Ana Fred
Terry Caelli
Robert P.W. Duin
Aurélio Campilho
Dick de Ridder (Eds.)

LNCS 3138

Structural, Syntactic, and Statistical Pattern Recognition

**Joint IAPR International Workshops
SSPR 2004 and SPR 2004
Lisbon, Portugal, August 2004, Proceedings**

Springer

Emotion Recognition from Dance Image Sequences Using Contour Approximation

Hanhoon Park[1], Jong-Il Park[1], Un-Mi Kim[2], and Woontack Woo[3]

[1] Department of ECE, Hanyang Univ., Haengdang-dong 17, Seongdong-gu, Seoul, Korea
hanuni@mr.hanyang.ac.kr, jipark@hanyang.ac.kr
[2] Department of Dance,Hanyang Univ., Haengdang-dong 17, Seongdong-gu, Seoul, Korea
kimunmi@hanyang.ac.kr
[3] U-VR Lab., K-JIST, Oryong-dong, Puk-gu, Kwangju, Korea
wwoo@kjist.ac.kr

Abstract. We present a novel approach that exploits shape context to recognize emotion from monocular dance image sequences. The method makes use of contour information as well as region-based shape information. The procedure of the method is as follows. First, we compute binary silhouette images and its bounding box from dance images. Next, we extract the quantitative features that represent the quality of the motion of a dance. Then, we find meaningful low-dimensional structures, removing redundant information but retaining essential information possessing high discrimination power, of the features using SVD (Singular Value Decomposition). Finally, we classify the low-dimensional features into predefined emotional categories using TDMLP (Time Delayed Multi-Layer Perceptron). Experimental results demonstrate the validity of the proposed method.

1 Introduction

For the last couple of years many researchers have focused on recognizing human emotion to achieve a more efficient and natural human-computer interface. The emotion recognition methods that extract emotional information from speeches or facial expressions have been extensively investigated [12-15]. However, gesture-based methods have been less explored except for a method that uses physiological signals [16]. This may be due to the fact that *gestures* are much too high dimensional, dynamic and ambiguous to analyze and recognize exactly.

Among some interesting exceptions are the following. To analyze the dynamically changing human motion, Kojima et al. defined "rhythm points" that means the time of the start and the end of a motion, so that the whole motion could be represented as the collection of partial motions (a unit of motion) with a period [10]. Wilson et al. also tried to identify temporal aspects of gesture [11]. They proposed a method that detects candidate rest states and gesture phases from gestures spontaneously generated by a person telling a story. Kojima's and Wilson's method are appropriate for analyzing a simple or well-regulated motion such as small baton-like movement. In

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 547–555, 2004.

general, however, human motions (especially in case of dance in this paper) are not so simple so we cannot easily determine the start/end of a unit of motion or rest state. Thus, their promising methods do not seem to be applicable to general human motions. In addition, these methods only aims at analyzing human motion but does not pay attention to emotions.

Recently, some methods that can recognize human emotion from modern dance image sequence have been introduced [3-6]. Dance is the most universal way of expressing human emotion. To represent the high-dimensional and dynamic change of gestures, these methods simplified the dynamic dance to the movement of rectangle surrounding human body and exploited several features related to the motion of rectangle using Laban's effort theory [9]. Then they analyzed the features and tried to find the relationship between the features and human emotion. They showed satisfactory results in most cases. However, they were confronted with difficulties in some cases. For example, when the bounding box associated with a human motion is the same as that associated with another one, they cannot discriminate the differences between the two human motions.

To resolve the problem mentioned above, this paper presents a method of exploiting new shape information for representing human motion. Note that the previous methods focus on only *region-based* shape information – the features such as bounding box or centroid etc. belong to the region-based shape information – but we additionally use the *contour-based* shape information, i.e. the number of dominant points on the boundary of human body. Thus our method can discriminate the subtle difference between the shape of human motion that have the same bounding box information.

