Analyzing Fine Motion Considering Individual Habit for Appearance-Based Proficiency Evaluation

Yudai MIYASHITA¹, Hirokatsu KATAOKA¹†, Nonmembers, and Akio NAKAMURAᵃ†, Member

SUMMARY We propose an appearance-based proficiency evaluation methodology based on fine-motion analysis. We consider the effects of individual habit in evaluating proficiency and analyze the fine motion of guitar-picking. We first extract multiple features on a large number of dense trajectories of fine motion. To facilitate analysis, we then generate a histogram of motion features using a bag-of-words model and change the number of visual words as appropriate. To remove the effects of individual habit, we extract the common principal histogram elements corresponding to experts or beginners according to discrimination’s contribution rates using random forests. We finally calculate the similarity of the histograms to evaluate the proficiency of a guitar-picking motion. By optimizing the number of visual words for proficiency evaluation, we demonstrate that our method distinguishes experts from beginners with an accuracy of about 86%. Moreover, we verify experimentally that our proposed methodology can evaluate proficiency while removing the effects of individual habit.

key words: dense trajectories, proficiency evaluation, removing individual habit, random forests

1. Introduction

Techniques of automatically evaluating proficiency and skill need to be developed for various applications, such as an expert teaching skills to students, sports coaching, and sports strategy analysis. These applications require analysis of the motion of experts. Methodologies of automatically analyzing human motion have been proposed [1], [2].

Many proposed methodologies of analyzing human motion use motion-capture systems or inertial sensors [3], [4]. Although attaching such sensors to the human body allows highly accurate observations to be made, the sensors sometimes obstruct the experts’ motion. Markerless human motion analysis techniques have thus been proposed in the field of computer vision [5], [6].

Methodologies of computer-vision-based analysis of human motion are broadly classified into appearance-based and pose-based approaches. Section 2.1 reviews conventional methodologies of classifying different actions. The present paper focuses on fine-grained action recognition including the recognition of similar actions for guitar picking. The use of dense trajectories (DTs) [5] is a state-of-the-art approach that can distinguish minute changes in activities, such as those of cooking and daily living [7]. The convolutional neural network [6], [8] has also been proposed for fine-grained action recognition. Although the convolutional neural network methodology is useful, its learning processes are not transparent. Explaining the processing and meanings of feature is therefore difficult. We thus select DTs to extract effective features in analyzing fine-grained motion in the present study.

Motion varies according to individual habit such that the results of evaluating proficiency can be erroneous. To robustly evaluate proficiency, it is important to remove the effects of individual habit. We address this problem, which has not been directly considered in existing approaches, in the present study. We propose an appearance-based methodology of proficiency evaluation based on the analysis of fine motion. Previously, we investigated removing individual habit from the evaluation of fine-motion proficiency [9]. The present paper makes an advanced evaluation for a larger dataset. We investigate the motion of guitar picking as described in Sect. 3. The contribution of this paper is threefold.

(i) We demonstrate that Dense Trajectories can be used for the effective analysis of fine motion skill.

(ii) We tune the motion features for proficiency evaluation.

(iii) We remove the effects of individual habit to robustly evaluate proficiency.

The remainder of the paper is organized as follows. Section 2 describes related work. Section 3 provides an overview of the proposed method. Section 4 explains the approach for extracting motion features. Section 5 describes the evaluation of proficiency having removed the effects of individual habit. Section 6 reports experimental results, thereby verifying the effectiveness of the proposed method, and discusses differences between experts and beginners. Section 7 describes an example of application of the proposed method. Section 8 presents conclusions drawn from the results of the study.

