A Method of Multi-factorization for Recognizing Emotions from Gestures

Masahide Naemura, Masaki Takahashi, Mahito Fujii and Nobuyuki Yagi

Science & Technical Research Laboratories
Japan Broadcasting Corporation (NHK)
1-10-11 Kinuta, Seatgayaku, Tokyo 157-8510, Japan
{naemura.m-ei,takahashi.m-iu,fujii.m-ii,yagi.n-iy}@nhk.or.jp

Abstract

We propose a new method of recognizing emotional factors from human gestures by analyzing motion capture (MoCap) data. It features multi-factorization processing combined with HMM recognition. The multi-factorization processing factorizes MoCap data into a third-order tensor that consists of spatial, statistical, and frequency-spatial components. This multi-factorization localizes the data in the factorized tensor space according to their mutual correlation, which results in helping data clustering. This means that the proposed tensor-shaped features have advantages over conventional features in recognizing emotions from gestures. The validity of the proposed method was confirmed using the results of experiments in which emotions from walking actions were analyzed.

1. Introduction

Demand has recently risen for more sophisticated semantic processing that can identify the internal states of content to achieve optimal goals in media processing. Media processing using video content mainly requires extremely semantic metadata to be efficiently extracted from media content such as TV programs. Therefore, we are carrying out extensive activities focusing on research on intelligent extraction of metadata [1][2][3]. One area we are tackling to extract more semantic metadata from video sequences is research on the recognition of emotions from gestures.

Studies on gesture recognition have been widely carried out since pattern-recognition research [4][5] first began. These have mainly aimed at recognizing different categories of human gestures such as walking or running, not taking the emotions that accompany them into account. Thus, emotion recognition from gestures is a field of research that has not been sufficiently exploited and many matters still bear further investigation. In fact, there have not been many previous studies on emotion recognition from gestures. However, it is becoming increasingly important to automatically deduce the internal states of people appearing in video content assisted by all possible means since these results can be applied to various application domains. In line with this scenario, the recognition of emotions from facial expressions has already attracted a great deal of attention and numerous novel methods are being developed [6][7]. The recognition of emotion from gestures is therefore expected to prevail in the future.

It is more difficult to recognize emotions from gestures than from our facial expressions. This is attributed to the complexity of all body parts moving interactively in three-dimensional space when emotions are being expressed with gestures, which interferes with efficient methods of recognizing them. To solve this problem, effective features have to be extracted and synthesized from the original data obtained from moving human-body parts. We developed a novel approach to achieving this aim, which was based on a combined method of multi-factorization of motion data and an HMM process that followed. Multi-factorization places the corresponding fractions of motion data into a third-order tensor, each index of which is spatially, statistically, and frequency-spatially factorized utilizing a subspace method and wavelet analysis. We did experiments to confirm the validity of the proposed method and encouraging results were obtained. This paper presents the details on our new approach.

2. Related work

Recent research on recognition of emotions from gestures has mainly been carried out from the psychological point of view. For example, Kobayashi and Naemura’s psychological study [8] confirmed that emotion can be discerned from differences in human-body movements through subjective estimates. However, there is not much previous work on emotion recognition from gestures that has been studied from the engineering point of view. This is due to the fact that there is a lack of fundamental knowledge about the mechanism responsible for what parameters are related to emotional expressions when gestures are made. In addition, large degrees of freedom (DOFs) in parametric space prevent motion features from being efficiently extracted.
Nevertheless, some trials are being done that classify human gaits using motion data [9][10]. These are not directly aimed at extracting human emotions, but they somehow have similar aspects to emotion-recognition research. Most of these first extract motion features that can accurately be identified from input-video content and they then identify humans or their categories of gait within an HMM framework. This research is expected to next be extended to emotion recognition. In fact, some work on emotion recognition from gestures [11][12][13] has already appeared. These researchers classified emotions using support vector machines (SVMs) and obtained relatively good results, although under rather constrained conditions. However, these previous researchers did not reach the level where subtle emotional movements within the same categorized gesture could be analyzed, which our research has aimed at.

