MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Feature extraction from Biological motion of human gait patterns for emotion discrimination

Hidenori Maruta
Information Media Center
Nagasaki University
1-14 Bunkyo, Nagasaki, 852-8521 Japan
hmaruta@nagasaki-u.ac.jp

Masahiro Ishii
Graduate School of Innovative Life Science
University of Toyama
3190 Gofuku, Toyama, 930-8555 Japan
ishii@eng.toyama-u.ac.jp

Abstract

We study a method of a feature extraction to discriminate emotions of human from a sensing data of human gait patterns as "Biological motion data". We assume that the high–dimensional biological motion data are generated by low–dimensional features whose components are statistically independent. So we use a method of independent component analysis to extract the features. The extracted feature is evaluated by a discriminated result of the given biological motion data which identified five types of categories, "Anger", "Grief", "Disgust", "Joy" and "Fear". We achieve 40% accuracy for 5–classes of emotion discrimination with 3 actors' biological motion data.

1. Introduction

Computers that interact with humans, such as nurse robots, pet robots, computer agents, etc., are pervading into our daily life. Smooth comfortable interaction between human beings and computers requires taking account of emotions while emotions are indispensable to communicate among human beings. A large number of studies have been made on building computers that discriminate human emotions from facial images, speeches, or other biological signals. What seems to be lacking, however, is use of physical movements that can be remotely captured with camera.

In this paper, we study the feature extraction to discriminate human emotions by computers with the biological motion[5] of dynamic human gait patterns, because it is one of the important cue to estimate one's emotion. A psychological experiment[2] shows that human beings can recognize the emotion of the dancing actor with 6–classes of emotions, "Disgust", "Grief", "Joy", "Anger", "Fear" and "Surprise", only from point-lights displays of biological motion. In that case, an accuracy of observers' reports achieved 63%. However, we don't use overacting human gait patterns such as dance that is extraordinary, but use usual one represented some emotions by actors. This is because we think usual one is suitable for human–machine interaction in life.

To extract the feature for the emotion discrimination, we assume that the biological motion (time series) data are generated from a few dimensional feature (time series) vector, whose elements are statistically independent. And we also assume that these features have similar properties in the same representing emotions and are not influenced by the difference of actors, and so they must have the same dimensionality. These assumptions have a very important role to reduce a computational complexity for their application such as the real–time human-machine interactive systems. On these assumptions, we use independent component analysis (ICA) to extract the feature vector. And we also use principal component analysis (PCA) not only as the dimensionality reduction technique but also as a pre–processing to determine the dimensionality of the feature vector(described in Section 2).

To show how the proposed feature can represent emotions, we categorized five different emotional biological motion data, "Anger", "Grief", "Disgust", "Joy" and "Fear" of 3 actors and achieved 40% accuracy (Section 3). Some discussion and conclusion are offered in Section 4.

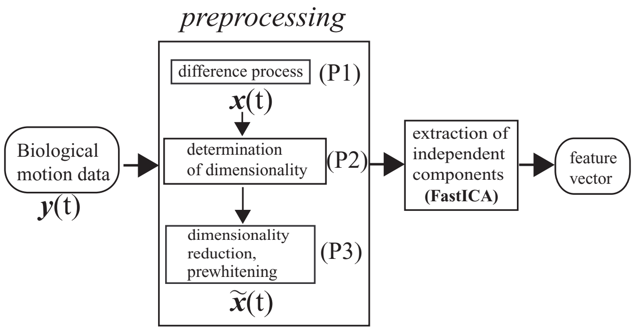
2. Feature Extraction from Biological motion

2.1 Procedure to obtain biological motion data

To obtain the biological motion data of actors, we put 12 positional sensors onto the actor's wrist/elbow/shoulder/waist/knee/ankle of both the left and the right side. A motion sensing system is electromagnetic and a sampling rate of the system is 30FPS. The actors are three male university students who have some experiences in acting. We label these actors as actor1, actor2 and actor3. To obtain the emotional gait pattern data, we indicate to actors to walk representing 5–classes emotions, "Anger", "Grief", "Joy", "Disgust", and "Fear". And we also indicate that during a walking, not to represent the emotion with static pose such as shrugging his shoulders or looking at a ground , and not to overact like dancing. We don't impose any time–limitation to actors, so each data has a different time length. As each positional sensor data is composed of 3–dimensional spatial time series, the biological motion data is the 36–dimensional time series. However, as a physical symmetry property of a human body, we use only the right side's information. So the biological motion data treated here becomes the 18–dimensional time series. We represent obtaining biological motion data as $\mathbf{y}_i(t) \in \mathbf{R}^{18}$($i$=1,2,$\cdots$,15; 3 actors, 5 emotion classes), where i means the index of the each data.

