Silhouette-Based Emotion Recognition from Modern Dance Performance

Hanhoon Park, Jong-Ill Park, Un-Mi Kim, and Woontack Woo

Abstract—We present a vision-based method that directly recognizes human emotion from monocular modern dance image sequences. The method only exploits the visual information within image sequences and does not require cumbersome attachments such as sensors. This makes the method an easy-to-use and human-friendly one. The procedure of the method is as follows. First, we compute binary silhouette images with respect to input dance images. Next, we extract the quantitative features, which represent the quality of the motion of a dance, from the binary silhouette images based on Laban’s movement theory and apply consolidate algorithms to them in order to improve their classifiability. Then, we statistically analyze the features and find meaningful low-dimensional structures, removing redundant information but retaining essential information possessing high discrimination power, using SVD. Finally, we classify the low-dimensional features into 4 predefined emotional categories (happy, surprised, angry, sad) using TDMLP. Experimental results show above 70% recognition rate outside training sequence and thus confirm that it is feasible to directly recognize human emotion using not physical entities but approximated features based on Laban’s movement theory.

Index Terms—Emotion recognition, modern dance analysis, noncontacting, Laban’s movement theory

I. INTRODUCTION

The ability to recognize, interpret and express emotions – commonly referred to as “emotional intelligence” [1] – plays a key role in human communication. Human-computer interaction follows the same principles, i.e. we have an inherent tendency to respond to media and systems in the ways that are common in human-human interaction [2]. Since emotions are an integral part of human-computer communication, computers with emotion recognition skill will allow a more natural and thus improved human-computer interaction.

For the last couple of years many researchers have focused on developing emotion recognition technology. It is understood that emotion has its expression in many areas such as one’s face, voice, and body movements. While the emotion recognition methods that extract emotional information from speeches or facial expressions have been extensively investigated [3-6] (although it is still a challenge [6]), the ones associated with body movements were less explored. This may be due to the fact that body movement is too high dimensional, dynamic, and ambiguous to analyze. There is no doubt that the need for a framework for emotion recognition from body movements has been increasing.

A. Related Work

1) Vision Based Human Motion Analysis: Most work in this area has been concentrated on body parts such as hands [28] and gait [26, 27] because they are relatively simple or well-regulated and thus easy to quantify and analyze. This type of method is not very applicable to other full body movements (neither body part nor gait) that change unexpectedly or dynamically. Methods that try to analyze more general human motions exist in the literature. However, most of them rely on motion-capture systems or are model-based approaches [29, 30, 31] of which both need a troublesome initialization step.

There are some methods that try to analyze body movements in time domain in the literature. Kojima et al. defined the word “rhythm points”, which indicates the time of the start and the end of a movement, so that the whole motion could be represented as the collection of partial movements (a unit of motion) with a period [7]. Wilson et al. also tried to identify temporal aspects of body movement [8]. They proposed a method that detects candidate rest states and motion phases from the movements spontaneously generated by a person telling a story. Kojima’s and Wilson’s methods are appropriate for analyzing simple and regular motions such as small buton-like movement or swing strokes. In general, however, human body motions (especially in case of dance in this paper) are not so simple or regular so we cannot easily determine the start/end of a unit of motion or rest state. Thus, their otherwise effective methods do not seem to be applicable to general and natural human motions.

All the methods explained in this section have only aimed at analyzing human motions but do not pay attention to the emotion communicated.

2) Emotion Recognition from Body Movement: Several methods aiming at recognizing human emotion from body movements exist in the literature.

Darwin described the motions associated with emotion and theorized on the relationship between emotions and their
expression [9]. Montepare et al. described the emotional cues that could be determined from the way someone walks [10]. For example, they found out that a heavy-footed gait and longer strides indicate anger while a faster pace indicates happiness. Barrientos suggested that some expressions of emotion should be expressed in writing since it is merely a stylized human movement [33]. And he described an interaction technique in which pen gesture controls avatar gesture. Picard et al. tried to recognize affective state by capturing the physiological signals, e.g. electromyogram (EMG), blood volume pressure, galvanic skin response, and respiration, etc. [11]. They considered emotion recognition as physiological pattern recognition. The aforementioned methods have limitation in that Darwin’s is heuristic, Montepare’s or Barrientos’s technique cannot be applicable to common human motions, and Picard’s requires complicated devices to capture the physiological signals and thus is not human-friendly.