2. Related Works

We describe related work on general analysis of human motion and vision-based analysis of human motion. We then outline the proposed method.
2.1 Human Motion Analysis

Many studies have used a motion capture system and/or inertial sensors to analyze human motion. Robinson et al. [1] investigated the analysis of the motion of a kayak using an acceleration sensor. They measured time metrics of kayaking to determine the skill of the kayaker. Ahmadi et al. [10] evaluated the proficiency of a tennis serve using a gyroscope and capturing a marker attached to the human body. They also measured the speed of a tennis racket by capturing a marker attached to the tennis racket. Their proficiency evaluation was based on the speed of the tennis racket and the speed of arm rotation.

These specialized sensor and motion-capture systems are able to extract fine motion. However, when sensors are attached directly to the human body, the expert may not present their natural unconstrained motion.

2.2 Vision-Based Human Motion Analysis

Motion feature extraction using markerless methodologies has been proposed in the field of computer vision. These methodologies are broadly classified into pose-based and appearance-based approaches.

2.2.1 Pose-Based Approach

Many studies have used joint positions or the skeleton as features. Du et al. [11] proposed bidirectional recurrently connected subnets, which entail a learned time-series variation of the human joint position using the deep learning approach of recurrent neural networks [12]. Wang et al. [13] analyzed a time series of variations of the human pose by examining partitioned joint locations, such as those of the hand, left arm, and right arm. They first estimated the human pose employing K-best estimation. They then divided the pose into five components (i.e., a hand, left arm, right arm, left leg, and right leg). They finally modeled the time-series variation employing machine learning.

Pose-based approaches are able to extract a rough human structure. They are useful for classifying different actions into action classes. However, in the case of evaluating proficiency, more fine-grained motion features are required when dealing with the same class of motion.

2.2.2 Appearance-Based Approach

Many studies have used features extracted from images in the appearance-based analysis of motion. Savarese et al. [2] proposed Spatial-Temporal correlations for the analysis of human motion via calculation of feature similarity in a three-dimensional (X, Y, time) region. Ma et al. [14] analyzed human motion employing graph theory. They assumed that human motion consists of combinations of basic motions, such as waving a hand and raising a foot. Wang et al. [5] proposed DTs for the extraction of fine-grained motion features. DTs can be used to distinguish minute changes in activities such as those of cooking and daily living. We thus use DTs to effectively extract features for proficiency evaluation in the present study.

3. Overview of the Proposed Method

We propose an appearance-based proficiency evaluation methodology based on fine-motion analysis. Figure 1 shows the framework of our proposed method.

First, we use DTs for which the state-of-the-art performance has been demonstrated on a dataset of cooking actions created by Rohrbach et al. [7].
Second, we produce a histogram from the motion features using the bag-of-words (BoW) model and determining the appropriate number of visual words.

Third, to remove the effects of individual habit, we extract the common principal histogram elements corresponding to experts or beginners according to discrimination contribution rates using random forests [15].

Finally, we calculate the similarity of histograms to determine the proficiency of the guitar-picking motion.

Guitar picking is a fine motion consisting of plucking individual guitar strings. It is a simple but important motion because the timbre changes according to small differences, such as the way the hand snaps. We select a target fine motion that is simple to perform but difficult to distinguish visually. Guitar picking is a simple motion of moving the hand up and down. To distinguish between experts and beginners, it is necessary to observe small differences in motion. We thus consider guitar picking to be a challenging subject of automatic proficiency evaluation.

4. Motion Feature Extraction

We first describe DTs. We then calculate histograms from the features using a BoW model to facilitate analysis.

4.1 Dense Trajectories

We extract features in three steps. (1) We generate an image pyramid. (2) We generate DTs for each scale in the image pyramid. (3) We extract features on the trajectories. Each of these steps is detailed below.

4.1.1 Generation of an Image Pyramid

We generate an image pyramid as a preliminary process in extracting features robustly in the event of scale changes (Fig. 2). The reduction ratio of the image size is \(1/\sqrt{2}\). The size reduction is terminated when the size of the image becomes \(32 \times 32\) pixels or we reach eight spatial scales.