2.2 Feature extraction

In this subsection, we propose a feature extraction method for the emotion discrimination from the biological motion data. The proposed method is summarized in Figure 1.



preprocessing
difference process
(P1)
x(t)
Biological motion data
y(t)
determination of dimensionality
(P2)
dimensionality reduction, prewhitening
(P3)
x̃(t)
extraction of independent components
(FastICA)
feature vector

Figure 1. Summary of the extracting process of the feature vector

As a first step, the biological motion data $\mathbf{y}_i(t)$ are converted to their difference process data, $\mathbf{x}_i(t) = \mathbf{y}_i(t) - \mathbf{y}_i(t-1)$, instead of the original one((P1) of Figure 1). This is because, as shown in the latter, we use the ICA method based on the *non-gaussianity* of the signals, and the difference process is more non-gaussian than the original one[3]. Accordingly, obtained positional data $\mathbf{y}_i(t)$ are converted to movement distance data, $\mathbf{x}_i(t)$. Next, we consider about pre–processing for a determination of the dimensionality of the feature vector((P2) of Figure 1). The feature vector is assumed that:

- whose elements are statistically independent
- it is not depend on actors, and it has same dimension among representing emotions

We also assume that $\mathbf{x}_i(t)$ are generated by the feature as the instantaneous linear mixture, this means that the mixing matrix is conserved while the actor has the same emotion. These assumptions are represented as eq. (1), where A_i is a time–independent mixture matrix and $\mathbf{s}_i(t)$ is the feature vector whose components are statistically independent.

$$\mathbf{x}_i(t) = A_i \mathbf{s}_i(t) \ \ (i = 1, 2, \cdots, 15) \qquad (1)$$

The problem here is to how to determine the dimensionality of $\mathbf{s}_i(t)$. To solve this, we examine contributions of the variance to see the redundancy of $\mathbf{x}_i(t)$ using PCA. For all i, the first two components taken together accounted for between 70.9% to 84.0% of the all variance. Because of this, we reduce the dimensionality of $\mathbf{x}_i(t)$ into a 2-dimensional vector by projecting it into a spanned orthogonal subspace by the first two principal components. And we also perform a *prewhitening* such that each elements of the 2-dimensional vector are mutually uncorrelated and all have unit variance((P3) of Figure 1). After the determination of dimensionality of the feature vector and prewhitening, $\mathbf{x}_i(t)$ is transformed to $\tilde{\mathbf{x}}_i(t) \in \mathbf{R}^2$. Thus eq. (1) has reduced to (2), where $B_i(t)$, due to the pre–processing, to be an orthogonal matrix.

$$\tilde{\mathbf{x}}_i(t) = B_i \mathbf{s}_i(t) \qquad (2)$$

At the same time, we determine the dimensionality of the feature vector $\mathbf{s}_i(t)$ which is 2, equals to the dimensionality of the pre-processed signal $\tilde{\mathbf{x}}_i(t)$.

Let us now consider how to recover the feature vector $\mathbf{s}_i(t) = (s_i^1(t), s_i^2(t))$ from $\tilde{\mathbf{x}}_i(t)$. In order to solve this problem, we use FastICA[4], which is one of the widely used method of ICA due to its simplicity and efficiency. In this case, it becomes a problem to search for a linear combination of the elements of $\tilde{\mathbf{x}}_i(t)$, say, $\hat{s}_i^j(t) = \mathbf{w}_j^T \tilde{\mathbf{x}}_i(t)$, such that it is the most *non-gaussian*. The procedure of FastICA is summarized in Figure 2. Note that recovered signals $\hat{s}_i^j(t)$ have a zero mean and a unit variance. Some examples of the feature vectors (independent component) are shown in Figure 3 and 4. We can see the difference of the two elements, which have different type of the periodic properties. We notice that a recovered signal has ambiguities in a sign and a order, this means that recovered signals are not ensured to become same each time. This becomes one of the problems when we compare the feature vectors, we consider this afterward.

