9 Accuracy and F_1 score of the proposed method compared against the conventional methods

standard deviation as error bars.

Figure 8(a) shows the processing time T_w in logarithmic scale. There we can confirm that the proposed method is drastically faster than the conventional methods in every case.

The variations in terms of the mean angular and offset error of the evaluated methods are negligible; therefore, we show their assessment based on more standard metrics used in binary classification tasks: precision γ, recall ζ and F_1 score.

Figure 8(b) shows the precision γ of the evaluated methods (see Eq. (11)). This metric measures how relevant were the detection results. Noticeably, the proposed method precisely detected appropriate planes in all tested datasets.

In Fig. 8(c), we show the recall ζ as defined in Eq. (12). This metric tells us how many of the ground truth planes were detected by the evaluated methods. The proposed method most of the time detected more ground truth planes than the evaluated methods. Only in the $\mathbb{R}$ noisy point cloud, its recall is paired with EFRANSAC; noticeably EFRANSAC had a very low precision (<0.6) as shown in Fig. 8(b).

In addition to the above metrics, we evaluated the F_1 score and efficiency E_{ff} of the evaluated methods as defined in Eq. (13) and Eq. (14) respectively. **Figure 9** shows the results of applying these metrics on the experiment results. Figure 9(a) shows the overall precision of the evaluated methods. The proposed method shows a superior precision in every case; it had better scores than the best execution of the conventional methods. Figure 9(b) shows the efficiency of the evaluated methods in logarithmic scale. Here, we confirm the proposed method detects planes more accurately in a drastically more efficient way.

Furthermore, it should be noted that the parameters set of the proposed method is smaller and easier to configure. It has only 5 parameters while EFRANSAC, CFRANSAC, and RHT have 6, 12, and 16 parameters respectively.

5. Conclusions and Future Works

Planes detection has numerous applications in 3D computer vision. However, its practical use was limited by the efficiency and accuracy of the detection methods. The conventional methods are relatively accurate but inefficient and non-deterministic, relying on random sampling which does not always guarantee good results. RHT achieved faster speeds by using random sampling and Hough Voting, but it has several parameters difficult to configure, its precision can be profoundly affected by the resolution of the accumulator, and according to the literature, its performance decreases in the presence of several planes models and non-planar points. Besides, RANSAC extensions need the computation of normal vectors to detect planes correctly, which drastically reduces their efficiency since normal vectors are not raw data that can be acquired when sensing point clouds.

Therefore, we focused on solving the efficiency problem while keeping the accuracy high without needing precomputed normal vectors or convergence mechanisms like Hough voting. The proposed method is deterministic and has fewer parameters than the conventional methods; hence it is easier to configure. Moreover, the key to its robustness and efficiency is the sliding voxel, which by utilizing its overlapping grid structure, it can locate coplanar regions and calculate curvatures drastically faster.

By experiments with ground truth data and realistic simulations, we confirmed that the proposed method is accurate and drastically faster than the conventional methods. Furthermore, it can achieve real time speeds with just a single CPU core. For future works, we are

considering to improve its recall even more by adding extra refinement steps and adaptive thresholds.

References

1) G. Atanacio-Jiménez, J.-J. González-Barbosa, J.B. Hurtado-Ramos, F.J. Ornelas-Rodríguez, H. Jiménez-Hernández, T. García-Ramirez, R. González-Barbosa: "Lidar Velodyne HDL-64E Calibration Using Pattern Planes", International Journal of Advanced Robotic Systems, Vol. 8, No. 5, pp.70–82 (2011).
2) C. Nguyen, S. Izadi, D. Lovell: "Modeling Kinect Sensor Noise for Improved 3D reconstruction and Tracking", Proc. of the 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), pp. 524–530 (2012).
3) 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).
4) 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).
5) Y. Zhang, W. Xu, Y. Tong, K. Zhou: "Online Structure Analysis for Real-Time Indoor Scene Reconstruction", ACM Trans. on Graphics (TOG), Vol. 34, No. 5, pp. 159:1–159:13 (2015).
6) J. Weingarten, R. Siegwart: "3D SLAM Using Planar Segments", Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3062–3067 (2006).
7) 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).
8) K. Pathak, A. Birk, N. Vaskevicius, M. Pfingsthorn, J. Poppinga: "Fast Registration Based on Noisy Planes with Unknown Correspondences for 3-D Mapping", IEEE Trans. on Robotics, Vol. 26, No. 3, pp. 424-441 (2010).
9) 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).
10) F. Tarsha-Kurdi, T. Landes, P. Grussenmeyer: "Hough-Transform and Extended Ransac Algorithms for Automatic Detection of 3D Building Roof Planes from Lidar Data", Proc. of ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007, Vol. 36, pp. 407–412 (2007).
11) G. Gordon et al.: "The Use of Dense Stereo Range Data in Augmented Reality" Proc. of the 1st International Symposium on Mixed and Augmented Reality, pp. 14 (2002).
12) K. Okada, S. Kagami, M. Inaba, I. Hirochika: "Plane Segment Finder: Algorithm, Implementation and Applications", Proc. of the IEEE International Conference on Robotics and Automation, Vol. 2, pp. 2120–2125 (2001).
13) S. Oßwald, J. Gutmann, A. Hornung, M. Bennewitz: "From 3D Point Clouds to Climbing Stairs: A Comparison of Plane Segmentation Approaches for Humanoids", Proc. of 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 93–98 (2011).
14) 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).
15) D. Holz, S. Holzer, R. Rusu, S. Behnke: "Real-Time Plane Segmentation Using RGB-D Cameras", Robot Soccer World Cup, pp. 306–317 (2011).
16) S. Holzer, R. Rusu, M. Dixon, S. Gedikli, N. Nassir: "Adaptive Neighborhood Selection for Real-Time Surface Normal Estimation from Organized Point Cloud Data Using Integral Images", Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2684–2689 (2012).
17) 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, No. 2, pp. 3 (2011).
18) 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).
19) L. Xu, E. Oja, P. Kultanen: "A New Curve Detection Method: Randomized Hough Transform (RHT)", Pattern Recognition Letters, Vol. 11, No. 5, pp. 331–338 (1990).
20) H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, W. Stuetzle: "Surface Reconstruction from Unorganized Points", Proc. of the 19th Annual Conference on Computer Graphics and Interactive Techniques, pp. 71–78 (1992).
21) 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).
22) A. Handa, T. Whelan, J. McDonald, A. Davison: "A Benchmark for RGB-D Visual Odometry, 3D Reconstruction and SLAM", Proc. of the IEEE International Conference on Robotics and Automation (ICRA), pp. 1524–1531 (2014).
23) Marela kitchen red&white, https://3dwarehouse.sketchup.com (2019).
24) C. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Natural Language Engineering, Cambridge University Press (2010).
25) C. Glasbey, G. Van der Heijden, V. Toh, A. Gray: "Colour Displays for Categorical Images", Color Research & Application, Vol. 32, No. 4, pp.304–309 (2007).
26) R. Rusu, S. Cousins: "3D Is Here: Point Cloud Library (PCL)", Proc. of IEEE International Conference on Robotics and Automation (ICRA), pp. 1–4 (2011).
27) CGAL: Computational Geometry Algorithms Library, https://www.cgal.org (2019).

(Received September 13, 2019)

Jaime SANDOVAL
(*Student 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. Currently, he is a Ph.D. student in the Interdisciplinary Graduate School of Science and Technology of Shinshu University with a major in Systems Development Engineering. His research interests are 3D Point Clouds Processing, Computer Vision and Image Processing. He received IEVC2019 Excellent Paper Award.

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. He received IEVC2019 Excellent Paper Award, and Excellent Journal Paper Award from IIEEJ in 2018.

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 during 2013–2016. 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 IEVC2019* --

Pairwise Registration of Low Overlapping Unorganized 3D Point Clouds Using Supervoxel Segmentation

Luis PERALTA†, Jaime SANDOVAL††, Munetoshi IWAKIRI†††, Kiyoshi TANAKA††††

† Graduate School of Science and Technology, Shinshu University,
†† Interdisciplinary Graduate School of Science and Technology, Shinshu University,
††† Department of Computer Science, National Defense Academy of Japan,
†††† Academic Assembly (Institute of Engineering), Shinshu University

<**Summary**> Since its introduction, the Iterative Closest Points (ICP) algorithm has led to developing a wide range of registration methods, most of these variations of ICP itself. Notwithstanding the efforts on improving the speed and accuracy of ICP, these variations cannot correctly align point clouds which overlapping ratio is considered low (under 40%) due to an inherited local minima convergence. Furthermore, more advanced registration techniques that rely on point descriptors also cannot overcome this problem because the tuning of their parameters tends to be volatile, which leads to making false point correspondences and consequently failing to perform an accurate registration. In order to solve this problem, we propose a pairwise registration approach that does not entirely rely on point descriptors and leverages the local minima convergence of ICP to correctly align 3D point clouds with overlapping ratios as low as about 20%. Our method uses the supervoxel segmentation technique to divide the point clouds into subsets and finds those which registration maximizes the overlapping ratio between correct correspondences in the full point clouds. We verified the effectiveness of the proposed method through tests in dense models and real-world scan datasets.

Keywords: 3D point clouds, registration, low overlapping ratio, supervoxel segmentation

1. Introduction

3D point clouds are three-dimensional data point sets that represent the shape of objects or scenes, which, in recent years, their use has gained popularity in the area of computer vision for the development of intelligent autonomous systems, such as indoor robots and self-driving cars. Besides, applications such as surface measuring and digital reconstruction for manufacturing, and geomechanics, to mention a few. The last two examples require the use of registration, one of the essential tasks in point cloud processing.

The registration task aims to find the Euclidean motion, expressed by a transformation matrix $\mathbf{T}$, that best aligns two or more point clouds with different reference frames to the same one. Hence, generating a more detailed and complete representation of a model or scene.

The ICP algorithm, introduced by Besl and McKay[1], is the most widely used method to perform registration of point sets. Its variations focus on improving the accuracy of alignment and expediting the convergence to the best alignment. However, since the core operation of ICP is an iterative local optimization, these get stuck in local minima. Hence, failing to correctly register point clouds which initial state is not close to the global minima, or have a low proportion of points in their shared surfaces. This proportion is known as the overlapping ratio ξ and has an essential role in the registration of partial overlapping point clouds. As ξ gets lower, it increases the probability of failure of conventional registration methods[2] such as ICP, normal to plane registration (also known as Normal-ICP or N-ICP)[3], and Levenberg-Marquardt ICP (LM-ICP)[4].

In order to address this issue, other approaches look for $\mathbf{T}$ between point correspondences obtained from 3D point descriptors[5]. Nonetheless, the success of finding correct correspondences depends on its descriptiveness, which at the same time depends on the normal vectors of the points. Besides, since the normal vectors are affected by ξ and no rule specifies what a suitable normal vector is, it is difficult to rely only on point descriptors.

To the best of our knowledge, the volume of work that addresses low overlapping ratio point clouds is very scarce. The only few existing approaches work based on

looking for the points or subsets that belong to the shared surfaces. Either by using point descriptors[6] or by computationally complex selection processes[2].

In this work, we propose a method capable of registering a pair of point clouds with overlapping ratios below 40%, which is not driven by point matches made from descriptors, and computational complexity allows it to find a solution considerably fast. The approach takes advantage of the local minima convergence of conventional methods, by iteratively clustering the point clouds with supervoxel segmentation[7], and looking among the produced subsets the pair which registration leads to the best alignment.

The remainder of this paper organizes as follows. In section 2, we introduce the mathematical representation for pairwise registration of 3D point clouds and review related work of partially overlapping point clouds registration. We present in detail our approach in section 3. Details of the experiments, results, and discussion are in section 4. Lastly, in section 5, we conclude and present the focus for future work.

2. Related Works

The volume of published work relevant to pairwise registration of 3D point clouds is very extensive. However, only a few approaches take into consideration the overlapping ratio. Thus, we only introduce related work that set the base for Euclidean registration, address the overlapping ratio, and inspired to define our approach. Furthermore, we refer the reader to surveys on registration[8)–11)] to know more about the different ICP variants and related approaches.

2.1 Conventional registration methods

ICP is the most straightforward registration algorithm, and its formulation is considered as the base for pairwise registration. For 3D point clouds, having two point clouds, $P = \{\mathbf{p}_i\}_{i=1}^{N_P}$ and $Q = \{\mathbf{q}_j\}_{j=1}^{N_Q}$, named as source and target respectively. ICP looks for the optimal $\mathbf{T}$ that applied to P best aligns it to Q minimizing the error

$$E = \sum_{i=1}^{N_P} w_i \left\| \mathbf{q}_{\phi_i} - \mathbf{T}\left(\mathbf{p}_i\right) \right\|^2 \tag{1}$$

given by the Euclidean distance between point correspondences $C_{pq} = \{(\mathbf{T}\left(\mathbf{p}_i\right), \mathbf{q}_{\phi_i})\}_{i=1}^{M}$. These correspondences are defined by a function ϕ that sets a weight w_i as 1 if a point $\mathbf{p}_i$ has a correspondence $\mathbf{q}_{\phi_i}$, or 0 otherwise. In ICP ϕ is the point-to-point nearest neighbor function; hence, $\mathbf{q}_{\phi_i}$ represents the nearest neighbor of $\mathbf{T}\left(\mathbf{p}_i\right)$ in Q. Moreover, $\mathbf{T}$ is a matrix of six Degrees Of Freedom (DOF), formed by rotation and translation parameters around and along the three axes.

Table 1 Conventional registration methods comparison

Difference	ICP	N-ICP	LM-ICP
ϕ function	Point-to-point nearest neighbor	Point-to-plane nearest neighbor	Point-to-point nearest neighbor
$\mathbf{T}$ update	Umeyama's least-squares estimation	Umeyama's least-squares estimation	Gradient descent and Gauss-Newton

ICP performs the registration process in iterations, and at each iteration two steps take place:

1. C_{pq} are estimated by a point-to-point nearest neighbor function.
2. $\mathbf{T}$ is updated based on Umeyama's least-squares estimation method[12].

The iterations are performed until the C_{pq} do not change, and there is no possible update in $\mathbf{T}$. Otherwise when E reaches a specific Euclidean distance ϵ_f known as the Euclidean Fitness Epsilon.

N-ICP performs registration in a very similar manner to ICP, but it computes the normal vectors and uses the spanning planes to estimate C_{pq} by a point-to-plane nearest neighbor function.

LM-ICP is an improved ICP robust to the initial positions of the point clouds. It estimates C_{pq} the same way as ICP, but updates $\mathbf{T}$ based on the Levenberg-Marquardt algorithm. **Table 1** shows a comparison between these three conventional registration methods.

2.2 TrICP

In conventional methods, the ϕ functions define C_{pq} taking into considerations the distance between all the points. However, when the point clouds do not have the same number of points or are partially overlapping these functions may estimate a high number of false C_{pq}. In order to estimate a more accurate error for these cases the Trimmed ICP (TrICP)[13),14)] restricts E by taking into consideration only the points that belong to the overlapping surface, proposing the Trimmed Mean Squared Error

$$TMSE = \frac{1}{N_P^{'}} \sum_{\mathbf{p}_i \in P_\xi} w_i \left\| \mathbf{q}_{\phi_i} - \mathbf{T}\left(\mathbf{p}_i\right) \right\|^2 . \tag{2}$$

Nevertheless, it defines the overlapping ratio as $\xi = \frac{N_P^{'}}{N_P}$, where $N_P^{'}$ is the number of points in the subset P_ξ, which represents the overlapping part of P to Q. It means that P_ξ defines the C_{pq} and the prior knowledge of the overlapping area is necessary.

2.3 Practical registration of LiDAR scans

Due to their partial overlapping and complex structures, 3D point clouds obtained from LiDAR scanners require practical considerations to ease and accelerate their registration. Cai et al.[15] consider that in practice relative rotations are constrained to the azimuth in terrestrial systems. This consideration removes two DOF and allows to look for a $\mathbf{T}$ defined only by translations along the three axes and a single rotation around the vertical axis.

Furthermore, to robustly estimate E and reject false C_{pq}, they proposed to measure the Euclidean distance error only between the points that lay within an inlier threshold ε that represents the scanning precision of the device. Nonetheless, this approach also relies on C_{pq} defined by descriptors in a pre-processing step.

2.4 Low overlapping ratio registration methods

Despite the lack of registration methods focused on low overlapping ratio point clouds, the few existing approaches have shown to obtain good results with or without relying on point descriptors. Wu et al.[6] rely on point descriptors to define the C_{pq} but proposes the idea that in low overlapping ratio point clouds the shared surfaces are more likely to be found at the boundaries, pruning the search for C_{pq} to only these areas. The approach proposed by Peralta et al.[2] follows the same boundaries ideas, but without relying on point descriptors. Instead, they proposed a random Hough voting search to look for the point subsets that best align between each other, and which registration correctly aligns the full point clouds. Nevertheless, due to its random search and inline construction of subsets for each point in the point clouds, the processing time required to analyze a pair of point clouds thoroughly is too long.

3. Proposed Method

3.1 Problem reformulation

In contrast to conventional registration methods, for low overlapping point clouds, the goal is to estimate the error of alignment considering only the points that belong to the overlapping surfaces. Thus, it is possible to say that Eq. (2) is a better error estimation than Eq. (1), making TrICP the ideal method for the problem. However, due to the way ξ is determined and its prior knowledge, delimit the practicality of the method.

Because of the lack of prior knowledge of ξ in a realistic situation, the overlapping ratio estimation method proposed Peralta et al.[2] becomes more suitable than the one in TrICP. This method measures ξ between a pair of already registered point clouds as the percentage of points that are close enough to be considered part of the same surface as

$$\xi = \left(\frac{2M}{N_P + N_Q} \right) 100 \ . \tag{3}$$

In order to define C_{pq}, the method looks for each point in the source the nearest neighbor in the target within a radius of a ε equal to two times the cloud resolution CR. That is, for each point in the source which nearest neighbor lays within ε, there is a correspondence. Thus, the total of points with a correspondence considering both point clouds is $2M$. We note that CR is the mean Euclidean distance between all the points in a point cloud and their nearest neighbor.

By combining a non-normalized TMSE and ε, it is possible to set an objective function that estimates the Euclidean distance error only between the correct overlapping points.

$$E\left(\mathbf{T} \mid C_{pq}, \varepsilon\right) = \sum_{i=1}^{M} w_i \left(\left\| \mathbf{q}_{\phi_i} - \mathbf{T}\left(\mathbf{p}_i\right) \right\|^2 \leq \varepsilon \right) \tag{4}$$

Then, the $\mathbf{T}$ that produces the best alignment is the one that solves the optimization problem:

$$E^* = \max_{\mathbf{T}} E\left(\mathbf{T} \mid C_{pq}, \varepsilon\right) \tag{5}$$

In order to find this transformation, we propose to exploit the local minima convergence of conventional methods by clustering the point clouds and look the pair of subsets (P_{ss_k}, Q_{ss_l}) that have a similar shape and size, and therefore its registration produces the $\mathbf{T}$ that applied to the full point clouds with a considerable small ε accomplishes Eq. (5). The core process of the proposed method is composed of the following four steps, detailed from section 3.2 to section 3.5:

1. Point clouds pre-processing.
2. Point subsets generation.
3. Point subsets combinations rejection and registration.
4. Transformations evaluation.

3.2 Point clouds pre-processing

As the first step, we compute the normal vectors of P and Q defined by a normal radius R_n. Then, these normal vectors are used to compute the FPFH descriptor[5] of all the points in both point clouds, parameterized by an FPFH radius R_{fpfh}.

Despite computing the FPFH point descriptors, unlike registration methods that make the C_{pq} by matching descriptors, our approach uses the FPFH descriptor to give a geometrical description to point subsets, as is introduced in section 3.4.

3.3 Point subsets generation

Next, we divide both point clouds into subsets $P_{sv} = \{P_{ss_k}\}_{k=1}^{SS_P}$ and $Q_{sv} = \{Q_{ss_l}\}_{l=1}^{SS_Q}$ using the Voxel Cloud Connectivity Segmentation, also known as supervoxel segmentation[7].

The supervoxel segmentation clusters point clouds similarly to the 2D superpixels but generating a fast volumetric over-segmentation of 3D point clouds that adheres to the boundaries. In detail, it divides the 3D space into a voxelized grid space with a resolution of seed radius R_{seed}, where each voxel has a resolution of voxel radius R_{voxel} and sets a point seed for each cluster. An expansion distance defined as

$$\Delta_P = \sqrt{\lambda \frac{D_c^2}{m^2} + \mu \frac{D_s^2}{3R_{seed}^2} + \epsilon D_{HiK}^2} \ , \tag{6}$$

computes the range of the clusters from the point seeds. Where λ controls the influence of color information as a Euclidean color distance D_c in the CIELab space, μ controls the spatial distance D_s between points, and ϵ controls the geometric similarity of the points measured by the FPFH descriptor.

Eq. (6) is the core that defines the point subsets in the supervoxel segmentation method. Nevertheless, our implementation, as we introduce in section 4, is built with the Point Cloud Library (PCL)[16] and its documentation [17] simplifies the expansion distance as

$$\Delta'_P = \sqrt{w_c D_c^2 + w_s \frac{D_s^2}{3R_{seed}^2} + w_n D_n^2} \ . \tag{7}$$

Unlike Eq. (6), the computation of Eq. (7) is in a feature space consisting of the color Euclidean distance D_c from the normalized RGB space, the space distance D_s normalized by R_{seed}, and the normal distance D_n that measures the angle between normal vectors. Moreover, w_c, w_s, and w_n replace λ, μ, and ϵ respectively, but still are user parameters which represent weights that control the effect of the distances.

We note that the actual implementation of the supervoxel segmentation available in PCL version 1.8 defines the expansion distance as

$$\Delta''_P = w_c D_c + \frac{w_s D_s}{R_{seed}} + w_n D_n \ , \tag{8}$$

Algorithm 1 Subsets combinations rejection

Require: Source point subset P_{ss_k}.
Require: Target point subset Q_{ss_l}.
Require: Corresponding descriptors P_{FPFHss_k} of P_{ss_k}.
Require: Corresponding descriptors Q_{FPFHss_l} of Q_{ss_l}.
for all $\mathbf{p}_{fpfh}$ in P_{FPFHss_k} **do**
  Find the 1st and 2nd nearest neighbors from Q_{FPFHss_l}.
  Compute the ratio between distances of the 1st and 2nd nearest neighbors.
  if ratio $< C_{thr}$ **then**
    Increase G_s by 1.
  else
    Continue.
if $G_s = 0$ **then**
  Do not register P_{ss_k} and Q_{ss_l}.
else
  Register P_{ss_k} and Q_{ss_l}.
  Save the corresponding $\mathbf{T}$ in V_T.

and no available documentation states the reasons for these changes, which are beyond the scope of this work. However, the implementation still works as asserted by the authors of supervoxel segmentation.

3.4 Point subsets registration and combinations rejection

Afterward, we look for transformations from the registration of subsets, which may be similar in size and shape. Therefore, all the possible subsets combinations (P_{ss_k}, Q_{ss_l}) are registered using LM-ICP, and the resulting transformations $\mathbf{T}$ are saved in a vector V_T, which size is $SS_P \times SS_Q$ if all the combinations are registered. However, registering all the combinations implies a probability of obtaining transformations which subsets do not correspond to geometrically related surfaces. Thus, it is necessary to avoid registering geometrically unrelated subsets. To achieve this, we set a degree of geometric similarity G_s, which measures how similar are a pair of subsets by counting the point descriptor correspondences between them. Hence, the higher the value of G_s, the more similar and geometrically related are the subsets.

The rejection of the point subsets process is summarized in **Algorithm 1**. Starting with a pair of subsets (P_{ss_k}, Q_{ss_l}) and since the points in P and Q are directly related their previously computed FPFH descriptors, it is possible to make clusters of these descriptors in the same order as resulted in the supervoxel segmentation for the point clouds. Thus, there is a one-to-one related descriptors subset pair $(P_{FPFHss_k}, Q_{FPFHss_l})$ that describes the original point subsets pair, and is used to perform the rejection of subsets combinations. First, descriptor correspondences are estimated in the same manner as Buch et al.[18] and Lowe[19], to measure the geometric similar-

ity between the subsets. For each point descriptor $\mathbf{p}_{fpfh}$ in P_{FPFHss_k}, the first and second nearest neighbors in Q_{FPFHss_l} are found. If the ratio between their distances is lower than a correspondence threshold C_{thr}, the first nearest neighbor is considered a correspondence, and G_s increases by one. Then, once all point descriptors in P_{FPFHss_k} are evaluated, the pair (P_{ss_k}, Q_{ss_l}) is rejected if G_s has a value of zero, or registered if otherwise.

For the registration of the subsets, we chose LM-ICP due to its robustness to the initial positions. A property that makes this algorithm ideal to obtain the registration transformation of subsets which may be virtually part of the same surface in the model, but have a different orientation in their respective point clouds.

3.5 Transformations evaluation

Lastly, the transformation $\mathbf{T}_b$ that achieves the best alignment is defined by applying one by one all the $\mathbf{T}$ saved in V_T to P and evaluating with Eq. (4). Then, $\mathbf{T}_b$ is the $\mathbf{T}$ that accomplish Eq. (5).

3.6 Iterative process

All the evaluated transformations come from the registration of considerably similar point subsets. Thus the size of these subsets plays a significant role in the proposed method. R_{seed} controls the size of these point subsets and should be more significant than R_{voxel}. The larger is R_{seed}, the less but bigger are the generated point subsets, and since there is no defined optimal value for R_{seed}, it is also necessary to find the value that generates the point subsets that lead to the optimal transformation $\mathbf{T}_o$.

In order to achieve it, we iterate over steps 2 to 4 of the core process by decreasing the value of R_{seed} at a step rate τ, from a maximum R_{max} to a minimum R_{min}. Which R_{max} can be as significant enough to at least generate a couple of subsets at the first iteration, but due to the spatial relationship between supervoxels and the voxel grid in the segmentation process $R_{min} = R_{voxel} + \tau$.

At each iteration, R_{seed} has a different value to generate point subsets in both point clouds and register them as described in section 3.4. Then as explained in section 3.5, the resulted transformations are evaluated to find $\mathbf{T}_b$. All the $\mathbf{T}_b$ from each iteration are saved in a vector V_{T_b} in order to perform the same evaluation with Eq. (4) and define the optimal transformation $\mathbf{T}_o$ as the best from the best. That is, $\mathbf{T}_o$ is the $\mathbf{T}_b$ that accomplish Eq. (5). **Algorithm 2** summarizes the complete iterative registration method.

Algorithm 2 Iterative registration method

Require: Source point cloud P.
Require: Target point cloud Q.
Compute the normal vectors of P.
Compute the normal vectors of Q.
Compute the FPFH descriptors of P.
Compute the FPFH descriptors of Q.
$R_{seed} = R_{max}$
while $R_{seed} > R_{voxel}$ **do**
    Divide P into subsets, $P_{sv} = \{P_{ss_k}\}_{k=0}^{SS_P}$.
    Divide Q into subsets, $Q_{sv} = \{Q_{ss_j}\}_{j=0}^{SS_Q}$.
    for all P_{ss_k} in P_{sv} **do**
        for all Q_{ss_l} in Q_{sv} **do**
            Run **Algorithm 1**.
    for all $\mathbf{T}$ in V_T **do**
        Evaluate $\mathbf{T}$ on Eq. (4).
        if $\mathbf{T}$ accomplishes Eq. (5) **then**
            $\mathbf{T}_b = \mathbf{T}$
            Save $\mathbf{T}_b$ in V_{T_b}.
        else
            Continue.
    $R_{seed} = R_{seed} - \tau$
for all $\mathbf{T}_b$ in V_{T_b}. **do**
    Evaluate $\mathbf{T}_b$ on Eq. (4).
    if $\mathbf{T}_b$ accomplishes Eq. (5) **then**.
        $\mathbf{T}_o = \mathbf{T}_b$
    else
        Continue.
Apply $\mathbf{T}_o$ to P.

4. Experimental Results

The proposed method was implemented in C++ 11 using the Point Cloud Library (PCL)[16] version 1.8. Furthermore, for parallel processing, we implemented the steps for the subsets rejection, registration, and evaluation of transformations using OpenMP.

4.1 Datasets and experimental setup

The utilized datasets consist of dense model point clouds, and sparse scene point clouds obtained from laser scanners. The dense point clouds are pairs of the models Bunny, Dragon, Happy Buddha, and Armadillo from the Stanford 3D Scanning Repository[20]. The sparse point clouds are pairs of the scene Stairs from the ASL Datasets[21]. Each pair of point clouds was chosen based on having an ξ below 40%, and the corresponding ground-truth alignment available. **Table 2** shows the specific datasets for each model and scene, the source-target arrangement, their number of points, and corresponding ξ.

We ran the proposed method with the different chosen datasets while recording the corresponding registration metrics E and ξ of $\mathbf{T}_b$ at each iteration, as well as the processing time. Additionally, in order to have a metric to estimate the error of alignment between the points that genuinely overlap the Mean Squared Error with Penalty was set as

Table 2 Datasets and corresponding ξ

Model/Scene	Datasets	ξ (%)
Bunny	P: bun000 N_P: 40,256 points Q: bun270 N_Q: 31,701 points	30.90
Dragon	P: dragonStandRight_0 N_P: 41,841 points Q: dragonStandRight_288 N_Q: 24,573 points	29.64
Happy Buddha	P: happyStandRight_0 N_P: 78,056 points Q: happyStandRight_288 N_Q: 72,346 points	19.88
Armadillo	P: ArmadilloStand_0 N_P: 28,220 points Q: ArmadilloStand_270 N_Q: 24,034 points	26.66
Stairs	P: Hokuyo_0 N_P: 181,077 points Q: Hokuyo_3 N_Q: 187,959 points	33.04

$$MSE_p = \frac{1}{N_P}\sum_{i=1}^{N_P} w_i \left(d_{pq} \leq \varepsilon\right) \tag{9}$$

$$d_{pq} = \begin{cases} \|\mathbf{q}_{\phi_i} - \mathbf{T}\left(\mathbf{p}_i\right)\|^2, & \text{if } w_i = 1 \\ 1, & \text{if } w_i = 0 \end{cases} \tag{10}$$

It measures the Mean Squared Error between the truly overlapping points, but unlike $TMSE$, instead of not considering the points of the not overlapping parts, MSE_p gives a penalty of 1 for each $\mathbf{p}_i \in P$ that does not have a correspondence $\mathbf{q}_{\phi_i} \in Q$ within ε (see Eq. (10)). Since in low overlapping conditions, the overlap is significantly small, MSE_p allows to estimate small values of error of alignment for point clouds that are correctly registered, and vice-versa.

For the dense model point clouds, the experiments were performed with the parameters shown in **Table 3** in terms of CR, except for C_{thr}, which represents a ratio $\in [0, 1]$. Parameter w_c was set as 0 because the datasets do not have color information, and C_{thr} as 0.1 to guarantee to find correct descriptor correspondences. R_{max} and τ, were defined large enough to generate at least two subsets at the first iteration, and small enough to gradually evaluate a wide range of sizes for the subsets. Moreover, ε was set with the same value proposed by Peralta et al.[2] to ensure the estimation of ξ between the overlapping points.

The experiments with the sparse scene point clouds kept the same values for most of the parameters, except for R_{voxel} and R_{max}. Since the scale of the stairs scene is several times larger than the models, by using the same values the supervoxel segmentation started generating a large number of subsets, when it is desired to generate only a few at the first iteration. Hence, these parameters were scaled-up by 10.

Table 3 Parameters values utilized in the experiments

Parameter	Value [CR]	Parameter	Value [CR]
R_n	10	ε	2
R_{fpfh}	10	R_{voxel}	2.5
w_c	0	R_{max}	100
w_s	2.5	τ	1
w_n	0.5	C_{thr}	0.1

Additionally, we ran the conventional registration methods with the same datasets to compare the results against the ground-truth and the proposed method.

4.2 Results and discussion

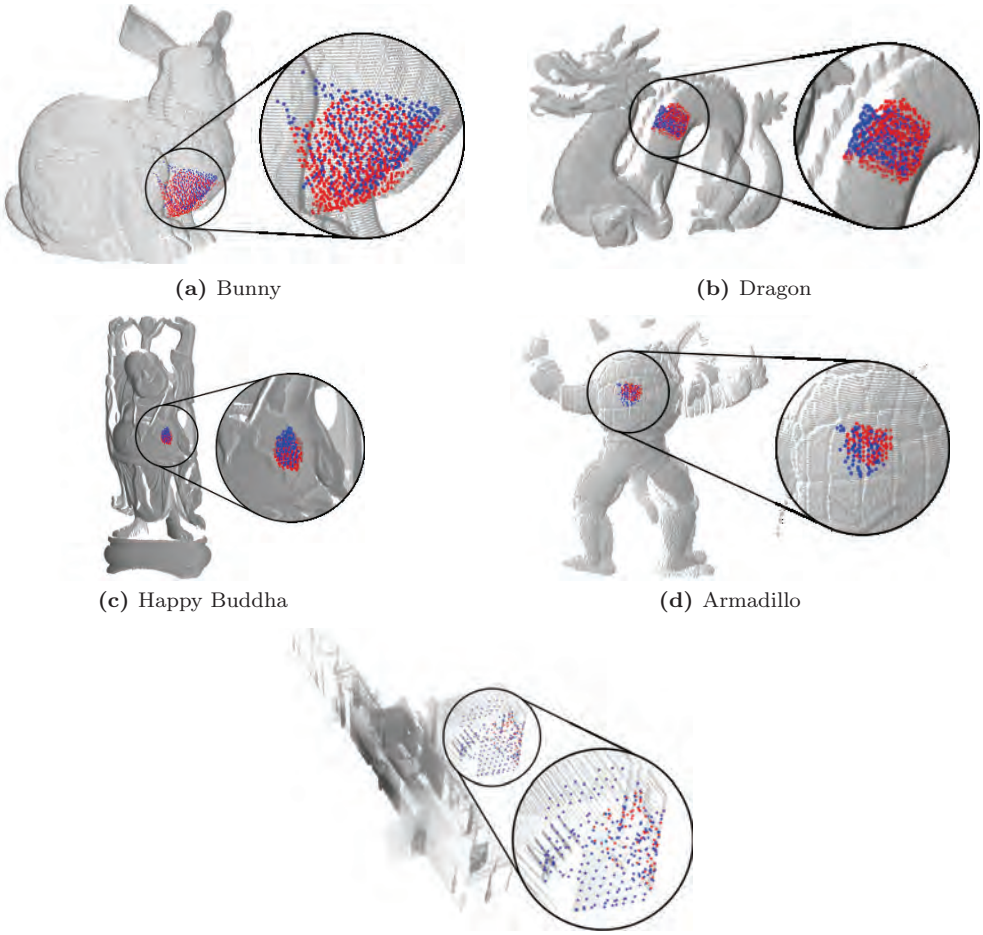
Being the source and target clouds depicted in red and blue, respectively, **Table 4** shows the registrations obtained by the ground-truth, the proposed method, and the conventional methods for all the datasets.

Visually, unlike conventional registration methods, the proposed method registers the point clouds keeping the shape of the models and the scene, generating an alignment close to the ground-truth. In the cases of Bunny, Dragon, Happy Buddha, and Stairs, the resulting alignments are almost perfect. Nevertheless, the lack of descriptiveness in the overlapping surfaces of the Armadillo datasets causes an alignment farther from the ground-truth. Using the same color pattern as in Table 4, **Fig. 1** depicts the subsets that lead to $\mathbf{T}_o$ for all the datasets. By comparing the Armadillo ground-truth in Table 4 and the position of the subsets in Fig. 1 (d) it is possible to notice the reason for the resulting alignment in this model. In the ground-truth, most of the overlapping surfaces are in the round part at the back of the Armadillo. This surface is mostly round and similar to a sphere; it does not have any other geometric description than being round. Thus, at step 3, the core process allows the registration of all the combinations coming from this part, without considering if these arise from opposite sides. An analogy to this effect is the registration of the two halves of a sphere since both do not have any other geometric description than their roundness, the final registration would be a bowl-like shape instead of a sphere.

In the ground-truth column of Table 4, the metrics E, ξ, and MSE_p have the values for the perfect alignment. Therefore, as the values of these metrics measured in the other registration methods become closer to the ones in the ground-truth, the more similar is the obtained registration to the perfect alignment. In this manner, the

Table 4 Registration comparison between the ground-truth, the proposed method, and conventional methods

Unregistered point clouds	Ground-truth	Proposed method	ICP	N-ICP	LM-ICP
 Bunny	 $E = 6.89$ $\xi = 30.99\%$ $MSE_p = 0.7239$	 $E = 4.74$ $\xi = 19.20\%$ $MSE_p = 0.8284$	 $E = 1.26$ $\xi = 5.06\%$ $MSE_p = 0.9547$	 $E = 1.55$ $\xi = 6.21\%$ $MSE_p = 0.9444$	 $E = 1.25$ $\xi = 4.99\%$ $MSE_p = 0.9554$
 Dragon	 $E = 5.48$ $\xi = 29.64\%$ $MSE_p = 0.7648$	 $E = 5.62$ $\xi = 30.18\%$ $MSE_p = 0.7605$	 $E = 1.41$ $\xi = 6.27\%$ $MSE_p = 0.9502$	 $E = 1.45$ $\xi = 6.24\%$ $MSE_p = 0.9504$	 $E = 1.43$ $\xi = 6.36\%$ $MSE_p = 0.9495$
 Happy Buddha	 $E = 6.08$ $\xi = 19.88\%$ $MSE_p = 0.8084$	 $E = 6.78$ $\xi = 22.70\%$ $MSE_p = 0.7813$	 $E = 3.11$ $\xi = 9.55\%$ $MSE_p = 0.9080$	 $E = 3.41$ $\xi = 10.67\%$ $MSE_p = 0.8971$	 $E = 2.99$ $\xi = 9.20\%$ $MSE_p = 0.9113$
 Armadillo	 $E = 4.56$ $\xi = 20.66\%$ $MSE_p = 0.7533$	 $E = 4.03$ $\xi = 23.63\%$ $MSE_p = 0.7813$	 $E = 0.64$ $\xi = 3.67\%$ $MSE_p = 0.9660$	 $E = 0.77$ $\xi = 4.61\%$ $MSE_p = 0.9573$	 $E = 0.64$ $\xi = 3.63\%$ $MSE_p = 0.9663$
 Stairs	 $E = 896.42$ $\xi = 33.04\%$ $MSE_p = 0.6682$	 $E = 812.27$ $\xi = 26.31\%$ $MSE_p = 0.7363$	 $E = 132.32$ $\xi = 4.09\%$ $MSE_p = 0.9589$	 $E = 96.11$ $\xi = 2.93\%$ $MSE_p = 0.9706$	 $E = 128.05$ $\xi = 3.87\%$ $MSE_p = 0.9611$



(a) Bunny **(b)** Dragon

(c) Happy Buddha **(d)** Armadillo

(e) Stairs

Fig. 1 Point subsets that lead to $\mathbf{T}_o$

Table 5 Optimal R_{seed} and processing times

Model/Scene	R_{seed} $[CR]$	t_T (sec)	t_M (sec)
Bunny	70	444.74	2.47
Dragon	58	136.88	1.04
Happy Buddha	31	552.60	3.08
Armadillo	31	94.75	0.73
Stairs	927	5,925.2	0.98

proposed method also quantitatively surpasses conventional methods. It is noticeable how the metrics in the proposed method are significantly closer to the ground-truth in all the datasets. Although there are still some noticeable differences, as in the case of Bunny, these do not indicate a wrong result, but that there is still room for improvement in the alignment, which can be achieved by a fine registration method. Thus, the proposed approach can be considered a coarse registration method.

Regarding the optimal R_{seed} and processing time, **Table 5** shows the resulting optimal sizes, as well as corresponding total t_T and median t_M processing times. In the same way as Peralta et al.[2)], this approach is based on a subsets registration principle. However, as opposed to that method, this one takes only a few minutes to process the model datasets fully and find a solution and does not require user intervention to compare the results with the ground-truth. On the other hand, due to the scale of the sparse scene datasets, the scaled-up values of R_{max} and R_{voxel} impact on t_T since more values of R_{seed} have to be processed.

5. Conclusions

By providing enough shape descriptiveness in the overlapping areas, the proposed approach is capable of finding transformations that register a pair of point clouds with a low overlapping ratio close to the perfect alignment, close enough to consider it a coarse registration method. Therefore, it may be useful to perform multi-view registration (i.e., registration of more than two point clouds), but using a fine registration method between every two datasets to avoid carrying out the residual error of misalignment.

For future work, we will focus on improving the dependence to the size and shape of the point subsets, as well as the indirect dependence to point descriptors to find re-

lated surfaces, by defining better techniques to describe the points that belong to the overlapping areas. Moreover, we will evaluate the adaptability of the method to real-time applications.

