A DMHI Technique for Human Motion Separation and Recognition

Md. Atiqur Rahman Ahad, T. Ogata, J. K. Tan, H. S. Kim, S. Ishikawa
Department of Control Engineering, Kyushu Institute of Technology
Sensuicho 1-1, Tobata, Kitakyushu 804-8550, Japan
Tel: +81-93-884-3191, Fax: +81-93-884-3183
E-mail: atiqahad@ss10.cntl.kyutech.ac.jp

Abstract:

Motion separation and recognition within a video sequence is one of the most challenging problems. This paper describes a new method of separating simple human motion by employing four-directional motion history images, called DMHI. It also covers human motion recognition utilizing DMHI technique. Four partial motion history images are created by exploiting the concept of Efros et.al.'s motion descriptors. The optical flow is separated into four components based on the four directions, namely, up, down, left and right. Based on this separation, four motion history images are developed and the corresponding brightness-levels are simultaneously integrated to decide the motion separation based on the four direction. The implementation results show that this new approach can separate motion promptly and properly.

人の動作の分割と認識のための DMHI 法

アハド アティクル ラハマン，緒方 健人，タン ジュークイ，金　亨燮，石川 聖二
九州工業大学大学院 工学研究科 機械知能工学専攻
福岡県北九州市戸畑区仙水町１－１〒804-8550
Tel: +81-93-884-3191, Fax: +81-93-884-3183
E-mail: atiqahad@ss10.cntl.kyutech.ac.jp

Abstract: 本論文は，人動作認識法を提案する．提案法は，従来のモーションヒストリーイメージ法に対し，動作を複数方向に分割して複数のモーションヒストリーイメージを生成することにより動作認識を行う．このため，本法の特徴として，従来法では難しかった複雑な動作の認識が可能である．提案法をラジオ体操の認識に適用した．実験では複数の被験者にラジオ体操を行ってもらい，そのビデオデータから体操の種類を認識した．提案法によれば，演技者の巧拙にかかわらずよい認識率が得られた．提案法をさらに機能強化して知能ロボットに実装すれば，人の生活上のさまざまな動作・活動の認識が可能になり，荷物を持つ，転びそうになったら手を貸すなど，知能ロボットによる人の行動・活動支援も将来可能になる．

1. Introduction

Motion recognition is a very promising and important research area in the field of computer vision, robotics and image processing. It has diverse application, for example, robotics, human-computer interaction, intelligent video surveillance, mixed reality, etc., to name a few. Human activity analysis in videos requires motion segmentation that facilitates the motion recognition easily and smartly to aid a robot to understand the semantics of motion to take its own decision instantly. Motion segmentation refers to grouping together pixels that undergo a common motion [1] or major motions in an image sequence [2]. In other words, it is the task of partitioning an image sequence to distinct regions based on different underlying coherent motion of each [3]. Motion segmentation and motion separation has been widely used in computer vision applications, such as in pattern recognition, environment visualization, analyzing medical images (medical imaging), face analysis, object tracking, scene text extraction, video-processing, video-conferencing, recognition, obstacle avoidance, etc. [3]. Motion separation and its recognition will be the motto of this paper. We will first present the multi-directional motion history image (DMHI) method for human motion recognition. Afterwards, we will present the motion separation based on the DMHI method.

We present the theory for motion representation for both the MHI and the development of DMHI algorithm after presenting some related works in the next section. Complex motion recognition methodology with some results and analysis by employing the DMHI is presented in Section 4. Section 5 illustrates the motion segmentation methodology. Finally, we conclude the paper with some recommendation for future work in Section 6.

2. Related Work

The DMHI method is the extension of MHI method [9]. In order to describe *how* the motion is moving in the image sequence, we form a motion-history image (MHI), and to represent *where* the motion or a spatial pattern is, we demonstrate this by creating motion energy image (MEI) [9]. The MHI expresses the recency of motion using intensity and it is a scalar-valued image where more recently moving pixels are brighter, and, in that way, it explains how the motion is moving in the image sequence. Bright pixels represent recent movement and dark pixels represent past movement and, therefore, it implicitly represents the direction of movement. Intensity of each pixel in MHI is a function of motion density at that location [10]. The basis of MHI is a *temporal template* - a static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence [9]. We extend basic motion history image (MHI) and the associated motion energy image (MEI). On the other hand, in motion segmentation arena, Kahol et.al. [5] focused on developing a gesture segmentation system, which is based on HMM states, called Hierarchical Activity Segmentation. Prior to their work, Adam Kendon showed in his work that humans first segment a gesture, and then recognize or *tag* it – a process analogous to segmentation of spoken words from a continuous stream of audio, and then recognition of the segmented word by the neuron-audio system in the humans [6]. Yuan et.al. [7] presented a method for detecting motion regions in video sequences observed by a moving camera, in the presence of strong parallax due to static 3D structures.