The number of the computed dominant points can explain how complex or star-like human motion is and thus be regarded as a contour-based shape descriptor [1]. It cannot explain the details of the shape context of human motion. However, it may suffice to utilize rough contour information because we do not aim to answer the question to "what is the gesture he/she has just made?" but to catch the overall mood i.e. emotion.

The structure of this paper is as follows: In Section 2, we briefly describe the overview of emotion recognition system. The contour approximation algorithm used in our method is presented in Section 3. We then demonstrate the result of a variety of experiments in Section 4. We conclude in Section 5.

2 Overview of the Emotion Recognition System

It is not easy to directly extract high-level (and qualitative) information like emotion from low-level (and quantitative) data like human motion. Recently, however, efforts to extract human emotion from dance have been made. Many researchers thought that it would be easier to recognize emotion from dance than from ordinary human motions because *dance* is what fully expresses human emotion.

Some introduced Laban's movement theory [9] to represent emotional dance quantitatively. Suzuki et al. [3] and Camurri et al. [6] presented a methodology for mapping dance into emotional categories that represent human emotion based on the Laban's theory for the first time. Suzuki et al. and Camurri et al. had shown the possibility of directly recognizing human emotion from dance image sequences. Woo et al. proposed a method that nonlinearly maps dance into emotional categories and emphasized the importance of *flow*[1] by introducing the Time Delayed Multi-Layer Perceptron (TDMLP) [4]. Recently, we extended the work of Woo et al. and proposed a statistical approach that recognizes the human emotion faster and more accurately [5].

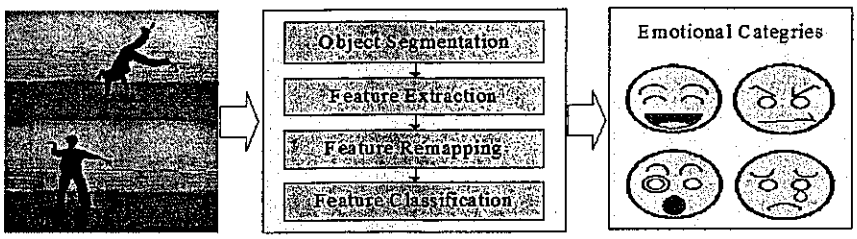

Fig. 1. Overview of emotion recognition system. The system directly maps low-level features extracted from dance images into predefined emotional categories.

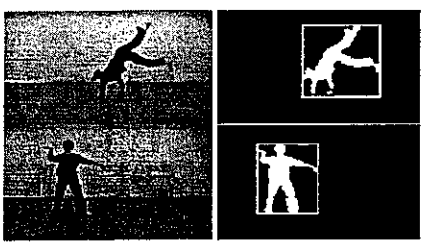

Fig. 2. Object segmentation. This shows the result of applying the background subtraction and the shadow elimination to an image. Refer to [7] for more details.

Table 1. Features extracted from binary images.

The aspect ratio of rectangle	H/W
The coordinate of centroid	(C_x, C_y)
The coordinate of the center of rectangle	(R_x, R_y)
The silhouette area	S_s
The rectangle area	S_r
The velocity of each feature	$f(\cdot)$
The acceleration of each feature	$g(\cdot)$
$f(x_n) = x_n - x_{n-1}$, $g(x_n) = x_n - 2*x_{n-1} + x_{n-2}$	

[1] *Flow* means the temporal variation of human motion, which is one of the factors that Laban adopted to quantitatively represent human motion.

Figure 1 shows the overview of our previous system [5]. First, the background and shadow in dance images are eliminated using the difference keying and normalized difference keying technique separately and then binary silhouette images and their bounding box are computed (Figure 2). Next, the features representing the quality of the motion of a dance are extracted (Table 1). These features are related to the factors i.e. *space*, *time*, *weight*, *flow* that Laban adopted to find out *effort*[2] from human motion. That is, Laban thought that they could quantitatively represent something (it may be emotion here) included in human motion. Next, Singular Value Decomposition (SVD) is applied to the extracted features and then the low-dimensional features associated with large eigen-value are selected. This has an effect of discriminating noisy information from the reliable features. Finally, the low-dimensional features are classified into predefined emotional space using TDMLP.