References

1) P.J. Besl, N.D. McKay: "A Method for Registration of 3-D Shapes", IEEE Trans. of Pattern Analysis and Machine Intelligence, Vol. 14, No.2, pp.239–256 (1992).
2) L. Peralta, J. Sandoval, M. Iwakiri, K. Tanaka: "A Preliminary Study on Low Overlapping Unorganized Point Clouds Registration Using Hough Voting", Proc. of IIEEJ Technical Meeting, Vol. 288, pp.8–13 (2019).
3) Y. Chen, G. Medioni: "Object Modeling by Registration of Multiple Range Images", Proc. of IEEE International Conference on Robotics and Automation (ICRA 1991), Vol. 3, pp.2724–2729 (1991).
4) A. Fitzgibbon: "Robust Registration of 2D and 3D Point Sets", Image and Vision Computing, Vol. 21, pp.1145–1153 (2002).
5) R. Rusu, N. Blodow, M. Beetz: "Fast Point Feature Histograms (FPFH) for 3D Registration", Proc. of IEEE International Conference on Robotics and Automation (ICRA 2009), pp.3212–3217 (2009).
6) Y. Wu, W. Wang, K. Lu, Y. Wei, Z. Chen: "A New Method for Registration of 3D Point Sets with Low Overlapping Ratios", Procedia CIRP (13th CIRP conference on Computer Aided Tolerancing), Vol. 27, pp.202–206 (2015).
7) J. Papon, A. Abramov, M. Schoeler, F. Wörgötter: "Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds", IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), pp.2027–2034 (2013).
8) J. Salvi, C. Matabosch, D. Fofi, J. Forest: "A Review of Recent Range Image Registration Methods with Accuracy Evaluation", Image and Vision Computing, Vol. 25, No. 5, pp.578–596 (2007).
9) J. Santamaría, O. Cordon, S. Damas: "A Comparative Study of State-of-the-art Evolutionary Image Registration Methods for 3D Modeling", Computer Vision and Image Understanding, Vol. 115, pp.1340–1354 (2011).
10) F. Pomerleau, F. Colas, R. Siegwart, S. Magnenat: "Comparing ICP Variants on Real-World Data Sets", Autonomous Robots, Vol. 34, pp.133–148 (2013).
11) S. Rusinkiewicz, M. Levoy: "Efficient Variants of the ICP Algorithm", Proc. of Third International Conference on 3-D Digital Imaging and Modeling, pp.145–152 (2001).
12) S. Umeyama: "Least-squares Estimation of Transformation Parameters Between Two Point Patterns", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 13, No. 4, pp.376–380 (1991).
13) D. Chetverikov, D. Svirko, D. Stephanov, P. Kresk: "The Trimmed Iterative Closest Point Algorithm", Object Recognition Supported by User Interaction for Service Robots, Vol. 3, pp.545–548 (2002).
14) S. Xu, J. Zhu, Z. Jiang, Z. Lin, J. Lu, Z. Li: "Multi-view Registration of Unordered Range Scans by Fast Correspondence Propagation of Multi-scale Descriptors", PLOS ONE, Vol. 13, No. 9, pp.1–18 (2018).
15) Z. Cai, T. Chin, A.P. Bustos, K. Schindler: "Practical Optimal Registration of Terrestrial LiDAR Scan Pairs", ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 147, pp.118–131 (2019).
16) R. Rusu: "Semantic 3D Object Maps for Everyday Manipulation in Human Environments", KI - Künstliche Intelligenz, Vol. 24 (2009).
17) pointclouds.org, Clustering of Point Clouds into Supervoxels, http://pointclouds.org/documentation/tutorials/supervoxel_clustering.php (2019).
18) A.G. Buch, D. Kraft, J. Kamarainen, H.G. Petersen, N. Krüger: "Pose Estimation Using Local Structure-specific Shape and Appearance Context", Proc. of IEEE International Conference on Robotics and Automation (ICRA 2013), pp.2080–2087 (2013).
19) D.G. Lowe: "Distinctive Image Features from Scale-invariant Key-points", International Journal of Computer Vision, Vol. 60, No. 2, pp.91–110 (2004).
20) M. Levoy, J. Gerth, B. Curless, K. Pull: "The Stanford 3D Scanning Repository" , http://graphics.stanford.edu/data/3Dscanrep (2005).
21) F. Pomerleau, M. Liu, F. Colas, R. Siegwart: "Challenging Data Sets for Point Cloud Registration Algorithms", The International Journal of Robotics Research, Vol. 31, No. 14, pp. 1705–1711 (2012).

(Received September 13, 2019)

Luis PERALTA (*Student Member*)
He received the B.Eng. degree in Mechatronics in 2014 from Instituto Tecnologico de Celaya, Mexico, and the M.Eng. degree in Electronic Information Systems in 2019 from Shinshu University, Japan. He is currently working towards an Eng.D. degree in Information Communication Systems at Shinshu University. His research interests include computer vision, virtual reality, machine learning, and robotics.

Jaime SANDOVAL
(*Student Member*)
He received the B.Eng. degree in Computer Systems in 2009 from Universidad del Valle del Fuerte, Mexico, and his M.Eng. degree in Electrical and Electronic Engineering in 2017 from Shinshu University, Japan. Currently, he is a Ph.D. student at the Interdisciplinary Graduate School of Science and Technology of Shinshu University with a major in Systems Development Engineering. His research interests are 3D point cloud processing, computer vision, and image processing. He received IEVC2019 Excellent Paper Award.

Munetoshi IWAKIRI (*Member*)
He received the B.Eng. degree in Computer Science, and the M.Eng. degree in Mathematics and Computer Science from the National Defense Academy of Japan in 1993 and 1998, respectively. In 1999, he joined the Department of Computer Science of the National Defense Academy of Japan, as a Research Associate. In 2002, he received the Dr.Eng. degree from Keio University, Japan. The institution where he became a lecturer and associate professor in 2005 and 2015, respectively. He is a member of the Information Processing Society of Japan and pursues research related to multimedia processing and information security.

Kiyoshi TANAKA (*Fellow*)
He received a B.Sc. and M.Sc. degree in Electrical Engineering and Operation Research in 1984 and 1989, respectively, from the National Defense Academy of Japan. In 1992, he received the Dr.Eng. degree from Keio University, Japan. In 1995, he joined the Department of Electrical and Electronic Engineering, Faculty of Engineering of Shinshu University, Nagano, Japan, and currently, he is a full professor at the academic assembly (Institute of Engineering) of Shinshu University. He is the Vice-president of Shinshu University as well as the director of the Global Education Center (GEC) of the same institution. 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 the JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation entitled Global Research on the Framework of Evolutionary Solution Search to Accelerate Innovation during 2013–2016. He is a member of IEEE, IEICE, IPSJ, and JSEC. He is the former editor in chief of the 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 IEVC2019* --

Weakly-Supervised Learning for Continuous Sign Language Word Recognition Using DTW-Based Forced Alignment and Isolated Word HMM Adjustment

Natsuki TAKAYAMA†(*Member*), Hiroki TAKAHASHI†,††(*Member*)

†Graduate School of Informatics and Engineering, the University of Electro-Communications,
††Artificial Intelligence Exploration Research Center

<**Summary**> The reduction of the manual work of annotation is an essential part of sign language recognition research. This paper describes one weakly-supervised learning approach for continuous sign language word recognition. The proposed method consists of forced alignment based on dynamic time warping and isolated word hidden markov model adjustment using 'embedded training'. While the proposed forced alignment only requires one manual annotation for each isolated sign language word, it can generate sufficient quality of the annotation to initialize isolated word hidden markov models. 'Embedded training' adjusts initial hidden markov models to recognize continuous sign language words using only ordered word labels. The performance of the proposed method is evaluated statistically using a dataset that includes 5,432 isolated sign language word videos and 4,621 continuous sign language word videos. The averaged alignment error of the proposed forced alignment was 4.02 frames. The averaged recognition performances of the initial models were 74.82% and 91.14% in the signer-opened and trial-opened conditions, respectively. Moreover, the averaged recognition performances of the adjusted models were over 65.00% for all conditions. The evaluation shows significant improvements compared to the previous weakly-supervised learning.

Keywords: DTW, embedded training, HMM, sign language recognition, weakly-supervised learning

1. Introduction

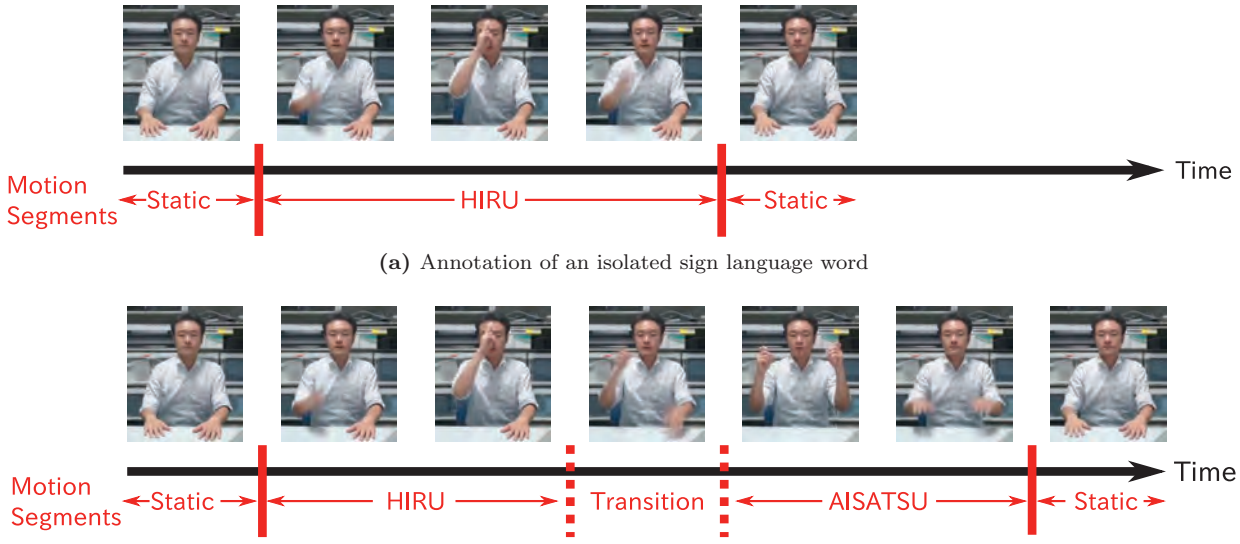
1.1 Background

Sign language recognition is an important research topic to improve communication between native signers who use sign language in their communication, and speakers. Continuous efforts of about thirty years in the field make it possible to recognize continuous sign language words[1)–4)]. These successes are, however, dependent on the availability of sign language corpora.

Sign languages are commonly represented by hand motions and shapes, as well as non-manual signals that include posture, facial expressions, gazes, and mouth motions. The representations and grammar of sign languages are, however, different in each community, and no unified rules have been established to describe them. Therefore, the sign language corpora for sign language recognition should be built individually.

The sign language corpora for recognition require large-scale videos and annotation. The annotation for sign language recognition is composed of defined time ranges based on motion units and their recognition labels. **Figure 1** shows examples of annotations to sign language videos. Figure 1 (a) shows an example of annotation for an isolated sign language word "HIRU," which means "afternoon" in Japanese Sign Language (JSL). Figure 1 (b) shows that of continuous sign language words "HIRU" and "AISATSU." "AISATSU" means "greeting," and these continuous sign language words make up the sentence "Good afternoon" in Japanese. The red bars indicate the borders of the motion units. Each label in the units describes the type of motion. "Static" indicates static posture without sign motions. Generally, the manual work of annotation is time-consuming. In particular, the annotation of continuous sign language word videos is a difficult task even to a professional annotator because there are "Transition" motions between words as shown in Fig.1 (b). For these reasons, it is essential to reduce the manual work of annotation to progress the research of sign language recognition.

As one of the solutions for the above problem, weakly-supervised learning, which trains models with a simple and limited amount of manual annotation, has received much attention in recent years. One of the standard approaches of weakly-supervised learning for sign language recognition is training with ordered word labels[2)]. The



(a) Annotation of an isolated sign language word

(b) Annotation of continuous sign language words

Fig. 1 Examples of annotation to sign language videos

ordered word labels are a word array without time range information. Previous weakly-supervised learning does not require the definition of time ranges; therefore, the annotation work can be reduced to a reasonable amount. The previous methods, however, rely on 'flat-start' initialization which segments a video uniformly according to the ordered word labels, and they train initial recognition models using the uniformly segmented videos. 'Flat-start' initialization often negatively affects recognition performance when the errors of the initial segmentation are significant.

1.2 Motivation and proposed method

Hence, in this paper, we propose one approach to weakly-supervised learning for continuous sign language word recognition using forced alignment based on Dynamic Time Warping (DTW)[5] and isolated word Hidden Markov Model (HMM) adjustment using 'embedded training' [6),7)]. The primary motivation of this research is building high-performance continuous sign language word recognition models with as little manual annotation as possible. The proposed weakly-supervised learning only requires one manual word-level annotation for each isolated sign language word, as well as ordered word labels of continuous sign language words. The word-level annotation defines the motion units based on the representing sign language words and gives them word labels, as shown in Fig.1 (a). The word-level annotation of the isolated sign language words is simple and easier than that of the annotation for continuous sign language words. This style of annotation can be expected to be feasible even for a non-professional annotator.

Forced alignment is a process to find an alignment of a time sequence to ordered word labels. Although the popular forced alignment in sign language recognition is Viterbi alignment based on HMM with ordered word labels[3], we employ that of DTW. While the HMM-based forced alignment requires a certain amount of manual annotation to build the pre-trained models, DTW-based forced alignment can be available with one manually annotated reference data for each word.

We apply the proposed forced alignment to isolated sign language word videos and train initial word HMM using the generated annotation. After the initial training, we apply 'embedded training', which is also known as concatenated training, to the initial HMM. 'Embedded training' requires only ordered word labels; therefore, difficult annotation is avoided.

The proposed initialization process provides a significant improvement of recognition performance over 'flat-start' initialization, and the recognition performance of the proposed method is close to that of 'bootstrap' initialization which utilizes all manual annotations of the isolated sign language word videos. This paper provides experimental comparisons to show the superiority of the proposed method over 'flat-start' and 'bootstrap' initialization.

The remainder of this paper is organized as follows. In Section 2, we introduce the previous weakly-supervised learning approaches of sign language recognition. In Section 3, we describe our research settings. In Section 4, we explain the proposed weakly-supervised learning. In

Section 5, we compare the results obtained by the proposed method and other initialization methods. Finally, in Section 6, we provide our conclusions and suggestions for future research.

In the following sections, the terms "isolated words" and "continuous words" indicate isolated sign language words and continuous sign language words, respectively, to avoid redundant representations.

2. Related Work

This section introduces the previous weakly-supervised learning approach of sign language recognition. Koller et al. applied 'flat-start' initialization as a pre-process of training a convolutional neural network to classify hand shapes during signs[2]. They define mappings between each word and hand shapes based on Sign-Writing[8], and 'flat-start' initialization divides video frames in words according to the mappings. We note that their 'flat-start' initialization is conducted on the sub-unit-level annotation. The sub-unit-level annotation depicts finer motion types than the word-level annotation[1], and the errors of the initial segmentation may be insignificant. Unfortunately, adequate mappings between each word and the hand shapes of many sign languages including JSL have not been established. Therefore, in many cases, 'flat-start' initialization has negative effects on recognition performance.

Koller et al. also proposed iterative forced alignment based on their CNN-BLSTM-HMM system[3]. Their method mutually updates the annotation and recognition models, and the recognition performance improves with each repetition. However, in their approach, the initial annotation should be given by manual or other pre-trained models. This limitation is common in HMM-based forced alignment.

Automatic model extraction using video subtitles is also an interesting approach in the field. Farhadi et al.[9] proposed a data mining method to automatically extract common pairs of words and signs from the subtitled videos. Buehler et al.[10] proposed a similar approach based on multiple instance learning. These methods have the advantage that manual annotation is not required, but it is difficult to train uncommon words.

3. Research Settings

In this section, we explain a database and a base sign language recognition method to depict the research settings in advance.

Table 1 Summary of the number of words in the sentences

Number of words	2	3	4	5	6	7	8
Sentences	14	17	12	4	8	5	2

3.1 Database

The specifications of the database are depicted in the following. We suppose a situation where a single signer signs in front of a camera. We recorded videos with a smart-phone camera and requested the signers to sit and sign in office environments. Moreover, we requested that the signers pose in the static posture at the beginning and the end of a sign, as shown in Fig.1. All video frames were recorded at 30 frames per second with 640×360 pixels. We note that each signer is recorded in different places. Therefore, each video has a unique background depending on the signer. The database includes 109 isolated JSL words that were signed by ten native signers and 62 types of continuous words that were signed by eighteen non-native signers and one native signer. The ten native signers include three males and seven females in their twenties to 50s. The one native signer of continuous words is a female in her twenties who is included in the ten native signers of isolated words. They use JSL in daily communication. The eighteen non-native signers include fourteen males and four females in their twenties to 50s. The recording of eighteen non-native signers was conducted after several hours of training because they are beginners of JSL. The continuous words are built from the 109 words to make short sentences. The number of words in the sentences are summarized in **Table 1**. We requested that the signers perform each isolated word five times. In this paper, we use the term "trials" to indicate these repetitions. As a result, 5,432 videos were obtained except for the recordings that failed. Moreover, we recorded four trials for each of the continuous words, and 4,621 videos were obtained except for recording and sign errors. While all isolated words have manually defined time ranges and those labels, ordered word labels without time ranges are defined for the continuous words.

All experiments in this study were conducted using the database. In this paper, we compare the three types of weakly-supervised learning, namely, 'flat-start' and 'bootstrap' initialization, and the proposed method. These methods are characterized by available annotations to train initial models as depicted in **Table 2**. We note that all methods use 'embedded training' with the ordered word labels to adjust the initial models.

Table 2 Available annotation for HMM initialization

Method	Annotation
'flat-start'	Ordered word labels
'bootstrap'	Full manual labels
Proposed	One manual label for each word

3.2 Base sign language recognition

Three types of the weakly-supervised learning are tested on the previously proposed sign language recognition[11]. The sign language recognition consists of body parts tracking using OpenPose[12)–15)], feature extraction, and parts-based multi-stream HMM. We provide a brief introduction to sign language recognition in the following.

(a) Body-parts tracking

In this research, we employ 7, 21, and 21 tracking points of OpenPose for a body, left hand, and right hand, respectively. The seven points of a body include the joints of the left and right arms, both shoulders, and the neck. We removed the points of a part of the torso, lower body, and face because the tracking points of these parts were almost static during signs in our settings.

(b) Feature extraction

The feature extraction converts the raw tracking data to abstracted sign features using base feature computation and feature abstraction. Three types of base features, normalized tracking positions, point connections according to the human skeleton, and first-order derivatives of positions and connections, are calculated from the raw tracking data. The normalized tracking positions are the transformed two-dimensional coordinates. The transformation is conducted for each body part. The point connections are distances and directions between the pairs of neighboring points. The first-order derivatives are calculated only for the body because the tracking of hands is sometimes unstable. The feature abstraction is composed of Z-score normalization and Principle Component Analysis (PCA). Z-score normalization adjusts the scales of each dimension. PCA removes the redundant dimensions.

As a result of these processes, the raw tracking data are converted to 96-dimensional sign features, which include 25, 37, and 34-dimensional vectors of the body and left and right hands. The forced alignment utilizes the 96-dimensional sign features that were obtained as a result of the feature extraction.

(c) Parts-based multi-stream HMM

We designed left to right parts-based multi-stream HMM to test weakly-supervised learning. $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ and $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ indicate model parameters of HMM and an observed time sequence, respectively. $\boldsymbol{A} = \{a_{ij}; i, j = 1, 2, \ldots, N\}$ and $\boldsymbol{B} = \{b_i(\boldsymbol{x})\}$ represent the state transition and output probabilities, respectively. These parameters are trained through initial and 'embedded training'. $\boldsymbol{\pi} = \{\pi_1 = 1, \pi_{i>1} = 0\}$ represents initial state transition probability, and we employed fixed $\boldsymbol{\pi}$ to construct left to right HMM. N is the number of hidden states of HMM, and we employed $N = 5$ and $N = 20$ for the models of the static posture and each isolated word, respectively. T is the number of observations in an observed time sequence.

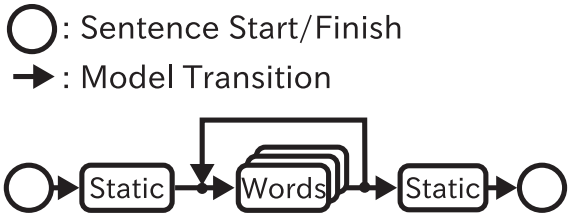


Fig. 2 Word network for sign language recognition

The output probability of the HMM is defined as:

$$b_i(\boldsymbol{x}_t) = \prod_{s=1}^{S} \left\{ \mathcal{N}[\boldsymbol{x}_{st}; \boldsymbol{\mu}_{is}, \mathrm{diag}(\boldsymbol{\Sigma}_{is})] \right\}^{\gamma_s}. \quad (1)$$

$S = 3$ is the number of streams associated with body parts. γ_s is a relative weight to a stream s. Although γ_s can be used to set the relative importance of each body part, we use $\gamma_s = 1$ in this paper to simplify the setting. $\mathcal{N}[\cdot; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\Sigma})]$ is a single multivariate Gaussian distribution with a diagonal covariance matrix. $\boldsymbol{\mu}_{is}$ and $\boldsymbol{\Sigma}_{is}$ are the multivariate mean and covariance, which are associated with the i_{th} hidden state.

We note that the first and last hidden states of HMM are non-emitted states, and these are used to connect HMMs in 'embedded training' and the classification.

The classification is based on the Viterbi algorithm using a linguistic model, which is shown in **Fig.2**. This linguistic model supposes that a sentence consists of "Static"→"Word"→"Static." The loop structure in Fig.2 allows for many word repetitions. The training of models and classification are conducted using Hidden Markov Toolkit[7].

4. Weakly-Supervised Learning

4.1 Process overview

In this section, we describe the proposed weakly-supervised learning. An overview of the proposed weakly-supervised learning is illustrated in **Fig.3**. The boxes and

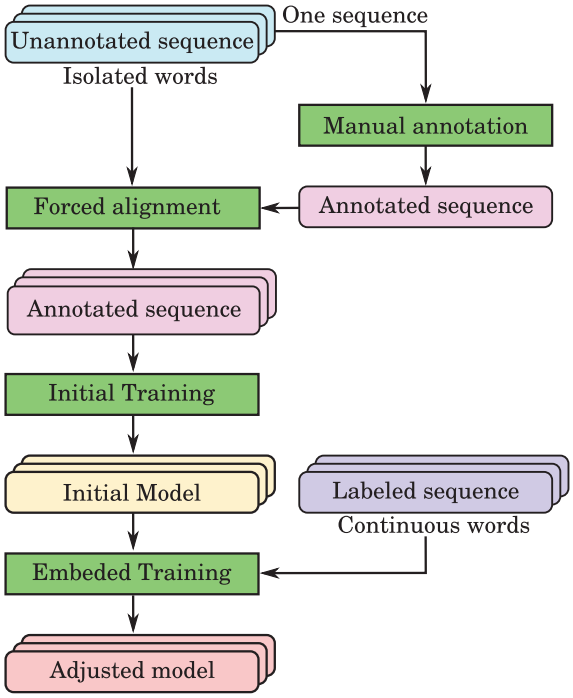


Fig. 3 Process overview

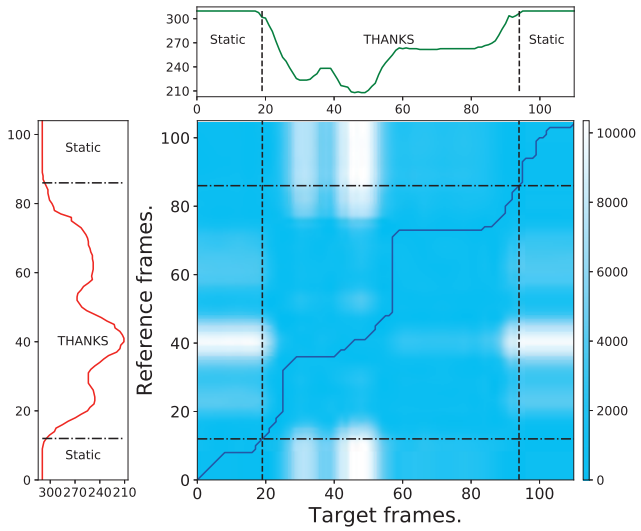


Fig. 4 Example of forced alignment using DTW

rounded boxes in Fig.3 are the processes and data, respectively. While the annotated sequence means that time ranges and those labels are defined, the labeled sequence only has ordered word labels. The unannotated sequence does not have any time ranges and labels, but it is assumed to have the same ordered word labels as manually annotated sequences. First, one sequence is sampled for each isolated word and annotated manually. This annotated sequence is used to assign annotations to the other sequences using forced alignment based on DTW. Next, initial word HMMs are trained based on the Viterbi algorithm and Baum-Welch (BW) re-estimation using the annotated sequences. Finally, the initial word HMMs are adjusted based on 'embedded training' using the labeled sequences of continuous words.

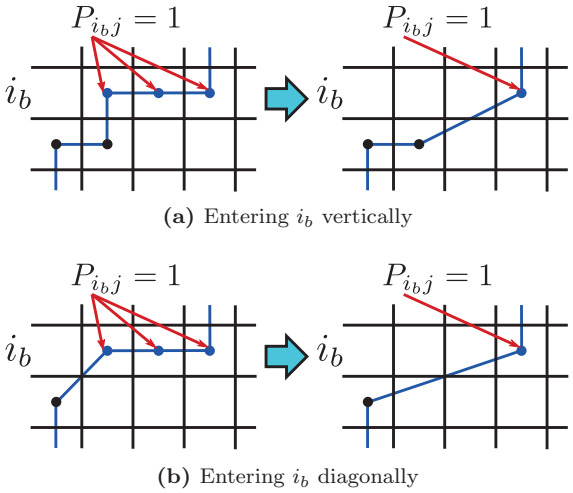
4.2 Forced alignment based on DTW

DTW[5] is one of the standard techniques to find a temporal alignment between a pair of time sequences. $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_{T_x}; \boldsymbol{x}_i \in R^n\}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_j, \ldots, \boldsymbol{y}_{T_y}; \boldsymbol{y}_j \in R^n\}$ indicate a pair of time sequences. $\boldsymbol{P}_{T_x,T_y} \subset \{0,1\}^{T_x \times T_y}$ is a set of binary alignment matrices. $\boldsymbol{P} \in \boldsymbol{P}_{T_x,T_y}$ is a $T_x \times T_y$ matrix that indicates an alignment path according to the DTW constraints, and $p_{ij} = 1$ indicates that $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are aligned. $\Delta(\boldsymbol{X}, \boldsymbol{Y})$ is a $T_x \times T_y$ cost matrix. Given these notations, the alignment matrix with the minimum cost $\boldsymbol{P}_{min} \in \boldsymbol{P}_{T_x,T_y}$ is defined as follows:

$$\boldsymbol{P}_{min} = \underset{\boldsymbol{P} \in \boldsymbol{P}_{T_x,T_y}}{\operatorname{argmin}} < \boldsymbol{P} : \Delta(\boldsymbol{X}, \boldsymbol{Y}) >, \qquad (2)$$

where $< \boldsymbol{M}_1 : \boldsymbol{M}_2 >$ is the Frobenius inner product of two matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, which have the same row and column dimensions. DTW efficiently finds $\boldsymbol{P}_{min}$ based on dynamic programming.

Figure 4 shows an example of forced alignment using DTW. The left and top graphs indicate y coordinates of a right wrist of a reference and target sequence during a sign, respectively. We note that Fig.4 shows a comparison of scalar values to visualize the sequences, but the 96-dimensional sign features are used in the experiments. The central colored matrix shows a cost matrix. The cost matrix has the squared Euclidean distances between the observations of the pair of time sequences as its elements. The vertical axes of the left graph and the cost matrix are the frame indices of the reference data. The horizontal axes of the top graph and the cost matrix are the frame indices of the target data. The blue line on the matrix shows the alignment path $\{(i,j); p_{ij} = 1\}$ with the minimum cost, which goes through the minimal cost points from the bottom left corner to the top right corner. The dotted lines and labels on the left graph show the annotation of the reference data. The boundaries of the reference annotation can be represented as horizontal lines in the cost matrix. When i_b is an index that coincides with a boundary of the reference annotation, an index of the target boundary can be found by seeking $p_{i_b j} = 1$. This means that the target boundary is determined as a vertical line that goes through a cross point of the reference boundary and the alignment path, as shown in the cost matrix of Fig.4.



(a) Entering i_b vertically

(b) Entering i_b diagonally

Fig. 5 Examples of the alignment path conversion

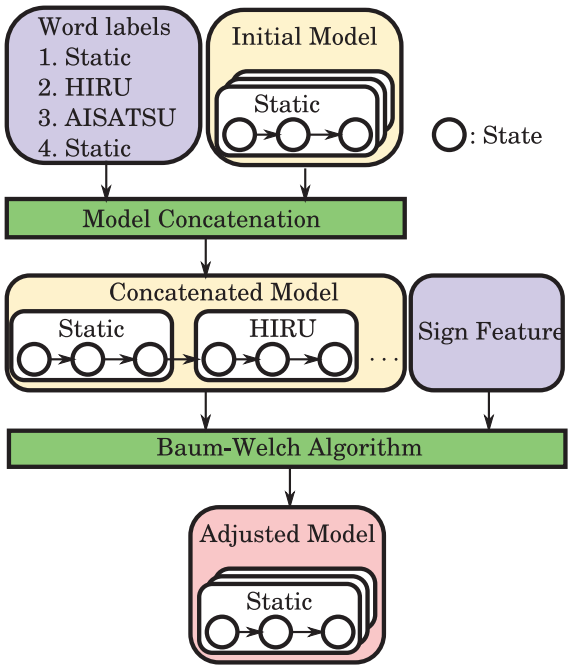


Fig. 6 Process flow of 'embedded training'

However, the index of the target boundary is not uniquely determined when the multiple target frames are aligned with the reference boundary. This is the case when the horizontal part of the alignment path goes along the dotted horizontal lines in Fig.4.

We avoid this case by using an alignment path conversion. We suppose two cases, as shown in **Fig.5**. The squares in Fig.5 show neighborhoods of $p_{i_b j} = 1$ on the alignment path. The blue dots indicate $p_{i_b j} = 1$ on the alignment path. Fig.5 (a) is the case when the alignment path enters i_b vertically, and Fig.5 (b) is the diagonal case. When the multiple candidates of $p_{i_b j} = 1$ exist, the alignment path is converted by employing the maximum index of j in both cases.

The path conversion is an empirical process. Although employing the minimum index of j and some interpolation techniques are also available, the effects of these selections on recognition performance were minor in our research settings.

4.3 Isolated word HMM adjustment

(a) Initial training

The initial training is composed of initial parameter estimation and the adjustment of parameters. First, training observation sequences of an isolated word HMM are sampled according to the annotation, which is generated by forced alignment. Next, each observation sequence is divided into equal segments according to the hidden states of the HMM, and the means and covariances of each output probability are initialized using the sample mean and covariances. Next, the observations are assigned to each hidden state using the Viterbi algorithm, and the parameters are updated. Finally, the BW algorithm adjusts the parameters based on expectation maximization. We applied the BW algorithm four times to adjust the parameters in this research.

(b) Model adjustment

Figure 6 shows the process flow of 'embedded training'. First, 'embedded training' concatenates isolated word HMMs according to the ordered word labels of continuous words. Next, 'embedded training' applies the BW algorithm to the concatenated HMM, and the parameters of the isolated word HMMs included in the concatenated HMM are simultaneously updated. We applied parameter adjustments using 'embedded training' on two occasions in this research.

5. Evaluation

This section describes evaluations of the proposed method. We report the alignment and recognition performances based on the forced alignment and classification by HMM, respectively. Moreover, we show comparisons to 'flat-start' and 'bootstrap' initialization. All evaluations in this section are based on the database and sign language recognition, which are described in Section 3.

5.1 Alignment performance

The performance of the forced alignment was evaluated based on the distances of annotation boundaries between the manual and aligned annotations. This evaluation was conducted using 5,432 videos of the 109 isolated words by ten native signers. The average duration of the 5,432

Table 3 Alignment performance [frames]

Method	Maximum	Minimum	Mean	SD
'flat-start'	24.58	9.83	16.05	3.12
Proposed	16.78	0.46	4.02	3.21

videos is 66.87 frames.

While the distances of 'flat-start' initialization are uniquely determined, those of the proposed method depend on the sampled reference annotation for DTW. Hence, we first divided the dataset into subsets for each isolated word and calculated the maximum, minimum, mean, and Standard Deviation (SD) of the distances for each isolated word. While the distances of the annotation boundaries of 'flat-start' initialization were calculated one by one with comparison to the manual annotations, those of the proposed method were calculated based on leave-one-out cross-validation for each isolated word. We then summarized the performance values based on macro averaging of all the isolated words.

The alignment performances are summarized in **Table 3**. The proposed method achieved distances about four times lower than that of 'flat-start' initialization.

5.2 Recognition performance

(a) Performance of initial models

We report the performance of the initial models using 5,432 videos of the 109 isolated words by ten native signers.

The recognition performance was calculated based on the word classification accuracy. The word classification accuracy is defined as:

$$Accuracy = \frac{N - D - S - I}{N} \times 100, \qquad 3$$

where N, D, S, and I indicate the total number of words, the number of deletions, substitutions, and insertions, respectively[7].

The performance of the initial models using 'flat-start' and 'bootstrap' initialization was evaluated using ten-fold and five-fold cross-validations, respectively. The ten-fold cross-validation uses the videos of nine subjects for training and those of one subject for testing. On the other hand, the five-fold cross-validation uses the videos of 4/5 trials of all subjects for training and those of 1/5 trials for testing. We refer to these experimental settings as the signer-opened and trial-opened conditions, respectively.

For the evaluation of the proposed method, we sampled one sequence for each word from the same subject; forced alignment was applied to other sequences. Next,

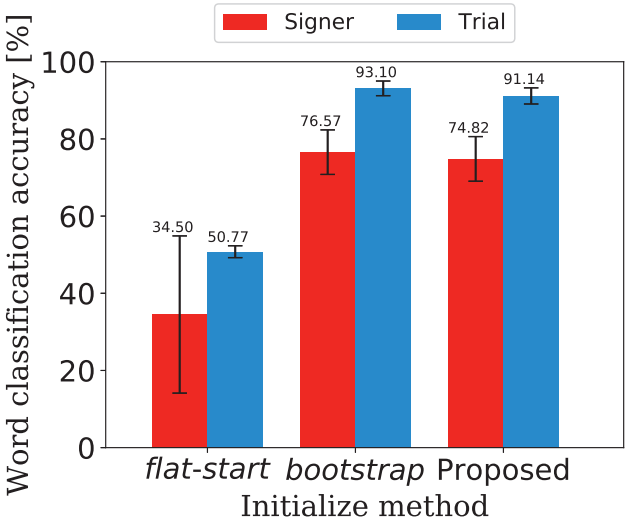


Fig. 7 Performance of the initial models

evaluations based on the signer-opened and trial-opened conditions were conducted. We repeated the sampling and evaluation for every subject, and the macro averages for all subjects were calculated.

The performance of the initial models is summarized in **Fig.7**. The red and blue bars in Fig.7 represent the word classification accuracy of the signer-opened and trial-opened conditions, respectively. The error bars represent the SD for the mean accuracies in the cross-validation. The values on the upper error bar denote the mean accuracies. We note that the summarized values of the proposed method are the macro averages for all patterns of the forced alignment.

As shown in Fig.7, the performance of the proposed method was superior to that of 'flat-start' initialization with sufficient margins, and comparable to that of 'bootstrap' initialization. At the same time, we found that the errors of 'flat-start' initialization resulted in degraded recognition performance.

(b) Performance of adjusted models

The performance of the adjusted models was evaluated using 4,621 videos of the 62 continuous words by eighteen non-native signers and one native signer. The performance of the adjusted models depends on the initial models and the composition of the training and test datasets. Therefore, we combined the cross-validation of initial and 'embedded training'. There are ten and five initial models respectively of the signer-opened and trial-opened conditions after the initial training. For the evaluation of the adjusted models, we sampled one initial model and applied 'embedded training'. Similar to

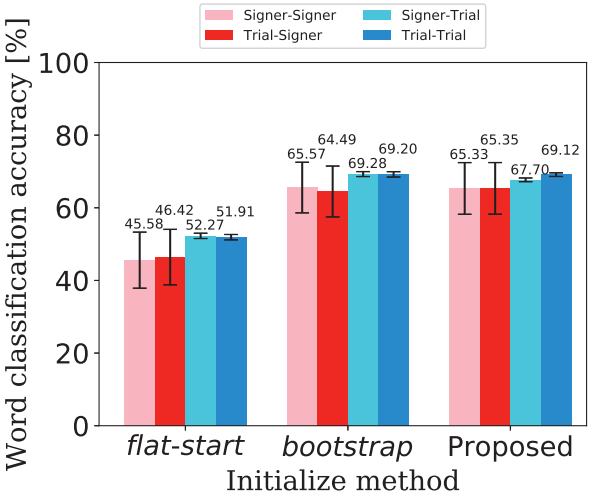


Fig. 8 Performance of the adjusted models

the evaluation of the isolated words, nineteen-fold and four-fold cross-validations were conducted based on signers and trials, respectively. We repeated the sampling of initial models and 'embedded training' for all initial models, and report four types of macro averages.

The summarized performance is illustrated in **Fig.8**. We denote the types of macro averages as Initial training condition - 'embedded training' condition in Fig.8. For example, Signer-Trial in Fig.8 denotes that the signer-opened and trial-opened conditions were applied to initial and 'embedded training', respectively. The pink, red, cyan, and blue bars in Fig.8 represent the word classification accuracy of Signer-Signer, Trial-Signer, Signer-Trial, and Trial-Trial, respectively.

The performance of the proposed method was superior to that of 'flat-start' initialization and comparable to that of 'bootstrap' initialization in every case, as shown in Fig.8.

6. Conclusions and Future Work

In this paper, we proposed weakly-supervised learning for continuous sign language word recognition using forced alignment based on DTW and isolated word HMM adjustment using 'embedded training'. As shown in the evaluation, the proposed forced alignment has achieved significant improvement in the alignment performance from that of 'flat-start' initialization. Moreover, we have found that the proposed weakly-supervised learning performed comparably to 'bootstrap' initialization with only one manual annotation for each isolated sign language word and the ordered word labels for the continuous words. Although the proposed method requires at least one manual annotation for each isolated word compared to 'flat-start' initialization, the performance improvement of the proposed method has been worth considering as a practical solution.

In spite of the improvements of the proposed method, there are still difficulties in contending with the variety of motions as indicated in the maximum and SD of the alignment distances in Table 3. The proposed forced alignment supposes that the target data are similar to the reference data. It is difficult to provide appropriate annotations when the difference between the target and the reference data is significant. Therefore, such a case should be annotated manually from the viewpoint of building corpora. Outlier detection of the time sequences may be available to exclude the expected failure cases and improve the overall efficiency of building the corpora. Moreover, some extensions of DTW[16),17)] can improve the stability of time sequence alignment.

While word-level annotation is a feasible solution to build corpora, detailed components of sign language, for example, "Transition" motions, sub-unit-level motions, and asynchronous motions of body parts, should be considered to establish high-level sign language recognition. The proposed method has avoided the annotation of "Transition" using 'embedded training'. This approach has limitations for large-scale continuous words because the many types of "Transitions" are difficult to learn as part of the word HMMs. The combination of the automatic generation of "Transition" models[2)] and soft boundary assignments[18)] are expected to lessen this limitation. The proposed method is independent of the fineness of motion units. Therefore, the proposed method is available for sub-unit-level annotation if manual annotation by experts is available. Moreover, DTW and 'flat-start' initialization can be combined to automatically generate sub-unit-level annotation when the adequate mappings between JSL words and motion units are established. For the asynchronous motions of body parts, extensions to asynchronous and multi-modal DTW[19)]and HMM[20)]may be available to our approach.

The current research settings have been relatively controlled. Weakly-supervised learning for sign language recognition in practice must be addressed.