(i) Take a random initial $\mathbf{w}_j$ of norm 1, let j=1
(ii)To converge $\mathbf{w}_j (j = 1, 2)$, iterate following steps

1. $\mathbf{w}_j^+ = E\left[\tilde{\mathbf{x}}_i(t) g\left(\mathbf{w}_j^T \tilde{\mathbf{x}}_i(t)\right)\right] - E\left[g'\left(\mathbf{w}_j^T \tilde{\mathbf{x}}_i(t)\right)\right]\mathbf{w}_j$, $g(u) = tanh(u)$
2. $\mathbf{w}_j = \frac{\mathbf{w}_j^+}{\left\|\mathbf{w}_j^+\right\|}$

(iii)$\mathbf{w}_{j+1} = \mathbf{w}_j - \Sigma_{k=1}^{p-1}\left(\mathbf{w}_k^T \mathbf{w}_k\right)\mathbf{w}_k, \mathbf{w}_{j+1} = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$ $(p = 2)$

Figure 2. The procedure of the independent component extraction with FastICA[4]

2.3 Comparison method of the feature

For the emotion discrimination, we have to compare the extracted feature. In this subsection, we consider the comparison method of the extracted feature for the discrimination of the input data. Because each $\mathbf{x}_i(t)$ has a different

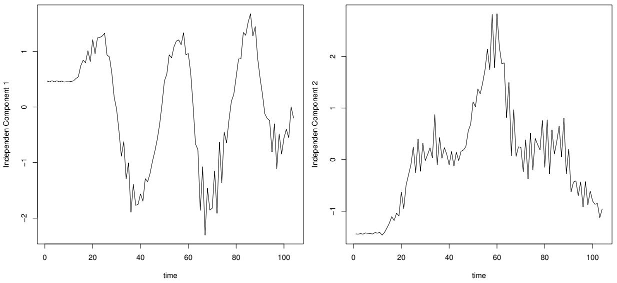


Figure 3. Example of two elements of the extracted feature(actor1, "Anger"). A horizontal axis is a time variable and a vertical axis is amplitude of the recovered signals.

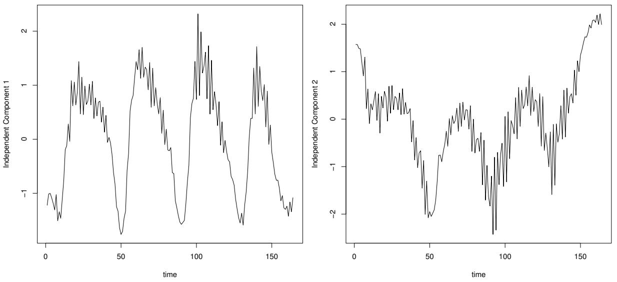


Figure 4. Example of two elements of the extracted feature(actor1, "Grief"). Horizontal and vertical axes are same as Figure 3.

time length, we extract a partial sequence from each data to compare. To extract the partial sequence, we select a start and an end point manually, based on the periodic property of the first principal component which appears in preprocessing for $\mathbf{x}_i(t)$. We select two local maximal points as the start and the end points so as to the extracted partial sequence which includes three local maximal points. Figure 5 shows some example of extracted partial sequences (shadowed). Table 1 shows the time length of extracted partial sequences of all data.

Following this, we determine the partial sequences $\tilde{\mathbf{s}}_i(t)$ from $\mathbf{s}_i(t)$. Table 1 reports that each $\tilde{\mathbf{s}}_i(t)$ has the different time length, so we must deal with variable length sequences. Because of this, we use symmetrical Dynamic Time Warping(DTW)(e.g. [1]) to normalize the variable time length. DTW algorithm used here is shown in (3), where $g(0,0) = d(\tilde{\mathbf{s}}_i(0), \tilde{\mathbf{s}}_j(0))$ and $d(\mathbf{x}, \mathbf{y})$ is Euclidean distance of $\mathbf{x}, \mathbf{y} \in \mathbf{R}^2$.

