Human Action Recognition Based on Non-negative Matrix Factorization

Chih-Yang Lin[1,2], Bo-You Chen[3], Wen-Chuan Wu[4], Wei-Yang Lin[3*] and Chia-Ling Tsai[5]
[1]Dept. of Computer Science & Information Engineering, Asia University, Taichung, Taiwan
Email: andrewlin@asia.edu.tw
[2]Dept. of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan
[3]Department of Computer Science & Information Engineering, National Chung Cheng University, Chiayi, Taiwan
Email: wylin@cs.ccu.edu.tw
[4]Dept. of Computer Science & Information Engineering, Aletheia University, Taipei, Taiwan
Email: au4387@au.edu.tw
[5]Department of Computer Science, Iona College, NY, USA
Email: ctsai@iona.edu
*Corresponding Email: wylin@cs.ccu.edu.tw

***Abstract*—In this paper, we propose a method to recognize human behavior by combining motion history images (MHI) and non-negative matrix factorization (NMF). The MHI preserves the temporal information of a behavior by holding the temporal motion appearance. Then, NMF is applied to extract the middle-level features of the moving object. The experimental results show that the proposed scheme can achieve robust recognition results using a public dataset.**

Keywords: Action recognition; non-negative matrix factorization; motion history image.

I. Introduction

Human behavior recognition is an important contemporary issue in computer vision, but it also poses a challenging problem because the human body is a non-rigid object and can change shape arbitrarily. Some researchers, like Bobick and Davis [3], have combined several motion history images (MHIs) from different angles to achieve motion recognition. Weinland and Boyer [4] have used 3D motion history to assist with recognition. Both methods require multiple MHIs from different sources. However, real-world cases may demand the ability to recognize human behavior from just one resource. Moreover, the 3D motion history volume proposed by Weinland and Boyer [4] necessitates heavy computations that pose barriers to mainstream implementation of their method. These drawbacks have contributed to our proposed method for human behavior recognition, which combines motion history images (MHI) and non-negative matrix factorization (NMF).

In this paper, we use MHI and motion energy image (MEI) to represent a human motion from one single source. Each MHI represents a pose with silhouette and temporal information. In other words, MHI can present several successive frames with motion information using a single 2D image. MEI is similar to MHI, but MEI lacks time correlations between frames. The 2D image is then used as the input for non-negative matrix factorization (NMF) [2], a middle-level feature extractor that can be used for a middle-level feature description. Since it is difficult to describe human behavior via low-level features, how to construct meaningful middle-level features for higher-level recognition becomes a challenging issue. Our goal in this paper is to test the potential of NMF for higher-level recognition. As the experimental results show, NMF can be effectively used in action recognition and middle-level extraction.

The following parts of this paper are organized as follows: Firstly, the proposed method is provided in Section 2. Next, the experimental results are presented in Section 3. Finally, conclusions are presented in Section 4.

II. Related Work and Proposed Method

A. Motion History Image

Motion history image (Fig. 1) is similar to motion energy image (Fig. 2); both of the two representations are able to present spatial information. However, unlike MEI, MHI can both carry spatial information, as well as keep temporal successive information. The definition of MHI is shown as follows:

$$\begin{aligned} H(x, y, t) = {} & t \qquad \text{if } D(x, y, t) = 1, \\ & \text{Or} = \max(0, H(x, y, t{-}1) - 1, \text{otherwise.} \end{aligned} \qquad (1)$$

According to Eq. (1), pixel intensity is based on the motion history at that location, where brighter values represent more recent motion. Therefore, the pixel with the most recent motion has the largest value. MHI can accurately capture the current motion status, but it makes it easy to overlook the information that has been passed across a period of time. In order to complement the strengths and compensate for the drawbacks of of MHI, MEI can be applied to generate a binary image, where each pixel in MEI represents whether this pixel has moved or not. Thus, the combination of MHI and MEI can provide more information that is better suited for some computer vision applications.

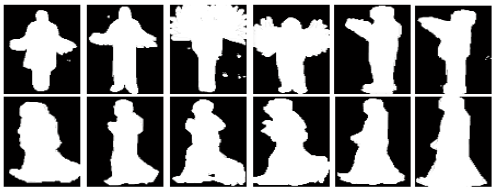

Fig 1. Motion Energy Image.

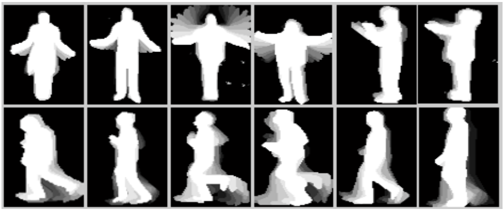

Fig 2. Motion History Image.