References

1) U. von Agris, J. Zieren, U. Canzler, B. Bauer, K.F. Kraiss: "Recent Developments in Visual Sign Language Recognition", Universal Access in the Information Society, Vol.6, No.4, pp.323–362 (2008).
2) O. Koller, H. Ney, R. Bowden: "Deep Hand: How to Train

a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3793–3802 (2016).
3) O. Koller, S. Zargaran, H. Ney: "Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3416–3424 (2017).
4) N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden: "Neural Sign Language Translation", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.7784–7793 (2018).
5) H. Sakoe, S. Chiba: "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.26, No.1, pp.43–49 (1978).
6) K.-F. Lee, H.-W. Hon: "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", Proc. of the International Conference on Acoustics, Speech, and Signal Processing, pp.123–126 (1988).
7) Cambridge University Engineering Department, Hidden Markov Toolkit Version 3.4.1, http://htk.eng.cam.ac.uk (2018).
8) SignWriting Web Site, http://www.signwriting.org (2018).
9) A. Farhadi, D. Forsyth: "Aligning ASL for Statistical Translation Using a Discriminative Word Model", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.1471–1476 (2006).
10) P. Buehler, A. Zisserman, M. Everingham: "Learning Sign Language by Watching TV (Using Weakly Aligned Subtitles)", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2961–2968 (2009).
11) N. Takayama, H. Takahashi: "Sign Words Annotation Assistance Using Japanese Sign Language Words Recognition", Proc. of the International Conference on Cyberworlds, pp.221–228 (2018).
12) S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh: "Convolutional Pose Machines", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4724–4732 (2016).
13) T. Simon, H. Joo, I. Matthews, Y. Sheikh: "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1145–1153 (2017).
14) Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh: "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.7291–7299 (2017).
15) Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh: "OpenPose: Realtime Multi-person 2D :Pose Estimation Using Part Affinity Fields", arXiv preprint arXiv:1812.08008 (2018).
16) M. Cuturi, M. Blondel: "Soft-DTW: a Differentiable Loss Function for Time-Series", Proc of the 34th International Conference on Machine Learning, pp.894–903 (2017).
17) J. Zhao, L. Itti: "ShapeDTW: Shape Dynamic Time Warping", Pattern Recognition, Vol.74, pp.171–184 (2018).
18) L. Ding, C. Xu: "Weakly-supervised Action Segmentation With Iterative Soft Boundary Assignment", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6508–6516 (2018).
19) M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, G. Rigoll: "A Multidimensional Dynamic Time Warping Algorithm for Efficient Multimodal Fusion of Asynchronous Data Streams", Neurocomputing, Vol.73, No.1-3, pp.366–380 (2009).
20) S. Bengio: "An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition", Proc. of the Advances in Neural Information Processing Systems 15, pp.1237–1244 (2003).

(Received September 11, 2019)
(Revised November 20, 2019)

Natsuki TAKAYAMA (*Member*)
From 2008 to 2013, he was with SIElectronics, Ltd., Tokyo, Japan. He received the D.E. degree from the University of Electro-Communications, Tokyo, Japan, in 2017. Since 2017, he has been with the Graduate School of Informatics and Engineering, the University of Electro-Communications, Tokyo, Japan, where he is now a Researcher.

Hiroki TAKAHASHI (*Member*)
He received the D.E. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2005. Since 2006, he has been with the Graduate School of Informatics and Engineering, the University of Electro-Communications, Tokyo, Japan, where he is now an Associate Professor.

IIEEJ **Paper** *-- Special Issue on Journal Track Papers in IEVC2019 --*

A Model Ensemble Approach for Few-Shot Learning Using Aggregated Classifiers

Toshiki KIKUCHI† (*Student Member*) , Yuko OZASA†

† Keio University

<**Summary**> Despite the recent success in deep neural networks on the visual domain, we need a large amount of data to train the networks. Previous works addressed this issue as the few-shot learning which is the task to identify the class of an example in new classes not seen in a training phase with only a few examples of each new class. Some methods performed well on the few-shot tasks, but need a complex architecture and/or specialized loss functions, such as metric loss, meta learner, and memory. In this paper, we evaluate the performance of the ensemble approach aggregating a huge number of simple neural network models (up to 128 models) on standard few-shot datasets. Surprisingly, although the approach is simple, our experimental results show that the ensemble approach is competitive with state-of-the-art methods among similar architecture methods in some settings.

Keywords: few-shot classification, model ensemble, deep learning

1. Introduction

Deep neural network-based approaches outperform other conventional approaches on object recognition tasks[1),2)]. However, we need a massive amount of labeled data to train deep neural network-based models. Simply training the model with a small dataset leads to over-fitting. The problem of learning using only a few examples is called one- or few-shot classification[3),4)]. One- or few-shot classification is the task to predict the class of the given example by comparing a few examples of each possible class. Because the class of the given example is not seen in a training phase and only a few examples of each of these classes are given, the task is challenging for deep neural networks.

In the common setting for few-shot learning, we can use prior knowledge from a training set. The training set and the dataset for evaluating few-shot classification, of course, have completely disjoint label spaces. However, the characteristic of the examples in both datasets are similar because they have the same or a similar domain (e.g., handwritten alphabet images[5)] or RGB object images[6)]). In recent work on few-shot learning with deep neural networks, many works use the training set to optimize the network weights. One notable method is Matching Nets proposed by Vinyals et al.[6)]. They simulate the one-shot classification setting on a training dataset and apply the trained model for one- or few-shot classification on the evaluation set. Their method and similar methods[7),8)] perform well, but they use specialized neural network architectures and loss functions.

One naïve method, without using specialized architecture and meta-learning techniques[9)–11)], is merely to use the feature representation extracted with a standard classifier. A convolutional neural network (CNN)-based classifier trained for the classification task can be used as a feature extractor for many tasks in computer vision[12)–14)]. However, the feature extracted with a single classifier is not useful enough for few-shot classification on an evaluation set, because the training set and the evaluation set have disjoint label space. Then, our fundamental issue is how we obtain enough feature for few-shot classification using the training set.

To overcome the issue in a simple way, we here adopt a manner of traditional *ensemble learning* that is the concept of combining several models to make a final decision. Basically, even when we train a classifier on the same dataset, there are some different features enough for the task (optimal solutions of the model). Therefore, we hypothesize that we can obtain a better feature for few-shot classification on the evaluation set by finding and aggregating the different features for the classification task on the disjoint training set. Therefore, we employ many classifiers in an ensemble manner, which are trained for classification task on the training set, and aggregate the features extracted with each classifier. Each

classifier is expected to provide a different feature because of the randomness that comes from the random initialization of each network's weight and the random sampling of mini-batches, even if all the classifiers have the same architecture and are trained with the same training set. Each extracted feature is rich enough for classification in the label space of the training set but not enough for few-shot classification on the evaluation set with disjoint label space. However, by concatenating the features, we can obtain richer features for few-shot classification. We call this ensemble approach "Classifier Aggregation" (ClassAgg). As the classifiers are trained to solve standard classification tasks, we do not need specialized architecture and loss function. In addition, the models do not have memory like the methods by Kaiser et al.[15], Santoro et al.[16], and Ravi and Larochelle[10]. However, we surprisingly found that the simple ensemble approach is competitive with state-of-the-art methods.

In this paper, we propose the method to adopt an ensemble fashion for few-shot classification, and evaluate the effectiveness of the ensemble approach, ClassAgg. Surprisingly, experimental results show that the approach is competitive with state-of-the-art among similar architecture methods for 5-way 1-shot classification on Mini-ImageNet[6] and achieved the same accuracies as the state-of-the-art, with 95 % confidence interval overlap, for 5-way 5-shot and 20-way 5-shot classification on Omniglot Previous Split[5),6]. We also show the evaluation of how the number of aggregated classifiers contributes to the performance of few-shot classification.

2. Related Work

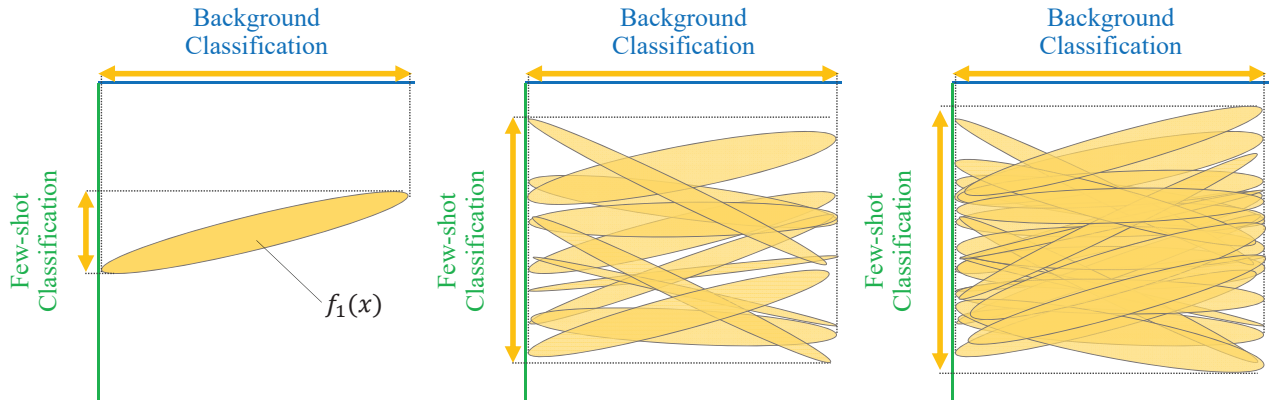
Many previous works are tackling few-shot classification[3),4)] which is a task to classify an example of the class not seen in training, with only a few examples for each class. Merely training the classifier with the few examples causes the model to overfit so that we cannot use this naïve approach.

Earlier approach for one-shot learning used the generative model[3),4)]. In recent work, Edwards and Storkey[17] used the variational autoencoder[18] to learn generative models for few-shot learning. The first approach for one-shot classification on the Omniglot dataset[5] uses specific knowledge for handwritten alphabet characters: pencil strokes. Their method, called Hierarchical Bayesian Program Learning, can perform well but mainly focuses on one-shot learning on only handwritten alphabet characters. Since the target domain of these methods is limited, we do not follow this approach.

Because deep neural network-based approaches outperform other conventional approaches in recent work on the object classification task[1),2)] with large-scale datasets, many recent works on few-shot learning are based on deep neural networks. Some works address the overfitting problem by simulating the few-shot classification with an additional dataset with disjoint labels and learn the metrics of the domain instead of training with a few examples. The first work following this approach is proposed by Koch[19]. Koch[19] used Deep Siamese Networks, which predict whether the classes of two input images are the same or not, for a one-shot classification task. Matching Nets[6] and Prototypical Nets[8] learn the embedding with CNN-based model by simulating the one-shot classification setting with a training set when they train the models. Sung et al.[7] also took a similar approach, and the model learns to compare the input images. Although these approaches demonstrated good performance for few-shot classification tasks, they need specialized neural network architectures and extraordinary loss functions. The ClassAgg approach instead does not simulate the few-shot classification in training phase so that we do not need them.

Another approach is employing a neural network with memory. Kaiser et al.[15] combined the Siamese Nets with long short-term memory (LSTM[20]). Santoro et al.[16] employed Neural Turing Machines[21] for one-shot learning. In the approach, they classify a given example using historical information in the memories. In contrast, the ClassAgg approach does not need complex architecture with memories like LSTM.

Finding the good initial condition of the network is an alternative approach proposed by Finn et al.[11]. They proposed the few-shot learning method for learning the initial condition of neural networks, as meta-information. The meta-learner provides a good initial condition to fine-tune with few-shot classification. MAML[11] performs well on few-shot classification, but the method needs fine-tuning. We do not refer the fine-tuning in the ensemble approach because of focusing on obtaining good representation for few-shot classification from the disjoint dataset. Meta Nets[9] and Meta-Learn LSTM[10] also focus on improving the optimization strategies. They use LSTM to learn the loss gradient of neural nets as meta-information for updating the parameters of the model for one-shot classification. These methods that use neural networks with memory have specialized architectures.



(a) Feature space with a single classifier (b) Feature space with some classifiers (c) Feature space with a large number of classifiers

Fig. 1 Concept of ClassAgg

In comparison to these previous approaches, since the ClassAgg approach employs simple CNN-based classifiers, our method does not need memory, fine-tuning or a meta-learner.

In terms of using multiple models, there is a traditional technique called 'ensemble learning' that is the concept of combining some models to improve the performance. In general classification tasks, ensemble learning combines classifiers for reducing the risk of misclassification by a poorly performing classifier[22]. In that setting, each classifier is trained for the same task as the final prediction task. The ClassAgg is a kind of ensemble learning, but in few-shot classification setting, the models are trained for the task different from the final prediction task. The ClassAgg aims to obtain richer feature representation with many classifiers.

3. Problem Setting

We consider the task of one- or few-shot learning. Following a task formulation by Vinyals et al.[6], we have three datasets: training set $\mathbf{B}$ (sometimes called the background set), support set $\mathbf{S}$, and test set $\mathbf{T}$. Support set $\mathbf{S}$ and testing set $\mathbf{T}$ have the same label space. In contrast, training set $\mathbf{B}$ has its own label space different from that of support set $\mathbf{S}$ and testing set $\mathbf{T}$. Therefore, the classes that we can see in a training phase are disjoint with those that we can see in an evaluation phase. Every example x in all datasets has its label y.

For performing one- or few-shot evaluation, we determine which example in $\mathbf{S}$ has the same class as the given test example $x^t \in \mathbf{T}$. We have to do this task only with prior information of examples in training set $\mathbf{B}$. If the support set consists of k-labeled examples for each of N different classes, the target few-shot problem is called N-way k-shot learning. In this case, we have $N \times k$ examples in support set $\mathbf{S} = \{x_i^s, y_i^s\}_{i=1}^{N \times k}$.

4. Classifier Aggregation

To bring an ensemble manner into few-shot classification tasks, we here use an aggregation of a large number of classifiers. Each classifier is trained to solve the classification task for the samples in training set $\mathbf{B}$. We call this task background classification. Although few-shot classification and background classification have disjoint label space, features useful enough for background classification can be useful for few-shot classification to some extent because these tasks use features extracted from similar domains (e.g., alphabet images and RGB object images). However, a feature extracted with a single classifier is not effective enough for few-shot classification because it is trained for the background classification task.

Figure 1 shows the feature space extracted from given image x which may have a label in training set $\mathbf{B}$, support set $\mathbf{S}$, or testing set $\mathbf{T}$. In Fig. 1, the vertical and horizontal axes indicate the effectiveness of the feature space for the few-shot classification and the background classification, respectively. Because the label spaces of these tasks are disjoint, the axes are orthogonal. Each yellow ellipse shows the possible feature space that the classifier can output. The ellipse can be projected to two basis vectors, the few-shot classification (vertical axis) and the background classification (horizontal axis) to measure the performance on each task. The range (length of the line segment) of the result of projecting the feature space to each axis represents the performance on the task related to the axis. If a broader range is covered, it means that the feature can be more useful for the task. Of course, the same feature has different performances on different tasks. Even if a feature is valuable for a task, the feature is not always beneficial for another task. Therefore, the

covered range of the classifier for each axis is different in Fig. 1.

If we train a single classifier for background classification as feature extractor f_1, the feature extractor provides a feature space as shown in Fig. 1(a). The feature is enough for background classification, but it covers a small range for the few-shot classification. Here, we suppose to train many classifiers that extract the different feature spaces and use these feature extractors. Figure 1(b) shows the feature space covered by a large number of classifiers. Of course, because all classifiers are trained for the background classification task, the range covered for the task does not change from Fig. 1(a). However, the aggregated feature covers a broader range for few-shot classification as long as each classifier learns a different feature space. Because of the randomness produced by the random sampling for mini-batches in the training phase and the random initialization of layer weights, each CNN-based classifier is expected to have a different optimal solution and learn a different feature space even if all the classifiers have the same network architecture. By aggregating them, we can obtain richer feature representation for few-shot classification than using a feature from a single classifier. Fig. 1(c) shows a covered feature space with a massive number of classifiers. As the number of classifiers increases, the coverage improvement for few-shot classification decreases. Because each classifier is trained for the same task (background classification), the probability of having a feature space similar to one of the feature spaces of previous classifiers increases with a more significant number of classifiers.

We define aggregated feature $f(x)$ as the concatenation of features extracted from each classifier:

$$f(x) = [f_1(x), f_2(x), ..., f_C(x)], \tag{1}$$

where $f_i(x)$ is the output of the penultimate layer of i^{th} classifier with given input x, and C is the number of classifiers. With the aggregated feature $f(x)$, we measure similarities between samples. We call this ensemble approach ClassAgg.

We use aggregated feature $f(x)$, which is extracted from a given image x, for the few-shot classification problem. The summary of the algorithm for few-shot classification with ClassAgg is outlined in **Algorithm 1**. In this algorithm, we predict which example in support set $\mathbf{S}$ has the same class as the test example x^t with feature extractors $\{f_c(x)\}_{c=1}^{c=C}$. For each support example index i, we compute the cosine similarity $sim[i]$ between aggregated features f^t and f^s which is extracted from x^t and a support example x_i^s, respectively, to find the $pred_i$ which maximize $sim[pred_i]$. i_{pred} is the index for the predicted label so the final predicted label is $y^s_{i_{pred}}$. In short, using the cosine similarities of the aggregated features of a test sample $x^t \in \mathbf{T}$ and support samples $\mathbf{S} = \{x_i^s, y_i^s\}_{i=1}^{N\times k}$, we perform N-way k-shot classification.

Algorithm 1 Few-shot classification with ClassAgg

Input: Test example $x^t \in \mathbf{T}$
  Support set $\mathbf{S} = \{x_i^s, y_i^s\}_{i=1}^{N\times k}$
  Feature extraction functions $\{f_c(x)\}_{c=1}^{c=C}$
Output: A predicted class label $y^s_{i_{pred}}$ for given test example x^t
  // Compute aggregated feature for test example
  $f^t \leftarrow [f_1(x^t), f_2(x^t), ..., f_C(x^t)]$
  // Compute similarities for all support examples
  for $i = 1, 2, \ldots, N \times k$
    $f^s \leftarrow [f_1(x_i^s), f_2(x_i^s), ..., f_C(x_i^s)]$
    $sim[i] \leftarrow$ cosine similarity between f^s and f^t
  end
  $i_{pred} \leftarrow \arg\max_i sim$
  return $y^s_{i_{pred}}$

As ClassAgg uses a standard classification model, we do not need to use any complex architecture, such as memory-augmented neural networks or meta-learning in reinforcement learning. Besides, we do not need to use a specialized objective function because each classifier is trained to solve the classification problem in a standard way: optimizing the network weights with cross-entropy loss function[23]. Because each classifier does not depend on other classifiers, we can train all classifiers in parallel to reduce computational time.

5. Experiments

To evaluate the performance of ClassAgg, we conducted experiments on three groups of datasets: Omniglot Previous Split[6], Omniglot Standard Split[5], and Mini-ImageNet[6]. For each dataset, we performed one- and few-shot classifications for 1000 episodes. Each episode has 10 examples in the test set $\mathbf{T}$, so 10000 trials are conducted for each dataset. With the result of the trial, we report the accuracy with 95 % confidence intervals.

5.1 Datasets

5.1.1 Omniglot Previous Split

Omniglot[5] is the dataset that contains 1623 characters from 50 different alphabets. Because every class has only 20 samples, we augmented the training set with random rotations and shifting. Consequently, we have 200 samples per class. Following previous few-shot learning

work[6),9)], all images are resized to 28×28 to reduce the computational cost.

Following Vinyals et al.[6)], we split the dataset into 1200 and 423 classes for training set **B** and evaluation set **E**, respectively. For each episode in the evaluation phase, we randomly sampled some examples from evaluation set **E** to generate support set **S** and testing set **T**.

For the dataset, we performed 5-way 1-shot, 5-way 5-shot, 20-way 1-shot, and 20-way 5-shot classification.

5.1.2 Omniglot Standard Split

Following the original condition of one-shot learning in the Omniglot dataset provided by Lake et al.[5)], we also conducted experiments on the Standard Split in the same way as the Previous Split, described in Section 5.1.1. The Standard Split provides 30 training (background) alphabets with 964 classes and 20 evaluation alphabets with 659 classes. The number of training classes in this setup is fewer than that of in the Previous Split setup. Therefore, this is a more difficult setup for one-shot learning.

Following Munkhdalai and Yu[9)], for the dataset, we performed 5-way 1-shot, 10-way 1-shot, 15-way 1-shot, and 20-way 1-shot classification.

5.1.3 Mini-ImageNet

Mini-ImageNet proposed by Vinyals et al.[6)] is the dataset that contains 100 classes from ImageNet[24)]. Because each class has 600 examples, the dataset consists of 60000 color images. We follow the class split proposed by Ravi and Larochelle[10)]. With the split, we obtain three groups: training with 64 classes, validation with 16 classes, and testing with 20 classes. Because we train the networks to solve classification in training set **B**, we do not use the validation set which has different class space with the training set. Following few-shot learning work[6),9)], all images are resized to 84×84 to reduce the computational cost.

Following Ravi and Larochelle[10)] and Munkhdalai and Yu[9)], for the dataset, we performed 5-way 1-shot and 5-way 5-shot classification.

5.2 Training details

We use a standard architecture for few-shot learning (e.g., References 6) and 10)). The model is a stack of four modules. The module consists of a 3×3 convolutional layer with 64 filters, a ReLU activation, batch normalization[25)], and a 2×2 maximum pooling layer. For Mini-ImageNet, the network is trained with dropout[26)] to avoid over-fitting at a dropout rate $p = 0.1$ and $p = 0.25$ after every convolutional layer and before the penultimate layer, respectively.

We trained 128 classifiers for each Omniglot setup and Mini-ImageNet. Therefore, we evaluated the performance of ClassAgg with $C = 1$ to 128 for each dataset. We train classifiers that have the same standard architecture. Each classifier provides the output of the penultimate layer as the feature vector $f_i(x)$ for the given x. The dimension of feature dimension $f_i(x)$ is 64.

Using training set **B** on each dataset, we train the classifiers to solve the 1200-way ($M = 1200$), 964-way ($M = 964$), and 64-way ($M = 64$) classification tasks for Omniglot Previous Split, Omniglot Standard Split, and Mini-ImageNet, respectively.

Before starting to train, every weight of the layers of classifiers was randomly initialized with Glorot uniform initialization[27)]. For optimizing the network, we use Adam[28)]. Because of the limitation of the memory, we used randomly sampled mini-batches containing 1024 samples and 512 samples for Omniglot and Mini-ImageNet, respectively.

To avoid over-fitting, we used 80 % of the training set **B** for training and 20 % for validation. Using validation-based early stopping, we stopped training a classifier when the metric for validation did not improve for more than P epochs. We used validation loss with $P = 2$ and validation accuracy with $P = 10$ for Omniglot-based datasets and Mini-ImageNet, respectively.

5.3 Results

For each dataset, we show the few-shot classification accuracies with various state-of-the-art baseline methods. For Omniglot Previous Split, the baseline methods include Siamese Nets[19)], MANN[16)], Matching Nets[6)], Siamese Nets with Memory[15)], Neural Statistician[17)], Meta Nets[9)], Prototypical Nets[8)], MAML[11)], and Relation Net[7)]. For Omniglot Standard Split, the baseline methods include Pixel kNN, Affine model, Deep Boltzmann Machines reported by[29)], Hierarchical Bayesian Program Learning[5)], Siamese Nets[19)], and Meta Nets[9)]. For Mini-ImageNet, Matching Nets[6)], Meta Nets[9)], Meta-Learn LSTM[10)], Prototypical Nets[8)], and Relation Net (Naive)[7)]. We also show the accuracy results for the methods using deeper architecture as the reference[7),30)].

Also, we show the relationship between the number of classifiers and the accuracy to demonstrate that aggregating a large number of classifiers is suitable for few-shot classification. It also shows the limitation of the ClassAgg approach: performance improvement decreases as

Table 1 Result of few-shot classification on Omniglot Previous Split ('-': not reported)

Method	Fine Tune	5-way Acc. (%)		20-way Acc. (%)	
		1-shot	5-shot	1-shot	5-shot
Convolutional Siamese Nets[19]	N	96.7	98.4	88.0	96.5
Convolutional Siamese Nets[19]	Y	97.3	98.4	88.1	97.0
MANN[16]	N	82.8	94.9	-	-
Matching Nets[6]	N	98.1	98.9	93.8	98.5
Matching Nets[6]	Y	97.9	98.7	93.5	98.7
Siamese Nets with Memory[15]	N	98.4	99.6	95.0	98.6
Neural Statistician[17]	N	98.1	99.5	93.2	98.1
Meta Nets[9]	N	99.0	-	97.0	-
Prototypical Nets[8]	N	98.8	99.7	96.0	98.9
MAML[11]	Y	98.7 ± 0.4	$\mathbf{99.9 \pm 0.1}$	95.8 ± 0.3	$\mathbf{98.9 \pm 0.2}$
Relation Net[7]	N	$\mathbf{99.6 \pm 0.2}$	$\mathbf{99.8 \pm 0.1}$	$\mathbf{97.6 \pm 0.2}$	$\mathbf{99.1 \pm 0.1}$
ClassAgg-1 ($C = 1$)	N	94.8 ± 0.5	98.8 ± 0.2	86.2 ± 0.4	95.9 ± 0.2
ClassAgg-64 ($C = 64$)	N	98.7 ± 0.3	$\mathbf{99.7 \pm 0.1}$	95.8 ± 0.2	$\mathbf{98.9 \pm 0.1}$
ClassAgg-128 ($C = 128$)	N	98.6 ± 0.3	$\mathbf{99.7 \pm 0.1}$	95.8 ± 0.2	$\mathbf{98.9 \pm 0.1}$

the number of classifiers increases.

5.3.1 Omniglot Previous Split

The results of few-shot classification on Omniglot Previous Split are shown in **Table 1**. We also show the result for ClassAgg-1, a method that uses a feature extracted with a single classifier, to demonstrate the performance of the ensemble approach. ClassAgg-128, which is the aggregation of 128 classifiers, achieved state-of-the-art level accuracy with 95 % confidence intervals for 5-way 5-shot and 20-way 5-shot. With a large number of classes in training set **B**, ClassAgg can perform as well as other state-of-the-art methods.

5.3.2 Omniglot Standard Split

We show the results for one-shot classification on Omniglot Standard Split in **Table 2**. Although ClassAgg-128 could not outperform the state-of-the-art method, we can see the effectiveness of the ClassAgg by comparing the results for ClassAgg-1 and ClassAgg-128. Whereas in 1-shot 5-way in Previous Split setting, ClassAgg achieved the same accuracy as Meta Nets[9] with confidence interval overlap, ClassAgg could not outperform Meta Nets in the Standard Split setting which has fewer classes in training set **B**. In the ClassAgg approach, since we use the classifiers, the performance of few-shot classification relies on the background set. Therefore, the ClassAgg approach can perform better when we can train with more background classes.

5.3.3 Mini-ImageNet

The results for few-shot classification on Mini-ImageNet are shown in **Table 3**. For 5-way 1-shot setting, ClassAgg-64 and ClassAgg-128 outperformed previous methods which follow the standard architecture proposed by Reference 6). As TCML[30] and Relation Net (Deeper)[7] are based on the deeper network than the standard architecture, we cannot compare the performance to these methods, similar to Grant et al.[31].

5.3.4 The number of classifiers

We show the relationship between the number of classifiers (horizontal axis) and the few-shot classification accuracy (vertical axis) for each dataset in **Fig. 2**. As the number of classifiers increase, the performances of the few-shot classification improve with some improvement decay. Obviously, the accuracies of the harder setting are lower from $C = 1$ to $C = 128$. For all datasets and all tasks, the average accuracy improves drastically as the number of classifiers increases from $C = 1$ to around 20. The first improvement between $C = 1$ and $C = 2$ is the highest one in all settings. As explained in Section 4., the accuracy improvement by adding one classifier decreases as the number of classifiers increases. Because all classifiers are trained for the same background classification, the possible feature space that the classifier extract is limited. Due to this, a classifier tends to have a feature space similar to one of the feature spaces of previous classifiers with a more significant number of classifiers.

5.3.5 Performance variance of single model

When using a single classifier for few-shot classification (in the case of $C = 1$), which classifier we use may cause divergent results because of the randomness of initialization of each network's weight and the random sampling of mini-batches. However, in the case of ClassAgg-1 ($C = 1$) on our experiments, we just used the first classifier we

Table 2 Result of one-shot classification on Omniglot Standard Split ('-': not reported)

Method	5-way (%)	10-way (%)	15-way (%)	20-way (%)
Human performance[5]	-	-	-	95.5
Pixel kNN[29]	-	-	-	21.7
Affine model[29]	-	-	-	81.8
Deep Boltzmann Machines[29]	-	-	-	62.0
Hierarchical Bayesian Program Learning[5]	-	-	-	**96.7**
Siamese Nets[19]	-	-	-	92.0
Meta Nets[9]	**98.45**	**97.32**	**96.4**	95.92
ClassAgg-1 ($C = 1$)	92.5 ± 0.6	87.0 ± 0.5	83.5 ± 0.5	80.8 ± 0.4
ClassAgg-64 ($C = 64$)	97.9 ± 0.3	96.4 ± 0.3	95.0 ± 0.3	94.0 ± 0.3
ClassAgg-128 ($C = 128$)	97.8 ± 0.3	96.5 ± 0.3	95.1 ± 0.3	94.1 ± 0.3

Table 3 Result of few-shot classification on Mini-ImageNet ('-': not reported)

Method	Base Architecture	Fine Tune	5-way Acc. (%)	
			1-shot	5-shot
Matching Nets[6]	Standard[6]	N	43.56 ± 0.84	55.31 ± 0.73
Meta Nets[9]		N	49.21 ± 0.96	-
Meta-Learn LSTM[10]		N	43.44 ± 0.77	60.60 ± 0.71
MAML[11]		Y	48.70 ± 1.84	63.11 ± 0.92
Prototypical Nets[8]		N	49.42 ± 0.78	**68.20 ± 0.66**
Relation Net (Naive)[7]		N	51.38 ± 0.82	**67.07 ± 0.69**
ClassAgg-1 ($C = 1$)		N	47.98 ± 0.99	58.52 ± 1.01
ClassAgg-32 ($C = 32$)		N	**54.43 ± 1.00**	64.89 ± 0.99
ClassAgg-64 ($C = 64$)		N	**54.55 ± 1.01**	65.38 ± 0.98
ClassAgg-128 ($C = 128$)		N	**54.82 ± 1.01**	65.68 ± 0.97
TCML[30]	Deeper	N	55.71 ± 0.99	68.88 ± 0.92
Relation Net (Deeper)[7]		N	57.02 ± 0.92	71.07 ± 0.69

trained as the feature extractor. Therefore, we evaluated the variance of the performance with the single classifier caused by the randomness.

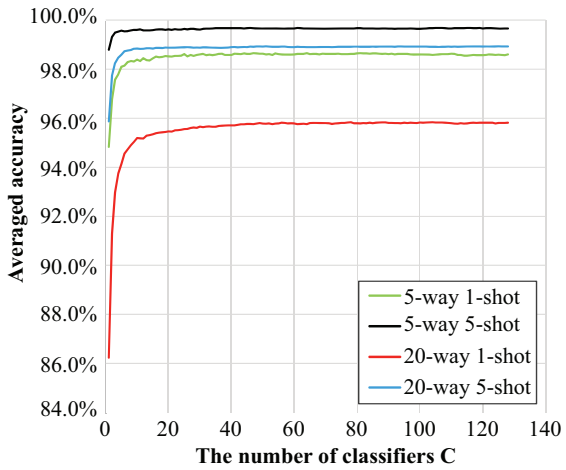
Table 4 show the results of few-shot classification with a different single classifier on Omniglot Previous Split, Omniglot Standard Split, and Mini-ImageNet, respectively. For each setting, these tables show the lowest accuracy (as Min.), the highest accuracy (as Max.), the difference between them (as Diff.), and the variance of the accuracy (as Var.) for each classifier. In this experiment, the used classifier is selected through the classifiers pool which consists of classifiers used for ClassAgg-128 ($C = 128$). Consequently, we conducted the few-shot classification evaluation for 128 classifiers.

As a result, for all settings, the performance on the harder setting has the more significant difference between the minimum and maximum performances and the higher variance. From these results, it appears that it is difficult to obtain richer representation required by the harder setting with the single classifier. Besides, even the highest accuracies with the single classifier are beaten by our results with aggregated classifiers, shown in Table 1, 2 and 3. This means that the ClassAgg approach works effectively to obtain rich representation.

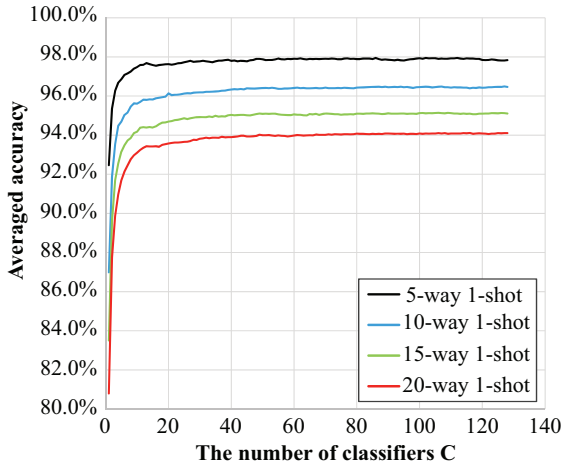
6. Discussion and Future Work

We propose the method to adopt an ensemble manner for few-shot classification, and evaluate the ensemble approach, ClassAgg. Just aggregating the trained classifiers achieved high accuracies for some tasks of one-shot classification. Because the ensemble approach is simple, we do not need to modify the architecture of a standard classification model, and we do not have to use specialized loss function to train networks.

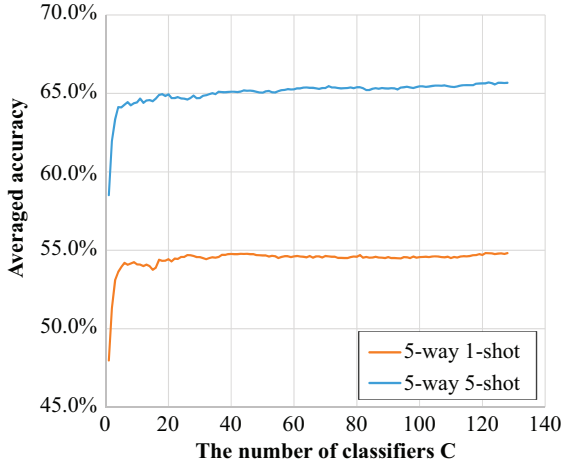
Each classifier is trained for the classification task of the training set **B**, which has a label space disjoint from support set **S** and test set **T**. Because of the randomness of the training (e.g., mini-batch sampling or random weight initialization), every classifier is expected to provide dif-



(a) Omniglot Previous Split



(b) Omniglot Standard Split



(c) Mini-ImageNet

Fig. 2 Effect of the number of classifiers

ferent features extracted from the same image although the label spaces are disjoint.

Our experimental results show that the one- and few-shot classification accuracy improves as the number of classifiers increases, as long as each classifier extracts a different feature. The results also show that it gets more difficult to extract a different feature, which is not similar to features extracted from previous classifiers, as the number of classifiers increases. Besides, the ensemble approach can not perform better when the number of the background classes is fewer because we employ classifiers for the background set. Therefore, we can conclude that our approach can work well if the data of the training set is diverse enough to train a classifier which can extract useful feature from disjoint evaluation set. Although we focused only on images, as long as the classifier can get well-trained for evaluation set, our approach may be extended to few-shot learning on the data from other modalities (e.g., audio and natural language).

Table 4 Result of few-shot classification using a single classifier

(a) Omniglot Previous Split

Setting	Min. (%)	Max. (%)	Diff. (%)	Var. (%)
5-way 1-shot	93.540	95.580	2.040	0.001
5-way 5-shot	98.240	99.070	0.830	0.000
20-way 1-shot	84.348	87.345	2.998	0.003
20-way 5-shot	94.858	96.160	1.303	0.001

(b) Omniglot Standard Split

Setting	Min. (%)	Max. (%)	Diff. (%)	Var. (%)
5-way 1-shot	91.290	93.430	2.140	0.002
10-way 1-shot	85.420	88.850	3.430	0.004
15-way 1-shot	81.667	85.617	3.950	0.005
20-way 1-shot	79.000	83.145	4.145	0.006

(c) Mini-ImageNet

Setting	Min. (%)	Max. (%)	Diff. (%)	Var. (%)
5-way 5-shot	56.330	63.040	6.710	0.018
5-way 1-shot	46.610	53.570	6.960	0.021

Another downside of the ensemble approach is higher computational cost in a prediction phase. We cannot merely compare with methods that perform without neural networks, but comparing other neural network-based methods, our approach has a higher computational cost depending on the number of classifiers we employed because we need to run feature extractor C times. While, in terms of computational *time*, since the computation of our ensemble approach is straightforward to be parallelized, the computational *time* cannot be a severe issue.

Our future work includes exploring a method that forces the classifier to learn different features so that we can achieve high performance with fewer models.

References

1) K. He, X. Zhang, S. Ren, J. Sun: "Deep Residual Learning for Image Recognition", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.770–778 (2016).
2) A. Krizhevsky, I. Sutskever, G. E. Hinton: "ImageNet Classification with Deep Convolutional Neural Networks", Communications of the ACM, Vol.60, No.6, pp.84–90 (2017).

3) L. Fei-Fei, R. Fergus, P. Perona: "One-Shot Learning of Object Categories", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.28, No.4, pp.594–611 (2006).
4) B. M. Lake, R. Salakhutdinov, J. Gross, J. B. Tenenbaum: "One Shot Learning of Simple Visual Concepts", Proc. of the 33rd Annual Conference of the Cognitive Science Society, pp.2568–2573 (2011).
5) B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum: "Human-Level Concept Learning Through Probabilistic Program Induction", Science, Vol.350, No.6266, pp.1332–1338 (2015).
6) O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra: "Matching Networks for One Shot Learning", Proc. of Advances in Neural Information Processing Systems, pp.3630–3638 (2016).
7) F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, T. M. Hospedales: "Learning to Compare: Relation Network for Few-Shot Learning", Proc. the IEEE Conference on Computer Vision and Pattern Recognition, pp.1199–1208 (2018).
8) J. Snell, K. Swersky, R. S. Zemel: "Prototypical Networks for Few-shot Learning", Proc. of Advances in Neural Information Processing Systems, pp.4077–4087 (2017).
9) T. Munkhdalai, H. Yu: "Meta Networks", Proc. of the International Conference on Machine Learning, pp.2554–2563 (2017).
10) S. Ravi, H. Larochelle: "Optimization as a Model for Few-Shot Learning", Proc. of the International Conference on Learning Representations (2017).
11) C. Finn, P. Abbeel, S. Levine: "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks", Proc. of the International Conference on Machine Learning, pp.1126–1135 (2017).
12) X. Wang, L. Lu, H. Shin, L. Kim, M. Bagheri, I. Nogues, J. Yao, R. M. Summers: "Unsupervised Joint Mining of Deep Features and Image Labels for Large-Scale Radiology Image Categorization and Scene Recognition", Proc. of the IEEE Winter Conference on Applications of Computer Vision, pp.998–1007 (2017).
13) A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson: "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.512–519 (2014).
14) J. Y. H. Ng, F. Yang, L. S. Davis: "Exploiting Local Features from Deep Networks for Image Retrieval", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.53–61 (2015).
15) Ł. Kaiser, O. Nachum, A. Roy, S. Bengio: "Learning to Remember Rare Events", Proc. of the International Conference on Learning Representations (2017).
16) A. Santoro, S. Bartunov, M. M. Botvinick, D. Wierstra, T. P. Lillicrap: "Meta-Learning with Memory-Augmented Neural Networks", Proc. of the International Conference on Machine Learning, pp.1126–1135 (2016).
17) H. Edwards, A. Storkey: "Towards a Neural Statistician", Proc. of the International Conference on Learning Representations (2017).
18) D. P. Kingma, M. Welling: "Auto-Encoding Variational Bayes", Proc. of the International Conference on Learning Representations (2013).
19) G. R. Koch: "Siamese Neural Networks for One-shot Image Recognition", Proc. of the ICML Deep Learning Workshop, Vol. 2 (2015).
20) S. Hochreiter, J. Schmidhuber: "Long Short-Term Memory", Neural Computation, Vol.9, No.8, pp.1735–1780 (1997).
21) A. Graves, G. Wayne, I. Danihelka: "Neural Turing Machines", The Computing Research Repository (CoRR), abs/1410.5401 (2014).
22) R. Polikar: "Ensemble Based Systems in Decision Making", IEEE Circuits and Systems Magazine, Vol.6, No.3, pp.21–45 (2006).
23) I. Goodfellow, Y. Bengio, A. Courville: "Deep Learning", MIT Press (2016).
24) J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei: "ImageNet: A Large-Scale Hierarchical Image Database", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.248–255 (2009).
25) S. Ioffe, C. Szegedy: "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", Proc. of the International Conference on Machine Learning, pp.448–456 (2015).
26) N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov: "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research, Vol.15, No.1, pp.1929–1958 (2014).
27) X. Glorot, Y. Bengio: "Understanding the Difficulty of Training Deep Feedforward Neural Networks", Proc. of the International Conference on Artificial Intelligence and Statistics, pp.249–256 (2010).
28) D. P. Kingma, J. Ba: "Adam: A Method for Stochastic Optimization", Proc. of the International Conference on Learning Representations (2015).
29) B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum: "One-Shot Learning by Inverting a Compositional Causal Process", Proc. of Advances in Neural Information Processing Systems, pp.2526–2534 (2013).
30) N. Mishra, M. Rohaninejad, X. Chen, P. Abbeel: "Meta-Learning with Temporal Convolutions", The Computing Research Repository (CoRR), abs/1707.03141 (2017).
31) E. Grant, C. Finn, S. Levine, T. Darrell, T. Griffiths: "Recasting Gradient-Based Meta-Learning as Hierarchical Bayes", Proc. of the International Conference on Learning Representations (2018).