4.1.2 Generating DTs at Each Scale

We densely extract feature points for each spatial scale in the generated image pyramid. We then generate trajectories by tracking the extracted feature points. For each spatial scale image, we sample grid points at five-pixel intervals. Here, to avoid false correspondences during tracking, we do not sample feature points in regions without texture. We then calculate the autocorrelation matrix and eigenvalues \((\lambda_1, \lambda_2)\) at each grid point.

If the eigenvalue is equal to or greater than the threshold \(T\) in Eq. (1), we extract the grid point as a feature point.

\[
T = 0.001 \times \max_{i \in D} \min(\lambda_1^i, \lambda_2^i) \tag{1}
\]

Figure 3 (a) shows the extraction of feature points from the input images. Feature points are extracted from each frame. The feature points extracted from one frame are continuously tracked in the next 14 frames using the Farneback algorithm. Thus, trajectories for 15 frames are densely generated. In Fig. 3 (b), the red points represent feature points and the green lines represent the generated trajectories.

4.1.3 Extracting Local Features on the Trajectories

We calculate local features within a space–time volume around the trajectories following the original work on DTs [5]. The size of the volume is defined by the dimensions of the patch \((N \times N\) pixels) and the number of frames \(L\). Here, \(N\) and \(L\) are respectively set to 32 and 15 as default parameters. The patch, which is set so that the trajectory passes through its center, is divided into four quadrants. Histograms of optical flow (HOF)[16] as motion information and motion boundary histograms (MBH; specifically, MBHx and MBHy) [17] as motion boundary information are adopted as local features and calculated in each quadrant. Local features obtained from three frames are combined to represent temporal variation. Figure 4 is a conceptual image of the above procedure.

Histograms of optical flow. The HOF has eight bins. Each bin corresponds to an angle range of 45 degrees of the argument of the optical flow vector. A bin is incremented by the length of the optical flow vector when the argument of
the vector corresponds to the bin. For example, if the optical flow vector has an argument of 50 degrees and length of 12, a value of 12 is added to the second bin (corresponding to an angle of 45–90 degrees). To remove noise, if the length of the optical flow vector is less than a threshold value, a ninth bin (for invalid measurements) is incremented by one. The same procedure is repeated for all optical flow vectors. The HOF is thus generated.

**Motion boundary histograms.** Derivatives are computed separately for the horizontal and vertical components of optical flow. Oriented information is quantized into eight bins in the histogram as was done for the HOF. We thus obtain an eight-bin histogram for each component (i.e., MBHx and MBHy). A bin is incremented by the gradient magnitude when the oriented information corresponds to the bin. We calculate the gradient magnitude and oriented information using the following equations.

\[
\begin{align*}
    f_u(u, v) &= I(u + 1, v) - I(u - 1, v) \\
    f_v(u, v) &= I(u, v + 1) - I(u, v - 1) \\
    m(u, v) &= \sqrt{f_u^2 + f_v^2} \\
    \theta(u, v) &= \arctan\left(\frac{f_v}{f_u}\right)
\end{align*}
\]

Here, \(I\) is a spatial image of optical flow in which a pixel value represents the gradient magnitude of the optical flow and coordinates \((u, v)\) represent the target pixel.

4.2 Bag-of-Words Model

The motion features extracted using DTs consist of more than 1000 points per frame. This large number of points raises challenges in the following procedures. To address this problem, we use the BoW model to generate histograms with constant dimensions.

We first construct a \(k\)-dimensional codebook (visual words, VWs) for the HOF, MBHx, and MBHy of each feature. We then generate the histogram by voting on features closest to VWs (Fig. 5). We finally normalize the histogram frequency according to the number of features.

5. Proficiency Evaluation with the Removal of the Effects of Individual Habit

We evaluate skill proficiency using the features extracted using DTs. We first calculate differences between experts and beginners according to the similarity of histograms. We then improve the robustness of the evaluation of skill proficiency by removing the effects of individual habit.