Recently, endeavors for recognizing emotion from dance have been made. Dance is close to common human motions that are the most typical – the basis of natural body expression has further developed in terms of dance – of expressing human emotion at the same time. Dittrich et al. tried to recognize emotion from dynamic joint-attached point light displays represented in dance [12]. They concluded biological-motion displays permit a large amount of emotional encoding to take place. Their method based on subjective evaluation using a questionnaire was not an automatic one based on vision techniques.

To resolve this problem, Park et al. proposed a method of exploiting new form feature for representing human motion in more detail [32]. They appended the contour-based shape information (concerned with silhouette boundary) to the region-based shape information (concerned with bounding box). Thus, they could discriminate the subtle difference between the shapes of silhouettes that have the same bounding box information. It had an effect that the recognition rate of each emotion category was improved and balanced.

B. Content of This Paper

In this paper, we present a new framework for noncontacting emotion recognition from modern dance performance. We consider emotion recognition as a direct 1-to-1 mapping problem between low-level features and high-level information (emotions). Notice that its feasibility has been verified in the previous works [13-15].

Human motions possess emotive meaning together with their descriptive meaning (Camurri et al. [18] used the terms, “prepositional” and “non-prepositional”, instead). Descriptive motions are a kind of signs to transmit meaning such as shaking one’s head to say “no”. In contrast, emotive motions are embodied in the direct and natural emotional expression of body movement based on fundamental elements such as tempo and force that could be combined in a vast range of movement possibilities [16]. In this sense emotive motions do not rely on specific motions, but build on the quality of movements i.e., how motions are carried through, for example whether it is fast or slow. In this paper, we pay attention to this property of the emotive movements.

The rest of this paper is organized as follows. In the next section we explain the method for recognizing human emotion from monocular dance image sequences in detail. In Section III, we provide the experimental results and conclude in Section IV.

II. Method

The overall procedure consists of four parts:

- Foreground (silhouette) extraction from each frame in dance image sequence (see Fig. 1);
- Reliable feature extraction from silhouette image;
- Transformation of the features into hidden space using Singular Value Decomposition (SVD);
- Classification of the hidden space features into 4 predefined emotional categories, i.e. happy, surprised, angry, and sad, using Time Delayed Multi-Layer Perceptron (TDMLP).

A. Foreground (Silhouette) Extraction

We remove the background and shadow out of dance images using difference keying and normalized difference keying technique separately and create binary silhouette images (see Fig. 1).

1) Difference Keying: In general, the background region in an
image is unnecessary and thus must be removed. Chroma-keying has been widely used for this purpose. While it has good performance when the color of background is uniform, it cannot be applied to natural (or cluttered) backgrounds.

We use an approach to separate objects from a static but natural background [19]. We use motion information that has been widely accepted as a crucial cue to separate moving objects from the background. The basic idea is to subtract the current image from a reference image. The reference image is acquired by averaging the images of a static background during a period of time in the well-known (R,G,B) color space. The subtraction leaves only non-stationary objects. For details, see [19]

2) Normalized Difference Keying: The conventional object separation techniques based on static background modeling are susceptible to both global and local illumination changes. Normally, moving objects create shadow according to lighting conditions. The (R,G,B) color space is not a proper space to handle shadow. Thus, we introduce a new color space, called the normalized (R,G,B) color space, which is able to cope with slight change of illumination conditions [19]. What we use normalized (R,G,B) color space means that we only consider chromaticity of pixels. It is based on the observation that shadow has similar chromaticity but slightly different brightness than those of the same pixel in the background. For details, see [19].

