INTERACTIVE MOTION SIMILARITY RETRIEVAL VIA THE LASSO

Hiromu Miyashita, Satoshi Shimada, Takashi Satou, Suguru Higashino

NTT Service Evolution Laboratories

ABSTRACT

Videos indexed according to subject motion are useful for review and study purposes. A video can include various motions, but only one is the movement of interest (MOI). In our approach, the user selects the desired motion by lassoing it on the screen; similar motions are retrieved based on an optical flow histogram of the selected region. We use two videos, tennis and pottery making, to verify the proposal's precision and to find the cause of detection failure.

1. INTRODUCTION

When learning a sport or a job that involves complex body movement, the trainee will improve her skill more quickly by watching a video of an expert. We are actively researching the incorporation of videos into an education and an e-learning [1, 2, 3]. And we need to examine the detection and usage of motor skills present in videos.

In this study, we focus on utilizing video material to enhance own skills. The trainee can discover and modify own motions effectively by perusing videos including those showing own motions. Unfortunately, creating readily accessible videos is very labor intensive. Raw videos tend to be overly long and contain unneeded scenes and obscured motions. Therefore, it takes time and labor to find the motions that the trainee should peruse.

It is clear that training videos must be indexed to permit MOI queries. In this paper, we define MOI as a human body movement that is of interest to the trainee. Our solution is an interactive system for motion similarity retrieval. This system allows the trainee to select an instance of MOI on the screen simply by using the lasso tool (Fig.1). The trainee is then able to check many instances and compare the motions rapidly and repeatedly.

The target of our proposal includes long videos containing iterated body movements such as repeated practice and sequenced trials by multiple trainees.

2. RELATED WORK

Attempts have been made to detect types of motion in videos and index characteristic scenes. For the surveillance event detection task of TRECVID, Kawai et al. developed a system that could automatically detect specific events (“Pointing”, “CellToEar”, and “ObjectPut”) in video sequences shot by fixed cameras installed in an airport [4]. The method is advantageous for the recognition of small predetermined motions but needs training datasets with target motion. Guanyu et al. proposed an action recognition approach for motion analysis of tennis players in broadcast videos [5]. Their system has functions for player tracking and action classification based on optical flow. They recognize two action categories: "right-swing" or "left-swing". Yu et al. recognized three group actions in soccer games captured by a moving camera [6]. They use the optical flow in regions to model a motion descriptor and set flow classes of "left side attacking", "stalemate", and "left side defending". These methods employ machine learning to train the classifier to recognize the actions.

The problem with conventional machine learning techniques is that the videos targeted here contain a wide variety of human motion and which is of interest depends on the trainee. Moreover, it is difficult to accumulate sufficient training data because environments within which the actions are performed can be quite different.
3. SYSTEM IMPLEMENTATION

We propose a video indexing system that allows the user to select the MOI and retrieves similar instances from a video or set of videos.

For practical reasons, we assume that each video is captured by a stationary camera that records continuously. The video is expected to include repeated instances of the MOI.

3.1. Approach

To index a video based on a MOI, we must calculate a motion descriptor from the video frames. We examined the regularity of characteristic motions presents in activities such as sports games and health care. To minimize costs, the system must offer flexibility terms of the MOI; the MOI cannot be fixed in advance.

In this study, we describe and prototype an interactive motion similarity retrieval system that well handles trainee selection of MOI. The area selected by the trainee and its associated optical flows are processed to yield the motion descriptor needed for subsequent retrieval.

3.2. Pre-Processing

Before viewing or while capturing the video, our system calculates and stores optical flow from all frames. In our system, block-matching based optical flow is detected at regular intervals (e.g. 10 pixels) from the entire frame images. All optical flow values are converted to lengths and angles and even small optical flows are deleted. An angle of optical flow is stored in an array for each frame and is associated with a region in the frame image. The resulting array of optical flow vectors is called the flow array below.

3.3. User Interface Components

Our system extends the basic video player interface to include a lasso function (Fig.1).

To help the viewer identify the MOI, the system provides a motion visualization function (Fig.2). For each video, the system calculates frame differences and then uses glow highlighting to show the areas of movement, both prior and future. Because this visualizes past as well as future motion, the trainee finds it easy to focus on those objects/people that will move in the immediate future (1 second in the current version).