Advanced work on computer graphics (CGs) is often seen. Brundelin and Williams [14] proposed a fundamental method of generating virtual movement in CG characters by tweaking the motion parameters. Unuma et al. [15] and Amaya et al. [16] did research on morphing CG characters according to pre-defined emotions. Emotional motion was virtually created by giving some pre-defined fluctuations to the parameters that defined motion dynamics in their work. Although emotional movements could seemingly be achieved, their methods were not based on real emotional movements that human beings make. By extending these classical approaches, more sophisticated methods that can take statistical observations of human gestures into account have attracted a great deal of interest in the field of CG research. Brand et al. [17] learned about differences in walking styles in the shapes of HMM parameters as a result of statistically training various examples of walking, and virtually created a new walking sequence by generating motion dynamics belonging to the required HMM class. Pioneering work on computer synthesis can be seen in the research by Rose et al. [18]. Various emotional poses were synthesized in their work by interpolating between multiple examples that were placed on an imaginary space specified as 'verbs' and 'adverbs'. Grochow et al. [19] synthesized smooth CG animation by interpolating a virtual pose with a statistical inverse kinematic model that they built from machine-learning techniques using a set of training pose data, even if the interpolated poses were not contained in the training set. These CG methodologies were concerned with the regeneration of subtle differences in styles the CG characters were supposed to have in making gestures.

This work is expected to be extended to a new algorithm that will enable further human-like movements to be created by combining research outcomes from the recognition of gestures from emotions, once motion features that are associated with emotional expressions in human gestures are clarified. Taking these insights into consideration, we are aiming at devising a new method that not only detects emotions from gestures but also makes it possible to reuse the extracted motion parameters to then generate virtual emotional movements. In the first step to achieving this goal, we developed a fundamental method in which original motion data were first factorized into a third-order tensor fitted to gesture-adaptive processing and the factorized data were recognized within an HMM framework.

3. Proposed method for emotion recognition

The algorithm for emotion recognition itself needs to be a generative model that can re-create the learned dynamics and enable the recognized emotional body movements to be applied to synthesizing the virtual gestures of a CG character. To accomplish this purpose of synthesizing virtual emotional movements as well as recognizing emotions from gestures in the future, the method we propose is based on an HMM that can regenerate stochastically resembling motion-feature sequences, which are equivalent to a virtual gesture, from the learned parameters. We will describe the details on the new method in this section.

3.1. Overview of algorithm

Fig.1 overviews the proposed method. The angular motion data for joints are obtained from a motion capture system. Then, the angular data for joints at frame $t$ are transformed into $M(t)$, which consists of quaternions and their time differentials in Eq. (1), and are used as input-motion data.

$$
M(t) = 
\begin{bmatrix}
    m_1(t) \\
    \vdots \\
    m_{BM}(t)
\end{bmatrix}
= 
\begin{bmatrix}
    (q_1(t), q'_1(t))^T \\
    \vdots \\
    (q_{BM}(t), q'_{BM}(t))^T
\end{bmatrix}
$$

where $q_i$ is the quaternion of the i-th body joint transformed from the angular data and $q'_i = q_i(t) - q_i(t-1)$. Here, the subscript, $BM$, represents the number of all joints. After this, we will omit time index $(t)$ from the expression of motion features for the sake of brevity. The reason we used quaternion-based data as motion data is that they have superior capabilities for expressing movements[20]. As can be discerned from Fig. 1, input motion data $M$ are spatially, statistically, and frequency-spatially factorized into third order tensor data in the multi-factorization process that consists of spatial-, statistical-, and wavelet-grouping processes. Then the optimum index elements are selected from the third-order tensor, and fed to the HMM learning/recognizing process. In the learning phase, HMM parameters are stored in the database as well as selected index information on the tensor. Emotion recognition is carried out using these learned parameters in the same way.
as the learning process. In parallel with the learning process, automatic temporal-motion segmentation, which is one of the main features of the proposed method (and will be explained in Section 4), is accomplished with the aid of wavelet analysis. This process helps us to amass a large volume of training data without excess effort.