5.1 Finely Tuning Visual Words for Proficiency Evaluation

In their original work on DTs, Wang et al. [8] investigated the appropriate number of VWs for different motion classification tasks. They set the number of VWs to 4000. However, as mentioned above, the evaluation of proficiency requires the discrimination of similar types of motion. The appropriate number of VWs may therefore differ from 4000. We thus vary the number of VWs from 800 to 8000 at intervals of 800, as detailed in Sect. 6.2, and tune the number to classify motions into expert and beginner classes.

5.2 Removing the Effects of Individual Habit

We calculate the similarity \(S\) between the expert histogram \(H_e\) and beginner histogram \(H_b\) using cross-correlation (Eq. (6)). We consider that the histogram similarity represents skill proficiency.

\[
S = \frac{\sum_{i=1}^{k} H_b[i]H_e[i]}{\sqrt{\sum_{i=1}^{k} H_b[i]^2 \times H_e[i]^2}} (0 \leq S \leq 1)
\]

However, the effects of individual habit cannot be ignored. The histogram varies with individual habit and the proficiency evaluation can be adversely affected.

To accomplish robust proficiency evaluation, we suppress the variation by removing the effects of individual habit. We consider valid VWs as those that can distinguish experts from beginners. The valid VWs are the principal elements of the histograms corresponding to experts or beginners. We thus remove the effects of individual habit by disregarding any motion other than these common principal elements of the histograms (Fig. 7).
We extract these common principal elements of the histograms using random forests. Random forests can be used to evaluate feature effectiveness[15]. In general, principal component analysis (PCA) is used for such multivariate analysis. PCA calculates the feature correlation using a linear model. It is therefore difficult for PCA to accommodate non-linear data. Features extracted using DTs are acquired from human motion and are thus possibly non-linear. In addition, the feature distribution varies according to the target motion. Kernel PCA, which is able to deal with non-linear data, has also been proposed; however, the result depends on the selected kernel. In contrast, random forests can be used to evaluate valid VWs regardless of the feature correlation. Random forests extract only valid components for classification. We can realize fine-grained motion analysis by referring to the valid components rather than compressed principal components, such as those of PCA. We therefore adopt random forests in this work.

Random forests can be used to extract valid VWs by calculating the error rate when classifying out-of-bag (OOB) features (i.e., features that are not used in learning) in the generated decision trees.

We extract the valid VWs in the following steps.

1. We calculate the error rate $E_1$ from OOB features to generate a decision tree.
2. We swap one of the OOB features with a random one.
3. We calculate the error rate $E_2$ from the swapped OOB features to generate a decision tree.
4. We compare the error rates $E_1$ and $E_2$.
5. We repeat steps (1)–(4) to process only a few of the features.
6. We compare the differences between the error rates of all the features.

6. Experiment on Proficiency Evaluation

We conducted two experiments for the optimization of BoW dimensions and evaluated the proficiency when removing individual habit. We first describe the dataset and experimental setup.

6.1 Experiment Setup

Twelve expert and ten beginner guitar players participated in our experiment. The beginners were totally inexperienced playing guitar while the experts had at least three years of experience playing guitar. Each subject performed six trials. We obtained a total of 132 trial videos (72 sequences for experts and 60 sequences for beginners) (Fig. 6). The video was captured under the following conditions.

- The location and parameters of the camera were fixed.
- Videos were captured for 10 seconds at 30 fps.
- The subject was captured straight from the front.
- The subject played guitar by picking only.
- The background was untextured.

Figure 8 shows the experiment environment.