3 Emotion Recognition Using Contour Approximation

While the features (Table 1) used in our previous system could be easily computed in real-time and showed a good performance, there was difficulty in discriminating between similar but contextually different human motions as shown in Figure 3. Therefore, we need to find new features to resolve this problem. In this paper, we compute the dominant points on the boundary of the silhouette area to represent human motion more exactly. To do this, we introduce the number of the dominant points as a new feature.

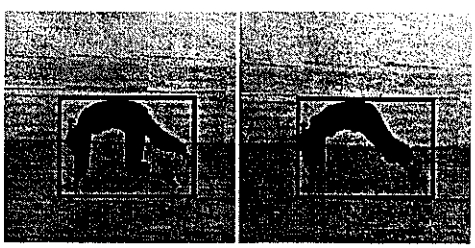

Fig. 3. Similar but contextually different motions. With only a bounding box, we cannot discriminate between two motions.

We use the Teh-Chin's algorithm [2] to detect the dominant points because it has shown reliable results even if the object is dynamically scaled or changed. The Teh-Chin algorithm is as follows: A measure of significance, e.g. curvature must be determined over some region of support. However, there is rarely a sound basis for choosing region of support. The chord length and the perpendicular distance can provide a basis for choosing the appropriate region of support. And the measure of

[2] Laban asserted that human motion include one's temper or propensity. In other words, human beings express themselves and transmit something (Laban called it *effort*) rising form their hearts through performing a motion.

significance of each point is determined by using the neighboring points within the extent of the region of support. The measure of significance and the region of support of each point are then used to guide the selection of points to be removed. The points remaining after the removal process are the dominant points.

Now we can figure out that the dominant points are those points that have a significant change of curvature. This means that an object that has many dominant points is star-like; on the other hand, an object that has few dominant points is circular. The number of dominant points represents the shape complexity of an object. Fig. 4 shows this relationship. The shape of an object has something to do with *space* in Laban's theory.

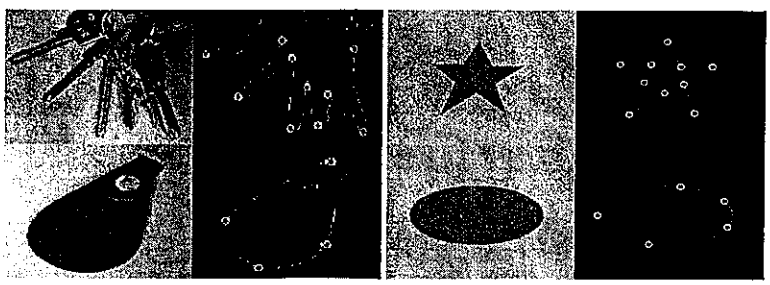

Fig. 4. Dominant point detection on real and synthetic images. The number of dominant points is associated with the complexity of the object.

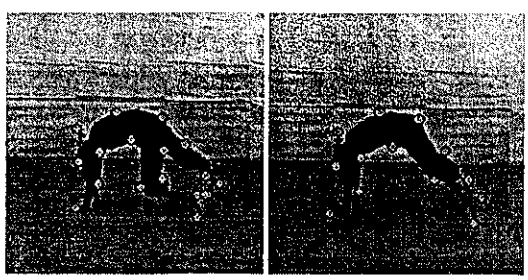

Fig. 5. The difference of the number of dominant points between similar but contextually different motions. In the left, a little motion of her left leg gave rise to more dominant points.

To exactly discriminate the similar but contextually different human motions, we must be able to describe the shape of the motions in detail. To do so, we will have to use the coordinates of respective dominant points. As mentioned earlier, however, we only need to use the number of dominant points because our aim is to catch the overall mood included in the motion of a dance. As shown in Fig. 5, the difference between similar but contextually different human motions can be detected easily even if we only use the number of dominant points. Notice that we could not discriminate the difference between similar but contextually different human motions using the bounding box information in Fig. 3.