$$
\begin{aligned}
& g(t_i, t_j) \\
& = min \left\{ \begin{array}{c} g(t_i, t_{j-1}) + d(\tilde{\mathbf{s}}_i(t_i), \tilde{\mathbf{s}}_j(t_j)) \\ g(t_{i-1}, t_{j-1}) + 2d(\tilde{\mathbf{s}}_i(t_i), \tilde{\mathbf{s}}_j(t_j)) \\ g(t_{i-1}, t_j) + d(\tilde{\mathbf{s}}_i(t_i), \tilde{\mathbf{s}}_j(t_j)) \end{array} \right. \\
& \qquad (1 \leq t_i \leq l_i, 1 \leq t_j \leq l_j) \qquad (3) \\
& D(\tilde{\mathbf{s}}_i(t), \tilde{\mathbf{s}}_j(t)) = g(l_i, l_j)/(l_i + l_j)
\end{aligned}
$$

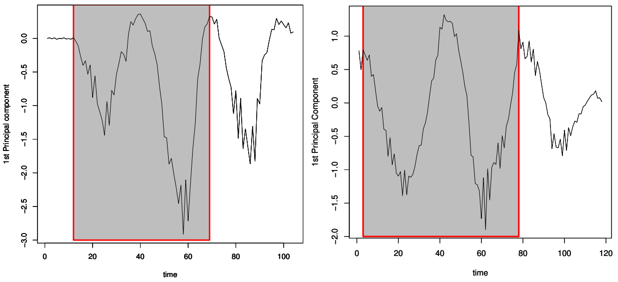


Figure 5. Examples of extracted partial sequences from the first principal component of the each data (left: the first principal component of actor1, "Anger"; right:the first principal component of actor3, "Joy")

However, there remains a problem noticed before, that the extracted feature has ambiguities in the sign and the order. To solve this problem, we redefine the distance $D_f(\tilde{\mathbf{s}}_i(t), \tilde{\mathbf{s}}_j(t))$ of $\tilde{\mathbf{s}}_i(t)$ and $\tilde{\mathbf{s}}_j(t)$ as the minimum value of $D(\tilde{\mathbf{s}}_i(t), \tilde{\mathbf{s}}_j(t))$ for all of the combinations of ambiguity in the sign and the order. Based on this redefined DTW D_f distance, we can compare and, as a result, discriminate the feature into classes of emotions.

3. Evaluation of the proposed feature extraction method

To show a validity of the proposed method, we compare the extracted features and classify the data into 5-class of emotions. We define a distance $\tilde{D}(\tilde{\mathbf{s}}_i(t), C)$ between the feature $\tilde{\mathbf{s}}_i(t)$ of $\mathbf{x}_i(t)$ and the class C which is one of the emotions as shown in (4). In this case, $C \in \{A(Anger), G(Grief), D(Disgust), J(Joy), F(Fear)\}$ means the class of emotions and $n = 3$.

$$
\begin{aligned}
& \tilde{D}(\tilde{\mathbf{s}}_i(t), C) \\
& = \left\{ \begin{array}{l} \frac{1}{n-1} \sum_{\tilde{\mathbf{s}}_j(t) \in C, \tilde{\mathbf{s}}_j(t) \neq \tilde{\mathbf{s}}_i(t)} D_f(\tilde{\mathbf{s}}_i(t), \tilde{\mathbf{s}}_j(t)) \\ \frac{1}{n} \sum_{\tilde{\mathbf{s}}_j(t) \in C} D_f(\tilde{\mathbf{s}}_i(t), \tilde{\mathbf{s}}_j(t)) \end{array} \right. \qquad (4)
\end{aligned}
$$

The distance $\tilde{D}$ defined above means:

1. if $\tilde{\mathbf{s}}_i(t) \in C$, $\tilde{D}$ is given as a mean of $D_f(\tilde{\mathbf{s}}_i(t), \tilde{\mathbf{s}}_j(t))$, where $\tilde{\mathbf{s}}_j(t) \in C$, except $\tilde{\mathbf{s}}_i(t)$ itself,

2. else, $\tilde{D}$ is given as a mean of $D_f(\tilde{\mathbf{s}}_i(t), \tilde{\mathbf{s}}_j(t))$, where $\tilde{\mathbf{s}}_j(t) \in C$.