3.2. Multi-factorization processing

3.2.1 Spatial (body parts) grouping

Motion-feature vector $\mathbf{M}$ is spatially factorized in the five cardinal body groups, which are the backbone (BB), right-arm (RA), left-arm (LA), right-leg (RL), and left-leg (LL) parts depicted in Fig. 2. The human body discussed in this paper has a total of 22 joints. Since there are eight five cardinal body groups, which are the backbone (BB), right-arm (RA), left-arm (LA), right-leg (RL), and left-leg (LL) parts depicted in Fig. 2. The human body discussed in this paper has a total of 22 joints. Since there are eight elements in each joint, the total number of elements in $\mathbf{M}$ is $22 \times 8 = 176$. Each of the body groups contains the joints below.

- $G_1 = \text{BB} = \{\text{Hips, Spine, Spine1, Spine2, Neck, Head}\}$
- $G_2 = \text{RA} = \{\text{RightShoulder, RightArm, RightForeArm, RightHand}\}$
- $G_3 = \text{LA} = \{\text{LeftShoulder, LeftArm, LeftForeArm, LeftHand}\}$
- $G_4 = \text{RL} = \{\text{RightUpLeg, RightLeg, RightFoot, RightToes}\}$
- $G_5 = \text{LL} = \{\text{LeftUpLeg, LeftLeg, LeftFoot, LeftToes}\}$

Equating an index number set $\{1, \ldots, |G_i|\}$ to the $i$-th body group set, this spatial grouping can be expressed as

$$\text{partM}[i] = \begin{bmatrix} m_{G_i[1]} \\ \vdots \\ m_{G_i[|G_i|]} \end{bmatrix} \quad (2)$$

where $i$ indicates the index of body groups and $|G_i|$ represents the cardinality of the $i$-th body group set. As a result of this grouping process, a 176-dimensional vector $\mathbf{M}$ is factorized into five-arrayed data $\text{partM}$ that are comprised of a $48 \times 8$ dimensional vector for BB and into a $32 \times (8 \times 4)$ dimensional vector for RA, LA, RL, and LL, so that we can deal with each element of $\text{partM}$ separately after this.

3.2.2 Statistical grouping

In the statistical-grouping process, spatially grouped data $\text{partM}$ are statistically factorized into uncorrelated data $\text{statM}$ using an eigen-space method [21] per body group. Fig. 1 outlines this statistical factorization. First, a set of all motion vectors $\text{partM}[i]$, which is comprised of the $i$-th spatially grouped motion data, is collected in all $N$ frames during which the specified gesture is made. At each frame, the element values of $\text{partM}[i]$ are normalized so that the total sum becomes zero and $\mathbf{M} \times N$ matrix $\mathbf{X}$ is generated separately per spatial group for each emotion by concatenating this normalized vector in $N$ units. $\mathbf{M}$ is assumed to be 48 for BB and 32 for RA, LA, RL, and LL. All matrices $\mathbf{X}$ are factorized by singular value decomposition (SVD) as

$$\mathbf{X} = \mathbf{U} \mathbf{W} \mathbf{V}^T \quad (3)$$
As a result of SVD, the column vectors of U are eigenvectors in ascending order of eigenvalues, which are diagonal values of W in Eq. (3). Using this matrix U of eigenvectors, arbitrary motion data on the spatial group at any frame belonging to emotion e, $x_e(t)$, is expressed as

$$ x_e(t) = c + \sum_{j=1}^{n} b_{e,j}(t) u_{e,j}(t) \quad (4) $$

Here, $b_e = (b_{e,1}, b_{e,2}, \ldots, b_{e,n})$, $n = \text{MIN}(M,N)$ is a projection vector of the eigenvectors and each value indicates the amount of deviation in the subspace generated by the eigenvectors. In estimating dominant motion, it is sufficient to estimate some elements in $b_e$ whose corresponding eigenvalues are sufficiently large because their contribution to movement is proportional to the total amount of the eigenvalues. For all spatial-grouped-data partM of the e-th emotion, projected data that satisfy these conditions are selected from the elements of $b_e$. This means that the spatial group data are statistically factorized into a second-order tensor, statM, whose dimensionality is reduced.

### 3.2.3 Wavelet grouping

Taking the previous work as a reference in which wavelet analysis was found to be extremely effective in analyzing the motion of subtle differences caused by body movements [22], we frequency-spatially factorized the data that were processed in the spatial- and statistical-grouping processes into a third-order tensor using a discrete wavelet transform.