4 Experiments and Results

To obtain experimental sequences, we captured the dance motion from 4 professional dancers using a video camera (Cannon MV1). The dancers freely performed the various movements of a dance related to pre-defined emotional categories (*happy*, *surprise*, *angry*, *sad*) within a given period of time. Fig. 6 shows the examples of dance images.

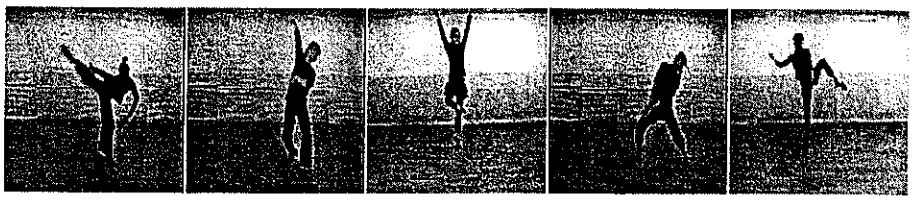

Fig. 6. Dance images. Each dancer performed the movements of a dance freely to express her emotion.

We eliminated the background and shadow of each frame in sequence and extracted binary images using the method previously explained. As shown in Table 2, we extracted 21 (7+7+7) features (of which we used in our previous method [5], we made S_s and S_r into one, i.e. ratio between them, and added N_d as a new one) representing dancing motion and applied SVD to them. We calculated the contribution coefficients (see Appendix) for 12 eigen-vectors having large eigen-values and extracted 12 features weighted with the contribution measure of every frame. Finally, we buffered 24 weighted features with one delay and exploited them as the input of the TDMLP. Then the TDMLP learned to map the weighted features to predefined emotional categories (*happy*, *surprise*, *angry*, *sad*) and was tested (= used) to recognize the emotion that an arbitrary dance image sequences represents. The TDMLP consists of 24 input nodes, 96 hidden nodes, and 4 output nodes. For more details about the experimental environments, refer to [5].

Table 2. Features used in the proposed method.

The aspect ratio of rectangle	H/W
The coordinate of centroid	(C_x, C_y)
The coordinate of the center of rectangle	(R_x, R_y)
The ratio between silhouette area and rectangle area	S_s/S_r
The number of dominant points on boundary	N_d
The velocity of each feature	$f(\cdot)$
The acceleration of each feature	$g(\cdot)$
$f(x_n) = x_n - x_{n-1}$, $g(x_n) = x_n - 2*x_{n-1} + x_{n-2}$	

Table 3. Recognition rate inside the training sequence.

Input \ Output	Happy	Surprised	Angry	Sad
Happy	88%(79%)[3]	2%(11%)	10%(8%)	0%(2%)
Surprised	2%(3%)	80%(78%)	12%(11%)	6%(8%)
Angry	12%(23%)	4%(0%)	84%(70%)	0%(7%)
Sad	0%(0%)	12%(7%)	2%(1%)	86%(92%)

Table 4. Recognition rate outside the training sequence.

Input \ Output	Happy	Surprised	Angry	Sad
Happy	75%(52%)	14%(16%)	11%(24%)	0%(8%)
Surprised	14%(16%)	70%(76%)	11%(0%)	5%(8%)
Angry	14%(14%)	15%(12%)	60%(64%)	11%(10%)
Sad	0%(0%)	11%(4%)	2%(2%)	87%(94%)

Table 3 and 4 shows the cross-recognition rate of four emotions to the dance image sequences inside and outside the training sequence respectively. In Table 3, we trained with sequences from all four dancers and tested with sequences again from all dancers. In Table 4, we tried to train with three dancers and test with the 4th dancer. The proposed method shows improved performance (higher and balanced recognition rate) than our previous system both inside and outside the training sequence. It is clear that this improvement results from using the new contour-based shape information, i.e. N_d because we used the same features excepted for this information in our previous method. Significantly, the proposed system can discriminate between some differences (e.g. the difference between *happy* and *surprise* in Table 3) that the previous method could not.