The result is shown in Table 3. This table reports:

- we achieve 40.0% accuracy that the represented emotion class by the data corresponds with the (emotion) class which has the minimal $\tilde{D}$

- we achieve 66.7% accuracy that the represented emotion class by the data corresponds with the (emotion) class which has the minimal or the second minimal $\tilde{D}$

Table 1. Time length of extracted partial sequences (The alphabet represents "A:Anger, G:Grief, D:Disgust, J:Joy, F:Fear", the number represents "1:actor1, 2:actor2, 3:actor3")

data	A1	A2	A3	G1	G2	G3	D1	D2	D3	J1	J2	J3	F1	F2	F3
time length	57	55	63	79	86	90	68	70	70	68	66	75	72	70	72

Table 2. Table of the comparison/discrimination result with eq. (4). The double underline value is the minimal $\tilde{D}$ for the each data, the single underline value are the second minimal $\tilde{D}$.

	A(Anger)	G(Grief)	D(Disgust)	J(Joy)	F(Fear)
A1	0.6960 (2)	0.7122 (4)	0.7635 (5)	0.7093 (3)	0.6516 (1)
A2	0.5716 (1)	0.6112 (3)	0.6918 (5)	0.6570 (4)	0.6025 (2)
A3	0.6062 (3)	0.4647 (2)	0.4247 (1)	0.6912 (5)	0.6105 (4)
G1	0.6769 (5)	0.5001 (2)	0.5525 (3)	0.5603 (4)	0.4921 (1)
G2	0.5507 (3)	0.4776 (1)	0.4985 (2)	0.6177 (5)	0.5553 (4)
G3	0.6260 (4)	0.4757 (2)	0.4530 (1)	0.6637 (5)	0.6165 (3)
D1	0.6725 (5)	0.3902 (1)	0.5755 (3)	0.6294 (4)	0.5507 (2)
D2	0.4819 (2)	0.5353 (3)	0.4662 (1)	0.6392 (5)	0.5717 (4)
D3	0.5226 (2)	0.5784 (4)	0.4874 (1)	0.6536 (5)	0.5544 (3)
J1	0.6784 (3)	0.7048 (5)	0.6950 (4)	0.5988 (2)	0.5535 (1)
J2	0.7750 (5)	0.5611 (2)	0.6321 (4)	0.6059 (3)	0.5588 (1)
J3	0.6064 (3)	0.6289 (4)	0.5950 (2)	0.5789 (1)	0.6853 (5)
F1	0.6324 (5)	0.5448 (2)	0.4844 (1)	0.6280 (4)	0.5659 (3)
F2	0.4770 (2)	0.4730 (1)	0.5430 (4)	0.5272 (3)	0.6426 (5)
F3	0.7552 (5)	0.6462 (3)	0.6495 (4)	0.6423 (2)	0.6044 (1)

4. Conclusion and Future work

We propose the feature extraction method from the biological motion data for the emotion discrimination. We start from supposing that the emotion discrimination can be realized by using the low dimensionality vector. And we use some preprocessing methods for the dimensionality reduction and the determination of the dimensionality of the feature vector. After that, we extract the feature vector with ICA. As for the extracted feature, each element of it has different properties that come from the difference of the emotion of the data. To see the efficiency of the proposed method, we compare the extracted feature to discriminate it into 5–class of emotions with eq. (4). From this comparison result, the extracted *low-dimensional* feature from the proposed method can be one of the important information resources for human–machine interactive systems, which use the emotional information of humans.

Further study is needed to improve the accuracy and the robustness of the emotion discrimination results. We think some possible directions of the study, for example, using other information like frequency components of the feature, or assuming the time–dependency of the mixture matrix in eq (1). These methods will extend the proposed method which can be used in the time–dependent human–machine interactive systems. We also think a larger dataset of biological motion data is needed for more conclusive studies.

Acknowledgments This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 17500140, 2005-06.

References

[1] Bristow, G., *ed.*, *Electronic speech recognition*, Collins, 1986.

[2] Dittrich, W., H., Troscianko, T., Lea, S., E., G., Morgan, D., "Perception of emotion from dynamic point-light displays represented in dance", Perception, 25, p727–738, 1996

[3] Hyvärinen, A., "Independent component analysis for time–dependent stochastic processes", Proc. Int. Conf. on Artificial Neural Networks, p135–140, 1998

[4] Hyvärinen, A., Oja, E., "Independent component analysis: algorithms and applications", Neural Networks, 13, p411–430, 2000

[5] Johansson, G., "Visual perception of biological motion and a model for its analysis", Perception & Psychophysics, 14, p201–211, 1973

[6] Troje, N., F., "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns", Journal of Vision, 2, p371–387, 2002