The proposed method uses a Harr wavelet filter for factorizing statM into the third-order tensor. The resulting transformed data were well grouped around similar levels of wavelet spaces due to the characteristics of wavelet transform. After all grouping processes were completed, we obtained third-order tensor TransM. In a nutshell, the original motion-vector temporal sequences of the original 176-dimensional motion vectors M were factorized into sequences of third-order-tensor data TransM using these three grouping processes. Here, the primary index of TransM indicates that there are five body parts and the third index indicates that there are six wavelet levels. The second index corresponds to the number of eigenspaces factorized by the statistical-grouping process and its dimensionality changes depending on the statistical properties of the original motion data. Fig.3 schematically illustrates the concept underlying this factorization process. In this figure, spatially, statistically, and frequency-spatially uncorrelated factors extracted from the original motion data are placed in positions corresponding to the box-like tensor to form sequences of tensor data. The final feature vector for the HMM training/learning process is made up by concatenating the element values in the selected tensor indices. The set of selected indices, Idx, is expressed as

$$ \text{Idx} = \{\text{body_idx, eig_idx, wav_idx,} \} \quad (5) $$

where $(\text{body_idx, eig_idx, wav_idx})$ indicates the index of the tensor data, and $\text{BG}$, $\text{WG}$, and $\text{EG}$ correspond to the index subspace selected from the entire tensor space $W = \{G1,..,G5\} \times \{\text{eigen1}..\text{eigenL}\} \times \{\text{level1}..\text{level6}\}$, respectively. Here, L is the number of dimensions reduced by statistical grouping.

![Fig.3: Schematic of multi-factorization processing. Optimum indices are selected from created tensor for emotion recognition depending on kind of gesture](image)

### 3.3. HMM training/recognition process

The third-order tensor sequence created through the factorization process was applied to an emotional learning/recognition process designed within the HMM framework. We used a continuous Gaussian mixture model as a type of HMM.

#### (1) Training HMM parameters

We used the conventional Baum Welch algorithm [23] to train the fundamental HMM parameters, $\lambda_{\pi} = (\pi, A, B)$, where $\pi$ is a set of the initial state probabilities, $A$ is a matrix of transition probabilities, and $B$ is a set of GMM parameters in each state. Here, $A$ subscript $e$ indicates the
class ID of emotions. In addition to the HMM parameters, we adopted a set of tensor indices selected as additional parameters; the final parameters we learned in the training phase being \( \Pi_e = (\lambda_e, Idx) \). The \( Idx \) is currently determined heuristically.

(2) Recognition using HMM

In recognition, an emotion whose likelihood is maximum in the learned HMM is detected as a recognized emotion, obeying the following:

\[
emotionID = \arg \max_e \{ P_e \} \tag{6}
\]

where \( P_e \) is the likelihood of emotion \( e \) in HMM recognition.

4. Experiments

4.1. MoCap Experiment

We collected motion data from gestures that three professional actresses performed with designated emotions using an optical MoCap apparatus (Vicon). When we instructed the actresses how to walk, we only demanded that they walk according to the emotions they felt. Therefore, they ended up by walking freely according to their own preferences.

The conditions for the MoCap experiment were as follows:

- Performers: Three females of average size
- Acting gesture: Walking
- No. of actions: Each actress repeated gestures twice
- Emotions: Anger, Happiness, Sadness, Indifference
- No. of joints for MoCap: 22 (see Fig. 2)

The acquired motion data were processed using MotionBuilder\textsuperscript{TM} to store them in FBX file format per action. The average FBX file was 600 frames long. Fig. 4 has skeletonized examples of the emotions performed.

![Skeletonized images of emotions in walkings](image)

Fig. 4 Skeletonized images of emotions in walkings

![Flowchart for motion segmentation algorithm](image)

Fig. 5: Flowchart for motion segmentation algorithm. An index of \( TransM \) is selected on the onset of this process so that the selected data sequences can be the most periodic among all index sets. In walking, we selected \( TransM[RL][0][6] \) for walking, meaning that a body part is the RightLeg, the eigen space is a principle, and the wavelet level is 6.