(Received August 30, 2019)
(Revised November 19, 2019)

Toshiki KIKUCHI (*Student Member*)
He received his B.Eng. degree in Information and Computer Science from Keio University, Japan, in 2018. He is currently a MS student in Machine Learning and Computer Vision at Keio University. His research interests include machine learning, audio-visual processing, and computer graphics.

Yuko OZASA
She received her PhD degree in engineering from Kobe University in 2015. She was a postdoctoral researcher at National Institute of Advanced Industrial Science and Technology (AIST) in 2015. Since 2015 she has been a research associate at Graduate School of Science and Technology, Keio University. Her research interests include object recognition and grounding, multimodal fusion, visual perception, and hyperspectral sensing.

IIEEJ **Paper** *-- Special Issue on Journal Track Papers in IEVC2019 --*

Visual Simulation of Tearing Papers Taking Anisotropic Fiber Structure into Account

Saeko SHINOZAKI† , Masanori NAKAYAMA†(*Member*) , Issei FUJISHIRO†(*Honorary Member*)

†Keio University

<**Summary**> Real paper gets deformed anisotropically due to a uni-directional placement of the inner fibers. Few existing CG researchers allow for such fibers to represent virtual papers, and thus anisotropic paper deformation has not been well represented. In this study, we present a two-dimensional visual simulation model for anisotropic papers, which abstracts mutual mechanical relationships among intersecting fibers by a network of filler and hinge springs, and incorporates connection points for keeping the shape of each bended fiber. By releasing the network connections when the filler and hinge spring extend above a certain limit, the paper provides a plausible tear. We succeeded in generating a different appearance of the torn-off line of papers according as the pulling direction.

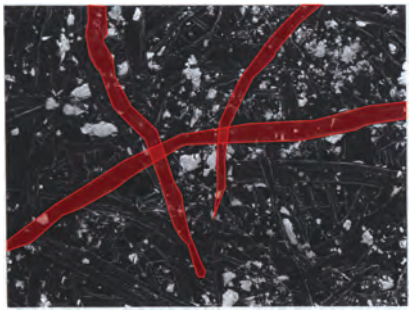
Keywords: visual simulation, microstructure, paper, tearing

1. Introduction

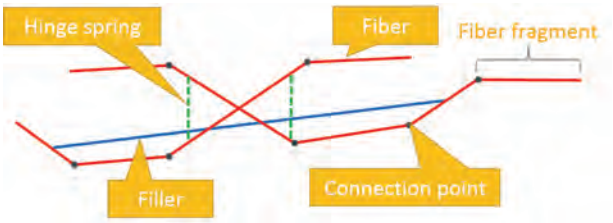
In the paper-forming process, the pulp, which has been dissolved in water, is ejected from the paper machine at a high speed. Even after dewatering and drying, the piece of paper retains its fiber direction. Thereby the deformation effect appears to differ depending on the direction of the external force applied. This direction, called 'grain', is essential for printing, decorating, wrapping, and bookbinding. For example, a newspaper is printed in such a way as not to fold in a vertical direction when it is open, and a book is printed in such a fashion so that its readers might turn its pages easily. The grain also influences the result of the tearing of the paper and, indeed, we often encounter the case where the tear tends to be irregular when the piece of paper is pulled in parallel with the fiber direction, or to be straight when pulled in a direction orthogonally to it. The fiber length of softwood is approximately 50 µm and that of hardwood approximately 20 µm, which is too thin for humans to discern. Fibers with such a minuteness running off from the tear cause the blurry outlines to be seen from a macro perspective.

There exist many CG studies on paper tearing. Several recent articles deal with interactive paper tearing, as seen in Schreck et al.[1] and Lejemble et al.[2], where they rely only on mesh-based paper models in order to reduce the cost of computation. Busaryev et al.[3] and Pfaff et al.[4] also apply mesh-based paper modeling schemes to simulate paper tearing, and attempt to balance the computational cost and the reasonable minuteness using adaptive remeshing. Directing our attention to the research on the simulation of a cloth, which is also comprised of fibers, Zhao et al.[5] use a model allowing for the microstructure, and thus improve the appearance of the entire object. However, any cost-effective approach to the simulated tearing of paper with fiber-level details has not been focused. For this reason, the simulated tear cannot represent the difference according to the direction of the fiber nor the elastic force caused by fiber connection. In addition, the minuteness obtained by the mesh-based modeling[1)–4)] is not so high that it can show the minute fibers running off the edge of the paper and the blurred tear specific to a paper.

In this research, therefore, we attempt to model a piece of paper with microstructure and tear the piece so as to generate realistic outlines reflecting its own feature dynamics. First, directed fibers are arranged on a two-dimensional plane, and neighboring fibers are mutually connected to form a sheet of fibers (simply referred to as 'paper fiber sheet' hereafter). Then, the deformation of the paper fiber sheet is simulated by an abstract model that incorporates into a single framework: filler filling the sheet; hinge springs repulsing against the fiber rotation; and connection points keeping the shape of bended fibers. Finally, the sheet is torn by considering the dynamic action caused by applying a given pulling force to it.



(a) An electron microscope photo of real paper: red area represents fiber, while white part filler



(b) Proposed model of paper fiber sheet

Fig. 1 Structure of real paper and proposed model

The remainder of this paper is organized as follows. The next section introduces prior work related to paper deformation and modeling. Section 3 explains the structure of paper fiber sheet and our main algorithm of generating and tearing the sheet. Section 4 explains an acceleration method for the fiber intersection detection and an associated data structure for the parallel processing using a GPU. Section 5 shows results of our paper fiber sheet tearing model. Section 6 concludes this paper and refers to our future work.

2. Related Work

A paper consists of pulp extracted from trees and is formed by intertwining fibers. Metaaphanon et al.[6] simulate the fray of yarn using a coarse yarn-based fabric. Zhao et al.[5] scan the fiber structure of cloth using a CT scanner and then convert to the volume data to reproduce the cloth with fiber microstructure. Jiang et al.[7] simulate the deformation operation to the object with a fiber structure, taking into account its plasticity, elasticity, and friction using the Material Point Method[8].

Inspired by the idea proposed in Takagi et al.[9], we have built upon the way of modeling paper with microstructure to generate plausible tears. In fact, they model fibers and filler as basic paper components in a volumetric manner. They arrange fibers so as to share a primary direction and also consider the local fiber deformation effects, such as constrictions and fluffs, generated during the paper-forming process. In contrast, we model a paper sheet with fibers so as to express the fluff seen in the paper tear, and introduce hinge spring and connection point to express the internal force caused by the deformation.

Kitani et al.[10] define a three-dimensional paper object as a mass-spring cubic mesh. They express a directional fiber by grouping neighbor cells to make the corresponding biased mass-spring parallel-piped. Furthermore, they realize a repulsion force of a bended paper by enlarging the cell groups through mutual connection of neighboring groups. However, the spring in the mesh group disappears when the end mass points are separated. Consequently, the deformed shape cannot return to its original state due to the lack of repulsion. In contrast, our method can transmit the elastic force without the loss of the spring property because, if one connection is broken, other components connecting to the same fiber can retain the elastic force.

3. Paper Model

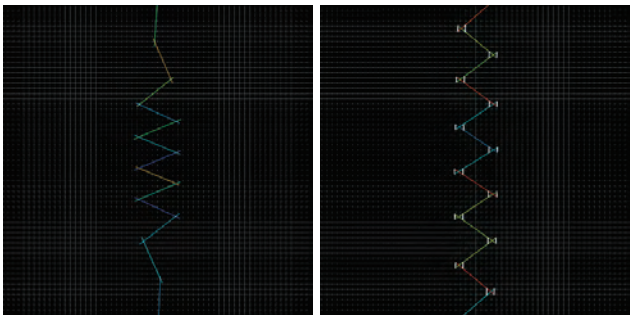
3.1 Structure of paper fiber sheet

In this study, based on observing the structure of real paper in **Fig. 1** (a), we model a paper fiber sheet by incorporating fibers and filler, and hinge springs, as shown in Fig. 1 (b). In addition, we connect a dozen 'fiber fragments' with connection points to express one bended fiber.

The length of each fiber fragment is set to 0.06 mm to 0.10 mm so that the length of fiber will be 1.0 mm to 1.5 mm, which is compatible with the average length of the pulp composing a real photocopy paper. Hereafter, the model for generating and tearing a paper fiber sheet is simply referred to as 'paper model'. Each of components is detailed below.

Filler is an inorganic pigment added for increasing the smoothness, whiteness, and printability of paper in the reality. In our method, filler is treated as the connector between fibers and has the spring property as with fiber fragments whereas it exhibits a different behavior. Filler is assumed to gain the bonding force between fibers and thus to increase or decrease the strength of the paper relative to its amount.

Hinge springs act to prevent the rotation of fiber fragments. By generating a hinge spring that is symmetri-



(a) Without hinge springs (b) With hinge springs

Fig. 2 Effect of hinge springs

cally centered around the intersecting point of fiber fragments, the restoring force from the elastic deformation of the hinge spring is transmitted to fiber fragments, and then the hinge spring restrains the fiber fragment rotation. **Figure 2** shows the effect of hinge springs when fiber fragments are alternately arranged. Figure 2 (a) demonstrates the result of pulling the upper and lower ends of the fiber fragment network connected with the single spring placed at the intersecting point in the vertical direction while Fig. 2 (b) the result with hinge springs. Without the hinge springs, the force is transmitted to the connecting fiber fragment after a fiber fragment starts to move. In contrast, hinge springs repulse the rotation of fiber fragment, so that the force is immediately transmitted to the neighboring fiber fragment. By repeating this, the force from the sheet edges can be evenly transmitted across the sheet.

In the paper model, a connection point connects ten to fifteen fiber fragments into one bended fiber. As in the case with filler and hinge springs, these points generate the repulsion according to the variation of their length and reproduce the action to keep the bended shape of fibers. The connection point itself does not break unlike filler and hinge springs because we do not take into account the break of fibers themselves.

3.2 Generation of paper fiber sheet

Our method has five steps for generating paper fiber sheet: determining the initial settings; generating fiber fragments; attaching fiber fragments to the corresponding cells; intersection detecting of fiber fragments and generating the inner components; and setting the area pulled by the external force.

First, the size of paper fiber sheet, the number of fibers, and the aspect ratio of the sheet are set to the user input.

Next, we generate the fiber fragments that vary in length, and form vertical and horizontal joints. The coordinate values of the middle point of the fiber fragment are set randomly in the generation area. The angle of the fiber fragment is determined by varying randomly within a range of ± 22.5 degrees from the 0 or 90 degrees depending on the direction of the fiber of the paper. The point vertex 0 at a distance of half of the fiber length at that angle from the midpoint is the starting point of the fiber fragment, and the point vertex 1 located on the opposite side is the ending point.

Then, ten to fifteen fiber fragments are rotated and connected into one bended fiber with reference to the appearance of real ones. Not all fiber fragments extend in the same direction but, if the fiber fragments are arranged vertically, they can extend randomly both in the upper and lower direction. At the same time, the information on how many fiber fragments are included in the fiber is attached to each fiber. There is a branch to the suspension of the extension caused by the fiber reaching the edge of the generating area when we form cuts, such as perforations.

After that, the intersections of these generated fibers are detected. In order to mediate between two intersecting fibers, a hinge spring is placed symmetrically with respect to the intersection point. At the same time, several intersecting fibers are mutually connected by filler. In addition, for each fiber the connection points are generated so as to connect the fiber fragments.

Finally, all vertices are scanned and if a vertex exists in the pulled area, then the vertex is added to the array. Note that the array is prepared separately on the pulling direction. For example, when the sheet is pulled in the Y-axis direction, the vertices with the Y coordinate value greater than a certain value are added to the array of the vertices pulled in a positive direction, while the vertices with less than the value are added to the array of pulled in a negative direction.

3.3 Repulsion computation

The tearing simulation is executed following **Algorithm 1**.

Each computation in the algorithm is executed in parallel by a GPU. Note that the iteration count $N = 200$ in the algorithm was set empirically to reproduce plausible tearing effects in our settings. If the simulation was not iterated, the fibers would tend to stretch almost infinitely. Hence, we decided to observe tearing results with a series of the iteration number from 100 to 500 by 100 increments. Consequently, the number 200 provided the

Algorithm 1 Tearing simulation

Move all vertices in the pulled area;
for N=1, ..., 200
  Initialize the force applied to all vertices
  Calculate the repulsion of fibers and add it to each vertex
  Calculate the repulsion of inner components and add it to each vertex of the connecting fibers
  Set a hiding flag when filler or a hinge spring extends above a certain limit
  Move all vertices according to the calculated force
end
Redraw the window

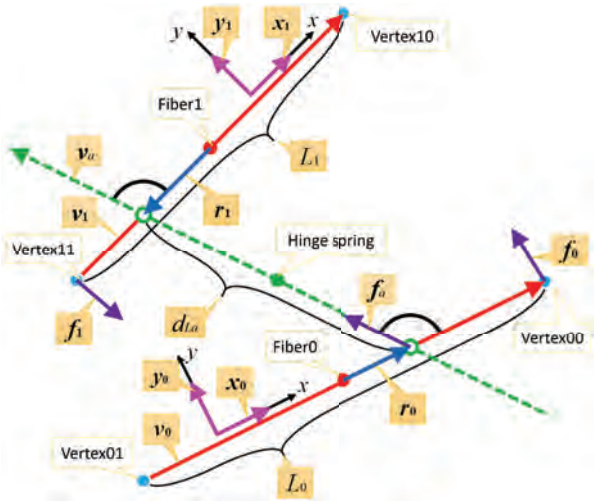


Fig. 3 Variables used for computing the repulsion caused by the components

best result in terms of the balance of the elongation and the shrinking of the fibers.

Sheet pulling is reproduced by moving the corresponding vertices and the repulsion is computed with each component, according to the moving width. The connections are released when filler or a hinge spring extends over the repulsion above the certain limit. The vanishment of torn filer and hinge springs is expressed by the suspension of rendering due to the restriction of the data structure. When the sheet is broken after repeating these processes, the tearing is deemed to have been completed.

The computation is different between fiber fragments and other components, as detailed in the following.

As an elastic body deforms, the force is generated to restore the body to its original shape. Since fiber fragments in our paper model have the spring property, the restoring force is generated according to the deformation on the both ends of the fibers.

The restoring force $\boldsymbol{f}$ caused by this property is given by multiplying the amount of the variation of their length d_L by the vector $\boldsymbol{v}$ of fiber fragment:

$$\boldsymbol{f} = 3.0 d_L \boldsymbol{v}. \tag{1}$$

Table 1 Variables and their meanings in Fig. 3

Variable	Description
d_{La}	The variation of length
$\boldsymbol{v_a}$	The unit vector
$\boldsymbol{f_a}$	The shrinking force
$\boldsymbol{v_0}, \boldsymbol{v_1}$	The vector of fiber 0 and 1
L_0, L_1	The length of fiber 0 and 1
$\boldsymbol{x_0}, \boldsymbol{y_0}$	The position vector for fiber 0 with respect to the X-axis direction and the Y-axis obtained by rotating the X-axis counterclockwise by 90 degrees on the plane on which the sheet is produced
$\boldsymbol{x_1}, \boldsymbol{y_1}$	The position vector for fiber 1 with respect to the X-axis direction and the Y-axis obtained by rotating the X-axis counterclockwise by 90 degrees on the plane on which the sheet is produced
$\boldsymbol{r_0}, \boldsymbol{r_1}$	The vector from the centroid of the fiber to the intersecting point of the components
$\boldsymbol{f_0}, \boldsymbol{f_1}$	The spinning force caused by the components pulling fibers

The forces $\boldsymbol{f_{p0}}$ and $\boldsymbol{f_{p1}}$ added to the vertex 0 as the starting point of the fiber vector and the vertex 1 as the ending point are calculated by the following equations:

$$\boldsymbol{f_{p0}} = +1.5\boldsymbol{f}, \tag{2}$$

$$\boldsymbol{f_{p1}} = -1.5\boldsymbol{f}. \tag{3}$$

Note that the coefficients in Formulae (1) to (3) are set to reproduce plausible tear with reference to the appearance of simulation results. They are the material constants of the fiber itself, and hence they will not change as long as the same fiber material is used.

Figure 3 shows the variables used for computing the repulsion caused by filler, hinge springs, and connection points. **Table 1** explains what each of the variables means. Note that the repulsion caused by filler, hinge spring, and connection point can be computed with the same formula described below by varying only the coefficients.

The shrinking force of the component $\boldsymbol{f_a}$ is given by:

$$\boldsymbol{f_a} = \alpha d_{La} \boldsymbol{v_a}, \tag{4}$$

where α is a constant to which a different value is assigned for each component and determines the strength of shrinking.

The moments caused by $\boldsymbol{f}$ are given by the outer product of $\boldsymbol{f_a}$ and $\boldsymbol{r_0}$, $\boldsymbol{r_1}$, as follows:

$$z_0 = \boldsymbol{r_0} \times \boldsymbol{f_a}, \tag{5}$$

$$z_1 = \boldsymbol{r_1} \times \boldsymbol{f_a}, \tag{6}$$

where that z_0 and z_1 are the scalar values in the direction of Z-axis in the $\boldsymbol{x_0}$, $\boldsymbol{y_0}$ and $\boldsymbol{x_1}$, $\boldsymbol{y_1}$ coordinate systems, respectively, and indicate the moments of the fiber pulling force by the inner components.

Thus, the forces of spinning fiber fragments $\boldsymbol{f_0}$ and $\boldsymbol{f_1}$ are given by:

$$\boldsymbol{f_0} = \frac{z_0}{L_0/2}\boldsymbol{y_0}, \tag{7}$$

$$\boldsymbol{f_1} = \frac{z_1}{L_1/2}\boldsymbol{y_1}. \tag{8}$$

The repulsion force is calculated by summing up $\boldsymbol{f_0}$, $\boldsymbol{f_1}$ and $\boldsymbol{f_a}$ on the vertices of fiber fragments. The forces of fiber fragments $\boldsymbol{f_{p00}}$, $\boldsymbol{f_{p01}}$, $\boldsymbol{f_{p10}}$, and $\boldsymbol{f_{p11}}$ which repulse the pulling by the connecting components added respectively to vertex 00, 01, 10, and 11 are obtained by taking into account the direction of the force as follows:

$$\boldsymbol{f_{p00}} = +\beta(\boldsymbol{f_a} - \boldsymbol{f_0}), \tag{9}$$

$$\boldsymbol{f_{p01}} = +\beta(\boldsymbol{f_a} + \boldsymbol{f_0}), \tag{10}$$

$$\boldsymbol{f_{p10}} = -\beta(\boldsymbol{f_a} - \boldsymbol{f_1}), \tag{11}$$

$$\boldsymbol{f_{p11}} = -\beta(\boldsymbol{f_a} + \boldsymbol{f_1}), \tag{12}$$

where β is a constant to which the different value is assigned for each of the components and determines how much the stretching of the components affects the vertices.

Finally, the force applied to the vertices of each fiber fragment is calculated by summing up the force to pull apart the sheet and the restoring force generated by each of the fibers, filler, hinge springs, and connecting points. The vertices of these fiber fragments get moved according to the calculated force, and if the displacement gets over a threshold, the fiber connection is released to have the sheet teared.

Table 2 lists the values of the parameters, including α and β for each component of the paper. Since hinge spring makes a pair, stretch limit value is set to a half of the one of filler. Meanwhile, the values of α and β of hinge spring and connection point are set empirically so that they connect fiber fragments more firmly than filler.

Table 2 The parameter values for filler, hinge spring, and connection point

Parameter	Filler	Hinge spring	Connection point
Stretch limit value	0.1	0.05	–
Shrinking strength α	1.5	3.0	1.5
Influence on vertices β	1.0	1.5	1.5

4. Accelerated Computation

A paper fiber sheet with a tremendous number of fibers takes an enormous computation time. Thus we introduced the accelerated computation for the sheet generation and the sheet tearing in our model.

4.1 Intersection detection

A paper fiber sheet with a tremendous number of fibers takes an enormous computation time for detecting the fiber intersections, which is necessary for generating filler and hinge springs. Thus, in this study, we attempt to accelerate the computation by imposing a grid on the fiber generation space.

First, the space of fiber generation and the number of grid cells are predetermined before the sheet generation. Note that some margin cells are added to the upper and lower and the right and left sides of the original divided area because without the margins, the fiber cannot stick out from the generating area after the extension of fiber fragments.

Next, at most two fiber fragments are generated for each of the cells. In the present paper model, the number of fiber fragments per cell is necessarily set a priori because the fiber intersection detection process is executed with respect to fiber fragment, which is linked together to form one bended fiber.

Then, the XY coordinate values of the rectangle, with a fiber fragment as a diagonal line are referred to and the fiber fragments are attached to all cells between the vertices, as shown in **Fig. 4**, where the red fiber fragment is attached to the upper two cells and the blue fiber fragment to all the four cells.

Finally, for each of the cells, the intersection of fiber fragments corresponding to the same cell is detected. In the case of **Fig. 5**, for green fiber, the intersection detection is executed only with two red fibers belonging to the cell to which the green fiber belongs. The intersection check with blue fiber in a distant position can be omitted. This algorithm is expected to localize the intersection detection. In comparison with a naive case where we check

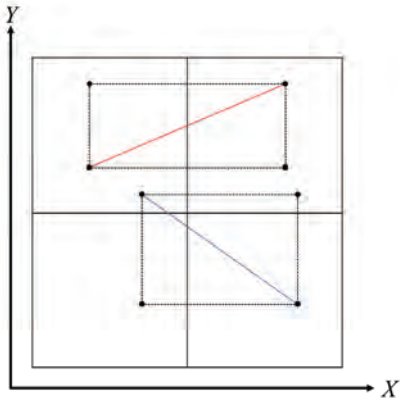


Fig. 4 Attachment of the fiber fragment to cell

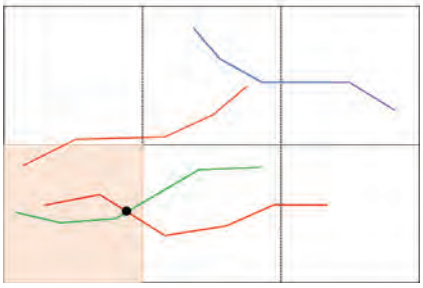


Fig. 5 Detecting intersection of neighboring fibers in a grid

a single fiber intersected with all the other fibers, the computation time for generating a paper sheet was substantially reduced, as shown in **Table 3**. Note that as the execution environment, we used a standard PC with two Intel Xeon E5-2687W 0 3.10GHz CPUs, a 64.00GB RAM, and an NVIDIA GeForce GTX 1070 Ti GPU.

4.2 Split-edge Data Structure

In our paper model, the internal repulsion generated by all components is computed, hence the higher the density of the fibers, the greater the computational complexity. To reduce the computation time for the simulation, the data structure described below is applied to our paper model and enables parallel processing on the GPU.

The repulsion generated by fibers, hinge spring, filler, and connecting points of fibers is calculated by summing up on the vertices of fiber fragments to which each component belongs. Fibers are needed to detect which type of components connects to themselves and to calculate the force according to the type to avoid the collision of the GPU memory writing. Thus, we propose a *split-edge data structure*, which is an extension of half-edge data structure[11] used for mesh processing.

In the split-edge data structure, filler, hinge springs, and connection points are divided into two, as shown in **Fig. 6**, and attached as plug to fiber fragments. Red lines

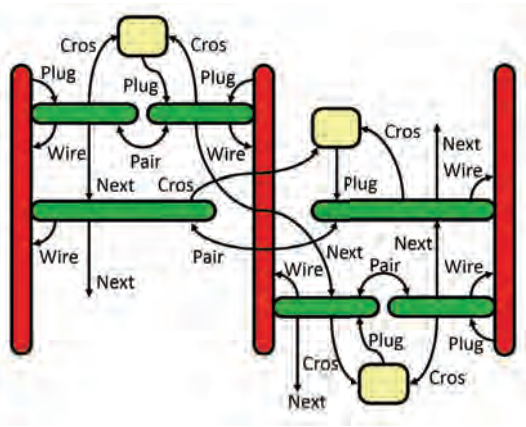


Fig. 6 The reference relationship in split-edge data structure

in Fig. 6 represent fibers, green lines plugs, and yellow boxes the components such as filler, hinge springs, and connection points. Each plug has four variables: *Cros* referring to which type of objects it comes from; *Pair* referring to another plug belonging to the same *Cros*; *Wire* referring to the fiber fragment which it connects to; and *Next* referring to the next plug connecting to the same fiber fragment. Dividing the components into two plugs enables fiber fragments to trace the components connecting to itself only with recognizing the first plug.

Table 4 compares the computation time required for one frame of simulation rendering of tearing a 1 cm × 8 cm size sheet of paper, which is pulled in parallel with the direction of fiber without and with the split-edge data structure.

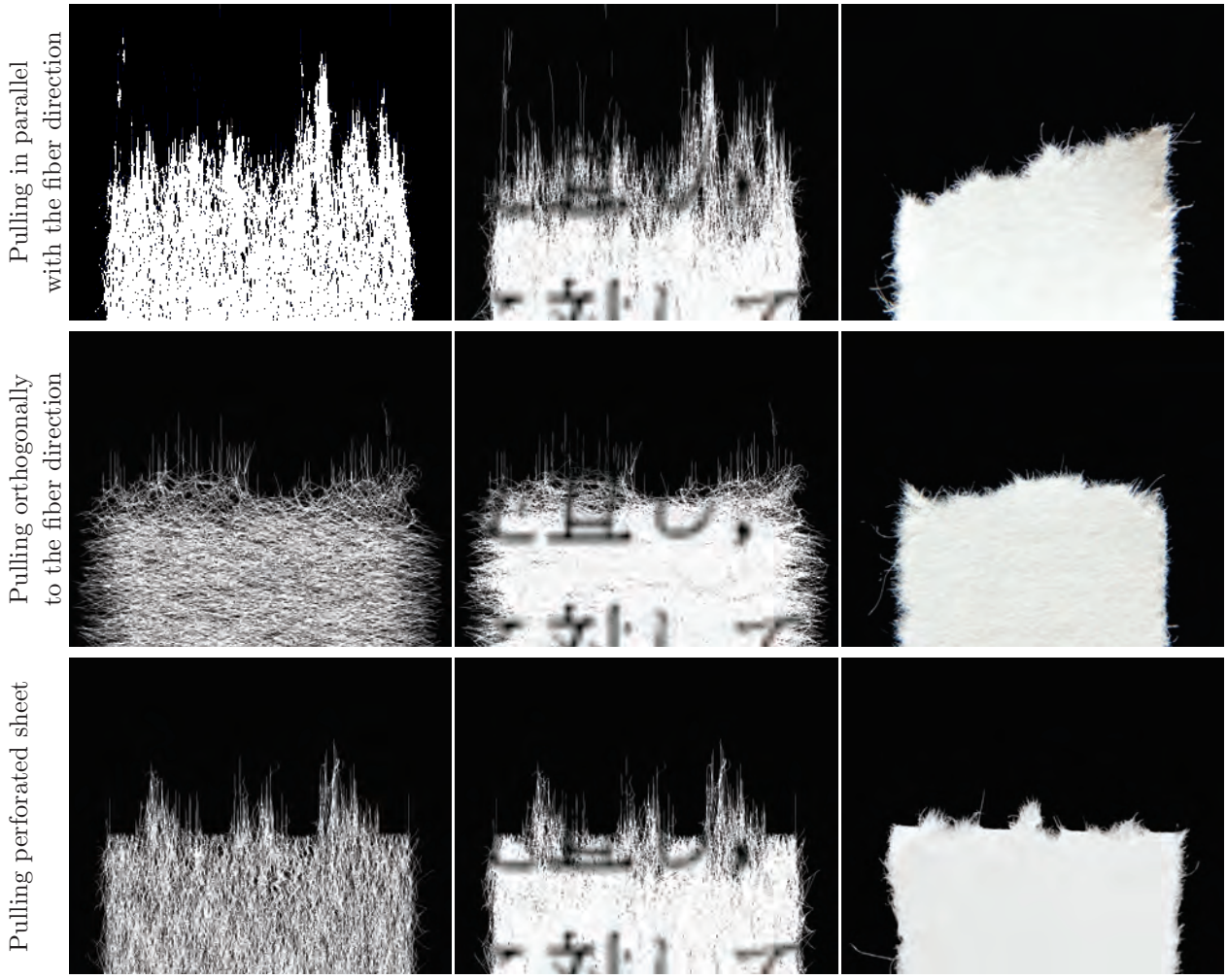
5. Results and Evaluation

Figure 7 shows the results of pulling apart a 1 cm × 8 cm sheet of paper in the vertical direction. The numbers of the components used for the experiments are tabulated in **Table 5**. All of the simulated papers consist of approximately one quarter of fiber density of real photocopy paper. Note that this density was judged to be maximum in our current execution environment due to its GPU memory constraints. Further results could be obtained by utilizing a higher-grade GPU. We succeeded in generating different appearances of torn-off line of a sheet of paper according as the direction of pulling.

In the case that the direction of pulling was in parallel with the fiber direction (top), the tear became composed of the bumped parts and the dented ones because the fibers are arranged in such a way as to stick out from the line of the tear. While in the case of being pulled orthogonally (middle), the shape of the tear is rounded

Table 3 Comparison of computation time for intersection detection (unit: msecs)

	Number of fibers	Fiber generation	Attachment to cell	Intersection detection	Drawing	Total
Round-robin	3,900	61	-	104,111	203	104,375
Our method	3,887	59	17	84	187	346



(a) Without texture (b) With filler and texture (c) Real paper

Fig. 7 Comparison of the tear of virtual paper with that of real paper

Table 4 Comparison of computation times for applying split-edge data structure (unit: msecs)

Without applying	With applying	Ratio
607,328	3	4.9×10^{-6}

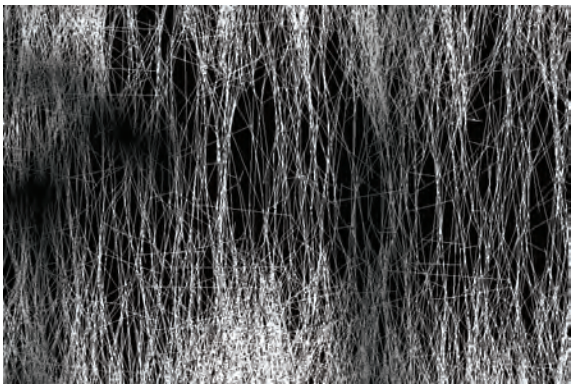
Table 5 The numbers of component occurrences in Fig. 7

	Fibers	Filler	Hinge springs	Connection points
Top	98,377	636,026	6,353,922	1,131,043
Middle	98,376	661,848	6,611,460	1,131,664
Bottom	98,054	627,542	6,289,268	1,124,802

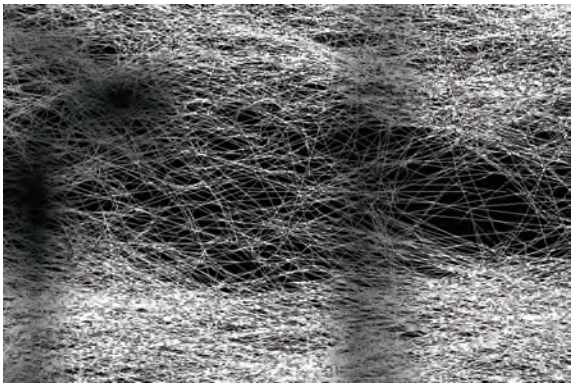
and smooth because the force just causes to expand the connection of fibers. When the perforated sheet is torn (bottom), the non-cut parts were torn in the area outside the perforation and look like tearing of the real paper.

Figure 8 gives the enlarged images captured when the sheet with texture pulled in parallel with the direction of fiber (Fig. 8 (a)) and when pulled orthogonally (Fig. 8 (b)). The sheet is applied the character texture which is the dark area in Fig. 8. When comparing the two, the fiber stretches straight in Fig. 8 (a), whereas the fiber network is expanded and the connection is released in order from the sparse part of the network in Fig. 8 (b).

The computation time is difficult to measure because the moment when all of hinge springs between upper and lower part of paper are released should be detected. To

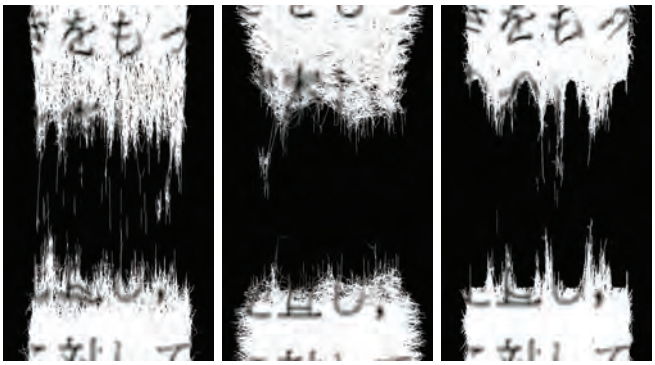


(a) In parallel with the fiber direction



(b) Orthogonally to the fiber direction

Fig. 8 Difference in the appearance of tears due to the fibers' direction



(a) (b) (c)

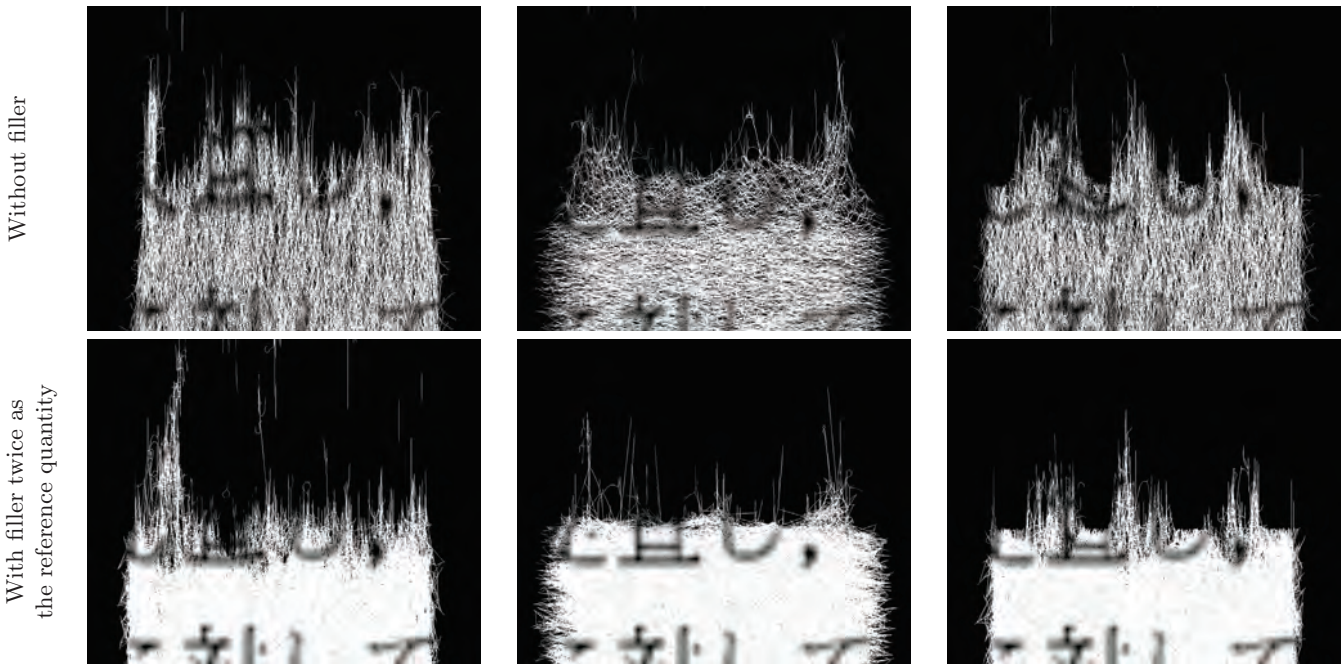
Fig. 9 Papers in Fig. 7 twenty minutes after starting simulation: (a) In parallel with the fiber direction, (b) Orthogonally to the fiber direction, and (c) Perforated paper

achieve that detection, at least two kinds of detection are needed: the tearing detection with all of the components belonging to the fiber network and whether that tear occurs the global paper tearing detection for one entire paper. In order to execute these checks, we need to trace the fiber network. It is needed to devise a graph search algorithm that solves the fiber network tracing, however it is beyond our instant development of the algorithm with tracing fibers in a reasonable simulation time. Hence the times for tearing simulation are not included in this paper.

In order to evaluate the physical validity of the simulation, we focused on the difference of the amount of paper movement among the three cases of Fig. 7. **Figure 9** shows the state of the paper after 20 minutes from the start of the simulation in Fig. 7. In the tearing simulation, the same amount of force is continuously applied to a model with the same density on the same time scale, and thereby the amount of movement is the same after the paper is completely torn. Enomae[12] shows that the tensile strength measured when a sheet of photocopy paper is pulled in parallel with the fiber direction is significantly greater than the one measured when being pulled orthogonally. We thought this could serve as a major criterion for delineating natural paper sheet deformation. In fact, on the paper sheet pulled orthogonally to the fiber direction in Fig. 9 (b), the position of printed characters is shifted more than the one on the paper sheet pulled in parallel with the fiber direction in Fig. 9 (a). The perforated paper in Fig. 9 (c) is torn in the same direction as in Fig. 9 (a), but the perforation weakened the paper fiber sheet and thus the position of printed character is shifted more than the one in Fig. 9 (a), and less than the one in Fig. 9 (b). This provides a good proof that the tensile strength of the paper sheet pulled in parallel with the fiber direction in Fig. 9 (a) is greater than the paper sheet pulled orthogonally to the fiber direction in Fig. 9 (b).

Figure 10 shows the tearing results varying the strength of paper fiber sheet by the change in the amount of filler. In the case of tearing the paper without filler (top), the sheet tends to be narrow in the horizonal direction of the figure and the tendency that the sheet stretches in the vertical direction is apparent from the distorted texture. In contrast, in the case of adding filler twice as the quantity in Fig. 7 (bottom), the density of the sheet is higher because the texture seems slightly darker. Compared with the tearing result in Fig. 7, the bulge in the vertical direction of the sheet is reduced. In addition, the influence to the inside of the tear caused by the break hardly occurs.

Figure 11 shows the result in tearing a postage stamp size sheet of perforated paper with 102,234 fibers, 702,184 filler occurences, 7,022,298 hinge springs, and 1,176,054 connection points. This sheet is torn orthogonally to the primary direction of the fibers. This example exemplifies



(a) In parallel with the fiber direction (b) Orthogonally to the fiber direction (c) Perforated paper

Fig. 10 Comparison of the difference between the amount of filler

(a) Original image (b) Resulting image

Fig. 11 A pair of torn postage stamps

the uneven tear compared to the machine cut part, and represents the feel of hand tearing.

6. Conclusion and Future Work

In this paper, we proposed a paper sheet tearing simulation model that takes into account the anisotropic microstructure and internal force generated by the paper components, so that we have the following two characteristic properties:

- Direction-dependent deformability and
- Irregularity of torn-off lines.

As a result, we succeeded in representing the uneven or rounded tear according to the direction of fibers. Furthermore, we provided a good example to show the possibility that the object with the microstructure makes the appearance from a macro perspective more realistic.

Contrary to our paper model, the real fibers are thick and branched, while the actual filler is a plane in such a way to fill the space in neighboring fibers. Thereby the simulated papers look a little too much transparent and stretched. We plan to reproduce the filler with mesh so as to fill the area surrounded by the fibers, in order to increase the density of paper fiber sheet. To this end, the radical improvement of the data structure and incorporation of a novel computation algorithm of the repulsion force which produces an appropriate result using that structure are needed. In addition, the physical validity for each of the parameters should be reconsidered. Since these changes require substantial design and implementation as well as evaluation, we decided to have them left as further research.

At this moment, the force is applied in such a way

that a sheet of paper is only pulled apart, because the force computation is limited in two-dimensions. However, in the reality, it is common to tear a sheet of paper by twisting it. To incorporate the twisting force into our model, the repulsion of the internal components against the three-dimensional force must be considered. This kind of extension would be needed to represent more daily interactions with paper products.