6.2 Evaluation Parameter for Proficiency Evaluation

We varied the number of VWs from 800 to 8000 at intervals.
Table 1 Evaluation of the number of visual words (with highest score shown in bold font)

<table>
<thead>
<tr>
<th>number of VWs</th>
<th>HOF</th>
<th>MBHx</th>
<th>MBHy</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>85.6</td>
<td>81.8</td>
<td>80.0</td>
</tr>
<tr>
<td>1600</td>
<td>84.8</td>
<td>79.5</td>
<td>68.2</td>
</tr>
<tr>
<td>2400</td>
<td>81.1</td>
<td>72.7</td>
<td>64.4</td>
</tr>
<tr>
<td>3200</td>
<td>77.3</td>
<td>67.4</td>
<td>61.4</td>
</tr>
<tr>
<td>4000</td>
<td>75.8</td>
<td>65.2</td>
<td>55.3</td>
</tr>
<tr>
<td>4800</td>
<td>72.7</td>
<td>61.4</td>
<td>54.5</td>
</tr>
<tr>
<td>5600</td>
<td>70.5</td>
<td>59.8</td>
<td>54.5</td>
</tr>
<tr>
<td>6400</td>
<td>70.0</td>
<td>54.5</td>
<td>54.5</td>
</tr>
<tr>
<td>7200</td>
<td>66.7</td>
<td>53.8</td>
<td>54.5</td>
</tr>
<tr>
<td>8000</td>
<td>64.4</td>
<td>54.5</td>
<td>54.5</td>
</tr>
</tbody>
</table>

Table 2 Variance of classification accuracy (×10⁻²)

<table>
<thead>
<tr>
<th>number of VWs</th>
<th>HOF</th>
<th>MBHx</th>
<th>MBHy</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>6.14</td>
<td>10.6</td>
<td>13.6</td>
</tr>
<tr>
<td>1600</td>
<td>7.30</td>
<td>12.9</td>
<td>16.9</td>
</tr>
<tr>
<td>2400</td>
<td>8.16</td>
<td>16.3</td>
<td>19.5</td>
</tr>
<tr>
<td>3200</td>
<td>10.7</td>
<td>17.5</td>
<td>20.6</td>
</tr>
<tr>
<td>4000</td>
<td>12.1</td>
<td>18.9</td>
<td>24.1</td>
</tr>
<tr>
<td>4800</td>
<td>14.0</td>
<td>20.6</td>
<td>24.8</td>
</tr>
<tr>
<td>5600</td>
<td>16.1</td>
<td>20.6</td>
<td>24.8</td>
</tr>
<tr>
<td>6400</td>
<td>16.6</td>
<td>23.5</td>
<td>24.8</td>
</tr>
<tr>
<td>7200</td>
<td>15.7</td>
<td>24.2</td>
<td>24.8</td>
</tr>
<tr>
<td>8000</td>
<td>17.0</td>
<td>24.8</td>
<td>24.8</td>
</tr>
</tbody>
</table>

In this experiment, we removed the effects of individual habit using the method proposed in Sect. 5.2. We compared the results obtained using all VWs with those obtained using the valid VWs obtained from random forests. The valid VWs were selected as the top 30 of 800 VWs. Here, we used the HOF, which had the highest discrimination performance in Sect. 6.2, as features.

We calculated the correlation matrix using Eq. (6) from the histograms corresponding to experts and beginners. Figure 11 shows the results. In the figure, a higher similarity between histograms is represented by lighter shading.

In the experiment, we captured six videos for each subject. One block thus corresponds to the proficiency evaluation for one subject. In Fig. 11 (a), the histogram similarity is high in the same block, while the histogram similarity is
low in different blocks. This result signifies the recognition of individuals because of the effects of individual habit.

In contrast, Fig. 11 (b) shows the evaluation of proficiency using the valid VWs with the removal of individual habit. Similarity values along a diagonal line are high, showing that experts are effectively discriminated from beginners.

Tables 3 and 4 show the average similarity values obtained from correlation matrices. The value in the second row and second column is the average similarity between Expert 1 and Beginner i (i = 1–60). The value the second row and third column is the average similarity between Expert 1 and Expert i (i = 1–72) except for Expert 1 (i.e., only Experts 2–72). Other values in the table are calculated in the same manner.