![Example of motion segmentation](image)

Fig. 6: Example of motion segmentation. Red dots indicate points of temporal segments in \( TransM \).
4.2. Motion Segmentation

Since the proposed method is based on an HMM framework, it is important to efficiently collect training data. A simple way of collecting useful data for HMM learning is by manually segmenting the intervals of primitive movements that accurately represent emotional gestures. However, this operation requires vast amounts of labor due to the necessity for large amounts of training data. We developed a new method of automatically segmenting the intervals of primitive movements of gestures using wavelet analysis. Fig.5 is a flowchart for the algorithm. The input is temporal sequences of data with high periodicity among all wavelet levels of the projected data in the principle eigenspace, which can be regarded as accurately representing the comprehensive movements of gestures. For example, in the walking gestures in our experiment, we used sequences of \[ \text{TransM}[k][i][j][l] \] for input because the fundamental movements in a walking gesture are contained in those of the right leg.

The basic period, \( T \), in the flowchart is computed using the FFT of data sequences in the wavelet-level layer for which there is the highest periodicity of all wavelet levels. An interval during consecutive local maximum points of the signal is output as the emotional segment for the training data. A search for the local maximum points is carried out by gradually changing the search window starting from basic period \( T \). We assumed that intervals whose length was close to that of the basic frequency contained in the original input signal would accurately represent the basic movements of the gesture. Fig.6 illustrates one of the results for this algorithm, which indicates that it works efficiently.

Thus, multiple segments of primitive walking motions automatically contained in the sequence were extracted. In the experiment, the numbers of extracted segments were: happiness: 40, anger: 40, sadness: 32, and indifference: 32.

Using these segments, we estimated the results of HMM recognition on the basis of k-fold cross validation (\( k=4 \)), in which each estimate of four trials was carried out on 30 training data and 10 test data for anger and happiness, and 24 training data and 8 test data for sadness and indifference.

4.3. Estimation Results

To demonstrate the validity of our new approach, we compared the proposed method with the conventional approach. We adopted the PCA-HMM (Principal Component Analysis-HMM) method as the conventional method since it is generally used for motion recognition tasks. PCA in the PCA-HMM method directly factorized the original motion data expressed in Eq. (1) into dimensionally reduced motion features without any grouping processes, and the motion features were fed to the HMM process to recognize emotions from the gestures. A 24-dimensional vector was created for this comparison of estimates by directly reducing all the 176-dimensional input motion data \( \mathbf{M}_i \) down to 24-dimensional data by PCA projection. The dimensionality of the reduced data was determined with the contribution rate of eigenvalues being 0.95. A 30-dimensional vector was created for the proposed method by concatenating a subspace of \( \text{TransM}[k][i][j][l] \) into vector form. The subspace index is:

\[
(k, j, i) \in \{(k, j, i) | k \in \{LA, RA, BB, LL, RL\}, j \in \{eigen1, eigen2, eigen3\}, i \in \{level5, level6\}\}
\]

Table 1(a)-(b) lists the results of the estimates. We can see the confusion matrices with the recall and precision rates. The total rate in the table was \( R/N \) where \( R \) equals the number of correctly recognized segments and \( N \) equals the total number of segments. It is interesting to see recall for happiness is very high with the conventional method while the rate for the other three emotions being misclassified as happy is also high. We considered that arm movements were so diverse in happy gestures that happy factors were dominant in the PCA-reduced data and affected the recognition results for the other three emotions. We can see from the results in Table 1(b) that the proposed method, for which the input signal was factorized into tensor form, was more balanced for all emotions than the conventional approach. This difference was considered to have been derived from the proposition that the dominance of arm movements in the factorized data was weakened by the spatial-grouping process. Moreover, we attributed the improvements in the other three emotions to the hypothesis that the distributed states of motion in the frequency domain affected the emotional expressions in making gestures and that the wavelet-grouping process in the proposed method accurately reflected this property of gestures. This corresponds to the assertions made in previous work [24].

Next, we selected \{BB, LL, RL\} as the body-part indices of the third-order tensor, excluding RA and LA from the selected body parts in the experiment since the RA and LA movements for walking were so diverse that it was difficult to extract meaningful statistical properties due to the small amount of training data from the previous estimation results. The motion feature used in this estimation was an 18-dimensional vector that was created by concatenating the sub-space of \( \text{TransM}[k][i][j][l] \) into vector form.