We also plan to extend the present model so as to handle other deformations. For example, folding is considered as one of the highest affinitive operations with our paper model because folded papers show different results depending on the fiber direction as well as tearing.

A real paper is made from multiple sheet layers and thus has a certain thickness. By considering the thickness effect in a thin volumetric model, we can express the inner layers without filler appearing at the top of paper sheet when the coated paper is torn.

ACKNOWLEDGMENT

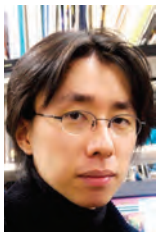
This work has been supported in part by JSPS KAKENHI under the Grant-in-Aid for Scientific Research (A) No. 17H00737.

References

1) C. Schreck, D. Rohmer, S. Hahmann: "Interactive Paper Tearing", Computer Graphics Forum, Vol. 36, No. 2, pp. 95–106 (2017).
2) T. Lejemble, A. Fondevilla, N. Durin, T. Blanc-Beyne, C. Schreck, P.-L. Manteaux, P. G. Kry, M.-P. Cani: "Interactive Procedural Simulation of Paper Tearing with Sound", Proc. of the 8th ACM SIGGRAPH Conference on Motion in Games, pp. 143–149 (2015).
3) O. Busaryev, T. K. Dey, H. Wang: "Adaptive Fracture Simulation of Multi-Layered Thin Plates", ACM Trans. on Graphics, Vol. 32, No. 52, pp. 1–6 (2013).
4) T. Pfaff, R. Narain, J. M. de Joya, J. F. O'Brien: "Adaptive Tearing and Cracking of Thin Sheets", ACM Trans. on Graphics, Vol. 33, No. 110, pp. 1–9 (2014).
5) S. Zhao, W. Jakob, S. Marschner, K. Bala: "Structure-aware Synthesis for Predictive Woven Fabric Appearance", ACM Trans. on Graphics, Vol. 31, No. 75, pp. 1–10 (2012).
6) N. Metaaphanon, Y. Bando, B.-Y. Chen, T. Nishita: "Simulation of Tearing Cloth with Frayed Edges", Computer Graphics Forum, Vol. 28, No. 7, pp. 1837–1844 (2009).
7) C. Jiang, T. Gast, J. Teran: "Anisotropic Elastoplasticity for Cloth, Knit and Hair Frictional Contact", ACM Trans. on Graphics, Vol. 36, No. 152, pp. 1–14 (2017).
8) D. Sulsky, Z. Chen, H. Schreyer: "A Particle Method for History-Dependent Materials", Computer Methods in Applied Mechanics and Engineering, Vol. 118, No. 1, pp. 179 – 196 (1994).
9) S. Takagi, M. Nakajima, I. Fujishiro: "Volumetric Modeling of Colored Pencil Drawing", Proc. of the 7th Pacific Conference on Computer Graphics and Applications, pp. 250–258 (1999).
10) K. Kitani, T. Tanaka, Y. Sagawa: "CG Expression of Paper Tearing", The Trans. of the Institute of Electrical Engineers of Japan, Vol. 128, No. 12, pp. 1735–1740 (2008).
11) M. Mäntylä, Introduction to Solid Modeling, New York, NY, USA: W. H. Freeman & Co. (1988).
12) T. Enomae, Basics and Printability of Paper –Structure, Physical Property, Processing, and Printing Quality Evaluation–, http://www.enomae.com/Paper%20Science%20seminar2/ (2019).

(Received April 29, 2019)
(Revised November 26, 2019)

Saeko SHINOZAKI
She received her B. E. and M. E. in computer sciences in 2017 and 2019 both from Keio University. She is currently a master student of Graduate School of Media and Governance, Keio University. Her research interests include paper tearing simulation.



Masanori NAKAYAMA (*Member*)
He is currently a Research Fellow in the Department of Information and Computer Science at Keio University. His principal occupation is chief bonze at Ankokuin, Chiba. 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 and SAS.

Issei FUJISHIRO
(*Honorary Member*)
He is currently a Professor at Department of Information and Computer Science, Faculty of Science and Technology, Keio University. He received his B. E. and M. E. in information sciences and electronics in 1983 and 1985 both from University of Tsukuba and his Doctor of Science in information sciences from the University of Tokyo in 1988. Before joining Keio University in 2009, he worked as a faculty member for the University of Tokyo, University of Tsukuba, Ochanomizu University, and Tohoku University. 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 the Japan Federation of Engineering Societies, a honorary member of IIEEJ, and a senior member of Information Processing Society of Japan.

IIEEJ System Development Paper -- Special Issue on Journal Track Papers in IEVC2019 --

aflak: Visual Programming Environment with Macro Support for Collaborative and Exploratory Astronomical Analysis

Malik Olivier BOUSSEJRA†, Rikuo UCHIKI†(*Member*), Shunya TAKEKAWA††, Kazuya MATSUBAYASHI†††, Yuriko TAKESHIMA††††(*Member*), Makoto UEMURA†††††, Issei FUJISHIRO†(*Honorary Member*)

†Keio University, ††Nobeyama Radio Observatory, †††Kyoto University, ††††Tokyo University of Technology, †††††Hiroshima University

<**Summary**> This paper describes an extendable graphical framework, "aflak", which provides a collaborative visualization environment for the analysis of multi-spectral astronomical datasets. aflak allows the astronomer to share and define analytics pipelines through a node editing interface, in which the user can compose together a set of built-in transforms (e.g. dataset import, integration, Gaussian fit) over astronomical datasets. Not only is aflak fast and responsive, but its macro can be conveniently exported, imported and shared among researchers using a custom data interchange format. aflak, while providing domain-specific features for astronomy, enables collaboration through shareability, guarantees end-to-end provenance management and achieves astronomer-in-the-loop. This paper compares aflak with conventional tools used in astronomy for specific use cases, such as the computing of equivalent widths. Finally, the paper demonstrates that aflak provides a reference implementation for the ProvenanceDM data model.

Keywords: astronomy, provenance, visual programming, visualization

1. Introduction

As all scientific disciplines, astrophysics requires deep collaborative support among researchers in order to make breakthrough results. The cliché of the lonely researcher hardly exists anymore in real life. This is not only shown by the ever increasing average number of co-authors in papers, but as well by the latest highly mediatized technological development that allowed the M87* black hole to be directly imaged. Imaging such a far away object requires a telescope roughly the size of the Earth to mitigate light diffraction on collecting the photons from an object with an angular size as small as M87*. Such feat can only be done by creating a virtual telescope through the cooperation of many telescopes scattered over the whole Earth—the paper by Akiyama et al. that revealed the black hole image boasts 143 different affiliations[1)].

Astrophysical data typically consists of multi-spectral images. Depending on the specific field of study an astrophysicist may dive in, one may encounter datasets with three to five dimensions. In addition to the two common spatial axes (usually right ascension and declination, but galactic coordinates may as well be used), one may find:

- Wavelength: The spacial period of a given ray received by a pixel's sensor, or a value derived from it (e.g. radial speed towards the Earth, derived thanks to the Doppler effect, as in the datasets used by Oka et al.[2)]).
- Polarization: A value characteristic of specific milieus or objects (e.g. blazars emit a polarized light[3)]).
- Time: Dynamic phenomenon may include time-dependent data (e.g. solar flare).

Various software toolkits are regularly used by astrophysicists. Some tools are mainly viewers used to visualize raw data or the result of an analysis, while other different tools are used for analytical purposes, with divergent tools being used in different sub-fields or by different researchers. In a word, an integrated environment filling the gap between all those divergent tools, akin to a "Swiss Army knife" for astronomical analysis and visualization, has yet to appear. Besides, whereas datasets may already be shared via open repositories, not the same can be said about the complete raw analysis process, which involves everything from the original data to the output that allows an astronomer to draw specific conclusions. This of course includes the analytical program devised by the astronomer, but as well all the steps by which this program was refined. Hence, we can foresee the need for

better sharing practices of analytical processes and provenance data to improve reproducibility and potential for collaboration.

The proposed framework, "aflak" (Advanced Framework for Learning Astrophysical Knowledge) aims at becoming the "Swiss Army knife" of multi-spectral data analysis in astronomy. Basically, aflak provides a visual programming environment that allows to load a dataset, to apply transformations on it and to visualize the outputs of these transformations in real time, thus providing a fast and smooth feedback loop to astronomers, significantly faster than the many tools used in astronomical research. aflak has built-in support for FITS (Flexible Image Transport System) files, which is the de facto standard for astrophysical images[4], and common data transformations specifically used by astronomers[5)–7]. The astronomer, by being able to systematically check the visualized output of the program as the visual program is being put together, can smoothly find interesting objects or rapidly prototype an analytics pipeline.

In this paper, we will focus on new features of aflak, especially the design decisions and the implementation of macro support, in such a way that interactive analysis time be reduced and collaborative use be simplified.

2. Related Work

2.1 Modular visualization environments

Several tools make use of a visual programming approach in order to achieve better accessibility for domain experts. AVS or IBM Data Explorer are the pioneers of visual programming systems for visualization applications, which have been released as commercial software from the 1990s[8]. Parker explains that SCIRun took this concept further by extending the dataflow visual paradigm to include numerical simulations[9]. This class of software is called "modular visualization environments." Such environments are highly detailed in a paper by Bungartz[10]. The base of such modular visualization environment systems is that, as the name indicates, they should be "modular" in that they include a user-extendable module library. "Visualizations" are used to represent the output of the program. Moreover, they provide the user with an integrated visual programming "environment."

These modular visualization environments process the data through four steps: data input, filtering, mapping and rendering. They have in common that they leverage visual programming in order to concentrate and allow heavy customizability on the mapping and rendering of acquired data (e.g. from measurements or numerical simulations). They use the concept of node to represent a module through which the data flows and is transformed. However, they are general tools and lack specific support and visualization primitives for the astronomical domain, and none provides re-usable, exportable and interactively editable macros, which aflak does. In the next subsection, we will focus on related software systems specifically developed to solve astronomy issues.

2.2 Viewers and analyzers for astronomical use

Astrophysicists use many different kinds of viewers, all of them with their own idiosyncrasies and specializations within a specific sub-field. Most of these tools are free and open source software, the apparently most famous and most used of which is SAOImage DS9[11], which can open FITS files and provide simple analytic tools. Kent et al.'s undertaking to re-use existing free modeling software such as Blender to image FITS files deserves notice[12]. We can as well identify new developments to visualize very large datasets that cannot be loaded in a single modern computer's running memory. As sensor technology improves, it can only be expected that such tools will become ever more important in the future[13),14].

Nonetheless, the tools above are viewers and they do not provide many data analytics features, if any. Other concurrent tools are in charge of data analytics, the oldest of which being IRAF[15]. Then, PyRAF was developed with the successful objective of providing a more user-friendly programming syntax to the IRAF environment[16]. PyRAF, with a Python-based shell syntax, includes a built-in image algebra and many transformations routinely used in astrophysics. PyRAF can execute scripts to enable code re-use. Then came Astropy, a Python library that solves most of the computing needs of astrophysicists[17]. Astronomical analysis can now be conducted using the Python scientific stack (NumPy, SciPy, etc.), at the cost for the user of writing the Python code on one's own.

Now, we can see the unfortunate divide between visualization and analysis tools in the astronomy software ecosystem. An astrophysicist's workflow usually consists in first manually analyzing datasets by applying and composing transformations on them; only then do they export the result, usually as a FITS file, to glance at it inside a viewer. Even for the widely acclaimed Astropy, exter-

nal libraries (e.g. "matplotlib") are necessary to visualize the results. In this fashion, the software ecosystem fails to provide a single integrate environment allowing rapid prototyping of analytics pipelines. There is too much lag between the composition of a program and the visualization of its output. Ideally, such system would stimulate the visualization discovery process, as defined by Johnson et al. with the "human-in-the-loop" terminology[18]. The human (i.e. the astronomer) must be kept in the loop during the designing of an analytics pipeline, which is exactly one of the goals that aflak strives to achieve.

3. System Overview

3.1 Criteria for a new astronomical analysis tool

We have surveyed general visual programming systems in section 2.1, and specific astronomical tools in section 2.2. Combining the modular features of general visual programming environments with the specific needs of astronomy is the success benchmark for the system this paper presents. From then, for the system to be of use for astronomers, it requires the four following criteria:

1. Enable collaboration through shareability;
2. Cover support for astronomical, domain-specific features;
3. Make human-in-the-loop a reality; and
4. Include end-to-end provenance management.

To achieve those requirements, we present aflak, which follows a visual programming approach similar to the tools mentioned in section 2.1, of combining nodes within a node graph to allow the user to compose algebraic transforms. Each node has input slots into which data flows, and a number of output slots from which the transformed data comes out. aflak provides an n-dimensional image algebra interface[19] similar to that of NumPy, which can be used to build analytics pipelines and smoothly visualize the resulting computations.

Then, to get around the shortcomings of conventional astronomical tools raised in section 2.2, aflak's objective[5)–7)] is to provide a universal collaborative and integrated environment to analyze and view astronomical data with very fast iterations. While "matplotlib" provides publishing quality graphs, it is far from suitable for fast iterations on relatively big datasets. Moreover, there is no built-in solution for code sharing among researchers. The best one can do is to share source files, but then no provenance is supported, reducing accessibility for convenient reproducible research. aflak shall be able to interactively reproduce the exact same numerical analyses than the conventional astronomy tools. The reader may refer to **Fig. 1** for a view of aflak in use.

3.2 Design goals

To fulfill the four criteria laid out in section 3.1, aflak is designed to meet the following requirements.

3.2.1 Re-usability and extendability

aflak can export and import the state of its interface. Moreover, aflak allows for the import and export of shareable composite nodes, referred to as "macro" hereafter. Convenient macros shall not impede the responsiveness requirements stated below. Macros can be shared among other users via a cloud platform or file exchange. What's more, nodes may be implemented as side-loaded shared libraries, allowing to create user-defined nodes in a language that supports C/C++ calling conventions.

Moreover, no research is done alone anymore. aflak aims at making collaborative development and code-sharing as easy as possible via its import/export features and compliance with standards in astronomy.

3.2.2 Specialized to tackle astronomical issues

Not only does aflak shows strict compliance with standards in astronomy (FITS, etc.), but it does as well adapt modular visualization environment concepts for more specific use cases. In section 2.1, we referred to modular visualization environments, and we saw that their visual programming interfaces concentrate on the downstream mapping and rendering phases of the visualization pipeline—which are input, filtering, mapping and rendering. Conversely, aflak concentrates on the upstream input and filtering phases. While the astronomer has complete interactive freedom over inputs and analyses, downstream visualizations use lightly parameterizable built-in features tailored for astronomy.

3.2.3 Ease of use and responsiveness: Achieving astronomer-in-the-loop

aflak is designed with ease of use in mind. Though using aflak requires astronomy-related domain-specific knowledge, it is designed to be intuitive. Dataflow is clearly indicated by connections between box-shaped nodes. The interface is responsive: the output of a visual program is refreshed in real time as the program is being updated, with very minor delay, so that the user not be confused about the provenance of the data being shown. Moreover, errors are shown using a stack trace to show at which exact node an error occurred.

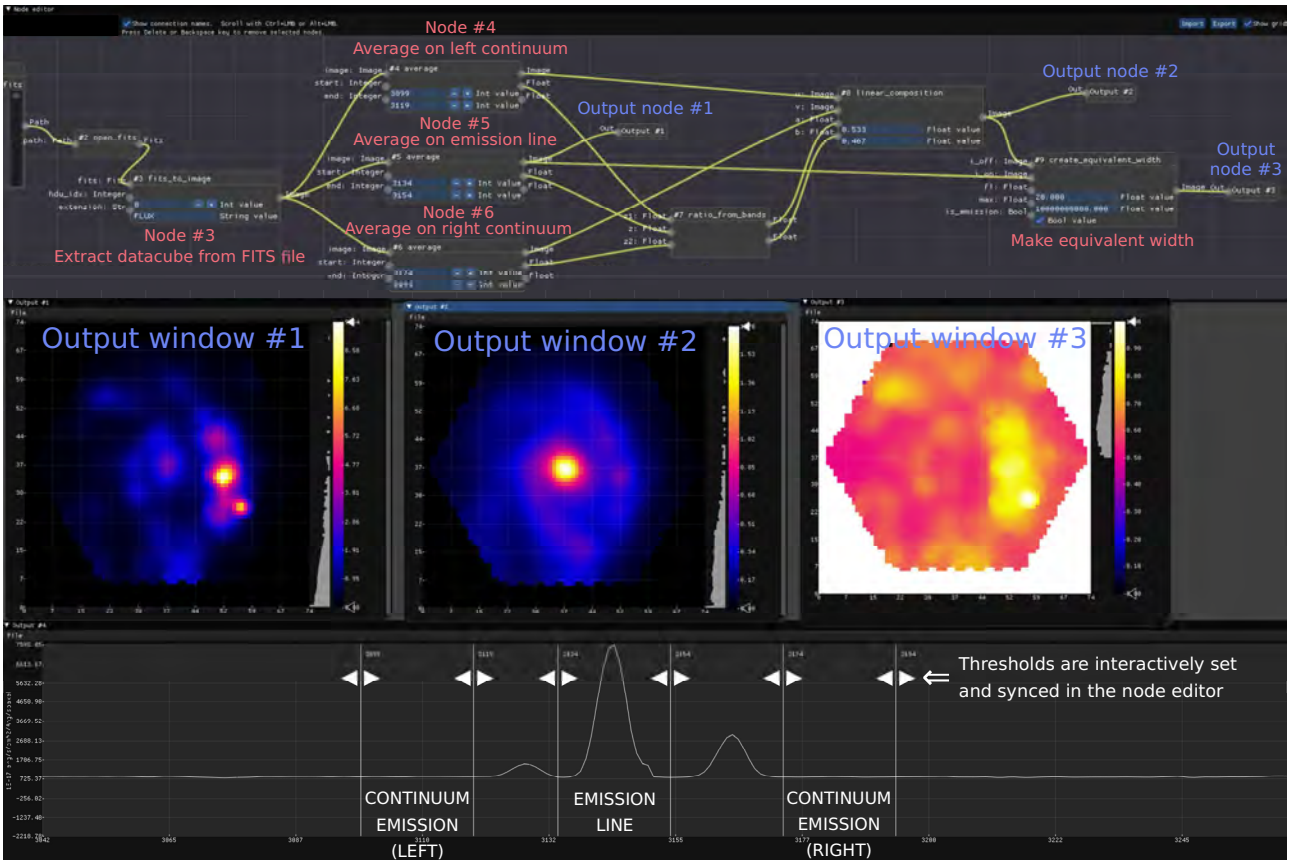


Fig. 1 Example of a semi-complex use of aflak to extract the equivalent width of a three-dimensional dataset

3.2.4 Provenance management

Some astronomers highlight the importance of provenance management. A provenance data model standard is currently in development by the International Virtual Observatory Alliance, whose name is ProvenanceDM[21]. aflak can be used to reproduce an analysis pipeline, from the data source to the final output, following a model very similar to that of ProvenanceDM. This data model is still a draft and does not have any reference implementation yet. aflak aims at successfully addressing each of the use cases that ProvenanceDM must fulfill, so that aflak provides an initial reference implementation of the ProvenanceDM data model.

3.3 Description

aflak's programming interface comprises a node editor, where nodes can be freely composed by linking a node's output slot to another node's input slot with the mouse cursor[5),6)]. The node editor's left pane contains the node list and the documentation of the currently selected node. By selecting a node in the node list, the view will directly jump and center onto that specific node in the node graph. The node graph basically reads from left (input) to right (output), in accordance with the left-to-right convention of most writing systems.

There are three different types of nodes:

- Transformation nodes: Composable transformation module. Contain a specific amount of input slots and output slots, each with a specific expected data type.
- Value nodes: Contain a parameter of a certain data type that can be input by the user, or that can be externally set by another tool.
- Output nodes: Final output of the flowing data. Any data type can be redirected to an output node. When a new output node is created, a new output window containing a visualization of the data arriving to this node is opened. The output window and the output node shares the same number. Such output windows can be seen in the middle of Fig. 1.

A node may as well be the result of the composition of several nodes. This feature is referred to as macro. Later, section 4.4.1 will expand further in details how macros are used and implemented within aflak.

4. Description of Algorithms

In order to achieve high speed and responsiveness, care was taken to design efficient algorithms and data structures. aflak is written in Rust, which allows C/C++-like fine-grained control on memory layout to maximize per-

formance, while providing a higher-level syntax and a safer paradigm to write highly concurrent programs, boosting the productivity of the implementer. All of computational tasks run by aflak are managed in an independent sub-crate called "aflak_cake" (a.k.a. cake, standing for Computational mAKE). First we will present the basic structure of cake, excluding macro support. Then we will explain how macros are implemented, while highlighting the choice of data structures.

4.1 cake: Computation mAKE

cake's main data structure is called "*DST*" (for Dynamic Syntax Tree). It is defined as shown in **Table 1**. What must be kept in mind is that a *DST* represents a directed acyclic graph of nodes. To represent such graph, we must keep a collection of transformation nodes (field "transforms" in Table 1). To do that, we assign a unique identifier to all nodes in the node editor. When a new node is added, an unused identifier—for a transformation node, an instance of $TransformId$ is used as identifier—is assigned to the new node. The actual node whose inner data is contained in a $MetaTransform$ data structure is then inserted into the sorted map containing all transforms.

But keeping the collection of transformation nodes is not enough: edges in the directed acyclic graph must as well be kept in memory. This is done in the "edges" field in Table 1. For aflak's use case, the data structure *Output* represents the n-th output slot of the node indexed by $TransformId$ as the tuple $(TransformId, n)$. Similarly, the data structure *Input* represents the n-th input slot of the node indexed by $TransformId$ as the tuple $(TransformId, n)$. The collection of edges of a graph can then be defined as a map from an output to a list of inputs, as shown in Table 1. Indeed, a single output slot can be attached to an indefinite number of input slots (one-to-many relationship). The other way around (single input slot attached to many output slot) does not make any sense as data can only flow from an output slot to (an)other node(s)'s input slot(s).

Finally, the collection of final outputs, which are defined as the actual values that are computed out of the graph, is stored in as a map in the field "outputs" in Table 1. Each output is assigned an identifier (an *OutputId*). Each *OutputId* is associated with an output slot—represented with the data structure *Output* defined above—, or with nothing (unattached). The concept "an output slot or nothing" is represented as the data structure $Option\langle Output\rangle$.

Table 1 Dynamic Syntax Tree data structure

DST	
transforms	$Map\langle TransformId, MetaTransform\rangle$
edges	$Map\langle Output, List\langle Input\rangle\rangle$
outputs	$Map\langle OutputId, Option\langle Output\rangle\rangle$
Input	
transform_id	$TransformId$
input_index	Unsigned Integer
Output	
transform_id	$TransformId$
output_index	Unsigned Integer

Table 2 MetaTransform data structure

$MetaTransform$	
t	$Transform$
input_defaults	$List\langle Option\langle T\rangle\rangle$ (where T is the type of the values passed around when a graph is computed)
updated_on	$Instant$

When a *DST* instance is constructed and every time a new node is added, the graph is checked for consistency. For example, no circular path can ever be created, cake will not allow that and gracefully abort the addition of a new node. aflak will then print the error message to the screen to let the user know what error was successfully avoided.

4.2 $MetaTransform$ data structure

As referenced in Table 1, the content of a node is stored as a $MetaTransform$ data structure, defined in Table 2. A $MetaTransform$ is a $Transform$ to which metadata is added. The $Transform$ contains the actual data describing how the computation for this node should be processed. $MetaTransform$'s metadata contains the "input_defaults" field, which is a list of optional values that represents the user-editable default input values for the node. Whenever the node is updated (e.g. by updating a default input value), the "updated_on" field is updated to contain the instant (in aflak's current implementation, an instant is not really a timestamp, but a measurement of a monotonically non-decreasing clock) on which the node was updated. As we will see later in section 4.3.2 on cache, storing the instant at which the node is updated is crucial so that cache be functional.

$Transform$'s data structure tabulated in **Table 3** is just a tuple of an *Algorithm* and the instant on which the *Algorithm* data structure was updated. *Algorithm*, as shown in **Table 4**, is an enumeration that can be one of the three following variants:

- Function: A pointer to a pure function that takes an array of data of type T and returns an array of data of type either T or E, where T is the data type on which computations are done (an enumeration over multi-spectral data, images, FITS files, etc. in aflak's context) and E is an error type, that can be represented as a human-readable string for ease of debugging.
- Constant: A constant of type T. This actually represents a constant node. Such a node has no input slot and has a single output slot that always returns T's value.
- Macro: Contains a handle to a macro $MacroHandle$ (see section 4.4.1).

All in all, we now have explained all the data structures used to store a node graph. We will then see how the computing is actually running.

4.3 Computing output with cache

4.3.1 Computation model

Given a DST representing a node graph, we iterate over each final output identified by an $OutputId$. For each final output attached to a node's output slot, the value at this output slot is computed and then returned.

A value at an output slot is computed as follows. Let N a node with n input slots and m output slots. For i and j two non-zero integers, I_i^N is defined as N's i-th input slot, and O_j^N is defined N's j-th output slot, The value at the i-th output slot of node N is noted as o_i^N, while value at the j-th output slot of node N is noted as o_j^N.

Let f the function so that $f(i_1^N, ..., i_n^N) = (o_1^N, ..., o_m^N)$, i.e. f is equivalent to the function that is evaluated when data $(i_1^N, ..., i_n^N)$ enters node N.

Let's say we want to compute one of the o_j^N for $1 \leq j \leq m$. For each i such as $1 \leq i \leq n$, we have either one of the three following cases:

Table 3 Transform data structure

$Transform$	
algorithm	$Algorithm$
updated_on	$Instant$

Table 4 Data structure of Algorithm enumeration

$Algorithm$'s variant list	
Function	Pure function pointer with some meta-data (name, description, version number, default values, input and output types etc.)
Constant	T
Macro	$MacroHandle$ (see section 4.4.1)

- I_i^N is not attached to any output slot and has no default value. In that case, o_j^n is uncomputable and an error indicating that a dependency is missing is raised and propagated.
- I_i^N has an associated default value x and is not attached to any output slot. Then $i_i^N \leftarrow x$.
- I_i^N is attached to an output slot, say $O_{j'}^D$ (designating j'-th output of node D, where D is a dependency of N attached to I_i^N). Then $i_i^N \leftarrow o_{j'}^D$.

With the above premises, we can then assuredly compute $f(i_1^N, ..., i_n^N)$, and thus we have o_j^N. The values of dependencies such that $o_{j'}^D$ can be recursively computed. The recursion will terminate, as the graph has no cycle and is finite, so there exists a starting node N_{start} such as no of N_{start}'s input slots are attached to another node's output slot.

4.3.2 Adding cache to the model

Recursively recomputing the value at an output slot from scratch every time the value must be displayed would take too much time and be extremely inefficient. For aflak to be responsive, we had to implement caching. The approach used is to keep track of the instants when values at output slots are first computable.

Let O_j^N the j-th output slot of node N. aflak appends the smallest instant in time t_{O^N} when the output slot's is theoretically computable to the value o_j^N in the current node graph's state.

We define an "updated_on" function that computes t_{O^N} for each node. It is recursively defined as:

$$\text{updated\_on}(N) = max \begin{pmatrix} N.\text{updated\_on} \\ N.t.\text{updated\_on} \\ \text{updated\_on}(D_1) \\ \vdots \\ \text{updated\_on}(D_i) \\ \vdots \\ \text{updated\_on}(D_n) \end{pmatrix} \quad (1)$$

where N is a node represented as an instance of the datatype $MetaTransform$ (see Table 2), $N.t$ is N's $Transform$, and D_i for $1 \leq i \leq n$ the n dependencies of N (n may be 0 if N has no dependency, in that case the recursion ends).

By definition, for a node N, updated_on(N) is the smallest instant in time when the outputs of N are computable in the current state of the full node graph. We then compute all the o_j^N and append updated_on(N) to the value o_j^N, storing them both into

memory. Considering that we are now at time t_0, let $t_{0O^N} = \text{updated\_on}(N)$. Next time o_j^N needs to be computed again say, at time t_1, if $t_{1O^N} = t_{0O^N}$ then there was no change in the graph state that would cause o_j^N to change, so we just retrieve the cached value. If on the contrary $t_{1O^N} > t_{0O^N}$, then o_j^N will be recomputed and the new value added to the cache alongside the new t_{1O^N}. The case $t_{1O^N} < t_{0O^N}$ cannot occur.

In addition, the cache data structure needs to support multi-threading, as all aflak's computing is done on several threads as a background process. This is aflak leverages a concurrent hash map implementation based on bucket-level multi-reader locks (implementation of the concurrent hash map used for cache lies in the "chashmap" crate[20]).

4.4 Macro Support for cake

4.4.1 Design decisions

The representation of a macro is a chunk of memory that has to be shared between several threads. aflak's runtime is composed of a UI thread that renders the user interface and handles the user's inputs, and several computation threads that compute the node editor's outputs. As a result, the UI thread requires read access to display the macro on the node screen, but requires as well write access to update the macro on user input. The computation thread requires read access to retrieve the macro's DST and run computation. As a result, it is clear that a macro must be shared between threads, behind a read-write lock. We call *MacroHandle* (which showed up in Table 4), an atomically reference-counted shared pointer to a read-write lock to the actual chunk of memory of a *Macro*, whose data structure we will define in the next section.

4.4.2 Data structures for macro support

A macro is a data structure defined as shown in **Table 5**. It includes a UUID version 4 (randomly generated Universally Unique IDentifier)[22], which is unique identifier that identifies a macro more specifically than its name. The UUID is generated on creation of a new macro according to the specification to guarantee its uniqueness, even after the macro is shared among users. Next, a macro contains a human-readable name (used for display) and an embedded DST that describes the behavior of the macro. In addition, while the outputs of a macro are determined by the output nodes of the embedded DST, it is necessary to keep the list of inputs for the macro. This list of inputs is represented by the "inputs" field

Table 5 Macro data structure

Macro	
id	*Uuid*
name	*String*
dst	*DST*
inputs	$List\langle MacroInput\rangle$
updated_on	*Instant*
MacroInput	
name	*String*
slot	*Input*
type_id	*TypeId* (where *TypeId* is a data type whose instance identifies the expected type of the variable that flows into the input slot)
default	$Option\langle T\rangle$

of type $List\langle MacroInput\rangle$. In the current implementation, the list of inputs is inferred and recomputed every time the macro's inner DST is updated: the inner DST's unattached input slots are considered to be the whole macro's input slots. Finally, as for all previously defined data structures, an "updated_on" *Instant* is appended.

4.4.3 Revamping the computation logic

The algorithm used for the whole graph is the same as the one explained in section 4.3.1, only the approach to computing the macro node differs. For a macro node N the "updated_on" function is defined as in Eq. (2). The only difference with Eq. (1) is that the macro's inner "updated_on" value is appended (in bold in the following equation).

$$\text{updated\_on}(N) = max \begin{pmatrix} N.\text{updated\_on} \\ N.t.\text{updated\_on} \\ \mathbf{\text{macro's updated\_on}} \\ \text{updated\_on}(D_1) \\ \vdots \\ \text{updated\_on}(D_i) \\ \vdots \\ \text{updated\_on}(D_n) \end{pmatrix} \quad (2)$$

When a macro node is evaluated, the macro is first deep-copied and sent to the computation thread to ensure that the macro's state does not change during computation. Each of the current values that flow into the macro's input slots is copied and rewired to the corresponding input slots in the macro's DST. Then the macro's DST is evaluated and the macro's output values are calculated.

4.4.4 Macro user interface

The macro editing user interface is integrated into aflak's UI. **Figure 2** shows a screen capture of aflak's macro editor. The macro named "See Manga" in this figure takes two inputs: a FITS file and an integer rep-

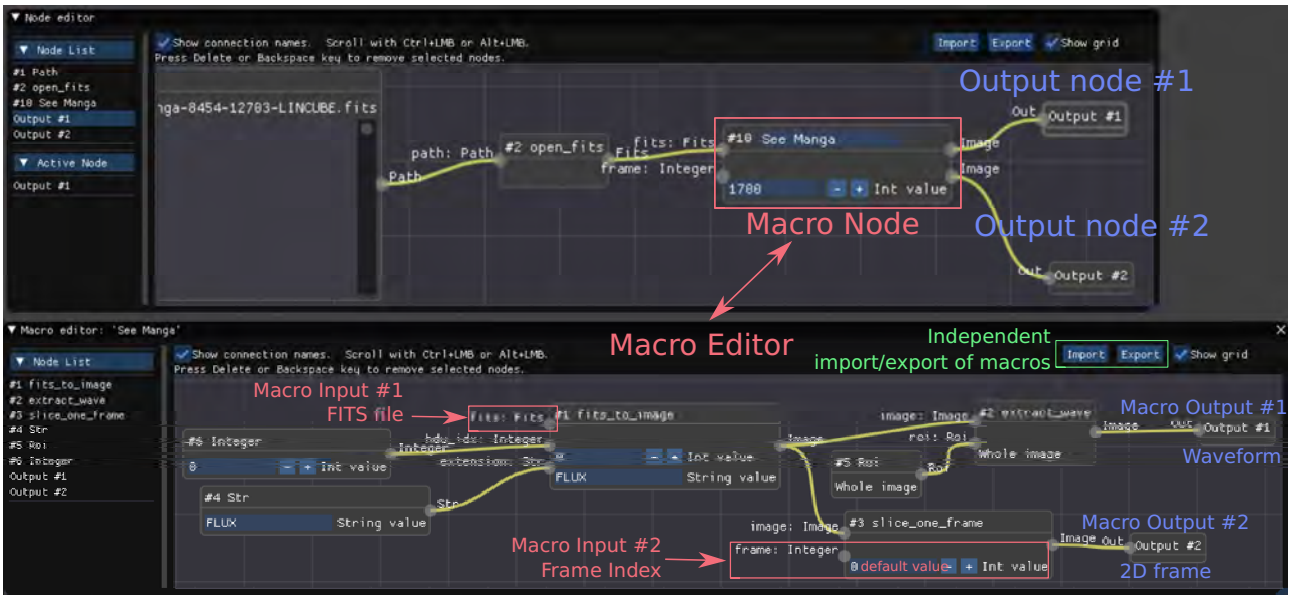


Fig. 2 Simple example of a macro that opens an image from the MaNGA dataset[23)]

resenting a frame number. The macro then outputs the waveform and the image data at the provided frame. This is the result that can be seen in output windows #1 and #2. The way the FITS file is open and the choice of extension (FLUX) is quite specific to the MaNGA dataset, hence the merit of defining a macro for such task.

A new macro may be created by a right click on the graph editor and then selecting the "Create new macro" option. On clicking, an empty macro will pop up. Once a macro has been created, it can be re-used and added as many time as one wishes into the node editor. By double-clicking on a macro, this macro's editor is opened (if it was not) and focused. The macro editor is mostly similar to aflak's main editor. It supports import and export of macros using a tailor-made data format. In addition, nested macros are supported. aflak implements some basic sanity checks to prevent endless loops from occurring. For example, if a macro is added within itself (i.e. recursion), aflak will show an error and prevent the user from proceeding and shoot themselves in the foot.

5. Evaluation

aflak was evaluated on a middle-end light laptop running Debian 9 with an Intel® Core™ i7-7560U CPU @ 2.40GHz processor and an integrated GPU chipset. aflak is supported on all mainstream operating systems (Linux, macOS and Windows 10). During evaluation, mid-size astronomical datasets (a few hundreds of megabytes) such as "manga-8454-12703-LINCUBE.fits" from the MaNGA survey were loaded. We will consider the extraction of equivalent width as an example.

5.1 Comparison with existing general systems

It should be noted that the ability to bring about shareability via macros is not new, and most conventional modular visualization environments such as those cited in section 2.1 already support macro-like features. However, aflak differs from them in that it concentrates on drafting an analysis pipeline from specific astronomical inputs, not on the visualization method, as already explained in section 3.2.2. aflak is novel in that it is a new kind of modular visualization environment that allows to specifically solve astronomical problems. In the next section, we will see how aflak compares to conventional tools used in astronomy.

5.2 Comparison with existing astronomical tooling

The listing on **Fig. 3** shows a commented Bash/IRAF script that computes the equivalent width of a datacube, using the same method as the node editor outlined in Fig. 1. The reader may refer to Carroll's book for precise definition of equivalent width[24)]. In Fig. 1, the "open_fits" node opens a FITS file, then several transformations are applied to the file to extract the equivalent width in the right-most output node (Output #3). The result of each output node is visualized in a corresponding output window. Continuum emission is computed by node #4 on the left side, and by node #6 on the right side. The average of the emission line is computed by node #5. Intermediary results are visualized via output nodes #1 and #2.

As the reader may infer from the code in the Fig. 3's listing, each operation is creating files on the file system,

```
# Sub-datacube for each band (in Fig. 1:  nodes #3, #4 and #5)
awk 'BEGIN{for (i = 3099; i < 3119; i++) {
    printf ("manga-8454-12703-LINCUBE.fits[1][,,%d]\n", i)}
  }' > file-off1.list
awk 'BEGIN{for (i = 3134; i < 3154; i++) {
    printf ("manga-8454-12703-LINCUBE.fits[1][,,%d]\n", i)}
  }' > file-on.list
awk 'BEGIN{for (i = 3174; i < 3194; i++) {
    printf ("manga-8454-12703-LINCUBE.fits[1][,,%d]\n", i)}
  }' > file-off2.list

 # Then compute average on each band
imcomb @file-off1.list off1-average.fits combine=average
imcomb @file-on.list on-average.fits combine=average
imcomb @file-off2.list off2-average.fits combine=average

 # Combine off-band images (in Fig. 1:  nodes #7 and #8)
echo "0.533333" > scale-off.dat
echo "0.466667" >> scale-off.dat
imcomb off1-average.fits,off2-average.fits off-average.fits \
  combine=average weight=@scale-off.dat

 # Make equivalent-width map (in Fig. 1:  node #9)
imarith off-average.fits - on-average.fits off-on-average.fits
imarith off-on-average.fits * 20 flux.fits
imarith flux.fits / off-average.fits equivalent-width.fits
```

Fig. 3 Extracting equivalent width with Bash/IRAF

while subsequent operations are using the created files. The content of the intermediary files can only be seen using dedicated viewers such as DS9[11]. With the example of equivalent widths, many constants must be precisely adjusted. Every time a constant is changed, all the next cascading operations must be re-run. Not only this process is time-consuming and error-prone—enlarging an already too long feedback loop—, but provenance is as well far from being managed. Out of the many files generated in the file system, how does the user remember how each individual file was generated? By comparison, aflak here shines by its fast feedback loop (less than a second is needed to refresh the visualizations in an output window after a change in the node editor) and automatic management of provenance. Computing speed is enough to responsively and interactively give feedback to the astronomer. Importantly, if we exclude negligible floating-point rounding errors, results obtained with IRAF and aflak are perfectly consistent.

5.3 Equivalent width with a macro

The previous section compares aflak with the existing tools. This section gives a concrete example of macro usage for equivalent widths. **Figure 4** shows an example of a macro implementing the computation of equivalent widths, the same processing as the one done in Fig. 1. The macro encapsulates all the logic implemented as shown in the node editor in Fig. 1, only exposing the relevant constants that the astronomers are expected to gradually adjust until they get a satisfactory outcome.

The advantage of using macros are clear: visual clutter on the original node interface is widely reduced, thus reducing cognitive load, while the computing speed is not impacted for a 423 MB input dataset. Only the input datacube and the constant parameters that are relevant in computing the equivalent width are exposed by the macro. The computed output is strictly identical whether a macro is used or not. Through the sharing and the export of macros such as this one given as example, common parameterized analytics possibilities are only a few clicks away.

aflak could serve as turnkey software if an administrating user judiciously designed a hierarchy of macro calls. Macros allow to hide details to concentrate on the look-and-feel and the features of importance for the astronomers. In fact, astronomers—our users—let us know that they expect a system that can give instant feedback. This is especially important for novice users. With the sharing of macros, there is no theoretical limit to the number of users aflak may spread to.

It should furthermore be noted that all values that can be selected and manipulated inside output windows as shown in Fig. 1 are redirected and bound to a value node, which then can be re-used within the node editor. The binding is bi-directional: editing the value of the node from the node interface will update the representation of the value in the output window as well. This allows the user to have fine-grained visual control on several parameters while designing an analytics pipeline. Clutter-free astronomer-in-the-loop is then achieved.

5.4 An implementation for ProvenanceDM

As we consider the full list of all use cases that ProvenanceDM must fulfill[21], we can see that aflak successfully addresses each of them, hence we can conclude that aflak provides a reference implementation for the ProvenanceDM data model defined by the International Virtual Observatory Alliance, whose requirements are listed below:

- Traceability of products: "Track the lineage of a product back to the raw material, show the workflow or the dataflow that led to a product." aflak's visual approach of representing dataflow with a visualized directed acyclic graph meets this use case.
- Acknowledgment and contact information: "Find the people involved in the production of a dataset, the people/organizations/institutes that one may want to acknowledge or can be asked for more information." aflak's nodes contain information about their author, thus aflak meets this use case.