A number of interactive systems have been successfully constructed using motion history template technology as a primary sensing mechanism. For example, using the MHI method, Davis et.al. [11] developed a virtual aerobics trainer that watches and responds to the user as he/she performs the workout. An interactive art demonstration can be constructed from the motion templates [11]. KidsRoom – an interactive, narrative play space for children [12] was developed using MHI method successfully. Yau et.al. [13,14] has developed a method for visual speech recognition employing the MHI method. The video data of the speaker's mouth is represented using grayscale images named as motion history image. Valstar et.al. [15] demonstrated motion history for facial action detection in video successfully to analyze the subtle changes in facial behavior by recognizing facial action unit (AU) that produces expressions. Automatically localizing and detecting moving object/vehicle for an automatic visual surveillance and tracking system was demonstrated [16], and, the same for road area in traffic, video sequences was illustrated by using the MHI method [10]. Kumar et.al. developed a system for hand gesture classification [17].

Most of these works have considered a simple MHI or some variants and none of these has considered the motion overlapping for complex nature of actions. One of the constraints of the MHI is that it erases past motion by overwriting new motion onto the past one, thereby creating a template that does not correspond to the motion properly, which causes poor recognition rate for natural motions, having complex natures and overlapping. We have solved this *overwrite* problem for complex video sequences where motions might have strength in various directions. We propose a technique called DMHI which separates motion flow into four directions and produces respective MHIs. From these directed templates, we extract feature vector and later do the recognition for various datasets.

Moreover, this paper demonstrates a noble methodology for motion annotation based on the directional motion history recognition methodology. It can parse continuous motion sequences (named in this paper as *complex motion*) into basic movements or into shorter motion sequences, based on four directions' (namely, left, right, up and down motion) motion history image (We name this method as '*Directional Motion Segmentation*', DMS). This process is different and simpler than Labanotation (Laban dance notation system – a system of movement notation, using symbols on a staff, that records the parts of a dancer's body, direction in space, dynamics, and tempo for all kinds of movement: used to record and reconstruct forms of dance and movement) or Benesh movement notation [4].

3. Motion Representations

3.1 MHI Method

The MHI is a view-based method and we can compute MHI $H_\tau(x,y,t)$ from update function $\Psi(x,y,t)$:

$$H_\tau(x,y,t)=\begin{cases}\tau & \text{if } \Psi(x,y,t)=1\\ \max(0, H_\tau(\mathrm{x},\mathrm{y},\mathrm{t}-1)-\delta) & \text{otherwise}\end{cases} \tag{1}$$

Here, x, y and t show the position and time, $\Psi(x,y,t)$ signals object presence (or motion) in the current video image, τ decides the temporal duration of MHI (in terms of frames), and δ is the decay parameter. This update function is called for every new video frame analyzed in the sequence.

We can generate MEI by thresholding the MHI above zero. Actually, motion energy image is cumulative binary motion images, which can describe where a motion is in the video sequence, computed from the start frame to the final frame. Let *D(x,y,t)* be a binary image sequence indicating regions of motion. Then, the binary MEI $E_\tau(x,y,t)$ is defined as:

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \tag{2}$$

3.2. Directional Motion History Image Method

This sub-section presents DMHI method by explaining the *overwrite* problem. In Fig. 1, images at top row (frame no. 22) and bottom row (frame no. 50) show that for down motion [y+] at 22nd frame, there is no motion upwards for DMHIy- for a 'sit down and up' motion. Therefore, his body first moves from up to down, and then returns to up. It is evident that in the original MHI, the first movement is overwritten when the second movement occurs. Therefore, we have extended the MHI by considering motion directions.