We considered that drawing on the screen displaying the image is most appropriate for indicating the MOI. We employ the standard lasso tool provided by most graphics editing applications. The point tool and line tool are not as effective because they can indicate only part of the MOI (Fig.3). The lasso tool allows the trainee to enclose an arbitrary region of the screen. This function can select multiple motions and is suitable for flow analysis and ROI retrieval.

The trainee stops the video and selects the area containing the MOI, which triggers video indexing and retrieval.

3.4. Video Indexing

After the lasso is used, the system determines the spatial extent of the clipped area. For each frame, the angles of the optical flow associated with the clipped area are used to create a histogram of MOI optical flow (Fig.4-(1)).

The histogram expresses the frequency of optical flow in the clipped area for each angle (Fig.4-(2)). The flow angles are quantized into 24 bins with overlap: 0 to 30 degrees, 15 to 45 degrees, 30 to 60 degrees, etc.

The histograms are converted into a grayscale image (Fig.4-(3)). Each line is 1 pixel wide x 24 pixels high. Each pixel in a line is associated with histogram values; e.g. the intensity of the top pixel indicates the number of optical flows with angles from 0 to 30 degrees. The color value increases (approaches white) with histogram frequency.
Fig. 4 Procedure to create histogram image based on optical flows in the ROI area

The system performs the above process for every frame; the lines are concatenated horizontally to form an image (Fig.4-(4)). The image is 24 pixels high and is width equals the number of frames. Patterns in the image show temporal changes in the angles of optical flow in the clipped area as a motion descriptor. Therefore, if the image has two similar patterns, we expected to find similar motions in similar frame regions.

The system then uses a 2 second period of the image centered on the clipped frame to conduct template matching against the complete image (Fig.5). Frames whose matching value exceeds a threshold are detected and indexed. The 2 second period was selected heuristically. We found that it is relatively easy to discriminate human motions in 2 second periods; longer periods can severely restrict the number of instances found.

4. EVALUATION

4.1. Overview

We evaluated the retrieval precision of our system by conducting motion search on two videos: tennis practice and pottery making.

(1) Tennis practice: Four players practiced their serves in turn. There were 53 serves in the 15 minute video (Fig.6).

(2) Pottery making: Two students tried to make a bowl on a potter's wheel with the teacher. Touching water is indication of a work break and occurred 23 times in the 32 minute video (Fig.7).

We tasked four participants to use our system. They were unaware of how many motions were contained in the videos and none were experts in tennis, pottery making, or video editing. First, they viewed each video. We asked them to select "serve" and "touching water" as the MOI. They were allowed to make another selection if they felt that some scenes that included the MOI had not been indexed (this is called additional selection below).

4.2. Results and discussion

We use Fig. 8 and 9 to discuss the factors impacting retrieval precision. We adjudged that it retrieval is correct if the retrieved scene has a two second period containing the MOI. X-axis is the number of MOI correctly retrieved (true positive), not retrieved (false
negative) and wrongly retrieved (false positive). Y-axis shows selection by four participants (A, B, C, D) and their additional selection (A', B', C', D').

For serve retrieval, precision was low with one selection but most scenes were found by the second retrieval. We compared all selections and found that the main cause of error was that some serve motion lay outside the clipped area. Differences in the style of motion didn't create false detection in this case.

For the pottery video, precision was 70% and recall was 66%. Retrieval errors were created when the system found some of the teacher's movements (right of Fig.7). Sequences in which the hand washing took longer than three seconds were also not found. In this case, precision was sensitive to motion style and the situation. Furthermore, participants failed to achieve complete indexing because the MOI (hand washing) happened at odd intervals, while the tennis MOI occurred at logical times.

An evaluation tested the ability of the proposed system to index tennis serves and the motion of touching water. The evaluation results indicate that making the system more robust against the differences in situation and motion style will raise retrieval precision.

6. REFERENCES


5. CONCLUSIONS

For learning complex movements, indexed video material is extremely useful in reviewing own motions and studying those of an expert. Indexing videos with sufficient accuracy is, however, an extremely difficult and costly process.

This paper proposed an interactive motion retrieval system that uses the lasso tool. Our system allows a viewer to select, via the lasso, a motion on the screen and then performs template matching based on the optical flow in the selected region.