5 Conclusion

We have presented a new approach for recognizing human emotion from dance image sequence. A key characteristic of our approach is the use of contour-based shape information together with region-based shape information. Thus our method could recognize the difference between similar but contextually different human motions that have the same bounding box information. Consequently, our method noticeably improved the performance of emotion recognition compared with the previous ones that only use the region-based shape information.

It was confirmed that recognizing human emotion using not physical entities but approximated features based on Laban's theory is feasible and shows acceptable performance (above 70% recognition rate in outside the training sequence).

[3] A parenthesized value presents the result of our previous method [5].

Currently, we are continuing this line of research by trying to recognize human emotions from traditional dance performance that are a little bit different from modern ones.

Acknowledgement

This work was supported by the research fund of Hanyang University (HY-2002-T).

References

1. Kim, Y.-S.: Shape Descriptor for Content-Based Image Retrieval. Ph.D. dissertation, Hanyang University. (2000)
2. Teh, C.-H., Chin, R.T.: On the Detection of Dominant Points on Digital Curves. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.11, no.8. (1989) 859-872
3. Suzuki, R., Iwadate, Y., Inoue, M., Woo, W.: MIDAS: MIC Interactive Dance System. IEEE Intl Conf. on Systems, Man and Cybernetics, vol.2. (2000) 751-756
4. Woo, W., Park, J.-I., Iwadate, Y.: Emotion Analysis from Dance Performance Using Time-Delay Neural Netwroks. Proc. of CVPRIP'00, vol.2. (2000) 374-377
5. Park, H., Park, J.-I., Kim, U.-M., Woo, W.: A Statistical Approach for Recognizing Emotion from Dance Sequence. Proc. of ITC-CSCC'02, vol.2. Thailand (2002) 1161-1164
6. Camurri, A., Ricchetti, M., Trocca, R.: Eyeweb-toward gesture and affect recognition in dance/music interactive system. Proc. IEEE Multimedia Systems. (1999)
7. Kim, N., Woo, W., Tadenuma, M.: Photo-realistic Interactive Virtual Environment Generation Using Multiview Cameras. Proc. of SPIE PW-EI-VCIP'01, vol.4310. (2001)
8. Open Source Computer Vision Library. http://www.intel.com
9. Laban, R.: Modern Educational Dance. Trans-Atlantic Publications. (1988)
10. Kojima, K., Otobe, T., Hironaga, M., Nagae, S.: A Human Motion Analysis Using the Rhythm. Proc. of IWRHIC'00, Japan. (2000) 190-193
11. Wilson, A., Bobick, A., Cassell, J.: Temporal Classification of Natural Gesture and Application to Video Coding. Proc. of CVPR'97. (1997) 948-954
12. Fuji, R., Matsumoto, K., Mitsuyoshi, S., Gai, L.: Researches on the emotion measurement system. Proc. of ICSMC'03, vol.2. (2003) 1666-1672
13. Cohen, I., Sebe, N., Cozman, F., Cirelo, M., Huang, T.: Learning Bayesian Network Classifier for Facial Expression Recognition using both Labeled and Unlabeled Data. Proc. of CVPR'03, vol.1. (2003) 595-601
14. Lee, K., Xu, Y.: Real-time Estimation of Facial Expression Intensity. Proc. of ICRA'03, vol.2. (2003) 2567-2572
15. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion Recognition in Human-Computer Interaction. IEEE Signal Processing Magazine. (2001) 32-80
16. Picard, R., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.23, no.10. (2001) 1175-1191

Currently, we are continuing this line of research by trying to recognize human emotions from traditional dance performance that are a little bit different from modern ones.

Acknowledgement

This work was supported by the research fund of Hanyang University (HY-2002-T).