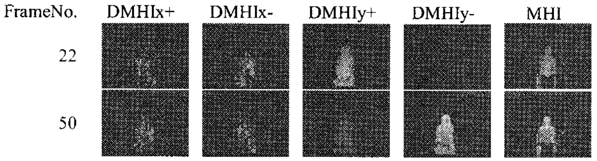

Fig. 1. Overwrite problem and multi-directional motion history images.

The proposed DMHI technique generates the MHI from directionally separated optical flow components. Based on Efros et.al.'s motion descriptors [25], the optical flow is separated into four components according to four directions: i.e., right, left, up and down (Fig. 2); and four individual motion history images are generated from them, named Directional Motion History Images (DMHIs).

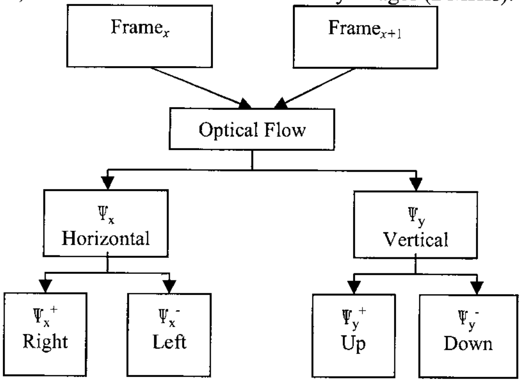

Fig. 2. For DMHI, optical flow is calculated from consecutive frames and split into four directions.

Based on the four directions, four separate MHIs are created after deriving the four optical flow components. For every case, based on a threshold th_p on pixel values, we compute the following four DMHIs:

$$H_\tau^{-x}(x,y,t) = \begin{cases} \tau & \text{if } \Psi_x^-(x,y,t) > th_p \\ \max(0, H_\tau^{-x}(\mathrm{x,y,t-1}) - \delta) & \text{otherwise} \end{cases} \tag{3}$$

$$H_\tau^{+x}(x,y,t) = \begin{cases} \tau & \text{if } \Psi_x^+(x,y,t) > th_p \\ \max(0, H_\tau^{+x}(\mathrm{x,y,t-1}) - \delta) & \text{otherwise} \end{cases} \tag{4}$$

$$H_\tau^{-y}(x,y,t) = \begin{cases} \tau & \text{if } \Psi_y^-(x,y,t) > th_p \\ \max(0, H_\tau^{-y}(\mathrm{x,y,t-1}) - \delta) & \text{otherwise} \end{cases} \tag{5}$$

$$H_\tau^{+y}(x,y,t) = \begin{cases} \tau & \text{if } \Psi_y^+(x,y,t) > th_p \\ \max(0, H_\tau^{+y}(\mathrm{x,y,t-1}) - \delta) & \text{otherwise} \end{cases} \tag{6}$$

After some post-processing, we compute the corresponding four motion energy images by thresholding the DMHIs above zero. By using the proposed DMHIs, the *overwrite* problem can be eliminated.

4. DMHI Recognition Methodology

4.1 Feature Vector Calculation

We employ Hu moments [18,19] to calculate the feature vectors for MHI, MEI and each component of the DMHIs and the DMEIs. For the MHI method, 14 feature vectors are calculated for both MHI and MEI; whereas, for the DMHI method, we also estimate eight normalized 0th order moments for the eight components (four for each DMHI and DMEI). Hence, we get a 64-dimensional feature vectors for every video sequences. We employ the nearest-neighbor (NN) classifier. The nearest-neighbor classifier labels an unknown image represented by an m-dimensional feature vector $\boldsymbol{x} = (x_1, x_2, \dots, x_m)$ with the label of the nearest neighbor of $\boldsymbol{x}$ among all the training samples [20]. The distance between $\boldsymbol{x}$ and a training sample is measured using Euclidean distance. This is a mapping from m-dimensional feature space onto a one-dimensional Euclidean space.

We employ leave-one-out cross-validation partitioning scheme. This means that out of N samples from each of the c classes per database, N - 1 of them are used to train (design) the classifier and the remaining one to test it [20]. This process is repeated N times, each time leaving a different sample out. Therefore, all of the samples are ultimately used for testing. This process is repeated and the resultant recognition rate is averaged. As experimental dataset, we opt for Japanese radio aerobics (‘*rajio-taiso*’ [21, 22] – an exercise program broadcast over the radio in Japan). For rajio-taiso, we consider six different exercise sets of seven inexperienced performers.