Fig. 4 Example of using an aflak macro to compute equivalent width

- Quality and reliability assessment: "Assess the quality and reliability of an observation, production step or dataset." The version and the unique UUIDs of the nodes and macros used in aflak to generate any dataset are logged, which allows to check quality and reliability at any time in the future.
- Identification of error location: "Find the location of possible error sources in the generation of a product." All errors that occur during processing on aflak are logged and a detailed stack trace is shown to thus user, thus fulfilling this requirement.
- Search in structured provenance metadata: "Use provenance criteria to locate datasets (forward search), e.g. finding all images produced by a certain processing step or derived from data which were taken by a given facility." It is theoretically possible to classify FITS files that were exported by aflak grouped by the exact node editor that generated them, as a serialized node editor is included in all generated FITS files[7].

6. Discussion and Future Work

aflak provides a fast and responsive modular visualization environment for defining custom analytical macros and sharing them with fellow astronomers. Macro support was a highly awaited feature that our dozen of users were eager to utilize. Indeed, astronomers expect to use turnkey-like software with the features they need already included.

However, aflak is not without its shortcomings: more memory-efficient use of cache is desired to reduce wasting of memory and have aflak deal with bigger datasets (over several gigabytes). Smarter probabilistic cache garbage-collecting resources that are very unlikely to be queried would be a huge improvement. Besides, on the UI side, being able to group several nodes into a single macro from the node editor screen would be extremely convenient and is a desirable feature. Finally, having a public aflak repository where anyone can contribute new nodes or macros is a planned feature. Using UUIDs to identify macros is a first step toward building an aflak repository.

7. Conclusion

This paper is a subsequent[7] report on aflak, a visual programming environment that provides fast and responsive macro support in the astrophysical domain. Collaboration, ease of use, responsiveness, incremental improvements and provenance are taken very seriously. Besides, by porting aflak to the browser using WebAssembly, astronomical analysis and access to data could be made accessible to a broader audience. Indeed, astronomy is a field where amateurs are far outnumbering professionals.

Acknowledgment

This work is supported by JSPS KAKENHI Grant Numbers 17K00173 and 17H00737.

References

1) K. Akiyama, A. Alberdi, W. Alef, K. Asada, R. Azulay, A.-K. Baczko, D. Ball, M. Baloković, J. Barrett, D. Bintley, et al.: "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole", The Astrophysical Journal Letters, Vol. 875, No. 1, p. L1 (2019).
2) T. Oka, S. Tsujimoto, Y. Iwata, M. Nomura, S. Takekawa: "Millimetre-Wave Emission from an Intermediate-Mass Black Hole Candidate in the Milky Way", Nature Astronomy, Vol. 1, No. 10, pp. 709–712 (2017).
3) I. Fujishiro, N. Sawada, M. Nakayama, H.-Y. Wu, K. Watanabe, S. Takahashi, M. Uemura: "TimeTubes: Visual Exploration of Observed Blazar Datasets", Journal of Physics: Conference Series, Vol. 1036, No. 1, p. 012011 (2018).
4) D. C. Wells, E. W. Greisen: "FITS: A Flexible Image Transport System", Image Processing in Astronomy, p. 445 (1979).
5) M. O. Boussejra, K. Matsubayashi, Y. Takeshima, S. Takekawa, R. Uchiki, M. Uemura, I. Fujishiro: "aflak: Pluggable Visual Programming Environment with Quick Feedback Loop Tuned for Multi-Spectral Astrophysical Observations", Proc. of 2018 IEEE Scientific Visualization Conference (SciVis), pp. 72–76 (2018).
6) M. O. Boussejra, S. Takekawa, R. Uchiki, K. Matsubayashi, Y. Takeshima, M. Uemura, I. Fujishiro: "aflak: Visual Programming Environment with Quick Feedback Loop, Tuned for

Multi-Spectral Astrophysical Observations", Proc. of Astronomical Data Analysis Software and Systems XXVIII, Vol.523, p.245–248 (2019),
7) M. O. Boussejra, R. Uchiki, Y. Takeshima, K. Matsubayashi, S. Takekawa, M. Uemura, I. Fujishiro: "aflak: Visual Programming Environment Enabling End-to-End Provenance Management for the Analysis of Astronomical Datasets", Visual Informatics, Vol. 3, No. 1, pp. 1–8 (2019).
8) G. Cameron: "Special Focus: Modular Visualization Environments (MVEs)", ACM Computer Graphics, Vol. 29, No. 2, pp. 3–60 (1995).
9) S. G. Parker, C. R. Johnson: "SCIRun: A Scientific Programming Environment for Computational Steering", Proc. of the 1995 ACM/IEEE Conference on Supercomputing, p. 52 (1995).
10) H.-J. Bungartz, A. Frank, F. Meier: "Design and Implementation of a Modular Environment for Coupled Systems", Preprint SFB-438-9802 (1998).
11) W. Joye, E. Mandel: "New Features of SAOImage DS9", Astronomical Data Analysis Software and Systems XII, Vol. 295, p. 489 (2003).
12) B. R. Kent: "Visualizing Astronomical Data with Blender", Publications of the Astronomical Society of the Pacific, Vol. 125, No. 928, p. 731 (2013).
13) S. Perkins, J. Questiaux, S. Finniss, R. Tyler, S. Blyth, M. M. Kuttel: "Scalable Desktop Visualisation of Very Large Radio Astronomy Data Cubes", New Astronomy, Vol. 30, pp. 1–7 (2014).
14) A. H. Hassan, C. J. Fluke, D. G. Barnes: "Interactive Visualization of the Largest Radioastronomy Cubes", New Astronomy, Vol. 16, No. 2, pp. 100–109 (2011).
15) D. Tody: "The IRAF Data Reduction and Analysis System", Instrumentation in Astronomy VI, Vol. 627, pp. 733–749 (1986).
16) M. De La Pena, R. White, P. Greenfield: "The PyRAF Graphics System", Astronomical Data Analysis Software and Systems X, Vol. 238, p. 59 (2001).
17) T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droettboom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf, et al.: "Astropy: A Community Python Package for Astronomy", Astronomy & Astrophysics, Vol. 558, p. A33 (2013).
18) C. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, T. S. Yoo: "NIH/NSF Visualization Research Challenges Report" (2005).
19) G. X. Ritter, J. N. Wilson, J. L. Davidson: "Image Algebra: An Overview", Computer Vision, Graphics, and Image Processing, Vol. 49, No. 3, pp. 297–331 (1990).
20) chashmap,https://docs.rs/chashmap/2.2.2/chashmap/ (2019).
21) M. Servillat, K. Riebe, C. Boisson, F. Bonnarel, A. Galkin, M. Louys, M. Nullmeier, N. Renault-Tinacci, M. Sanguillon, O. Streicher: "IVOA Provenance Data Model Version 1.0. IVOA Proposed Recommendation 2019-07-19", http://www.ivoa.net/documents/ProvenanceDM/20190719/ProvenanceDM-1.0-20190719.pdf (2019).
22) P. Leach, M. Mealling, R. Salz: "A Universally Unique IDentifier (UUID) URN Namespace", Technical Report, https://www.rfc-editor.org/rfc/pdfrfc/rfc4122.txt.pdf (2005).
23) K. Bundy, et al.: "Overview of the SDSS-IV MaNGA Survey: Mapping Nearby Galaxies at Apache Point Observatory", The Astrophysical Journal, Vol. 798, No. 1, p. 7 (2015).
24) B. Carroll, D. Ostlie: An Introduction to Modern Astrophysics, Pearson International Edition, Pearson Addison-Wesley (2007).

(Received April 26, 2019)
(Revised August 22, 2019)

Malik Olivier BOUSSEJRA
He graduated from École Centrale de Nantes, a French engineering school, then completed a double Master's Degree with Keio University in 2016. In 2019 he then proceeded to complete a Ph.D. in computer science at Keio University. He is mainly driven by curiosity and willingness to contribute.

Rikuo UCHIKI (*Member*)
He received a B. E. in 2019 from Keio University. Currently, he is a master student at Keio University. His research interests include astrophysical data visualization.

Shunya TAKEKAWA
He received a Ph.D. degree from Keio University in 2018. Currently, he is a project research staff of the Nobeyama Radio Observatory, National Astronomical Observatory of Japan. His research interests include radio astronomy and observational studies of the galactic center.

Kazuya MATSUBAYASHI
He is a Program-Specific Assistant Professor at Okayama Observatory, Kyoto University. He received his Ph.D. degree (Science) in 2011 from Kyoto University. His research topics are astronomical instrumentation and galaxies.

Yuriko TAKESHIMA (*Member*)
She is currently a Professor at School of Media Science, Tokyo University of Technology. She received her B. Sc., M. Sc., and Ph. D. degrees in Computer Sciences from Ochanomizu University in 1994, 1996, and 1999, respectively. Her research interests include volume visualization, toplogy-based visualization, and visualization environments.

Makoto UEMURA
He is an associate professor at Hiroshima Astrophysical Science Center, Hiroshima University. He received his Ph.D. degree from Kyoto University in 2004. His research interests include optical and near-infrared astronomy.

Issei FUJISHIRO

(*Honorary Member*)

He is currently a Professor at Department of Information and Computer Science, Faculty of Science and Technology, Keio University. He received his B. E. and M. E. in information sciences and electronics in 1983 and 1985 both from University of Tsukuba and his Doctor of Science in information sciences from the University of Tokyo in 1988. Before joining Keio University in 2009, he worked as a faculty member for the University of Tokyo, University of Tsukuba, Ochanomizu University, and Tohoku University. 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 the Japan Federation of Engineering Societies, a honorary member of IIEEJ, and a senior member of Information Processing Society of Japan.

IIEEJ **Paper**

An Efficient Entropy Coding of Sparse Coefficients Based on Sparsity Adaptation and Atom Reordering for Image Compression

Ji WANG† , Yoshiyuki YASHIMA† (*Member*)

† Graduate School of Information and Computer Science, Chiba Institute of Technology, Narashino, Japan

<**Summary**> Sparse coding is a technique that represents an input signal as a linear combination of a small number of atoms in the dictionary. When sparse coding is applied to image compression, it is necessary to perform efficient code assignment taking into account the statistical properties of weighting factors for each atom. In this paper, we analyze in detail the position indices and magnitude of non-zero coefficients in a dictionary designed by K-SVD. Based on the analyzed results, we propose an efficient entropy coding method introducing sparsity adaptation and atom reordering. Simulation results show that the proposed method can reduce the amount of generated bits by up to 6.2% compared to the conventional methods.

Keywords: image compression, sparse coding, entropy coding, K-SVD

1. Introduction

With the recent increase in image resolution and increase in the amount of images by SNS, efficient image coding technology is indispensable in the field of communication services and many home appliances. As image compression coding standards that are widespread in the world, there are JPEG[1] for still pictures and H.264/AVC[2], H.265/HEVC[3] for moving pictures. In addition, studies of the next-generation video coding standard VVC, including further improvement of coding efficiency, have already been started[4]. In most conventional international standards, DCT (Discrete cosine transform) is mainly used as transformation methods for efficiently representing images. The DCT gives a good image representation for various images, but is not the best transform for each individual image. In order to solve this problem, sparse coding is attracting attention. Sparse coding can design a dictionary by training the local features of the image to be coded and can represent the target image efficiently. In sparse coding, a dictionary can be designed to minimize the reconstruction error when approximating the original image using a linear sum of a predetermined number of atoms. Since sparse coding allows most of the weight coefficients to be zero and image representation can be performed with very few nonzero weight coefficients, its application to image compression can be expected[5),6]. When the sparse coding is applied to image compression, the problem is how to encode the nonzero coefficients distributed in sparse. The statistical properties of sparse nonzero coefficients have been analyzed in some previous studies. In Reference 7), it has been experimentally reported that the atom indices to indicate the occurrence position of nonzero coefficient can be approximated by uniform distribution, and nonzero coefficient levels can be approximated by Laplacian distribution. However, it is not clear how the atom indices and the nonzero coefficient levels in a block are related to the number of nonzero coefficients in the block. Also, a detailed analysis of the relationship between a nonzero coefficient level and its corresponding atom's feature has not been performed. For more efficient entropy coding design, it is necessary to analyze statistical properties of nonzero coefficients in more detail.

In this paper, we analyze the statistical properties of nonzero coefficients in detail from theoretical and experimental viewpoints, and propose an efficient entropy coding method of sparse coefficients based on the analysis. Section 2 reviews the dictionary design method by K-SVD and describes the application method to image coding using the designed dictionary. In addition, we describe typical conventional code assignment methods to transform coefficients (weight coefficients), and point out their problems when applying to sparse coefficients encoding. Section 3 analyzes the statistical properties of the sparse coefficients in detail. First, we measure the occurrence probability of atom indices and coefficient levels for nonzero coefficients, and clarify the distribution char-

acteristic of zero run length between nonzero coefficients. Based on the distribution characteristics, we propose a context adaptive code assignment method to zero run length and nonzero coefficient level based on the number of nonzero coefficients in the block. Next, we show that the distribution characteristics of nonzero coefficient levels differ depending on features of atoms, and clarify that context adaptive coding to nonzero coefficient levels based on feature of atoms is effective. Furthermore, we show that the zero run length can be coded efficiently by rearranging the atoms by their features. In Section 5, we summarize the results obtained and discuss further works. The main focus of this paper is a research on symbol generation for efficient entropy coding, rather than actual code design method itself such as variable length code tables or arithmetic coding. Therefore, the amount of generated information is discussed mainly based on entropy.

2. Related Works

2.1 Review of K-SVD

In this section, we review the dictionary learning procedure based on K-SVD[8]. The set of sample vectors for dictionary learning is indicated as the matrix $\mathbf{Y}$, and each column of $\mathbf{Y}$ corresponds to N sample vectors $\mathbf{y}_i (i = 1, 2, \cdots, N)$. For image representation, $\mathbf{y}_i$ is often set as a vector whose elements are the pixel values in the i-th small block obtained after dividing the image. Let $\mathbf{d}_k (k = 1, 2, \cdots, K)$ be the k-th atom vector, and let dictionary $\mathbf{D}$ be a matrix in which these atoms are arranged as columns. The dimension of these atoms equals that of $\mathbf{y}_i$. We represent signal $\mathbf{y}_i$ as a linear combination of these atoms as expressed by Eq.(1).

$$\mathbf{y}_i = \sum_{k=1}^{K} a_{ik} \mathbf{d}_k \tag{1}$$

a_{ik}, which denotes the k-th element of vector $\mathbf{a}_i$, is the representation coefficient of the sample $\mathbf{y}_i$. Using coefficient matrix $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N)$, Eq.(1) can be written as

$$\mathbf{Y} \approx \mathbf{DA} \,. \tag{2}$$

We consider an optimization problem with sparsity constraints that expresses input vector $\mathbf{y}_i$ with as few atoms as possible. Approximations with greater sparsity and smaller error can generally be obtained by using a dictionary learned from samples having characteristics similar to the samples to be represented. Therefore, it is desirable to co-optimize both dictionary $\mathbf{D}$ and coefficient $\mathbf{A}$. This problem can be formulated as

$$\min_{\mathbf{D},\mathbf{A}} ||\mathbf{Y} - \mathbf{DA}||_F^2 \quad \textit{subject to} \quad \forall i, ||\mathbf{a}_i||_0 \leq T_0 \,. \tag{3}$$

Here, notation $\| \cdot \|_F$ stands for the Frobenius norm and T_0 is the sparsity constraint threshold.

K-SVD solves Eq.(3) by iterating two stages, sparse coding stage and dictionary update stage. The former applies orthogonal matching pursuit (OMP)[9]to determine $\mathbf{a}_i$ for each $\mathbf{y}_i$ while fixing $\mathbf{D}$; and the latter updates $\mathbf{D}$ together with the nonzero coefficients of $\mathbf{a}_i$. The algorithmic description of K-SVD is detailed in Reference 7).

2.2 Image compression by multiclass K-SVD

A dictionary designed by K-SVD under the constraint that the number of nonzero coefficients falls to T_0 or less can minimize the reconstruction error when the image used to design the dictionary is represented by T_0 or fewer nonzero coefficients. However, the characteristics of K-SVD derived dictionaries are highly dependent on the feature of the images used in training. A dictionary trained for a specific image is optimum for that image, but not necessarily for other images. When applying K-SVD to image coding, the decoder has to use the same dictionary as the encoder, so the designed dictionary itself must be encoded and transmitted to the decoder. However, the coding and transmission of dictionaries for each image every time incurs large overheads for information transmission, and is not practical from the viewpoint of rate distortion performance.

To solve this problem, we consider the use of the multiple dictionary approach. That is, first, the image is divided into small blocks to calculate local features, then a set of blocks having similar feature is created as a class, and finally, K-SVD is executed for each class so as to design multiple dictionaries. Local features such as DSIFT[10], intraframe/interframe prediction residual power in H.264/AVC and H.265/HEVC[7),11)] have been utilized for classification. A set of dictionaries designed for each class is shared in advance by the encoder and decoder, and these are adaptively switched when encoding. This eliminates the need to encode dictionaries, and makes it possible to represent more images efficiently.

In this paper, we calculate the local features of the block based on the idea of classification shown in Reference 9). After classifying the blocks used for training based on the local feature, a dictionary is designed for

each class. DSURF is used instead of DSIFT in this study for high speed processing. **Figure 1** shows some examples of dictionaries designed by multiclass K-SVD. It is important to note that the atoms in a dictionary designed by K-SVD are not necessarily arranged in frequency order like DCT, and the atoms with different properties appear randomly.

In order to effectively represent the actual image, each dictionary should contain one DC atom, as has been confirmed[1),12)]. Therefore, one DC atom is included in the initial dictionary for each class, and the DC atom is not changed during K-SVD iterative processing. Also, all atoms other than DC are normalized so that the mean value is zero and the standard deviation is one.

A block diagram of the encoder and decoder using dictionaries designed by the multiclass K-SVD algorithm is shown in **Fig. 2**. All dictionaries are prestored in both encoder and decoder. In the encoding process, an image to be coded is divided into small blocks of the same size as the used in the training process. Then, OMP is performed for each target block $\mathbf{t}_i$ under sparsity condition T_0; the squared errors $\mathbf{e}_c (c = 1, 2, \cdots, C)$ are calculated as follows:

$$\mathbf{e}_c = \|\mathbf{t}_i - D^c \mathbf{x}_i\|_F^2 , \tag{4}$$

then the class index c and sparse coefficients $\mathbf{x}_i$ that minimize squared error $\mathbf{e}_c$ are determined. Quantized coefficients $Q(\mathbf{x}_i)$ and class index c are encoded and transmitted.

In the decoding process, the dictionary is adaptively selected block by block based on the decoded class index, and pixel values in the block are reconstructed as the sum of atoms weighted by the decoded sparse coefficients.

2.3 Entropy coding for transform coefficients

In image coding, it is necessary to make symbols to be coded into binary codes. This procedure is called entropy coding, and various kinds of variable length codes based on the occurrence probability of symbols are utilized. By assigning fewer bits to encode more frequently occurring symbols, the total amount of bits used to encode the all symbols can be reduced. The Huffman code and the arithmetic code are typical ones.

Here, we review some code assignment techniques for transform coefficients. Discrete Cosine Transform (DCT) is one of the most popular techniques used today in video compression schemes. **Figure 3** (a) shows the 8x8 array of atom images for the two dimensional DCT. DCT converts a set of pixels in a block into the weighted sum of DCT atoms. The weighting factors are called DCT coefficients. Statistically, the magnitude of the DCT coefficients for low frequency atoms are greater than that for high frequency atoms. Also, by quantizing the coef-

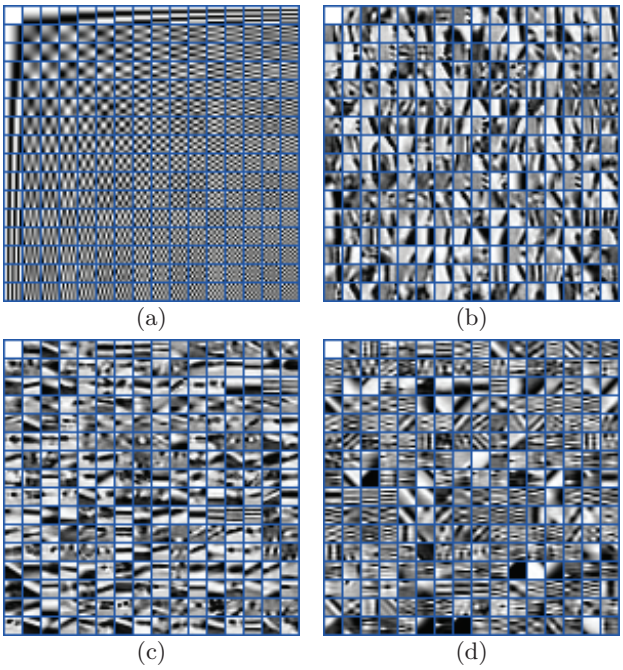


Fig. 1 Examples of multiclass dictionaries designed by K-SVD; (a) overcomplete DCT, (b, c, d) Dictionaries designed by K-SVD. The upper left corner represents the DC atom

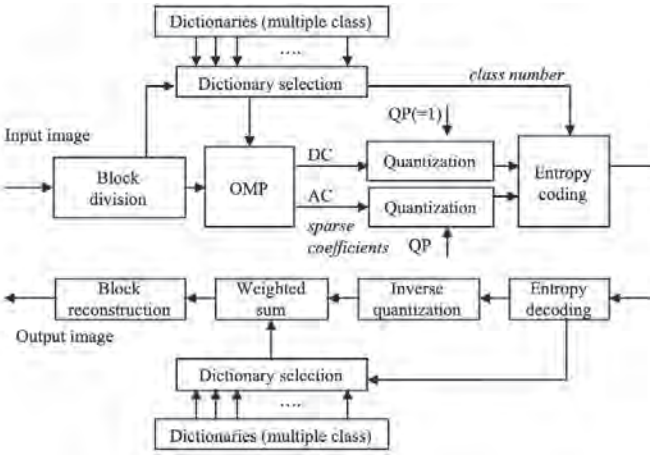
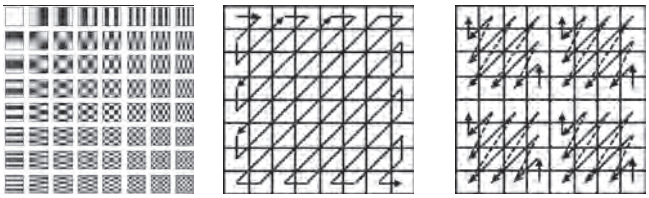


Fig. 2 Encoder and decoder configuration



(a) DCT atom (b) Scan order in JPEG (c) Scan order in H.265/HEVC

Fig. 3 DCT atom and scan order

ficients, many DCT coefficients for high frequency atoms will be zero.

This property is used to perform efficient entropy coding by appropriately setting the scan order of DCT coefficients. The order of zigzag scan in JPEG and MPEG-2 is shown in Fig. 3(b). The first coefficient of each block obtained as a result of zigzag scan is called the DC coefficient while the other coefficients are called AC coefficients. For AC coefficients, a variable length code is assigned for the pair of a nonzero coefficient and its preceding zero run length[1),13)]. An End-of-Block (EOB) at the end of each block indicates the rest of the coefficients of the block are all zero, and it enables to represent long consecutive zeros effectively. In H.265/HEVC, quantized DCT coefficients are coded as follows. They are scanned diagonally to form a 1D array as shown in Fig. 3(c). The context adaptive binary arithmetic coder (CABAC) encodes the last position of nonzero coefficients, a significance map indicating the positions of nonzero coefficients, and the quantized coefficient level values[3),14)].

In the case of complete DCT, the frequency characteristics of each atom are known, and they are regularly arranged. The relative relationship between the characteristics of each atom and the magnitude of the transform coefficient corresponding to each atom is also clarified. Therefore, by setting the scan order as shown in Fig. 3 based on these characteristics in advance, the number of occurred bits can be reduced effectively. Also, for the atoms based on complete DCT, the international standard methods H.264/AVC and H.265/HEVC have adopted a method of switching the code table for each block using the number of non-zero coefficients as a context[2),3)]. On the other hand, each atom of the overcomplete dictionary designed by K-SVD does not have regular frequency characteristics like DCT. Thus, it has not been clarified what kind of atom has a large non-zero coefficient. Also, it has not been clarified how the coefficient quantization level distribution changes with the number of non-zero coefficients in the block. Therefore, in order to perform entropy coding for sparse coefficients as efficient as the conventional method, we need to clarify the statistical properties of the sparse coefficients and to clarify how to reorder the sparse coefficients based on the findings.

Several entropy coding for sparse representations have also been studied. In image coding using sparse representation, OMP is performed using a dictionary for each block to be coded, and at most T_0 nonzero coefficients are calculated. All other coefficients are zero. For the entropy coding of the sparse representation, the indices of atoms corresponding nonzero coefficients after quantization and the nonzero coefficients levels are encoded. In the conventional studies, it has been clarified that the histogram of the atom 's indices is approximated to a uniform distribution, the histogram curve for the quantized coefficient levels is approximated to the Laplacian distribution[7),15)]. In Reference 15), it is shown that the nonzero coefficient in the case of sparse representation by overcomplete ICT becomes Laplacian distribution. Based on these features, a fixed length code was assigned for the atom index coding, and Huffman code or a truncated unary code combined with an Exponential-Golomb code was employed to encode the quantized coefficient level[7),15)]. On the other hand, instead of assigning a code directly to an index, a method of assigning a Huffman code to a zero run length (i.e. the number of consecutive zero coefficients between nonzero coefficients) has also been studied [17)].

However, in the conventional researches, the relationship between the atom indices corresponding the nonzero coefficients in a block and the number of nonzero coefficients of the block has not been clarified. Also, the relationship between the probability distribution of the nonzero coefficient levels and the number of nonzero coefficients of the block has not been clarified. In addition, the detailed analysis of the relationship between the magnitude of nonzero coefficient level and the feature of the corresponding atoms has not been conducted. Therefore, there is room for improving the conventional code assignment procedure by using the number of nonzero coefficients and feature of the atoms as a context. In the next section, we analyze the statistical properties of nonzero coefficients in detail from theoretical and experimental viewpoints for sparse representation of images, and we propose an efficient entropy coding scheme for sparse coefficients.

3. Statistical Feature Analysis of Sparse Coefficients

In this section, we analyze the statistical properties of the sparse coefficients in detail for the entropy coding scheme design. The analysis in Section 3 is carried out theoretically and experimentally. A set of small blocks extracted from six types of images, "BQTerrace", "BasketballDrive", and "Cactus", "ChristmasTree", "Kimono1" and "ParkScene" from the MPEG test sequence are used for statistical analysis, where these images are also used

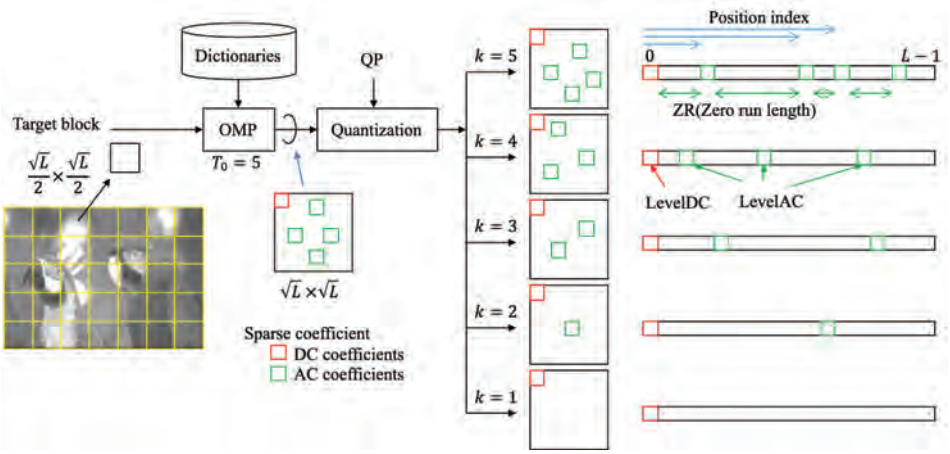


Fig. 4 Sparse coefficients to be coded

as test data for the experiments in Section 4. The sparse coefficients to be encoded can be illustrated as in **Fig. 4**. First, the image is divided into small blocks of $\sqrt{L}/2 \times \sqrt{L}/2$. Next, for each small block, OMP is performed on the dictionary designed by K-SVD to obtain T_0 sparse coefficients. The dimension of a dictionary is L. After quantization, we obtain sparse coefficients to be encoded for each small block. Let k be the number of nonzero coefficients after quantization. Here, all DC coefficients are quantized by step one and they are always encoded. Also, nonzero AC coefficients are quantized by the quantization step QP. The number of nonzero AC coefficients after quantization is at most $T_0 - 1$. Set the number of blocks in which the number of nonzero coefficients to be coded becomes k among all blocks of the image as $N(k)$. $N(1)$ means the number of the blocks represented by DC coefficients only. The total number of blocks in the whole image, N, is $N = \sum_{k=1}^{T_0} N(k)$, and the number of DC coefficients in the image, N_{DC}, is equal to N. In addition, the number of nonzero AC coefficients in the whole image, $N_{nonzeroAC}$, is expressed by

$$N_{nonzeroAC} = \sum_{k=2}^{T_0} (k-1)N(k)\,. \tag{5}$$

3.1 Syntax of sparse coefficients coding

Figure 5 shows the sparse coefficient coding syntax analyzed in this study. The information required for each block to be encoded are, *class No.*: a class number indicating which of dictionaries is used, k: the number of nonzero coefficients in the block, $coef_{DC}$: a weighting factor for DC atom, and $coef_{AC}$: weighting factors for AC atoms. Also, the number of nonzero AC coefficients is $k-1$, and it is necessary to encode atom indices and quantized coefficient levels for each nonzero AC coefficient. In this study, in order to perform code allocation adaptively by the number of nonzero coefficients for each block, the number of nonzero coefficients in a block, k, is encoded prior to the atom indices of the nonzero coefficients and the quantized coefficients level. For the syntax of AC coefficients, the atom indices for nonzero coefficients and the nonzero quantized coefficients level are encoded. Regarding the atom indices, we will consider two kinds of methods: direct encoding of indices and a method of using zero run length between indices of nonzero coefficients.

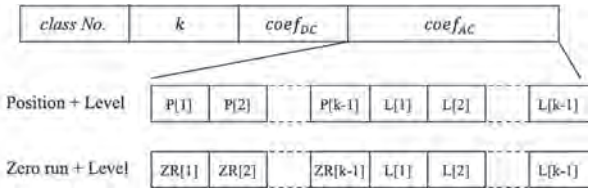
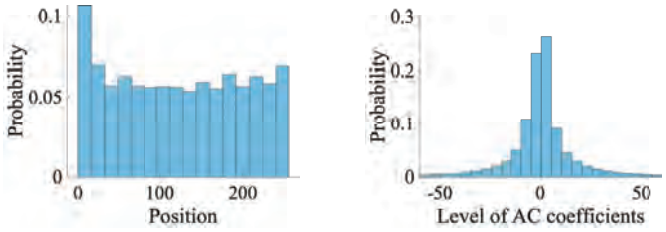


Fig. 5 Bit stream structure for sparse coefficients

3.2 Nonzero coefficients distribution and entropy

In the conventional research[7),15)], the atom indices for nonzero coefficients after quantization and the nonzero quantized coefficient levels are coded independently, and any adaptation of code assignment depending on the number of sparse coefficients in the block and the feature of the atoms has not been studied. In this subsection, we first unify the symbols of all blocks based on the conventional method and analyze the statistical properties of the nonzero coefficients and the amount of generated bits. In this study, the amount of generated information is analyzed based on the entropy calculated from the occurrence probability of the symbols to be coded. The total amount of generated bits for the whole image is expressed as

$$Bit_{all} = Bit_{class} + Bit_k + Bit_{DC} + Bit_{AC}\,. \tag{6}$$



(a) (b)

Fig. 6 Probability histograms of (a) position index and (b) magnitude of nonzero quantized AC coefficients

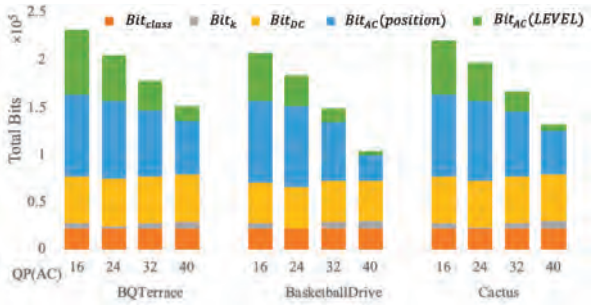


Fig. 7 Number of bits generated

Here, Bit_{class}, Bit_k, Bit_{DC} and Bit_{AC} are the amount of generated bits for class number, the number of nonzero coefficients, DC coefficient, AC coefficient, respectively. The amount of each code bits can be calculated as follows.

First, a class number can be expressed as a fixed-length code of $\log_2 C$ bits per block, where C is the number of classes. The amount of generated bits in the whole image, Bit_{class}, can be calculated as $Bit_{class} = N \times \log_2 C$.

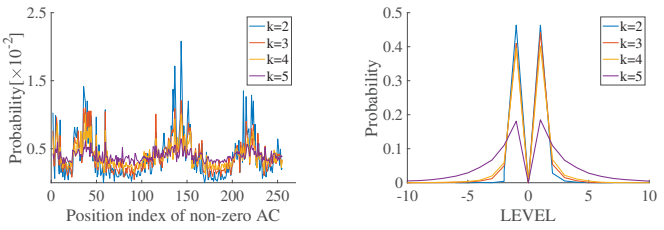
Next, to calculate the number of bits for the number of nonzero coefficients, it is necessary to consider the distribution of the occurrence probability $p(k)$. $p(k)$ changes with the quantization step QP for the coefficients. When the QP becomes smaller, the occurrence probability of large k increases, and as the QP becomes coarser, the occurrence probability of small k increases. The amount of bits for the number of nonzero coefficients in the whole image is calculated as $Bit_k = E_k \times N$, where E_k is the entropy of $p(k)$ as shown following equation.

$$E_k = -\sum_{k=1}^{T_0} p(k) \log_2 p(k) \tag{7}$$

The amount of bits generated for the DC coefficient is calculated as follows. Since the DC coefficients reflect the average value of the block, there is a high correlation between the DC coefficients of adjacent blocks. Therefore, DPCM is performed based on the difference with the previous block. Since the probability distribution of the difference signal is approximated as a Laplacian distribution centered at zero, the total amount of bits is calculated as $Bit_{DC} = E_{DC} \times N$, where E_{DC} is an entropy based on the occurrence probability of differential DC values.

The amount of bits generated for nonzero AC coefficient is calculated from the distribution of their atom indices and coefficient levels. **Figure 6** (a) shows a histogram of atom indices for nonzero AC coefficients measured when sparse coding is performed on the test images by setting $T_0 = 7$ and QP$= 16$. From the results, the occurrence probability of atom indices for nonzero coefficient is almost uniform. Similar measurements were performed for various combinations of T_0 and QP, and a chi-square test was performed for each case. As a result, we could confirm the uniformity of the probability distribution of atom indices as in the conventional study[7]. When uniformity of the occurrence probability distribution of atom indices can be assumed, $\log_2 L$ bits are needed per one atom index theoretically. Therefore, the total amount of bits for atom indices in the whole image, Bit_I, is $Bit_I = \log_2 L \times N_{nonzeroAC}$. Also, Fig. 6(b) shows the distribution of nonzero quantized AC coefficient levels, which can be approximated by the Laplacian distribution centered on zero. Note that there is no zero coefficient. Coarse quantization concentrates the occurrence probability distribution to smaller levels and increases the number of zero coefficients. The amount of bits for nonzero coefficient levels in the whole image, Bit_L, is $E_L \times N_{nonzeroAC}$, where E_L is the entropy of the nonzero AC coefficient levels. The total amount of bits for nonzero AC coefficient in the whole image, Bit_{AC}, is calculated as the sum of Bit_I and Bit_L.

Figure 7 shows the amount of bits generated in the whole image measured by changing QP. The coefficient level becomes smaller when the coarse quantization step is used, so the amount of bits for AC coefficient levels decreases. Similarly, when coarse quantization step is used, the number of nonzero quantized AC coefficients decreases, so the amount of bits for atom indices decreases. Since the quantization step for DC coefficients is always one, Bit_{DC} is constant regardless of the quantization parameter QP for AC coefficients. Bit_k shows a slight increase or decrease because the distribution of the number of nonzero AC coefficients changes depending on the magnitude of QP. Bit_{class} is constant because it is determined only by the number of classes. From Fig. 7, it is clear that reducing the amount of bits for expressing the AC coefficient is very significant.



(a) (b)

Fig. 8 Probability histograms of (a) position index and (b) magnitude of nonzero quantized AC coefficients, after categorizing based on k

In order to reduce the amount of generated bits for the nonzero AC coefficients, it is possible to divide the nonzero AC coefficients into multiple categories according to the number of nonzero coefficients in the block and perform code allocation suitable for each category. Theoretically, if the symbols can be separated into multiple categories so that their occurrence probability distributions are as different each other as possible, the total code amount can be reduced. **Figure 8** shows the distribution of atom indices for nonzero AC coefficients and the distribution of AC coefficient levels, after categorizing based on the number of nonzero coefficients in the block. As shown in Fig. 8, it is clear that the information symbols separation by k has little effect because the probability distribution of the atom's indices corresponding nonzero coefficients is almost same regardless of the value of k. On the other hand, since the occurrence probability of nonzero quantization level numbers show different distributions depending on k, it is considered to be significant to perform the symbol separation by k.