3) Combining Two Keying Techniques: In 2), The result binary image computed by 1) is used as a mask image. Fig. 2 (c) shows the separation result using both difference keying and normalized difference keying.

![Figure 2](image)

Figure 2. Moving silhouette extraction. (a) Image when 4 directional lights were used, (b) only using difference keying, (c) after shadow elimination. Shadow was clearly eliminated.

B. Reliable Feature Extraction

After we create binary silhouette images, we extract the features representing the quality of the motion of a dance based on Laban’s movement theory (see Appendix for details) (Table 1).

Many researchers have used their features based on Laban’s movement theory to analyze human motion and have supported the theory lasting previous years [13-15, 17, 18, 20, 32, 33]. Among them, Suzuki et al. redefined the time-space-weight concept introduced by Laban to be as generally and easily extractable from video images as possible [13].

They omitted flow and redefined space as the wideness or an amount of space instead of design or form because they thought that it was a minor feature that could be derived from the redefined time-space-weight features. Afterward, Woo et al. introduced TDMLP to take flow into consideration [14]. Suzuki and Woo used similar features as shown in Table I except $N_d$.

<table>
<thead>
<tr>
<th>Features Extracted from Binary Silhouette Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>The aspect ratio of rectangle</td>
</tr>
<tr>
<td>The coordinates of centroid</td>
</tr>
<tr>
<td>The coordinates of the center of rectangle</td>
</tr>
<tr>
<td>The ratio between silhouette area and rectangle area</td>
</tr>
<tr>
<td>The number of dominant points on boundary</td>
</tr>
<tr>
<td>The velocity of each feature</td>
</tr>
<tr>
<td>The acceleration of each feature</td>
</tr>
</tbody>
</table>

Where $W$, $H$, $C_x$, $C_y$, $R_x$, $R_y$, and $S$ represent the width and height of bounding box, the $x$-, $y$-coordinates of centroid, the $x$-, $y$-coordinates of rectangle, the area of silhouette, and the area of rectangle, respectively. $f(\cdot)$ represents the velocity of each feature, $g(\cdot)$ represents the acceleration of each feature.

![Figure 3](image)

Figure 3. Similar but contextually different motions. (a) With only a bounding box, we cannot discriminate between two motions. (b) On the right, a little motion of her left leg gives rise to new dominant points. Then we can discriminate between the two motions.

Without the feature $N_d$, the difference between similar but contextually different human motions cannot be detected as shown in Fig. 3. Park et al. introduced $N_d$ to resolve this problem [32]. The improvement resulted from using $N_d$ is clear in their experimental results. In the same purpose, we use $N_d$ as a feature. And we use Teh-Chin’s algorithm to compute $N_d$ because it has shown reliable results even if the object is dynamically scaled or changed [21].

Dominant points represent the points that have significant change of curvature. This means that an object with a lot of dominant points is star-like, on the other hand, an object with few dominant points is circular. In other words, the number of dominant points represents the shape complexity of an object (see Fig. 4). Needless to say, the shape complexity of an object is related to space in Laban’s movement theory.

At this point in time, a question arises, “Why is the number of dominant points used as a feature, not the dominant points themselves or more detailed representations?” The answer is that it may suffice to utilize rough contour information because we do not aim to answer the question “what is the gesture he/she made?” but to catch the overall mood, i.e. emotion.

In general, the features extracted from dynamically changing dance images include some outliers. Outliers are main culprit of
decreasing recognition rate. To remove outliers, we calculate the mean and deviation of each feature in advance and clamp the outliers that deviate from the mean more than the allowed deviation. The value of each feature in Table I does not have same scale. To alleviate the scale problem, we normalize the features to be distributed evenly between their maximum value and minimum value. This normalization procedure can resolve the problem resulted from the physical difference between dancers at the same time [15].

Figure 4. Dominant point detection on (a) real images and (b) synthetic images. The number of dominant points is associated with the complexity of an object.