References

1. Kim, Y.-S.: Shape Descriptor for Content-Based Image Retrieval. Ph.D. dissertation, Hanyang University. (2000)
2. Teh, C.-H., Chin, R.T.: On the Detection of Dominant Points on Digital Curves. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.11, no.8. (1989) 859-872
3. Suzuki, R., Iwadate, Y., Inoue, M., Woo, W.: MIDAS: MIC Interactive Dance System. IEEE Intl Conf. on Systems, Man and Cybernetics, vol.2. (2000) 751-756
4. Woo, W., Park, J.-I., Iwadate, Y.: Emotion Analysis from Dance Performance Using Time-Delay Neural Netwroks. Proc. of CVPRIP'00, vol.2. (2000) 374-377
5. Park, H., Park, J.-I., Kim, U.-M., Woo, W.: A Statistical Approach for Recognizing Emotion from Dance Sequence. Proc. of ITC-CSCC'02, vol.2. Thailand (2002) 1161-1164
6. Camurri, A., Ricchetti, M., Trocca, R.: Eyeweb-toward gesture and affect recognition in dance/music interactive system. Proc. IEEE Multimedia Systems. (1999)
7. Kim, N., Woo, W., Tadenuma, M.: Photo-realistic Interactive Virtual Environment Generation Using Multiview Cameras. Proc. of SPIE PW-EI-VCIP'01, vol.4310. (2001)
8. Open Source Computer Vision Library. http://www.intel.com
9. Laban, R.: Modern Educational Dance. Trans-Atlantic Publications. (1988)
10. Kojima, K., Otobe, T., Hironaga, M., Nagae, S.: A Human Motion Analysis Using the Rhythm. Proc. of IWRHIC'00, Japan. (2000) 190-193
11. Wilson, A., Bobick, A., Cassell, J.: Temporal Classification of Natural Gesture and Application to Video Coding. Proc. of CVPR'97. (1997) 948-954
12. Fuji, R., Matsumoto, K., Mitsuyoshi, S., Gai, L.: Researches on the emotion measurement system. Proc. of ICSMC'03, vol.2. (2003) 1666-1672
13. Cohen, I., Sebe, N., Cozman, F., Cirelo, M., Huang, T.: Learning Bayesian Network Classifier for Facial Expression Recognition using both Labeled and Unlabeled Data. Proc. of CVPR'03, vol.1. (2003) 595-601
14. Lee, K., Xu, Y.: Real-time Estimation of Facial Expression Intensity. Proc. of ICRA'03, vol.2. (2003) 2567-2572
15. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion Recognition in Human-Computer Interaction. IEEE Signal Processing Magazine. (2001) 32-80
16. Picard, R., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.23, no.10. (2001) 1175-1191

Appendix: Contribution Coefficient Vector

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

$$\mathbf{F} = \mathbf{U}\Sigma\mathbf{V}^T$$

The r eigen-vectors (u_1, u_2, ..., u_r) associated with large eigen-values are represented by linear combination of the feature vectors (f_1,f_2, ...,f_n). That is:

$$\begin{aligned}
\mathbf{u}_1 &= \alpha_{11}\mathbf{f}_1 + \alpha_{12}\mathbf{f}_2 + \cdots + \alpha_{1n}\mathbf{f}_n \ , \\
\mathbf{u}_2 &= \alpha_{21}\mathbf{f}_1 + \alpha_{22}\mathbf{f}_2 + \cdots + \alpha_{2n}\mathbf{f}_n \ , \\
&\vdots \qquad\qquad\qquad \vdots \\
\mathbf{u}_r &= \alpha_{r1}\mathbf{f}_1 + \alpha_{r2}\mathbf{f}_2 + \cdots + \alpha_{rn}\mathbf{f}_n \ .
\end{aligned}$$

This is rewritten as follows:

$$\mathbf{u}_i = \mathbf{F}\boldsymbol{\alpha}_i \quad for\ i = 1, 2, \cdots, r,$$

Here, we call α_i contribution coefficient vector and it's solved as follows:

$$\boldsymbol{\alpha}_i = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{u}_i .$$