3.3 Sparsity adaptive sparse coefficient coding

Another way to represent atom indices of nonzero coefficients is to use the number of zero coefficients (i.e. zero run length) preceding nonzero coefficients[17]. We analyze the statistics of zero run length when L coefficients are divided by k nonzero coefficients as shown in **Fig.9**. This problem can be solved theoretically as a consequence of the broken stick problem[18] which is an analysis problem concerning the probability distribution of the length for a piece of sub segments when the line segment of length 1 is divided by $n-1$ random points. The probability density function of the length $r(0 \le r \le 1)$ of any divided segments is

$$g(r) = (n-1)(1-r)^{n-2}\,. \tag{8}$$

The probability $P(r_0)$ that the length of each segment becomes $[r_0, r_0+\epsilon)$ is obtained by integration of equation (8) as

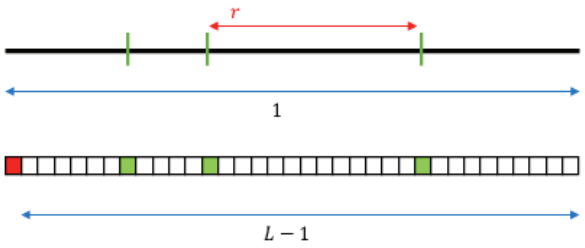


Fig. 9 The probability density function of r, the length of any divided segments

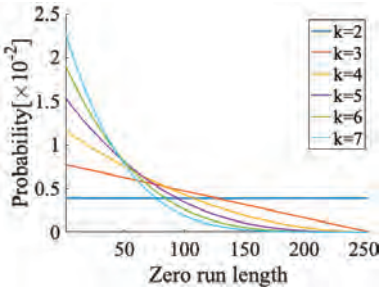


Fig. 10 The theoretical probability distribution of zero run length

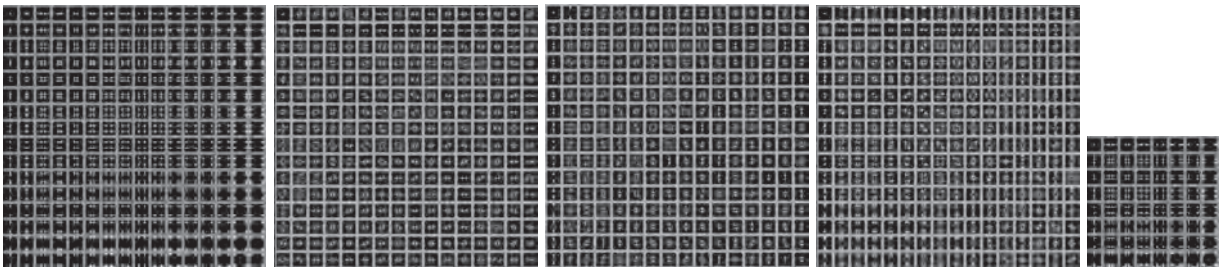
$$\begin{aligned} P(r_0) &= \int_{r_0}^{r_0+\epsilon} g(r)\mathrm{d}r = [-(1-r)^{n-1}]_{r_0}^{r_0+\epsilon} \\ &= (1-r_0)^{n-1} - (1-(r_0+\epsilon))^{n-1} \end{aligned} \tag{9}$$

Applying the above analysis to the distribution of sparse coefficients, the length of the original line segment corresponds to the dimension L of a dictionary, and the length of each divided segment corresponds to the zero run length. **Figure 10** shows the theoretical probability distribution of zero run length when the length of the original line segment is set to $L = 256$. The occurrence probability is found to be a distribution based on an exponential function. In Reference 17), code design is performed by integrating the occurrence probabilities, that is, without classification by the number of nonzero coefficients. However, from Fig. 10, since the parameters of exponential function clearly differ depending on the number of nonzero coefficients in the block, it can be expected that more efficient code assignment for zero run length becomes possible by categorizing nonzero coefficients by k. The entropy of the zero run length is

$$E_{run}(k) = \sum_{i=0}^{L-1} P(i/L)\log_2 P(i/L) \tag{10}$$

where $P(i/L) = (1-i/L)^k - (1-(i/L+1/L))^k$. The amount of bits to represent the atom indices in the whole image is calculated as

$$Bit_I = \sum_{k=2}^{T_0} E_{run}(k-1)N(k)(k-1) \tag{11}$$



(a) (b) (c) (d) (e)

Fig. 11 Power spectrum for atoms; (a) overcomplete DCT (Fig.1(a)), (b) K-SVD(Fig.1(b)), (c) K-SVD(Fig.1(c)), (d) K-SVD(Fig.1(d)), (e) complete DCT

3.4 Adaptive coding by atom features

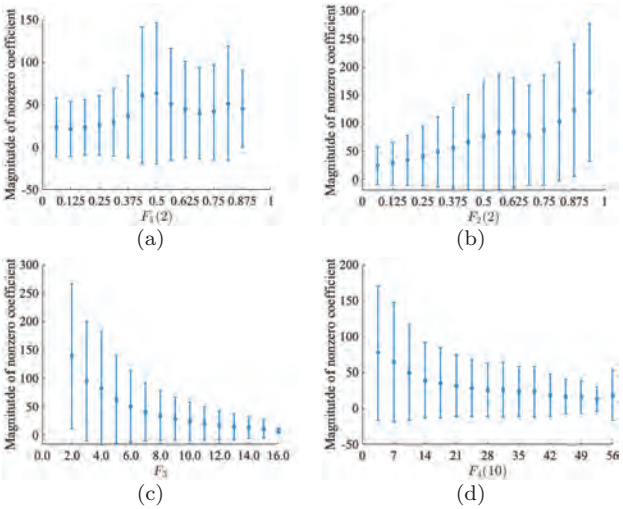
It is known that the features of the atoms appearing in the dictionary designed by K-SVD are strongly influenced by the features of the training samples, and they are different from general atoms such as DCT. **Figure 11** shows the Fourier power spectrum of each atom for the four dictionaries shown in Fig. 1. The center of each spectral image corresponds to the DC component, and the longer the distance from the center, the higher the frequency. For comparison, the power spectrum for complete DCT was added as shown in Fig. 11(e). From Fig. 11, in the dictionary consisting of atoms with regular frequency arrangement such as DCT and overcomplete DCT (Fig. 11(a), (e)), each atom complements each other so as to cover all frequency bands. On the other hand, the overcomplete dictionary designed by K-SVD (Fig. 11(b), (c), (d)) does not necessarily consist of atoms that cover all frequency bands. It can be seen that it is composed of atoms that can express a specific frequency band in more detail. When expressing images using a training-based dictionary, there have been no studies investigating the dependency between the characteristics of the weighting factors and the features of the atoms. If there is a correlation between some features of atom and weighting factors, it is possible to reduce the amount of generated bits by adaptively changing the code assignment to the weighting factors with the atom's features as the context.

Therefore, we first investigated the relationship between some features of atoms and the magnitude of the weighting factors. Let $b(i,j)$ be an $M \times M$ atom in a dictionary designed by K-SVD. The following four features are investigated as the features of each atom,

- Fourier transform:

$$F_1(th) = \frac{\sum_{|u|+|v|\le th} |B(u,v)|^2}{\sum_{u,v} |B(u,v)|^2},$$

where $B(u,v)$ is the Fourier power spectrum of $b(i,j)$, $-M/2 \le u \le M/2$, $-M/2 \le v \le M/2$.



(a) (b) (c) (d)

Fig. 12 Correlation between atom feature and magnitude of nonzero coefficients

- Discrete cosine transform:

$$F_2(th) = \frac{\sum_{(u+v)\le th} |C(u,v)|^2}{\sum_{u,v} |C(u,v)|^2},$$

where $C(u,v)$ is the DCT coefficients of $b(i,j)$, $0 \le u \le M-1$, $0 \le v \le M-1$.

- Total variation:

$$F_3 = \sum_i \sum_j (|b(i+1,j) - b(i,j)| + |b(i,j+1) - b(i,j)|)$$

- Number of strong edge:

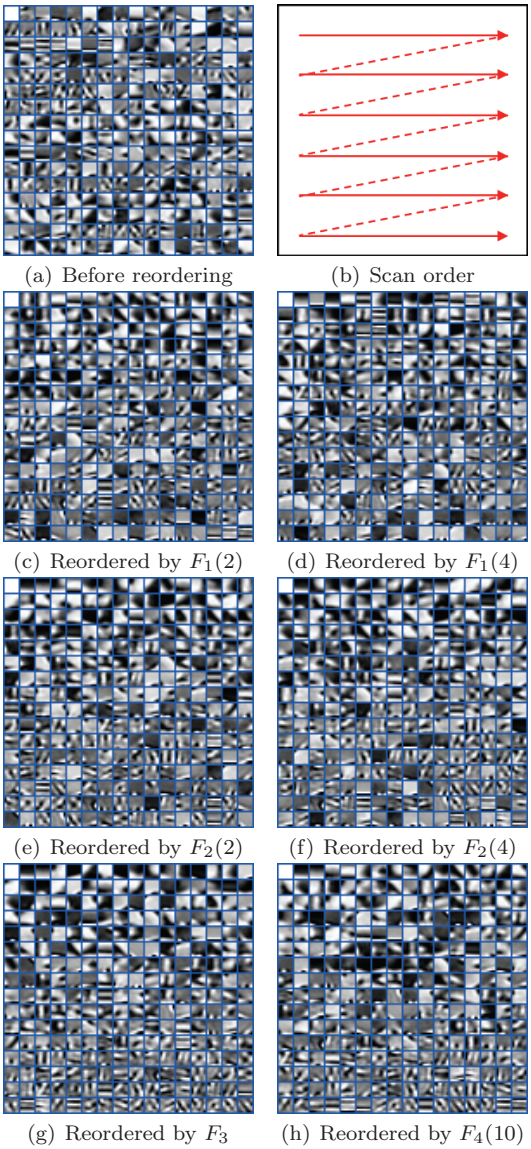
$$F_4(th) = \sum_i \sum_j (m_H(i,j) + m_V(i,j)),$$

where

$$m_H(i,j) = \begin{cases} 1, & if |b(i+1,j) - b(i,j)| > th \\ 0, & otherwise \end{cases}$$

$$m_V(i,j) = \begin{cases} 1, & if |b(i,j+1) - b(i,j)| > th \\ 0, & otherwise \end{cases}$$

Figure 12 shows the correlation between each feature of atoms and the magnitude of the nonzero AC coefficient. Figure 12 also shows the average and the standard



(a) Before reordering (b) Scan order

(c) Reordered by $F_1(2)$ (d) Reordered by $F_1(4)$

(e) Reordered by $F_2(2)$ (f) Reordered by $F_2(4)$

(g) Reordered by F_3 (h) Reordered by $F_4(10)$

Fig. 13 Examples of the atoms reordered by their features

deviation of the absolute value of nonzero AC coefficients generated for the atoms in each section after dividing the feature quantity into 16 sections. In Fig. 12, the results show the case where the parameter th for each feature value is set so that the correlation coefficient becomes the highest. There is a significant correlation between these four feature values and the magnitude of nonzero AC coefficients. Therefore, if we adapt the code assignment to the nonzero AC coefficient levels according to the feature of their corresponding atoms, the amount of generated bits can be reduced. Also, from the observation in Fig. 12, we can consider that more efficient code assignment for the length of zero runs is performed by reordering the atoms so that the coefficients with large absolute values are scanned first. **Figure 13** shows the examples of the atoms reordered by their features. Because the reordering of atoms concentrates nonzero AC

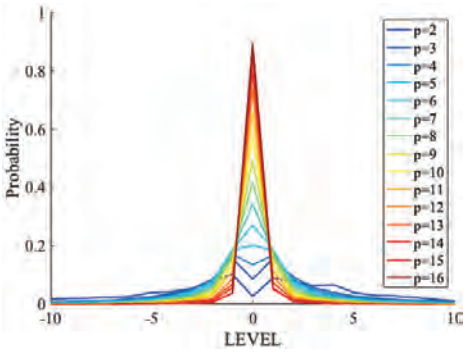
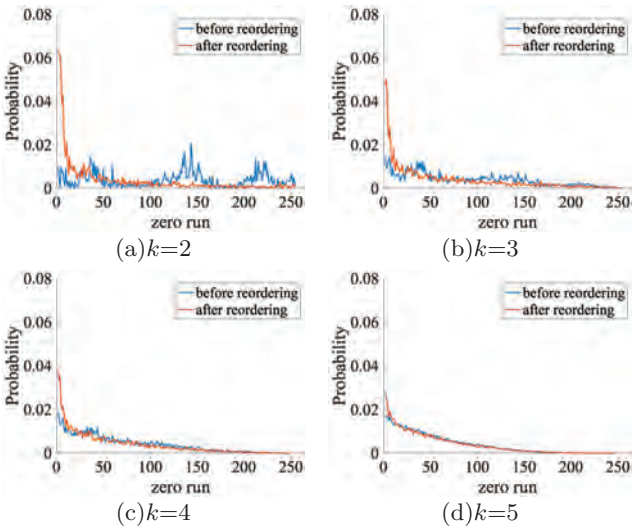


Fig. 14 Probability distribution of level after reordering



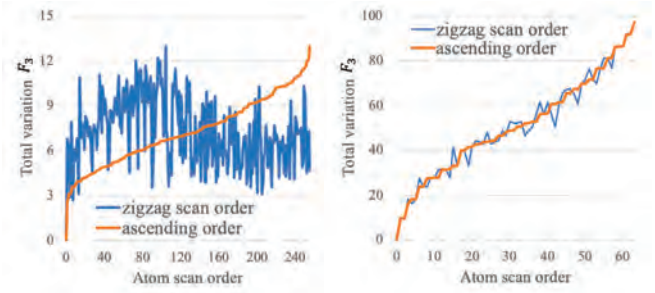
(a)k=2 (b)k=3

(c)k=4 (d)k=5

Fig. 15 Probability distribution of zero run before and after reordering

coefficients at the start of the scan, so the probability of having a short zero run length becomes high. This results in more efficient code assignment to zero run length.

Figure 14 shows the occurrence probability of nonzero AC coefficient levels for the feature value of atom. Here, the feature value of atom utilized is F_3 which showed the strongest correlation from the measurement results shown in Fig. 12. After defining $p = \mathrm{int}(16 \times F_3/max(F_3))$, we measure the probability distribution for each p. This measurement is performed under the condition of $T_0 = 5$ and QP = 32. It is clear that the probability distribution is different depending on the feature value of atoms. In addition, **Figure 15** shows the comparison between occurrence probability of the zero run length under original order and that after reordering the atoms using the feature value F_3. We can find that the zero run length has a distribution that concentrates on smaller values for all k compared to before reordering the atoms. Therefore, it was verified that the adaptive code assignment by considering the atom feature is very significant for reducing both the amount of nonzero AC coefficient level and zero



(a) Dictionary designed by K-SVD (b) Complete DCT

Fig. 16 Atom feature F_3 according to scan order

run length.

Figure 16 (a) shows the measured feature value F_3 for each atom in the dictionary designed using K-SVD. The blue line in Fig. 16 is the result by the conventional zigzag scan order, and the red line is the result by scanning in ascending order of F_3. When arranged in the conventional zigzag scan order, the feature value F_3 fluctuates drastically. As a result, the probability that a coefficient with a large magnitude and a coefficient with a small magnitude will randomly occur becomes higher, and code assignment to zero runs becomes inefficient. If the coefficients are scanned in ascending order of F_3, the probability that coefficients with large magnitude will be concentrated at the beginning of the scan becomes higher, and efficient code assignment can be realized. On the other hand, the measurement results for complete DCT under the same conditions in Fig. 16(a) are shown in Fig. 16(b). We can see that even if the proposed method is applied to complete DCT, the scan order is almost unchanged from the zigzag scan used in the conventional method, and the effect of increasing the coding efficiency is small. The reason why the scan order hardly changes even when the proposed method is applied is that zigzag scan itself is already setting effectively for complete DCT whose atom features are already known. Note that the results in Fig. 16 was confirmed to be the same when not only the feature value F_3 but also other feature values F_1, F_2, and F_4 are used.

4. Experiments

4.1 Experimental conditions

In this section, based on the analysis in section 3, simulation experiments are performed under various conditions to verify the coding efficiency. The experimental conditions are shown in **Table 1**. In order to design the dictionary, a total of 1.2 million 8×8 blocks were extracted from the images of the ITE/ARIB HDTV test materials database[19] as training data, and they were classified by DSURF. The multiclass dictionaries was designed under the number of classes of 16, 32, 64 and 128. In each class, a dictionary is designed by K-SVD with an overcomplete DCT of dimension 16×16 as the initial dictionary. The sparse constraint parameter T_0 was set to 3, 5, and 7.

Table 1 Simulation conditions

Training data	1.2M 8×8 blocks from 18)
Feature for classifier	DSURF
Number of classes C	$16, 32, 64, 128$
Initial dictionary	16×16 overcomplete DCT
T_0	$3, 5, 7$

If the number of samples used for training is too small, the image representation performance by the designed dictionary will be degraded, and as a result meaningful experiments for this study will not be possible. If the number of samples used for training is large enough and various features of general images are well-balanced in them, the dictionary created by learning will converge to a versatile optimal solution. In this study, the ITE test image database was used for training. This is because it is composed of images with various features targeted for codec evaluation, etc., and it is considered that the features of images that are generally used can be captured sufficiently by using all these images for training. In the conventional studies, training is performed using samples of tens of thousands of blocks (for example, about 68000 blocks in Reference 7) and one hundred thousand blocks in Reference 17)). On the other hand, the number of 1.2 million blocks used in this study is sufficiently large compared to the number of blocks used in the conventional studies. So, it is considered that an appropriate dictionary is designed for entropy coding research, which is the focus of this paper.

When encoding, first, the image to be encoded is divided into 8×8 small blocks, and their class number are determined by the k-means method according to the DSURF feature of the small block. Next, using OMP and the selected class dictionary, we obtain T_0 sparse coefficients that approximate the small block to be encoded. In OMP, DC atom is always used. Therefore, the number of AC coefficients is $T_0 - 1$. The obtained DC coefficients are quantized with quantization step 1 (i.e., rounded to the nearest integer). On the other hand, AC coefficients are quantized with quantization step QP. In this experiment, we set QP = 16, 24, 32, 40. Under these pa-

rameters, the effectiveness of introducing zero runs, the effectiveness of adaptation with the number of nonzero coefficients, and the effectiveness of adaptation by feature of atoms feature are examined, in comparison with conventional entropy coding. Note that, in this research, the image quality does not change in case the same QP is used, so the effectiveness can be verified only based on the amount of generated bits. As shown in Section 3, the amount of generated information is calculated by the entropy based on the occurrence probability of the symbols to be coded. In the original documents (References 7) and 17)) of the conventional methods to be compared, Huffman codes and Golomb-Rice codes are assigned to the generated symbols. However, for the conventional methods in this experiment, instead of actually assigning a code bit, the amount of information is calculated based on the entropy of the generated symbol in order to make a fair comparison.

4.2 Experimental results and discussion

First, under the conditions fixed at $T_0 = 5$ and QP=24, we measured the effectiveness of *k-adaptation*, i.e. the adaptive encoding by the number of nonzero coefficients. **Figure 17** (1) shows the result of the conventional method based on Reference 7), in which the index of the atom corresponding to the nonzero coefficient is directly encoded. Figure 17(2) shows the result of the conventional method based on Reference 17), in which the zero run length between nonzero coefficients is encoded. The k-adaptation is not performed in both Fig. 17(1) and Fig. 17(2). On the other hand, Fig. 17(3) shows the result of applying k-adaptation to Fig. 17(2). We found from Fig. 17 that the introduction of zero run length can reduce the amount of information generated for the indices, so the total amount of information decreases accordingly. However, it should be noted that the amount of information for the quantized level number of the nonzero coefficients has not been reduced. On the other hand, introduction of k-adaptation can not only reduce the amount of information for zero run length but also reduce the amount of information for level number, as a result it is possible to reduce the total amount of information up to 11.0% compared to Reference 7) and up to 4.7% compared to Reference 17). These characteristics were also found to be similar when using different T_0 and QP.

Next, we verified the effectiveness of adaptation based on the feature of the atoms. The experiment was performed under the condition that the zero run length and the nonzero AC coefficient level were classified according to the number of nonzero coefficients in each block, and they were encoded independently. We created a dictionary in which the atoms were reordered using the four features defined in section 3, and compared the amount of bits using the new dictionary with the amount of bits using the original dictionary. The measured results are shown in **Table 2**. The column (a) in Table 2 shows the amount of information generated by the conventional method shown in Reference 17), the column (b) in Table 2 shows that generated when k-adaptation is applied to the conventional method, and the column (d) of Table 2 shows the amount of generated information when atom reordering is performed in addition to k-adaptation method. The column (c) and (e) in Table 2 show the reduction rate of the amount of generated information for the column (b) and (d) based on the column (a), respectively. Table 2 shows that reordering of atoms by any of the four features makes it possible to reduce the amount of generated information compared to before reordering. In particular, it can be confirmed that the amount of generated bits can be minimized when using the atom feature value F_3. The reason is that, the feature value F_3 is highly correlated with the nonzero AC coefficient level as described in section 3. As a result, the amount of bits nonzero AC coefficient levels can be reduced by adopting different code assignment rules for them according to the feature of atoms. Also, the reordering of atoms can concentrate the distribution of zero run length closer to zero, and leads to a reduction the amount of bits for zero run length.

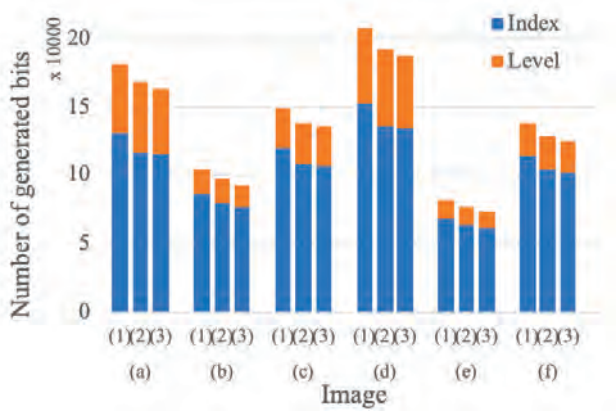


Fig. 17 Number of generated bits by (1) conventional[7], (2) conventional(zero run)[17] and (3) k-adaptive, for (a) BQTerrace, (b) BasketballDrive, (c) Cactus, (d) ChristmasTree, (e) Kimono1 and (f) ParkScene. $T_0 = 5$, QP=24

Table 2 Number of generated bits (kbit); (a) conventional[17], (b) k-adaptation, (c) bit saving ratio(%): ((b)-(a))/(a), (d) k-adaptation+atom reordering with the feature F_1, F_2, F_3, F_4, (e) bit saving ratio(%): (F_3-(a))/(a)

Image		(a)	(b)	(c)	(d)				(e)
					F_1	F_2	F_3	F_4	
BQTerrace	Index	117.2	115.8	-1.2	115.7	115.3	115.1	113.8	-1.8
	Level	50.5	47.6	-5.7	46.3	44.0	44.2	45.5	-12.5
	Total	167.7	163.4	-2.5	162.0	159.3	**159.2**	159.3	-5.0
BasketballDrive	Index	79.8	76.9	-3.6	77.6	73.4	73.6	73.6	-7.7
	Level	17.7	16.0	-9.7	15.1	12.9	12.9	13.4	-27.1
	Total	97.5	92.9	-4.7	92.7	**86.4**	86.5	87.0	-11.3
Cactus	Index	108.4	106.5	-1.8	104.9	104.0	103.4	103.7	-4.6
	Level	29.9	28.7	-3.8	27.3	24.5	24.7	25.8	-17.5
	Total	138.3	135.2	-2.2	132.2	128.5	**128.1**	129.5	-7.4
ChristmasTree	Index	136.4	134.8	-1.1	134.2	134.5	134.0	133.7	-1.7
	Level	55.5	52.0	-6.3	51.0	49.6	49.4	50.4	-11.0
	Total	191.9	186.8	-2.6	185.2	184.1	**183.5**	184.2	-4.4
Kimono1	Index	63.6	61.6	-3.2	60.1	58.6	58.4	59.3	-8.2
	Level	13.1	12.3	-6.2	11.5	9.5	9.6	10.5	-26.9
	Total	76.7	73.9	-3.7	71.5	68.1	**68.0**	69.8	-11.4
ParkScene	Index	104.6	102.3	-2.2	100.0	99.5	98.8	99.7	-5.5
	Level	24.0	23.0	-4.2	21.7	18.9	18.9	20.3	-21.2
	Total	128.6	125.3	-2.6	121.7	118.4	**117.7**	120.0	-8.4

Table 3 BD-rate[%] between proposed method and Reference 17)

Image	T_0	Number of class			
		16	32	64	128
BQTerrace	3	-1.43	-1.35	-1.37	-1.31
	5	-5.50	**-6.27**	-5.56	-6.10
	7	-4.27	-4.21	-4.16	-4.11
BasketballDrive	3	-2.03	-2.07	**-3.45**	-1.33
	5	-0.48	-0.72	-1.08	-0.81
	7	0.44	0.23	0.17	-0.35
Cactus	3	-2.37	**-3.73**	-2.94	-3.67
	5	-3.12	-3.27	-3.53	-3.26
	7	-3.72	-3.36	-3.41	-3.28
ChristmasTree	3	-1.13	-0.83	-0.83	-1.06
	5	-1.26	**-4.97**	-4.27	-1.17
	7	-3.02	-3.00	-3.26	-2.41
Kimono1	3	**-3.13**	-3.08	-2.22	-2.70
	5	-3.00	-2.90	-2.86	-2.61
	7	-3.02	-2.72	-2.62	-2.58
ParkScene	3	**-4.90**	-1.94	-3.87	-3.01
	5	-3.30	-3.80	-3.41	-3.37
	7	-3.53	-3.57	-3.43	-3.50

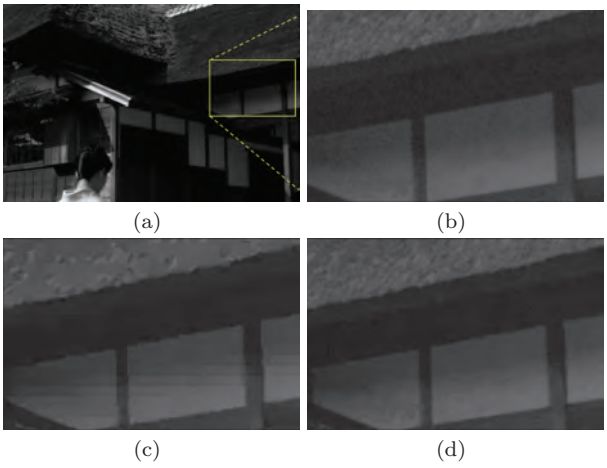


Fig. 18 Image quality comparison (0.30 bit/pel); (a) original, (b) enlargement of partical original image, (c) decoded image of (b) (conventional[17]), and (d) decoded image of (b) (proposed)

Finally, we measured the overall performance under setting the feature value used to reorder the atoms to F_3 which is the most effective to reduce the amount of bits. The class number C was set to four types of 16, 32, 64 and 128. For each C, the sparse parameter T_0 was set to 3, 5 and 7, and the quantization parameter QP was set to 16, 24, 32 and 40. The total amount of bits generated was measured as the sum of the amount of bits for the class number, for the number of nonzero coefficients, for coefficients of DC atom and for coefficients of AC atom. The measured average performance gain, BD-rate, between the proposed method and the conventional method is shown in **Table 3**. Table 3 shows that the proposed method can reduce the total amount of bits up to 6.2% compared to the conventional method.

K-SVD is a block based processing similar to DCT-based coding, so the block noise occurs when the compression ratio becomes high. Using the proposed entropy coding method, a smaller quantization step can be used in comparison with the conventional entropy coding methods under the same compression ratio. As a result, block noise can be reduced as shown in **Fig. 18**.

In the experiments by combining the number of classes (C = 16, 32, 64, 128), quantization step (QP = 16, 24, 32, 40) and sparsity (T_0 = 3, 5, 7) as experimental pa-

rameters, we clarified that the proposed entropy coding is effective at any bit rate from high compression to low compression. When K-SVD is applied to actual compression coding, multiple parameters of the number of dictionary classes, the quantization step and the sparsity parameter must be controlled in order to keep the amount of generated bits within a predetermined compression ratio. It is considered that the proposed entropy coding method can be utilized for the rate-distortion optimization control for image compression with K-SVD, it will be addressed as a future work.

5. Conclusions

In this paper, we focused on an efficient entropy coding for sparse coefficients when sparse coding is applied to image coding. First, the statistical properties of the sparse coefficients under various sparsity parameter and quantization step were analyzed in detail. Next, based on the analysis, we proposed two methods of the adaptive code assignment with the number of nonzero coefficients in the block and reordering of atoms by their features. The proposed methods enable to encode the indices and quantized levels of the nonzero sparse coefficients efficiently. Finally, by experiments using various sparsity parameters and quantization width, it was clarified that the amount of generated bits can be reduced up to 6.2% compared with the conventional method. The application to sparse coding for intra-frame/inter-frame prediction error, and the application to ultra-high definition video such as 4K and 8K will be studied as interesting researches in the future.

References

1) ISO/IEC 10918-1 | ITU-T Recommendation T.81: "Information Technology - Digital Compression and Coding of Continuous-tone Still Images: Requirements and Guidelines" (1994).
2) ISO/IEC 14496-10: "Information Technology - Coding of Audio-Visual Objects - Part 10: Advanced Video Coding" (2014).
3) ISO/IEC 23008-2: "Information Technology – High Efficiency Coding and Media Delivery in Heterogeneous Environments – Part 2: High Efficiency Video Coding" (2017).
4) Versatile Video Coding (VVC) | JVET, https://jvet.hhi.fraunhofer.de (2019).
5) O. Bryt, M. Elad: "Compression of Facial Images Using the K-SVD Algorithm", Journal of Visual Communication and Image Representation, Vol.19, No.4, pp.270–282 (2008).
6) M. Kalluri, M. Jiang, N. Ling, J. Zheng, P. Zhang: "Adaptive RD Optimal Sparse Coding with Quantization for Image Compression", IEEE Trans. on Multimedia, Vol.21, No.1, pp.39–50 (2019).
7) Je-Won Kang, M. Gabbouj, C.C.J. Kuo: "Sparse/DCT (S/DCT) Two-Layered Representation of Prediction Residuals for Video Coding", IEEE Trans. on Image Processing, Vol.22, No.7, pp.2711–2722 (2013).
8) M. Aharon, M. Elad, A. Bruckstein: "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", IEEE Trans. on Signal Processing, Vol.54, No.11, pp.4311–4322 (2006).
9) Y.C.C. Pati, R. Rezaiifar, P.S.S. Krishnaprasad: "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition", Proc. of 27th Asilomar Conference on Signals, Systems and Computers, pp.1–5 (1993).
10) A. Vedaldi, B. Fulkerson: "Vlfeat: An Open and Portable Library of Computer Vision Algorithms", Proc. of the International Conference on Multimedia, p.1469 (2010).
11) J.W. Kang, C.C.J. Kuo, R. Cohen, A. Vetro: "Efficient Dictionary Based Video Coding with Reduced Side Information", IEEE International Symposium of Circuits and Systems (ISCAS), pp.109–112 (2011).
12) B. Olshausen, D. Field: "Natural Image Statistics and Efficient Coding", Network: Computation in Neural Systems, Vol.7, No.2, pp.333–339 (1996).
13) ISO/IEC 13818-2: "Information Technology – Generic Coding of Moving Pictures and Associated Audio Information" (2000).
14) J. Sole, R. Joshi, N. Nguyen, T. Ji, M. Karczewicz, G. Clare, F. Henry, A. Duenas: "Transform Coefficient Coding in HEVC", IEEE Trans. on Circuits and Systems for Video Technology, Vol.22, No.12, pp.1765–1777 (2012).
15) N. Pati, A. Pradhan, L.K. Kanoje, T.K. Das: "An Approach to Image Compression by Using Sparse Approximation Technique", Procedia Computer Science, Vol.48, pp.769–775 (2015).
16) A. Budillon, G. Schirinzi: "Low Bit Rate SAR Image Compression Based on Sparse Representation", Digital Image Processing, pp.51–70 (2012).
17) R. Vinith, A. S. Aswani, K. Govindan: "Medical Image Compression Using Sparse Approximation", International Journal of Advanced Computer and Mathematical Sciences, Vol.6, No.2, pp.30–39 (2015).
18) D. Webb: "The Statistics of Relative Abundance and Diversity", Journal of Theoretical Biology, Vol.43, No.2, pp.277–291 (1974).
19) HDTV Test Materials for Assessment of Picture Quality, https://www.ite.or.jp/content/test-materials (2018).

(Received May 31, 2019)
(Revised August 27, 2019)

Ji WANG
He received the B.E. and M.E. degrees from Chiba Institute of Technology, Chiba, Japan, in 2012 and 2014, respectively. Currently he is studying in Graduate School of Information and Computer Science, Chiba Institute of Technology. His research interests include next generation video coding, image coding.

Yoshiyuki YASHIMA (*Member*)
He received the B.E., M.E., and Ph.D degrees from Nagoya University, Nagoya, Japan, in 1981, 1983 and 1998, respectively. In 1983 he joined the Electrical Communications Laboratories, Nippon Telegraph and Telephone Corporation (NTT), Kanagawa, Japan, where he has been engaged in the research and development of high quality HDTV signal compression, MPEG video coding algorithm and standardization. He was also a visiting professor of Tokyo Institute of Technology during 2004-2007. In 2009, he moved to Chiba Institute of Technology. Currently, he is a professor at the Faculty of Information and Computer Science, Chiba Institute of Technology. His research interests include next generation video coding, pre- and post-processing for video coding, processing of compressed video, compressed video quality metrics and image analysis for video communication system. He was awarded the Takayanagi Memorial Technology Prize in 2005, and received the FIT Funai Best Paper Award in IEICE in 2008. Professor Yashima is a Fellow member of the IEICE, a senior member of the IEEE, and a member of ACM, IPSJ and ITE.

IIEEJ Paper

Value Estimation of SpO2 Using a Non-Contact Method : Comparison with a Contact Method

Yoshimitsu NAGAO[†‡] (*Member*), Yanan GAO[†‡], Jiang LIU[†‡] (*Member*), Shigeru SHIMAMOTO[†‡]

[†] Department of Computer Science and Communications Engineering, WASEDA University
‡ Graduate School of Fundamental Science and Engineering, WASEDA University

<Summary> It is impossible to estimate arterial oxygen saturation (i.e., SpO2) for individuals by using conventional approaches unless the given sensor of the pulse oximeter is attached to an individual's finger. This study introduces a novel method to solve this problem. This study has focused on realizing SpO2 measurements by using non-contact space measurements, and the success of the approach is validated through experiments. Finally, despite a few problems including the susceptibility of the proposed approach to other light interference, the study offers an initial method to utilize laser wavelengths for the fore-mentioned purposes. As the characteristic of elderly individuals involves the hardening of the fingertips' skin, it is difficult for the light of a probe to enter the same. Therefore this study can be applied to medical care, elder care, and other related fields. .Additionally, there are cases that symptoms are unmeasurable. The light receiving property to other light interference in space constitutes a problem for the fore-mentioned method.

Keywords : SpO2, saturation pulse, non-contact space measurement, light sensing, biological information, LMM, pulse oximeter, blood oxygen saturation

1. Introduction[1)-4)]

1.1 Background of starting this research

Initially, a trigger of this research is our following experience. When the author visited a considerable elderly in the hospital, the nurse started a measurement by attaching a probe to the old man's finger to measure SpO2 (Oxygen saturation from the pulse). Unfortunately, the saturation value was not able to be read. The reason was that as the human beings get aged, the wall thickness increases and becomes harder, so it is difficult for the light emitted into the probe to enter the finger, which makes it impossible to measure the value. This kind of thing is likely to happen often. Under such circumstances, if we cannot know biological information as to whether the necessary oxygen has been brought into the human body or not, it is considered to be a problem related to life and death. Knowing this kind of situation, we started this research, looking for a non-contact measurement method to read the SpO2 values.

1.2 Spatial measurements of blood oxygen saturation SpO2

Existing probes are equipped with excited individual light-emitting elements of red and infrared LEDs. They use the different absorption rate of red light and infrared light due to the binding of oxygen and hemoglobin in human blood to measure blood oxygen saturation degree (i.e., SpO2 of the artery (oxy-hemoglobin) and vain (Deoxy-hemoglobin) values) by using contact-based methods. Furthermore, only contact-based methods have been used for these types of measurements.

Only a few impractical approaches were proposed by extant studies, and there is paucity of studies examining contactless approaches. Therefore, in this research, a light emitting element transmits light to penetrate human fingers and measures it by using a non-contact method with two different wavelengths, and it will be make possible to calculate the values of SpO2. Thus, the study proposes a completely different approach to measure the target SpO2 values. Light is measured by detecting transmitted light in the red and infrared spectra.

1.3 Measurement principles and types of SpO2 values

The light-receiving element transmits light to the measurement object medium and extracts the component of light that is not absorbed after penetrating the measurement object, with respect to the molar extinction coefficient of the molecule and wavelength of the light, the artery, vein and transmittance wavelength characteristics. The transmittance exhibits a wide absorption band range from 640 [nm] to 1600 [nm] or more. The standard deviation corresponds to the number of received data and is in the interval between the peak of the heartbeat's waveform and the time of the peak. It is crucial to obtain information with respect to the amount of oxygen that is supplied to the blood to measure the state of the living body. The index is termed as the arterial blood oxygen saturation or SpO2. The device that is currently used can continuously and non-invasively measure SpO2 and is called a pulse oximeter. The SpO2 values are calculated by sandwiching the finger between irradiance devices to irradiate two types of lights with different wavelengths, and then by measuring the amount of transmitted light to get the result. Although the methods used in extant studies correspond to the first method that uses laser light, they are facing an inherent problem as they are susceptible to interferences from other light sources in the environment. This is considered as a primary difficulty in realizing contactless approaches today. The human body is mainly composed of water. The

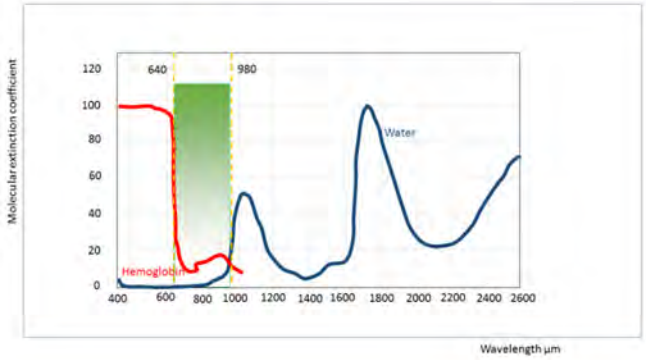


Fig. 1 Characteristics of the transmittance wavelength of contact sensors

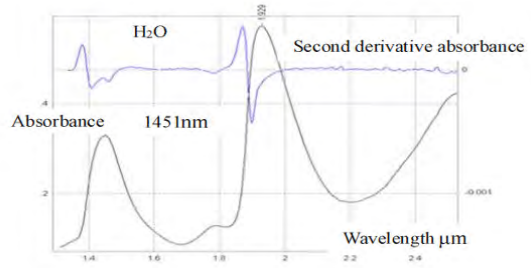


Fig. 2 H_2O spectrum as obtained by near-infrared spectroscopy

percentage of water corresponds to approximately 90% for fetuses, 75% for newborns, 70% for children, 60% to 65% for adults, and 50 to 55% for elderly individuals. This indicates that the moisture ratio reduces with aging, and thus it is difficult to measure the SpO2 values with aging. The wavelength and absorbance of blood and water are shown in **Fig. 1** and **Fig. 2**.

2. System Architecture (PPG and LMM)

2.1 Current SpO2 value measurement method for measuring blood oxygen saturation

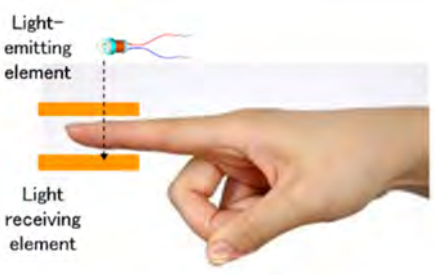
Currently, contact sensors comprise of photo plethysmography (PPG) that uses conventional photoelectric volume pulse waves. These types of sensors are incorporated into a probe and clip-on device that should be attached to an earlobe or a fingertip. As shown in Fig.1, and Fig.2, there are two measured wavelengths, namely a wavelength that approximately corresponds to 650 [nm] in the red light range and another wavelength that approximately corresponds to 940 [nm] in the infrared range. As shown in **Fig. 3**, using a pulse oximeter, it will be performed that the finger interposes between two probes as shown later. Then the measurement value and the output value of the acquired data are compared. The size of these devices approximately corresponds to 6.6 [cm] × 2.9 [cm] (Fig. 3 (a)).

2.2 Laser beam measurement method (LMM)

2.2.1 Fear of danger during operation by using laser

Lasers and amplifier light sources generate light that

(a)



(b)

Fig. 3 A pulse oximeter; (a): product, (b):the principle of the contact SpO2 value measurement

does not exist in nature. When a stimulus such as high intensity light is introduced into the laser cavity, the laser is radiated and light is generated. Although the output is light, the laser and amplified light are greatly different from sunlight and bulb light. Laser and amplified light have special characteristics, so there is a danger for the operation of the equipment, during servicing. A laser or amplified light source generates visible light or invisible light of monochromatic light with very high intensity. Laser and amplified light are coherent, meaning that light waves are in phase with each other.

At present, standards for correctly manufacturing and using laser products are defined globally by IEC 60825-1. EU countries also comply with this (EN-60825-1), Japan complies with IEC 60825-1 compliant JIS C 6802-1:2005 "Radiation safety standard of laser products"[7] (revised January 20, 2005) of radiation safety standards.

Among them, from the viewpoint of safety, it is a class defined for a visible laser and visible light with a single pulse (pulse width 10^{-7}[s]) of 1 [mW] or less, and 2 [W] or less in He - Ne, against a blinking time of 0.25[s]. For the safety to ensure that the retina is not damaged in this research, we are taking into consideration within the scope of Class 2[7].

2.2.2 A mechanism for performing space measurement with two wavelengths

In this study, two lights with different wavelengths are emitted in a non-contact manner by a SpO2 spatial measurement transmission method that is proposed based on the laser transmission method. Specifically, as depicted

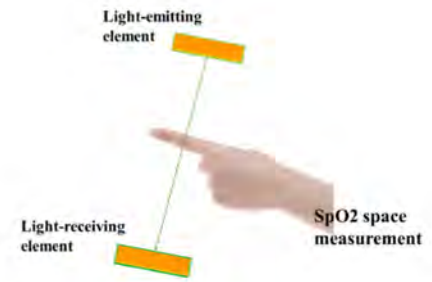


Fig. 4 Illustration of the manner in which SpO2 value is measured in the proposed approach

Table 1 Measurement principles of different light emitting systems

	Conventional method Photoelectric volume pulse wave sensor PPG: (Photo plethysmography) Contact and transmissive (mainstream), reflective		Proposed method (the present study) Laser beam measurement method(LMM) Non-contact and transmissive	
Emission system	Red LED(R)	IR LED(IR)	Red laser LM-101-A2 Diode-pumped solid-state laser	Infrared laser PPLS Diode-pumped solid-state laser
The light-receiving element sensitivity Sensitivity wave-length range Maximum sensitivity wavelength	Phototransistor 1000~4300μA	Phototransistor 145~580μA	Si PIN Photodiode 320~1100nm 960nm	Si PIN Photodiode 320~1100nm 960nm
Emission wavelength output Power supply	660±3nm	900~940±10nm	640nm 1mW or less DC3V 40mA or less	980nm 5mW or less DC3~5V 40mA or less
Measurement object	artery	vein	artery	vein

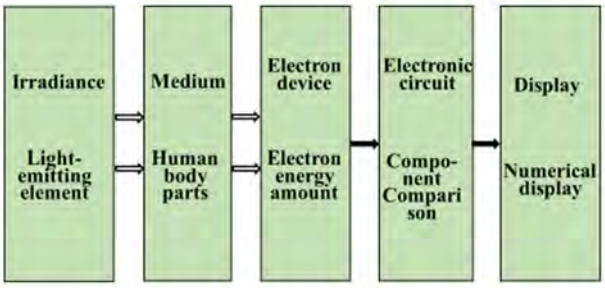


Fig. 5 Conceptual diagram of our non-contact SpO2 measurement system

in **Fig. 4**, light is transmitted through a finger that serves as a medium and is measured by the light-receiving element, and the target SpO2 value is obtained without using a probe. In this method, the oxygen saturation of arterial blood and venous blood is measured by utilizing the different absorption rates between red light and infrared light due to the binding of hemoglobin in the oxygen to the oxygen in the blood to detect the transmitted light. Given the non-contact approach, by measuring two lights with different wavelengths that penetrated after the finger, it will help to measure the target SpO2 value.