The features are inherently noisy and need to be smoothed to obtain a reliable result. Thus, we take frame-by-frame median filtering. The window size (= the number of frames) for median filtering heavily affects the recognition rate [15]. If the window size is too small, it tends to be noisy. On the contrary, if the window size is too large, valid information for effective recognition is blurred (= deteriorated). We set the window size to 10.

C. Hidden Space Transform

After applying the preceding algorithms to the low-level features, they are still high dimensional and hard-to-classify. We take into account statistical properties of the features. We apply SVD to the features and select the eigen-vectors having large eigen-value. This has an effect to remove redundant information out of the features while maintaining high recognition rate by classifying them in hidden space [22]. In the section, we give details of our hidden space transform for a reader with little experience in computer science. These details can be skipped by a reader in the field of computer science.

In the case of extracting n features from the dance sequence having m frames, we apply SVD to F matrix having m×n features as its elements. That is:

\[
F = U \Sigma V^T
\]

where

\[
F = \begin{bmatrix}
    f_1^1 & f_1^2 & f_1^3 & \ldots & f_1^n \\
    f_2^1 & f_2^2 & f_2^3 & \ldots & f_2^n \\
    f_3^1 & f_3^2 & f_3^3 & \ldots & f_3^n \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    f_m^1 & f_m^2 & f_m^3 & \ldots & f_m^n
\end{bmatrix}
\]

\[
\Sigma = \begin{pmatrix}
    \sigma_1 & 0 & \cdots & 0 \\
    0 & \sigma_2 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & \sigma_r
\end{pmatrix},
\]

\[
D = \begin{pmatrix}
    \sigma_1 & \cdots & 0 \\
    \vdots & \ddots & \vdots \\
    0 & \cdots & \sigma_r
\end{pmatrix}
\]

\[
\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \quad r \leq m, n.
\]

Here, \( f_i^j \) represents the value of the i-th feature in the j-th frame. \( \sigma_i \) represents the singular value and k-th column of \( U \) is eigen-vector associated with \( \sigma_i \). The \( r \) eigen-vectors (\( u_1, u_2, \ldots, u_r \)) associated with large \( \sigma \) are represented by linear combination of the feature vectors (\( f_1, f_2, \ldots, f_n \)). That is:

\[
u_1 = \alpha_1^1 f_1^1 + \alpha_1^2 f_2^1 + \cdots + \alpha_1^n f_n^1,
\]

\[
u_2 = \alpha_2^1 f_1^1 + \alpha_2^2 f_2^1 + \cdots + \alpha_2^n f_n^1,
\]

\[\vdots \]

\[
u_r = \alpha_r^1 f_1^1 + \alpha_r^2 f_2^1 + \cdots + \alpha_r^n f_n^1.
\]

Equation (6) is rewritten as follows:

\[
u_i = F \alpha_i \quad \text{for } i = 1, 2, \ldots, r,
\]

where

\[
\alpha_i = \begin{bmatrix}
    \alpha_i^1 & \alpha_i^2 & \cdots & \alpha_i^n
\end{bmatrix}^T,
\]

\[
F = \begin{bmatrix}
    f_1^1 & f_2^1 & \cdots & f_n^1 \\
    f_1^2 & f_2^2 & \cdots & f_n^2 \\
    \vdots & \vdots & \ddots & \vdots \\
    f_1^n & f_2^n & \cdots & f_n^n
\end{bmatrix}.
\]

Here, \( \alpha \) represents contribution coefficient vector and it is solved as follows:

\[
\alpha_i = (F^T F)^{-1} F^T u_i.
\]

Given \( \alpha \), the hidden space features are represented by weighted sum of the original features extracted from dance image sequence:

\[
p_1 = \alpha_1^1 f_1^1 + \alpha_1^2 f_2^1 + \cdots + \alpha_1^n f_n^1,
\]

\[
p_2 = \alpha_2^1 f_1^1 + \alpha_2^2 f_2^1 + \cdots + \alpha_2^n f_n^1,
\]

\[\vdots \]

\[
p_r = \alpha_r^1 f_1^1 + \alpha_r^2 f_2^1 + \cdots + \alpha_r^n f_n^1.
\]

where \( f' \) represents the original features extracted from each frame of dance sequence.