Here,

\[
(k, j, i) \in \{(k, j, i) | k \in \{BB, LL, RL\}, j \in \{eigen1, eigen2, eigen3\}, i \in \{level5, level6\}\}
\]

Table 1(c) lists the results of estimates for the second proposed method. We can see that total performance is improved by excluding RA and LA from the spatial grouping for this data set. This suggests that the proposed
method can be flexible in selecting motion features according to their properties.

In addition, we compared the proposed method with another PCA-HMM method in which PCA is applied only to the same body parts as the proposed method in Table 1(c), and we confirmed that the proposed method excels by about 0.4 in the total rate. From this result, it is considered that the multi-factorization of the propose method can extract the core elements that represent emotional movements well while using the specific body parts ambiguates the emotional recognition in the conventional PCA-HMM.

Finally, we selected several different sets of tensor indices, from which the final motion features were concatenated, and applied them to the HMM process. The total rate reached nearly 0.8 at the index set {RL, RL, BB} and was saturated beyond this index as plotted in Fig. 7. This suggests that there is an optimal set of indices for recognizing emotion from gestures.

Table 1 Estimation results

(a) PCM-HMM method

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>Happy</th>
<th>Indifferent</th>
<th>Sad</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>19</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0.46</td>
<td>0.48</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>1</td>
<td>0.53</td>
<td>0.98</td>
</tr>
<tr>
<td>Indifferent</td>
<td>19</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0.53</td>
<td>0.98</td>
</tr>
<tr>
<td>Sad</td>
<td>3</td>
<td>6</td>
<td>23</td>
<td>0</td>
<td>0.96</td>
<td>0.72</td>
</tr>
<tr>
<td>Total rate=</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Proposed method 1

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>Happy</th>
<th>Indifferent</th>
<th>Sad</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>24</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0.55</td>
<td>0.6</td>
</tr>
<tr>
<td>Happy</td>
<td>12</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0.61</td>
<td>0.7</td>
</tr>
<tr>
<td>Indifferent</td>
<td>5</td>
<td>2</td>
<td>26</td>
<td>1</td>
<td>0.93</td>
<td>0.81</td>
</tr>
<tr>
<td>Sad</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>25</td>
<td>0.96</td>
<td>0.78</td>
</tr>
<tr>
<td>Total rate=</td>
<td>0.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(c) Proposed method 2

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>Happy</th>
<th>Indifferent</th>
<th>Sad</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>29</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>Happy</td>
<td>4</td>
<td>35</td>
<td>0</td>
<td>1</td>
<td>0.73</td>
<td>0.88</td>
</tr>
<tr>
<td>Indifferent</td>
<td>4</td>
<td>1</td>
<td>25</td>
<td>2</td>
<td>0.93</td>
<td>0.78</td>
</tr>
<tr>
<td>Sad</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>25</td>
<td>0.89</td>
<td>0.78</td>
</tr>
<tr>
<td>Total rate=</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The total rate is defined as RR/NN, where RR is the correct recognized number, and NN is the total number of test samples.

5. Conclusion and future directions

We proposed a new method of recognizing emotions from gestures, which is a research topic that has not been exploited extensively from an engineering point of view. It features multi-factorization processing, in which MoCap data are factorized into a third-order tensor, with a combination of HMM recognition. Multi-factorization processing factorizes MoCap data into third-order-tensor data that include spatial, statistical, and frequency-spatial axes. The components on these axes were obtained by respectively grouping body parts, an eigen-space method, and wavelet grouping. The experiments we carried out, in comparison with other method without multi-factorization, revealed the superiority of the proposed method. We conjectured that the multi-factorization of motion data shows promise in effectively recognizing emotions from gestures even if the research is at a preliminary stage. The new approach needs to be enhanced so that it can deal with more complicated movements using hands. We therefore need to collect more examples of gestures that underlie emotions.

Moreover, the way the proposed method was evaluated has room for improvement. We assessed it on the basis of the actresses’ intentions in their choice of emotions regardless of their skills. Estimates should be made on the basis of people’s general opinions about the emotional expressions in the gestures they saw to meet research objectives. We also intend to model an emotion-recognition framework from more complicated gestures, establish a framework for hierarchically recognizing both gestures and emotions, and synthesize virtual emotional gestures in CG.
References