Compared with the conventional methods, this study proposes a non-contact approach that uses an infrared laser beam of 980 [nm] (1[mW]) and a laser light of 640 [nm] (1[mW] or less). Recent studies confirmed that near-infrared light (i.e., 700 [nm] to 1,500 [nm]) exhibits high permeability in living tissues.

Oxygen concentrations in living tissues are measured by using light in this region. Thus, advancements are expected in terms of new non-invasive measurement techniques (as shown in **Table 1**).

LMM is radiated by red and infrared diode laser excited elements (as shown in **Fig. 5**). In the study, due to non-contact, the difference in light transmitted through the finger (which is the medium over which light of two different wavelengths of output is radiated) is measured, and the target SpO2 value is also measured. An infrared laser beam of 980 [nm] (1 [mW]) is used as the laser light of 640 [nm] (1[mW] or less). Recent studies indicated that near-infrared light (from 600 to 1500 [nm]) has a high permeability with respect to living tissues and that oxygen concentration in living tissue is measured by using light in this region. Thus, extant studies endeavored to use technological advancements to establish a new noninvasive measurement technique. In this study, LMM is used to realize the measurement in space (as shown in **Fig. 6**). Experiments indicate that the light receiving element is significantly affected by the surrounding light environment. Additionally, the study involved developing a structure that reduces the influence of the surrounding light environment by attaching a special cover and fabricating a light receiving element module to the element[8].

3. Deriving SpO2 Values by Irradiance Passing through a Medium

3.1 Measurements of laser light

Firstly, SpO2 is defined as the SpO2 value that is derived by a calculation formula if the measured value is obtained. It is necessary to accurately analyze waveforms and numerical values via measurements of laser light. The principle of deriving SpO2 value is as follows: A fingertip is irradiated with two lights, i.e., Oxy-hemoglobin and Deoxy-hemoglobin, that correspond to red light and infrared light, respectively, based on the absorption characteristics of light and from the ratio of the magnitude of the pulse wave of transmitted light to arterial blood and vein to calculate absorption characteristics. Theoretically, oxygen saturation corresponds to 100% when the total hemoglobin contains oxygen and changes to Oxy-hemoglobin. The SpO2 value, an indicator of the amount of oxygen present in blood is then determined from the amplitude ratio of the pulse wave due to light at wavelengths of 640 [nm] and 980 [nm] that are irradiated based on the oxygen saturation of arterial blood and venous blood, respectively.

With respect to the oxygen saturation of arterial blood and venous blood, 70% of the gas in venous blood consists of red darkened with oxygen. It should be noted that the numerical value of SpO2 is expressed as a percentage. The

reference value typically ranges from 97% to 99% for a healthy individual, and it is largely considered that pulmonary function declines if the reference value of an individual equals or is lower than 90%.

3.2 Mechanism of derivation of SpO2 value using Lambert-Beer's law[4), 9), 10)]

This section compares incident light to transmitted light in a specific concentration of a solution. If the absorption coefficient of the given solution at a specific wavelength is determined in advance, then it is possible to obtain the concentration of the solution by measuring the incident light, transmitted light, and the distance of the solution. Thus, arterial blood oxygen saturation of SpO2 is expressed by *B* will be given as follows:

$$B = \frac{\Delta Coxy \cdot L^{P1-P2}}{\Delta Coxy \cdot L^{P1-P2} + \Delta Cdeoxy \cdot L^{P1-P2}} \quad . \quad (1)$$

This is followed by obtaining the maximum amplitude of each hemoglobin change within a heartbeat:

$$M = \Delta Coxy \cdot L^{P1-P2} + \Delta Cdeoxy \cdot L^{P1-P2} \quad , \quad (2)$$

where, $\triangle$Coxy $\cdot$ $L^{P1\text{-}P2}$ represents the red laser and amount of the change, and $\Delta Cdeoxy \cdot L^{P1-P2}$ represents the infrared laser and amount of the change.

Furthermore, a state in which hemoglobin protein which is present in erythrocytes, a protein consisting of four peptide chains and heme is bound, oxygen is bound to the iron atom at the center of heme.

The Deoxyhemoglobin estimation Hb (vein) involves checking the waveform with an oscilloscope by using red laser light and infrared laser light that are not coupled with oxygen.

Additionally, two changes, namely *ΔCoxy* and *ΔCdeoxy* are observed in the waveform of the oscilloscope when a finger is held between the light-emitting element and the light-receiving element. In the fore-mentioned case, the two components are transmitted normally, and thus detection is possible.

As illustrated in Fig. 6, the Lambert-Beer law is used to obtain the incident light *Iin* to penetrate a solution with the constant concentration to determine the transmitted light by measured absorbance of the solution. This is expressed as follows:

$$A = -\log(I\,out / I\,in) = \varepsilon C L, \quad (3)$$

where *Iin* is the initial light intensity, *Iout* is the light intensity after passing through the solution, *A* represents the absorbance, *ε* represents the extinction coefficient of the solution, *C* represents the concentration of the solution, *L* represents the distance (i.e., the average optical path length). The extinction coefficient of the solution is previously determined at a specific wavelength epsilon.

Subsequently, *Iin* is used to determine concentration *C* of the solution by measuring *Iout* via *L*. the modified Lambert-Beer Law, if Lambert-Beer law is applied to the scattered medium, will be as follows:

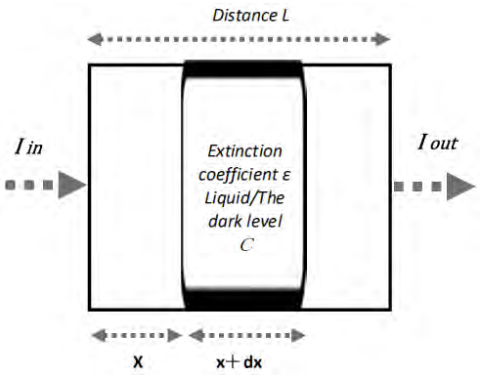


Fig. 6 Vital signs obtained from the density of the given liquid

$$A' = -\log(I\,out / I\,in) = \varepsilon C L + S, \quad (4)$$

where, *S* represents the attenuation of the light intensity due to scattering.

When the concentration of the solution changes from *C* to *(C + ΔC)*, the quantity of transmitted light changes to *Iout + ΔIout*, and the relation is given as follows:

$$Ar = -\log[(I\,out + \Delta I\,out) / I\,in] = \varepsilon (C + \Delta C) L + S, \quad (5)$$

The attenuation (*S*) of the light intensity due to scattering is not changed in Eq. (2) and Eq. (3), and this fact results in the following expressions:

$$R = -\log[(I\,out + \Delta I\,out) / I\,in] = \varepsilon \Delta C L \quad , \quad (6)$$

$$R(\lambda) = -\log[(I\,out(\lambda) + \Delta I\,out(\lambda)) / I\,in(\lambda)] = (\varepsilon\,oxy(\lambda) \cdot \Delta Coxy + \varepsilon\,deoxy(\lambda) \cdot \Delta Cdeoxy) \cdot L \quad . \quad (7)$$

From the incident light that illuminates living tissue of a specific wavelength *λ*, it is possible to know the amount of absorption due to scattering in the living body and the amount of permeation to the outside of the living body.

In Eq. (6) as shown above, the concentration variation *C oxy* of OxyHb determines the concentration amount of change *ΔCdeoxy* of DoxyHb. Specifically, given that two variables are used to determine *ΔCdeoxy*, with respect to *ΔCoxy* the red-infrared extinction coefficient is used at the two wavelengths corresponding to 640 [nm] and 980 [nm], and this results in the following expressions of Eq. (8) and Eq. (9), respectively :

$$R(980) = -\log[(I\,out(\lambda_{980}) + \Delta I\,out(\lambda_{980}))\,/\,I\,out(\lambda_{980})] = (\varepsilon\,oxy\,(\lambda_{980}) \cdot C\,oxy + \varepsilon\,deoxy\,(\lambda_{980}) \cdot \Delta C\,deoxy) \cdot L\,, \quad (8)$$

$$R(640) = -\log[(I\,out(\lambda_{640}) + \Delta I\,out(\lambda_{640}))\,/\,I\,out(\lambda_{640})] = (\varepsilon\,oxy\,(\lambda_{640}) \cdot \Delta C\,oxy + \varepsilon\,deoxy\,(\lambda_{640}) \cdot \Delta C\,deoxy) \cdot L\,. \quad (9)$$

If optical path length *L* cannot be set, *L* is left as *L·ΔCoxy* or *L·Δcdeoxy*, for *D·ΔCoxy and L·ΔCdeoxy*, i.e., *ΔCoxy+L·ΔCdexoy=L·ΔCtotal*, and the concentration variation of the total hemoglobin is given as *ΔCtotalHb*.

Oxygen is present in blood in a form that is bonded to hemoglobin. Oxy-hemoglobin (artery), which is termed

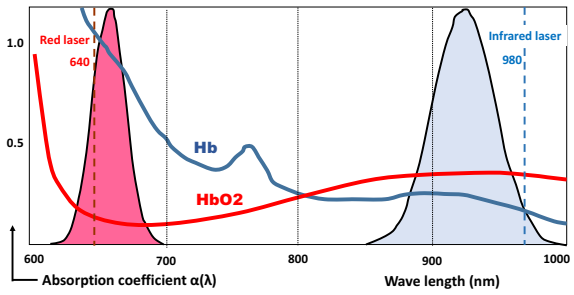


Fig. 7 Ttransmittance wavelength characteristics of red light and infrared ray

as HbO2, is bound to oxygen. Arterial blood that contains significant amount of oxygen exhibits a bright red color. Conversely, deoxy-hemoglobin (vein) that consists of Hb that is not bound to oxygen, and venous blood after releasing oxygen into the body exhibits a dark color. A conventional method involves superimposing the emission spectrum of an LED (which is a typical light emitting diode) and the relative light absorption spectrum of oxygenated and deoxygenated hemoglobin. The transmittance wavelength characteristics of red light and infrared ray are shown in **Fig.7**.

3.3 Arterial blood oxygen saturation (SaO2 value)[11)]

The 50% saturated blood spectrum is considered, and the relative red absorbance is compared with the IR indicated by the black circle. The broken line shows the spectrum of 50% saturated blood and the relative absorbance of red. Additionally, R is indicated by a circle. Based on the absorption characteristics of light, two fingertips of Oxy-hemoglobin and Deoxy-hemoglobin red light and infrared light are irradiated to the fingertip. The absorption characteristic of only arterial blood is calculated from the ratio of the magnitude of the pulse wave of the transmitted light. An oxygen saturation corresponding to 100% is obtained when the total hemoglobin contains oxygen and it changes to Oxy-hemoglobin. Arterial blood oxygen saturation (SaO2) is determined by the oxygen saturation of arterial blood from the amplitude ratios of the pulse waves due to the two irradiated lights of 640 [nm] and 980 [nm].

In the oxygen dissociation curve (**Fig. 8**), the arterial blood oxygen saturation (SaO2) coupled with hemoglobin is plotted on the vertical axis, and the arterial oxygen partial pressure (PaO2) is plotted on the horizontal axis. In the solid line of the bell, the oxygen partial pressure inside the alveolar becomes normal at 100 [mmHg], and the oxygen saturation corresponds to approximately 98%. With respect to an oxygen partial pressure of 60 [mmHg], oxygen saturation reaches approximately 90%. Oxygen saturation is maintained despite a slight decrease in the oxygen partial pressure. Furthermore, the oxygen carrying capacity with respect to the periphery is high. A large amount of oxygen is used in peripheral tissues from a mixed blood level of oxygen partial pressure of 60 [mmHg] or less, and thus the oxygen saturation level significantly reduced. Oxygen carrying capacity is

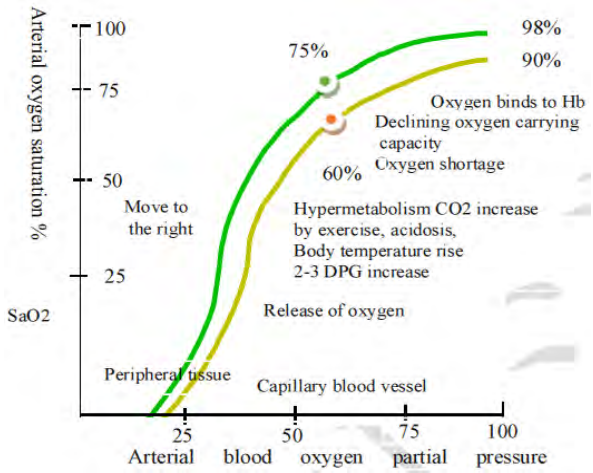


Fig. 8 Oxygen dissociation curve

reduced. The oxygen partial pressure corresponding to 60 [mmHg] and the Oxygen saturation degree below 90% results in a condition that necessitates oxygen inhalation, which is termed as respiratory insufficiency. Thus, 40 [mmHg] released venous blood contains approximately 23% of oxygen although 75% of the oxygen is bound. Hemoglobin binds tightly to oxygen in places with high oxygen partial pressure and is released immediately in places in which oxygen partial pressure is low (peripheral tissues). An increase in CO_2 due to body movements tends to acidosis. Increased hyper metabolism that corresponds to an increase in body temperature due to diseases leads to the production of 2-3 DPG in hypoxia (increased number of erythrocytes due to glycolysis). This increases the demand for oxygen, and oxygen is used in large quantities. With respect to 40 [mmHg], the saturation level significantly decreases from 75% to 60%. The released oxygen gradually increases. With respect to P50 (saturation degree: 50%), which corresponds to the half-life of oxygen saturation, it causes a state in which oxygen hypoxemia is created and lasts long. It is assumed that the dissociation curve moves to the right and the curve He moves oxygen to the right by increasing the oxygen.

4. Production of SpO2 Value Spatial Instrument Prototype[12), 13)]

4.1 Sensor module installed in equipment[14)]

Initially, the study involved measuring SpO2 data uses a non-contact method through irradiating light by attaching light-emitting elements of two wavelengths to a gantry (**Fig. 9**).

4.2 Equipment[15)-17)]

As depicted already, a device is used to detect SpO2 values by using red laser (i.e., artery) and infrared laser (i.e., vein). The device is used to detect normal values in a stable state. A relationship exists between the oxidized

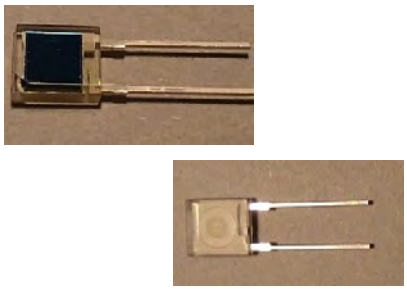


Fig. 9 Dimensions of light receiving element

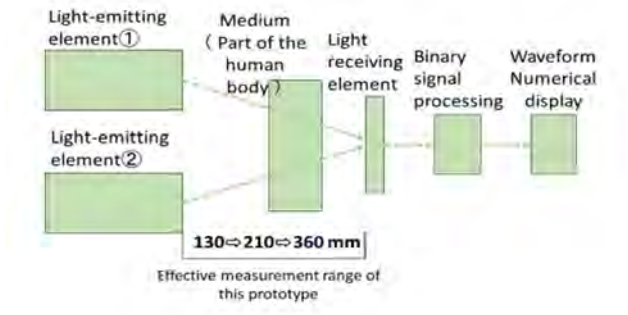
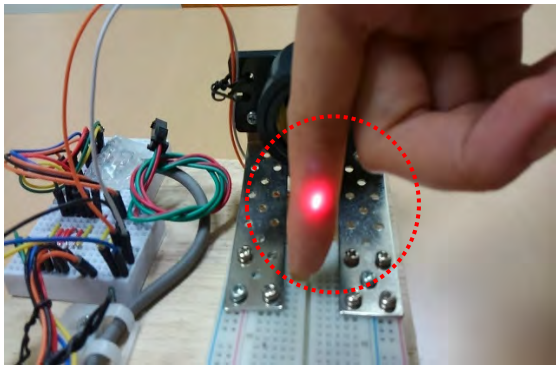
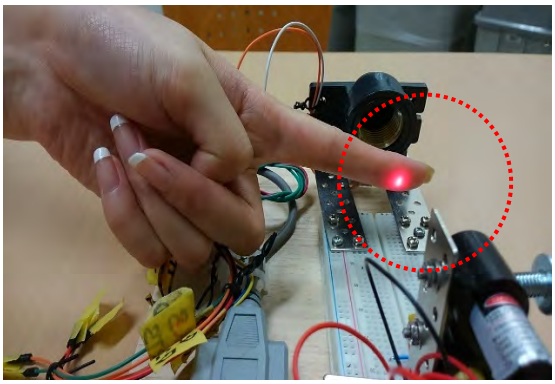


Fig. 10 Conceptual diagram of our signal processing workflow



(a)



(b)

Fig. 11 Prototype of non-contact SpO2 measuring device in operation and measuring with (a) right and (b) left index finger

hemoglobin concentration of the artery at 640 [nm] and the reduced hemoglobin concentration of the vein at 980 [nm] in relation to the molecule (i.e., molar extinction coefficient) and wavelength. Therefore, the standard deviation is a method of grasping the peak of the waveform and averaging the number of received data at each time interval by a fixed number. This is used to identify the quality of the data and by using a numerical value of 25 or less. Specifically, the value of the ratio between the AC component and the DC component is calculated from the visible light and infrared light as the R-value of the SpO2 value. With respect to the study specifications, the light-receiving surface size corresponded to 5.5 [mm] × 4.8 [mm] (**Fig. 10**), the effective light-receiving area corresponded to 26.4[mm^2]. The reverse voltage (RV) displayed an absolute maximum of 35 [V], the sensitivity wavelength range (λ) from 320 [nm] to 1100[nm], and the maximum sensitivity wavelength (λp) receiving a good component and the signal by 980[nm].

As shown in the signal processing concept in Fig.10, this constitutes a device for detecting SpO2 value by using red laser (artery) and infrared laser (vein). A relationship exists between oxidized hemoglobin concentration, Oxy-hemoglobin (640[nm] for artery) and reduced hemoglobin concentration Deoxyhemoglobin (vein 980[nm]), the molecule (molar extinction coefficient) and the wavelength.

This is used as the quality of the data and reliability is determined with a numerical value of 25 or less. The value of the ratio between the AC component and the DC component is calculated as visible light and infrared light as corresponding to the R value of the SpO2 value to be calculated. The SpO2 value is calculated from the R value and is expressed in terms of a percentage value. Experiments on the effectiveness of the proposed method were performed by linking the created program and the fabricated board to the laser beam module (as shown in **Fig. 11**). The results indicated that the SpO2 value was effective and confirmed the success of the spatial measurement by non-contact method.

4.3 Proposal method of performance evaluation[18)-19)]

4.3.1 How to check the infrared light hit on the finger

The wavelengths of two kinds of light used here are 640 [nm] (visible light region) and 980 [nm] (infrared region) as shown in the photograph (Fig.11). Since 640 [nm] emits red light, it can be grasped visually so it's easy to hit light on your fingers. However, 980 [nm] cannot be grasped visually. You can check the light of 980 [nm], which should not be seen, by looking at the light while operating the smart camera or the CCD camera.

Two kinds of light are reflected in the photographs (Fig.11) because they have high sensitivity in the same wavelength region due to the spectral sensitivity characteristics in the near-infrared region.

4.3.2 Can the SpO2 value be estimated even when the finger is set at any angle?[20)]

Although the angle of the finger to the light will cause a problem, in the experiment, when changing the angle of the finger hitting the light with the finger, the SpO2 value changes. Even if the fingertip is changed by about 30 degrees, even if there is a change in the numerical value due to reflection or refraction of the light transmitted through the finger, there are no influences on the final derived numerical value because the two numerical values are contrasted.

Table 2 Report data with respect to a commercially available pulse oximeter

Commercial pulse oximeter data report

User Information **Name:** yn **Date(Y/M/D):**2017/02/07
Age: **Sex:** M **Height/Kg:** **Nationality:**
Time Length: 00:03:40 **Time:** 00:01:55 **Comments:**

Event Data	SpO2	PR	%SpO2 Level	vents	Below(%)	Time(%)
Total Event	0	0	99-95	0	100	100.0
Time in Event(min)	0.0	0.0	94-90	0	95	0.0
Avg. Event Dur.(sec)	---------	------	89-85	0	90	0.0
Index(1/hr)	0.0	0.0	84-80	0	85	0.0
Artifact(%)	21.8	21.8	79-75	0	80	0.0
Adjusted Index(1/hr)	0.0	0.0	74-70	0	75	0.0
SpO2 Data			69-65	0	70	0.0
Basal SpO2(%)	98.4		64-60	0	65	0.0
Time (min)<88%	0.0		59-55	0	60	0.0
Events<88%	0		54-50	0	55	0.0
Minimum Spo2(%)	97		49-45	0	50	0.0
Avg. Low SpO2(%)	---------		44-40	0	45	0.0
Aglow SpO2<88%	---------		39-35	0	40	0.0
PR Data			34-30		35	0.0
Avg. Pulse Rate(bpm)	75.9				30	0.0
Low Pulse Rate(bpm)	68					

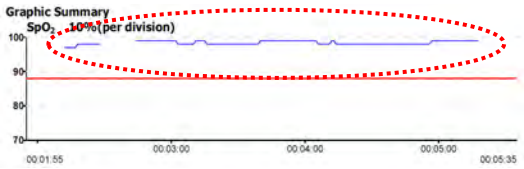
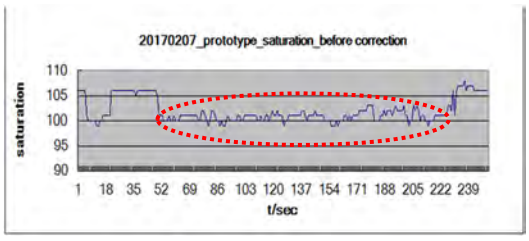
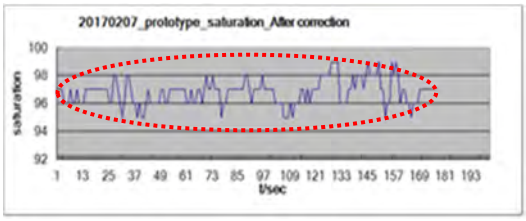


Fig. 12 Measured waveform obtained from a commercially available pulse oximeter



(a) Before correction



(b) After correction

Fig.13 Measured waveform obtained from the proposed prototype SpO2 device

4.3.3 Can the SpO2 value be estimated when the distance between the laser and the finger takes any value

The experimental distance of 110 [mm] is between the light source and the light receiving part, but the distance to the finger is 110 [mm] minus finger thickness. In the waveform data by the oscilloscope, it shows that the scattering is large when the finger is close to the light source, and the scattering becomes smaller as it becomes farther from the light source and becomes closer to the light receiving portion.

5. Verification of the Prototype Method Proposed for Contactless Measurement Data[21)-22)]

5.1 Comparison of a traditional contact-based method and the proposed non-contact method

Data comparison was performed by using a conventional type SpO2 measuring instrument. It should be noted that the authors served as human subjects in the experiments. **Table 2** and **Fig. 12** show the results of measuring SpO2 values by using a commercially available device. In the figures, the basal SpO2 value corresponded to 98.4% while the minimum SpO2 value corresponded to 97%. In the measurement waveform as shown in **Fig. 13**, the part surrounded by the dotted line denotes SpO2 from 99% to 97%. Additionally, the prototype SpO2 measurement output is shown in Fig.13 (a) and (b) from 52 [s] to 222 [s].

Its waveform was analyzed, and it was determined as stable in saturation from the highest 100% to the lowest 99%.When the interval between 52 [s] and 222 [s] that served as the actual measurement time range was extracted, it resulted in a time range ranging from 0[s] to 165[s]. Therefore, the use of an output parameter variable corresponding to -3 allows the use of the commercially available one as shown in **Table 3**. The output numerical values in those two figures are combined to show that the measurements in the space measurement prototype were normal.

6. Conclusion

Many reports have been issued on studies of direct oxygenity and reflectivity of blood oxygen levels. In this paper, we aim to find the saturation value of SpO2 by contactless space measurement method for knowing biological data for applications such as medical care, nursing care, everyday life, field work and so on. We investigated the effectiveness and characteristics of laser light and the effectiveness of biometric measurement. The data measured with the model proposed in this study is almost the same as the value of SpO2 obtained from a commercially available pulse oximeter (Fig.3), and it was confirmed that constant and stable numerical display was produced.

Furthermore, when considering a wide range of applications, you can also use the method proposed in

Table 3 Spatial measurement of SpO2 and corrected saturation value

(i)	(ii)	(iii)	(iv)	(v)	(vi)	(Vii)	(Viii)
124450	18	1978	0	0.37	101	−3	98
124810	17	1973	0	0.39	100	−3	97
124990	19	2013	0	0.42	100	−3	97
125207	20	2050	0	0.39	100	−3	97
125599	77	1210	0	0.38	101	−3	98
125913	87	77	0	0.38	100	−3	97
126117	89	81	0	0.39	100	−3	97
126591	83	108	0	0.36	101	−3	98
127240	66	154	0	0.38	100	−3	97
127804	57	149	0	0.39	100	−3	97
128270	55	148	0	0.4	100	−3	97
128973	48	161	0	0.38	101	−3	98
129339	46	115	0	0.35	101	−3	98
129681	48	136	0	0.37	101	−3	98
129947	55	147	0	0.35	101	−3	98
130160	63	160	0	0.36	101	−3	98
130358	71	170	0	0.37	101	−3	98
130794	81	85	0	0.38	101	−3	98
131174	81	87	0	0.37	101	−3	98
131716	73	125	0	0.37	101	−3	98
131954	74	128	0	0.37	101	−3	98
132323	68	116	0	0.36	101	−3	98
133048	55	153	0	0.39	100	−3	97
134170	44	293	0	0.4	100	−3	97
134613	43	298	0	0.34	102	−3	99
135319	41	291	0	0.33	102	−3	99
135643	40	275	0	0.34	101	−3	98
136456	36	258	0	0.42	100	−3	97
136741	40	297	0	0.43	99	−3	96
137241	48	191	0	0.4	100	−3	97
137614	49	197	0	0.33	102	−3	98
138097	53	175	0	0.32	102	−3	99
138363	54	184	0	0.37	101	−3	98
(i)	(ii)	(iii)	(iv)	(v)	(vi)	(Vii)	(ix)
138895	61	105	0	0.38	100	−3	97
139279	58	91	0	0.39	100	−3	97
139885	56	112	0	0.43	99	−3	96
139986	62	170	0	0.41	100	−3	97
140435	63	168	0	0.42	99	−3	98
140581	67	187	0	0.45	99	−3	96

Notes: (i) Observed time, (ii) Heart rate, (iii) Standard deviation (contact type-R), (iv) no data, (v) Non-contact type-R, (vi) The original S-value, (vii) Parameter, (viii) Output parameter / corrected with variable / saturation value, (ix) Finger angle about 30 degrees / in refractive index change

this research to obtain SpO2 values for pets and wild animals, and can also be used for zoo health management. Therefore, we believe that this research represents an important and potentially pioneering contribution in the field of biological information.

References

1) C. Andrews, R. L. Phillips, Laser Beam Propagation through Random Media, Second edition, pp. 57–73, The International Society for Optical Engineering (2005).
2) Y. Ozaki, S. Kawata, Near Infrared Spectroscopy, 2nd edition, Measurement Method Series 32, Academic Publication Center, The Spectroscopically Society of Japan, pp. 11–21 (1998).
3) S. Hashimoto, Introduction to Biomedical Engineering, pp.153–160, Corona Publishing Co., Ltd. (2013).
4) Y. Ozaki, H. Iwahasi, Introduction to Biomolecule Spectroscopy, First edition, pp.13–17, pp. 45–51, Kyoritsu Shuppan Co., Ltd. (1992).
5) T. Sawada, S. Oda, K. Washio, Photoacoustic Spectroscopy and Its Application-PAS, Measurement Method Series 32, Academic Publication Center, The Spectroscopically Society of Japan, pp.170–178 (1982).
6) H. Ohnishi, I. Tsukahara, Absorptiometry Inorganic, Japan Society for Analytical Chemistry, Second edition, pp. 15–36, pp. 39–47, Kyoritsu Shuppan Co., Ltd. (1988)
7) JIS 6802: 2014 (IEC 60825-1: 2014), Safety of Electrical Instrumentation, Control and Laboratory Equipment – General Requirements : Safety of Laser Products – Safety of Optical Wireless Communication Systems for Information Transmission (2014).
8) Y, Nagao, M. Hatsuda, J. Liu, S. Shimamoto: "Using Light Sensing to Acquire SpO2 Biological Information via a Non-Contact Approach", Proc. of International Conference on Mobile and Wireless Technology 2017 (ICMWT2017), pp. 259–268 (2017).
9) Y. Yoshimura, Determination of chemical species in solution using light absorption, Second edition, http://kuchem.kyoto-ac.jp/ubung/yyosuke/uebung/light_abs03.htm. (Last accessed December 6, 2019).
10) University of Iowa Health Care , Pulse Oximetry Basic Principles and Interpretation, pp. 2-3, https://medicine.uiowa.edu/iowaprotocols/pulse-oximetry-basic-principles-and-interpretation (Last accessed December 6, 2019).
11) M. Kanazawa, Pulse Oximeter Handbook, p.20, The Japanese Respiratory Society (2014).
12) K. Itoh, Body Intelligence System Theory – Control of Movement and Learning with Human Robotics – , pp. 52–78, Kyoritsu Shuppan Co., Ltd. (2005).
13) Y. Ohba, S. Yamauchi, Edition of the Chemical Society of Japan, Electron Spin Resonance Spectroscopy, First edition, pp. 106–150, Kyoritsu Shuppan Co., Ltd. (2017).
14) Y. Ozaki, Invitation to Spectroscopy – New measurement technology developed by light – , pp.22–35, pp.62–78, Sangyo Tosho Co., Ltd.(1997).
15) H. Kurosawa, Laser Theory, First edition, pp. 46–53, pp. 58–71, pp. 233–263, Optronics Inc. (2011).
16) H. Sano, Introduction to Mossbauer Spectroscopy, First edition, pp.91–102, Kodansha (1972).
17) H. Ushio, Engineering and Laser Basics and Mechanism – Properties and Application of Light –, Second edition, pp. 54–72, pp. 290–311, Shuwa System (2010).
18) K. Ando, Semiconductor Laser, First edition, pp. 86–89, Technical Criticism Company (2011).
19) N. Hosimiya, Biometric Measurement, First edition, pp.81–89, Tohoku Publishing Co., Ltd. (2002).
20) Y. Nagao, Patent Application, International Patent Classification, October 2019, "Blood Oxygen Saturation Measuring Device" (2019).
21) T. Iijima, et al., Visualization Techniques Biological Information, First edition, pp.107–134, Biological Information Visualization Technology Editorial Committee, Corona Publishing Co., Ltd. (1997).
22) M. Saitoh, Basics of Medical Engineering, First edition, pp.69–87, pp.140–155, Shokodo Co., Ltd. (1990).

(Received September 5, 2017)
(Revised November 14, 2017)

Yoshimitsu NAGAO (*Member*)

He received the Master degree from WASEDA University in Graduate School of International Information and Telecommunications in 2003, and is with doctoral course from 2015. He has been Invited researcher at GITS research center from 2005, at GITS faculty of Science and Engineering Information Technology. His research field is to measure non-contacting space such as blood oxygen concentration by light sensing and information analysis technology in non-contact high precision human respiration monitoring by DRM using radio wave to support healthy longevity society. He is a staff of Building Monitoring Application Subcommittee engaged in log house design and construction. He is a member of Architectural Institute of Japan, IEICE, IET, and IEEE. He has been Chairperson of common area study group on architecture and image electronics (AIM) of IIEEJ from 2015. He was IIEEJ financial director in 2018.

Yanan GAO
She received the Bachelor degree in information engineer from Qingdao University of Science and Technology, China, in 2012. She received Master degree in signal and information process from The Communication University of China and in wireless communication field from Waseda University, Japan, in 2015 and 2018 respectively.
Her research interests include glucose measurement, machine learning and human body communication for body sensor network, near field communication and some applications about human healthcare and so on.

Jiang LIU (*Member*)
She received the M.S. and Ph.D. degrees in information and telecommunications from Waseda University in 2006 and 2012, respectively. In the academic years 2009 to 2012, she was a research associate at Waseda University. In 2012 she joined Faculty of Science and Engineering at Waseda University as an assistant professor, and since 2017 she has been an associate professor affiliated with the International Center for Science and Engineering Programs. Her research focuses on the optical wireless communications, wireless network systems, near field communication, and their applications on 5G network design and healthcare industry. She is a member of the Institute of Image Electronics Engineers of Japan (IIEEJ), and the Institute of Electronics, Information and Communication Engineers (IEICE). She also serves as the secretary of the Japan Division for the Institution of Engineering and Technology (IET).

Shigeru SHIMAMOTO
He received the B.E. and M.E. degrees from the University of Electro-Communications, Tokyo, Japan, in 1985 and 1987, respectively. He received the Ph.D. degree from Tohoku University, Sendai, Japan, in 1992. He was with NEC Corporation from 1987 to 1991. From 1991 to 1992, he was a research associate in the University of Electro-Communications, Tokyo, Japan. He was a research associate in Gunma University, Gunma, Japan, from 1992 to 1993, and was an associate professor, from 1994 to 2000. In 2000, he was an associate professor in the Graduate School of Global Information and Telecommunication Studies (GITS), Waseda University, Tokyo, Japan. Since 2001, he has been a professor in the Graduate School of GITS, Waseda University. He was a visiting professor of E.E. at Stanford University in 2008. Currently, he is a professor of department of Communication and Computer Engineering in Waseda University. His main research interests include satellite communications, mobile communications, optical wireless communications, ad-hoc networks, sensor networks, and body area networks.

IIEEJ Transactions on

Image Electronics and Visual Computing

Call for Papers

Special Issue on the 6th IIEEJ International Conference on Image Electronics and Visual Computing (IEVC2019)

(Vol.8, No.1, 2020)

Editorial Committee of IIEEJ

The 6th IIEEJ International Conference on Image Electronics and Visual Computing (IEVC2019) will be held in Bali, Indonesia, on August 21-24, 2019. The aim of the conference is to bring together researchers, engineers, developers, and students from various fields in both academia and industry for discussing the latest researches, standards, developments, implementations and application systems in all areas of image electronics and visual computing.

The Conference already solicited Journal Track (JT) Papers to be included in December 2019 issue of Transactions on Image Electronics and Visual Computing (Vol.7, No.2), and the deadline of the submission has been announced as September 13, 2019. The editorial committee plans to publish the June 2020 issue of the IIEEJ Transactions on Image Electronics and Visual Computing as the special issue on ``Extended Papers Presented in IEVC2019". The editorial committee will widely ask for submissions for the papers of the following area by extending the material presented in all sessions of IEVC2019. We hope you would submit your high-quality original papers, after checking the electronic submission guidelines on our site.

Topics covered include but are not limited to:

Image and Video Coding, Transcoding, Coding Standards and Related Issues, Image and Video Processing, Image Analysis, Segmentation and Classification, Image Recognition, Image Restoration, Super-Resolution, Color Restoration and Management, Computer Vision, Motion Analysis, Computer Graphics, Modeling, Rendering, Visualization, Animations, Interaction, NPR, Virtual Reality and 3D Imaging, Data Hiding, Watermarking and Steganography, Content Protection, Bioinformatics and Authentication, Computer Forensics, Image database, Image and Video Retrieval, Digital museum, Digital Archiving, Content Delivery, Image Assessment, Image Quality, Printing and Display Technologies, Imaging Devices, Digital Signage, Electronic Paper, Visual Communication; Human Interfaces and Interactions, Mobile Image Communication, Networking and Protocols, Optical Communication, Hardware and Software Implementation, Image Related Applications, LSI, Understanding of Human Vision and/or Human Tactile Sense, Web-Related Techniques, Personalization Technique, Interaction between Human and Computer, Usability, Accessibility, Image Processing Technique Considered Emotion, Other Fundamental and Application Technique, International Standardization.

Paper Submission Due Date: January 8, 2020

The Institute of Image Electronics Engineers of Japan

Call for Papers

Special Issue on CG & Image Processing Technologies for Automation, Labor Saving and Empowerment

IIEEJ Editorial Committee

A declining birthrate and a rapid aging population are common problems in developed countries. Particularly in Japan, the working-age population is declining continuously, after reaching its peak in 1995, resulting in a shortage of labor. On the other hand, per capita labor productivity in Japan is very low, ranking 21st out of 36 OECD countries in 2018. In order to solve the labor shortage, it is necessary not only to expand the workforce but also to improve labor productivity.

Low productivity may be due to Japan-specific personnel systems and social customs, but improvements in productivity can also be expected through automation and labor saving technologies such as autonomous driving that has been actively studied in recent years and white collar work automation using RPA (Robotic Process Automation). Specifically, automation applications can be expected to expand into a wide range of fields in the future by using evolving image recognition technology, increasing information from IoT devices, and less expensive robots.

Productivity can also be improved by adding value to the product or empowering workers, such as by using the Cloud. A wide range of image processing technologies can contribute to productivity, such as decision making support with image analysis, time saving with remote processing using IoT devices, and improvement of customer service through interactive video processing systems.

Based on this background, we look forward to receiving your papers, system development papers, and data papers in this special issue.

1. Topics covered include but not limited to
   Image Processing, Image Recognition, Image Detection, Pattern Recognition, Machine Learning, Computer Vision, IoT, Ubiquitous, Big Data, Autonomous Driving, RPA, Automation, Robotics, Usability, Interface, Interaction, Other related fundamental / application / systemized technologies.
2. Treatment of papers
   Submission paper style format and double-blind peer review process are the same as an ordinary contributed paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as an ordinary contributed paper. We ask for your understanding and cooperation.
3. Publication of Special Issue:
   IEEJ Transactions on Image Electronics and Visual Computing Vo.8, No.2 (December 2020)
4. Submission Deadline:
   Friday, May 29, 2020
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 Technologies for the Realization of Future Society

IIEEJ Editorial Committee

There is a great expectation for an advanced and comfortable society brought about by economic development and the solution of social issues through the introduction and spread of ICT technology. For this expectation, the government advocates and promotes Society 5.0 as a new future society, following hunting society (Society 1.0), agricultural society (Society 2.0), industrial society (Society 3.0), and information society (Society 4.0). It is clearly stated that this purpose is to build a system that coalesces cyber space (virtual space) and physical space (real space) at a high level, and integrates drones, AI devices, medical / nursing care, smart work, smart management and autonomous driving etc.

Not only image recognition and visualization, but also XR that integrates virtual reality (VR), augmented reality (AR), mixed reality (MR) is also necessary to make cyber space more familiar. In addition to visual effects, cross-modal sensory presentation that appeals to the human senses is emphasized. Therefore, technological innovation in computer graphics, computer vision, user interface, user experience, etc., which form these technological foundations, is important, and practical application of technology that appeals not only to vision but also to other senses through images and video.

In this special issue, we look forward to receiving your papers, system development papers, and data papers that will realize a future society through images and video.

1. Topics covered include but not limited to

   VR, AR, MR, Computer graphics, Image processing, Interaction, Realtime processing, Cross-modal sensory, Computer vision, Machine learning Image analysis, Object detection, Image recognition, User interface, User experience

2. Treatment of papers

   Submission paper style format and double-blind peer review process are the same as an ordinary contributed paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as an ordinary contributed paper. We ask for your understanding and cooperation.

3. Publication of Special Issue:

   IEEJ Transactions on Image Electronics and Visual Computing Vo.9, No.1 (June 2021)

4. Submission Deadline

   Friday, October 30, 2020

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. 7, No. 2 2019

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