In practice, we calculated the contribution coefficients for 12 eigen-vectors having large eigen-values and extracted 12 hidden space features associated with the eigen-vectors every frame. The number of eigen-vectors can be changed according to the characteristics of data. We used 12 eigen-vectors because the eigen-values related to the eigen-vectors were much larger than the other ones. In addition, the hidden space features associated with few largest eigen-values cannot account for all of the three kinds of features i.e. space-time-weight. The ones associated with smaller eigen-values can account for the velocity and acceleration features (see Fig. 5). Finally, these 12 hidden space features are used as an input of TDMLLP that will be explained in the next section.

D. Classification of Hidden Space Features

The distribution of the features is very complicated and over-populated. It is almost impossible to classify them linearly.
We introduce a neural network to classify the features nonlinearly. Multi-Layer Perceptron (MLP) is a feedforward neural network that has hidden layers between input layer and output layer. MLP can be used to classify arbitrary data because its internal neurons have nonlinear properties [23].

The number of neurons of hidden layer is not fixed but can be determined by multiplying the number of inputs by the number of outputs. We obtain a satisfactory performance in this way.

When a sigmoid function is used as active function, the value of output is not rounded off to 0 or 1 and it is distributed uniformly between 0 and 1. If we use these values as is, the sum of errors becomes very large. Applying the Winner Take All (WTA) rule, the values of output are rounded and we can reduce the result error-sum.

In the case of dynamically changing data like dance, it is of great importance to find out how they vary temporally (Laban introduced the factor, flow, to explain it). However, MLP cannot deal with such a changing pattern of data. As shown in Fig. 6, TDMLP stores the input data in a buffer for some period and uses all the delayed data as input. Because putting together data acquired at different times has an effect on coping with the change among them, TDMLP can effectively analyze not only an instant value but also a changing patterns of input data [14].

To find the optimal number of delays is necessary because the performance of TDMLP is influenced by the number of delays [14, 15]. Through experiments, we confirmed that zero-delayed TDMLP (simply MLP) cannot cope with temporal changes of data and TDMLP with lots of delays lose its separability for data with different characteristics in spatial domain. In practice, we used 1-delayed MLP.

We used 12 hidden space features as an input of TDMLP as explained previous section. Thus, the TDMLP used in our experiments has 24 (=12+12) input nodes, 96 hidden nodes, and 4 output nodes.

![Figure 6. The structure of our TDMLP. It has three layers and a buffer in input layer.](Image)

III. EXPERIMENTAL RESULTS

To obtain experimental sequences, we captured the motions of a dance from 4 professional dancers using a static video camera (Cannon MV1). The images have 320×240 pixels. The dancers freely performed the various movements of a dance related to pre-defined emotional categories, keeping within about 5m×5m size of space with static background, within a given period of time. Fig. 7 shows the examples of the dance images. Some sequences were used for training the TDMLP and the others for testing the TDMLP, not duplicated.

The frame-rate of each sequence used in our experiments was 30Hz. We obtained 3 recognition results per second because we took the frame-by-frame median filtering every 10 frames as explained before. Strictly speaking, 1/3 sec is too short to estimate an exact emotion. Therefore, the recognition results
were saved into a buffer with a given size\(^1\) and we took the major one as a final result.

We implemented most of the algorithms using the functions in OpenCV library [24] that aims at real-time computer vision.

Because insufficient training data may result in erroneous recognition, we tried to capture the dance image sequence including various motions of a dance for a long time. If the sequence is too short, it has few specific patterns. The TDMLP learned with such a sequence can recognize only the specific patterns and thus does not have a good recognition rate. As shown in Fig. 8, when the number of frames is greater than 5000, satisfactory performance was obtained.