B. *Non-negative Matrix Factorization*

Although MHI and MEI can preserve the motion history required for human behavior recognition, suitable features must be extracted from them to help action recognition. This task is particularly difficult because low level features, such as intensity, color information, edge information, and motion vectors, are not sufficient for action recognition, which requires much more complicated descriptions.

In our proposed method, NMF is applied to extract middle-level features based on the part-based characteristics of NMF. The model of NMF is formulated as follows, where all three matrices have no negative elements:

$$V \approx WH. \qquad (2)$$

In this paper, each column in matrix V represents an MHI data, matrix W is a basis matrix, and the column matrix H represents a coefficient matrix. Initially, W and H are random matrices. By iterating the updating rules shown in Eqs. (3) and (4), the result of multiplying W and H gradually approximates V. This is the training process of obtaining middle level features.

$$W_{i\alpha} \leftarrow W_{i\alpha} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{\alpha\mu} \qquad (3)$$

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \sum_{i} W_{i\alpha} \frac{V_{i\mu}}{(WH)_{i\mu}} \qquad (4)$$

After the training process becomes convergent, the matrix W is called the basis, and the matrix H becomes the corresponding weights. Any column in V can be regarded as a linear combination of columns of W using coefficients by H. Therefore, the matrix W can also be regarded as middle-level features.

C. *Proposed Method*

In the training phase, each type of action, like jogging, will be put through the process of NMF. Each training video of the same type will be transformed into an MHI and regarded as a column of V. After training, the output, W, is the basis matrix for a type of action. Different types of actions will have their own W's. Then, we concatenate all types of W's to form a larger basis called Wa to involve possible actions. The basis image of Wa is shown in Fig. 3.

In the test phase, the test data will be also transformed to an MHI and treated as a column of vi as shown in Eq. (5).

$$v_i \approx W_a h_i. \qquad (5)$$

Based on the NMF process, the given v_i and W_a can obtain h_i. The largest coefficient of h_i determines what type of action v_i belongs to.

To improve the recognition rate, the test data can be partitioned into several pieces, each of which generates an MHI. Using Eq. (5), an MHI will obtain a vote to determine its action type; the final decision is made by its voting result.

Similar types of actions, such as walking and jogging, are tough to distinguish using the above process. A two-layer recognition process was built in order to solve for this. The first layer is the same as before. However, when the test data belongs to an action and its votes are lower than or equal to 50% of the total votes, the testing process moves down to the second layer. In the second layer, the number of columns of V is increased to involve more different MHI's. The sampling rate of this layer for constructing MHI adopts every 2 frames and every 3 frames as shown in Fig. 4. By doing this, the similar pose but different speeds can be differentiated.

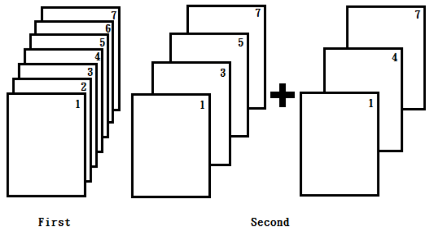

Fig. 4: Different sampling rates for the second layer.

III. EXPERIMENTS

Our experiments leveraged the public dataset (http://www.nada. kth.se/cvap/actions/) with 6 motion types and each MHI was composed of 20 frames. The experimental results are shown in Table 1. The second column of Table 1 is the first-layer recognition results, where over half of the testing data can be recognized correctly. When the second-layer recognition is applied, the recognition rate can be greatly improved as shown in the third column of Table 1. The result demonstrates that the proposed multi-layer structure can actually enhance recognition and correct judgment errors.

Table 1: Precision of human behavior recognition using 6 motion types

	1st layer	2nd layer
Waving	85%	100%
Boxing	100%	100%
Clapping	80%	100%
Jogging	65%	100%
Running	80%	100%
Walking	55%	85%

IV. Conclusions

In this paper, we combine NMF and MHI for a new action recognition approach. MHI was chosen for its ability to preserve rich motion history information, while NMF was selected for its middle-level feature generation benefits because the descriptions of corresponding MHI were not easily represented by its low level features. The higher level features of NMF are more suitable for complicated recognition needs, like those involving human behavior. The experimental results show that the combination of NMF and MHI, and the multi-layer recognition structure pose a promising approach for human action recognition.

Acknowledgment

This work was supported by Ministry of Science and Technology, Taiwan, under Grants MOST 103-2221-E-468-007-MY2 and 103-2221-E-194 -034.

References

[1] A. Bobick and J. Davis, "Real-time recognition of activity using temporal templates," in *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision*, 1996, pp. 39-42.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, pp. 788-791, 1999.

[3] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, 2001.

[4] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding,* vol. 104, no. 2, pp. 249-257, 2006.