When using all VWs (Table 3), the difference between the average similarities of experts and beginners (0.33) and average similarities of experts and experts (0.49) is relatively small at 0.16. We consider that the similarity between experts is lower than it should be owing to the effects of individual habit.

In contrast, when using only valid VWs (Table 4), the difference between the average similarities of experts and beginners (0.20) and experts and experts (0.72) is relatively large at 0.52. This means experts are effectively distinguished from beginners. We consider that using the principal motion features excluding features corresponding to individual habit improves the evaluation of the proficiency.

Through this experiment, we demonstrated the effectiveness of our proposed method in removing the effects of individual habit.

7. Application Example

Figure 12 shows an application example of the proposed method. The modifications for beginners are visualized using the beginners’ common principal elements in the histogram extracted by random forests.

In the BoW voting explained in 4.2, if a motion feature of MBHx is voted as a common principal histogram element
MIYASHITA et al.: ANALYZING FINE MOTION CONSIDERING INDIVIDUAL HABIT FOR APPEARANCE-BASED PROFICIENCY EVALUATION

for beginners, the origin of the trajectory corresponding to the motion feature is plotted in green as a modification for beginners (Fig. 13).

We manually analyzed the differences between experts and beginners. In Fig. 12, beginner modifications occur densely in the strumming region. As mentioned in Sect. 5.2, the beginner’s hand moves through translation while the expert’s hand moves by snapping. We consider this to be a major difference extracted by the proposed method. A comparison of the expert and beginner motions shows that the beginner’s hand is oriented inside the strumming region while the expert’s hand is oriented outside the strumming region. The final aim of the present research is that the system is able to analyze differences between experts and beginners automatically and make suggestions to the learner. At the present stage of research, the system is able to evaluate the proficiency level simply and indicate inappropriate points of motion. However, the system is not able to demonstrate the content of modification since interpretation of the VWs is difficult. The number of VWs is conceptually demonstrated in Fig. 13.

8. Conclusions

We proposed an appearance-based proficiency evaluation approach that removes the effects of individual habit. We adopted guitar picking as the objective of motion evaluation and used DTs to analyze human motion. The study made three main contributions.

(i) We demonstrated that DTs can be used for the effective analysis of fine motion skill.

(ii) We evaluated differences in fine-motion skill between experts and beginners using DTs. We customized the number of VWs \( k \) in a BoW model for proficiency evaluation. In experimental trials, \( k = 800 \) allowed the discrimination of experts from beginners with average accuracy of approximately 86%.

(iii) We effectively removed the effects of individual habit. In experimental trials, we verified the effectiveness of the proposed method in removing individual habit.

Acknowledgments

This work was in part supported by the JST RISTEX Service Science, Solutions and Foundation Integrated Research Program.

References


Yudai Miyashita received B. Eng. and M. Eng. degrees from Tokyo Denki University in 2014 and 2016, respectively. He has researched activity recognition with computer-vision techniques. He is currently with CSE Co., Ltd.

Hirokatsu Kataoka received a Ph.D. in Engineering from Keio University in 2014. He is currently a research scientist at the National Institute of Advanced Industrial Science and Technology. His research interests include computer vision, pattern recognition, and machine learning. He is mainly engaged in human sensing including action recognition, action prediction, and pedestrian detection.

Akio Nakamura is an associate professor at the Department of Robotics and Mechatronics, Tokyo Denki University, Japan. He received B. Eng., M. Eng., and Ph. D. degrees from the University of Tokyo in 1996, 1998, and 2001, respectively. From 2001 to 2005, he was an assistant professor at Saitama University. In 2005, he moved to Tokyo Denki University as an associate professor. He is engaged in education on mechatronics and research on robotics, especially computer vision and man-machine interface systems. He is a member of IEICE, IEEE, ACM, RSJ, JSPE, IEEJ, JSME, and SICE.