![Figure 7. Dance images.](image)

![Figure 8. Recognition rate according to the number of frames in training sequence. When the number of frames is greater than 5000, satisfactory performance was obtained.](image)

<table>
<thead>
<tr>
<th>Output</th>
<th>Happy</th>
<th>Surprised</th>
<th>Angry</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>88%</td>
<td>2%</td>
<td>10%</td>
<td>0%</td>
</tr>
<tr>
<td>Surprised</td>
<td>2%</td>
<td>80%</td>
<td>12%</td>
<td>6%</td>
</tr>
<tr>
<td>Angry</td>
<td>12%</td>
<td>4%</td>
<td>84%</td>
<td>0%</td>
</tr>
<tr>
<td>Sad</td>
<td>0%</td>
<td>12%</td>
<td>2%</td>
<td>86%</td>
</tr>
</tbody>
</table>

\(^1\) We set the buffer size to 6

Table II and III shows the cross-recognition rate of four emotions to the dance image sequences inside and outside the training sequence respectively. In both sequences, we could obtain more advanced performance than the previous methods [13, 14] and more balanced performance than the previous method [15].

The recognition rate outside the training sequence is less than that inside the training sequence since the method of each professional dancer’s expressing his/her emotion is different. This can be alleviated if we acquire the sample sequence from more dancers and have the TDMLP learn to adapt to various expression methods included.

We obtained above 70% recognition rate outside the training sequence using the proposed method. This is really nice. Even people cannot recognize other person’s emotion with that precision when they watch only his/her natural motion. In practice, we confirmed through a subjective evaluation on college students that people have the correctness of approximately 50-60%.

### IV. CONCLUDING REMARKS

We presented a framework to recognize human emotion from modern dance performance based on Laban’s movement theory. Through experiments, we obtained acceptable performance and confirmed that it was feasible to directly recognize human emotion using not physical entities but approximated features based on Laban’s movement theory.

We expect that the proposed method be practically used in entertainment and education fields that need humanlike, real-time human-computer interaction on which we are intensively investigating.

Currently we are trying to recognize human emotion from traditional dance performance that differs from modern dance.

As interesting future work, it is necessary to focus on how to categorize emotions. In this paper, we categorized human emotions into just four common ones. However, they are not sufficient for describing all emotional categories and might not be familiar to those who are from other (not Korea) countries. Anthropologists report that there are enormous differences in the ways that different cultures categorize emotions and the words used to name or describe an emotion can influence what emotion is experienced [25].
APPENDIX: LABAN’S MOVEMENT THEORY [16]

While it is abstract, human movement connotes a physical law and a psychological meaning. As Laban made use of this characteristic, he tried to observe, analyze and classify human movement. While the movement of animals is instinctive and responds to external stimulus normally, human movement is filled with human temperament. People express themselves and transmit something (effort, named by Laban) occurred to heart by their movement. To analyze the effort, Laban adopted the following elements:

1) **Weight:** The element of effort, “firm”, consists of a powerful resistance to the weight or depressed motion or heaviness as a sense of motion. The element of effort, “soft”, consists of a weak resistance to the weight or light feeling or lightness as a sense of motion.

2) **Time:** The element of effort, “sudden”, consists of a rapid speed or a short time or a short length of time as a sense of motion. The element of effort, “consistent”, consists of a slow speed or endlessness or long time of length as a sense of motion.

3) **Space:** The element of effort, “straight”, consists of a straight direction or narrowness or droop like a thread as a sense of motion. The element of effort, “pliable”, consists of a wavelike direction or curved shape or a bend toward different direction as a sense of motion.

Additionally, Laban adopted flow to represent the temporally changing pattern of human movement. It means that he tried to observe the tendency of human movement within a specific period as well as instant human movement.

Laban’s movement theory has been used by many group of researchers, dance performers, etc., to observe and analyze how human movement changes in weight, time and space.

REFERENCES