4.1. Recognition Results and Analysis

We investigate the performance of the MHI and the DMHI methods. Using the DMHI technique, the recognition rate was found satisfactory (the average recognition rate is 87%), whereas, employing the basic motion history image method, the average recognition rate is only 53% (Table 1). From manual check on the video action per person for which the recognition was found as ‘false’ recognition, we have found that the accuracy of that subject’s action was poor, comparing to the accurate action and this produced false recognition.

Table 1. Recognition rates for various actions between MHI and the proposed method (DMHI).

No.	Action Name	MHI	DMHI
1	Straightening back	50%	100%
2	Wave arms, bend & straighten leg	17%	67%
3	Turn arms	67%	100%
4	Bend chest	17%	67%
5	Bend body aside	84%	84%
6	Bend body to front & back	84%	100%
Average recognition rate		53%	87%

5. Directional Motion Segmentation

This section presents a *Directional Motion Segmentation* (DMS) technique by employing the above directional motion history image. The DMS is the technique for the intermediate interpretation of complex motion into four directions, namely, right, left, up and down. We calculate the directional motion history images for a complex motion having several directions in it. Then for consecutive frames, we calculate the volume of pixel values (v_τ) after summing up the DMHIs’ brightness levels as follows:

$$v_t^\ell = \sum_{x=1}^{M}\sum_{y=1}^{N} H_\tau^\ell(x,y,t) \tag{7}$$

We can decide the label (ℓ – which can be one of the four directions) of the segmented motion after some threshold values (to determine the starting point for a motion Θ_α and to determine that the motion has stopped Θ_β) as shown in the following equations. Here, Δ_t^ℓ is the difference between two volume of pixel values (v_τ) for two frames. Variables k are the frame number, where the value of k might be 1 or more. When k=1, then we are calculating consecutive two frames in the video sequences.

$$\Delta_{t+k}^\ell = v_{t+k}^\ell - v_t^\ell \tag{8}$$

$$\ell = \begin{cases} \ell & if \;\; \Delta_t^\ell > \Theta_\alpha \\ \Phi & if \;\; \Delta_t^\ell < \Theta_\beta \end{cases} \tag{9}$$

Variable ℓ defines the four different labels (namely, left, right, up and down) in our experiment. When the difference Δ_t^ℓ is more than a starting threshold value Θ_α, we can decide the label of the segmented motion. But when the Δ_t^ℓ reduces to $\Theta_\beta > \Delta_t^\ell$ – we can say that there is no motion or the motion is no longer present (Φ). We choose Θ_α and Θ_β empirically based on the performed experiment. Therefore, based on this mechanism from the directional MHIs, we can easily segment a complex motion sequence into four directions. This is very useful for an intelligent robot to decide the directions of the human movement. Thus an action can be understood based on some consecutive left-right-up-down combination.

The following graph explains the segmentation for a complex motion. Fig. 3(i) presents some consecutive frames for action-2 of rajio-taiso. In this exercise, initially both hands are stretching in both directions (Fig. 3(i): a, b) – and hence, we see in the graph for action-2 – *left* and *right* motions are increasing in parallel (see Fig. 3(ii)). Both hands are going down along with the body (legs are bending down a bit here; Fig. 3(i):b). So we see the motion for *downward* and just after this, the legs are going up slowly while the hands are waving downwards (c). Till now, we do not see any significant motions for *up* motion. After frame no.16 (shown in x-axis in Fig.3(ii)), we can see the *up* motion is increasing as legs are going upwards (c) and after this both hands are waving upward (d, e). Around Frame no. 25 in Fig. 3(ii), we notice the crossing point when the *down* motion is saturating and *up* motion is increasing (in the graph, saturation in one direction means there is no significant motion change in that direction). At the end (g, h), we notice hands are waving down a bit and then over the chest, crossing and hence we get a bit increment of pixel volume for both *up* (i) and *down* motions. From the graph, we notice that the lines are approaching towards saturation. Due to the presence of noise, we still get a bit increasing tendency and therefore, we can note that after a steep change in graph, if we notice a slow slop in the graph, we consider this as saturation. In this way, it becomes easy for a robot to take a decision. We tried with other action sets as well and have found satisfactory results. Once the program starts with some complex motion, it prints the directions of the motion one after another based on the switching and, in this way, a robot can get help to find its decision.

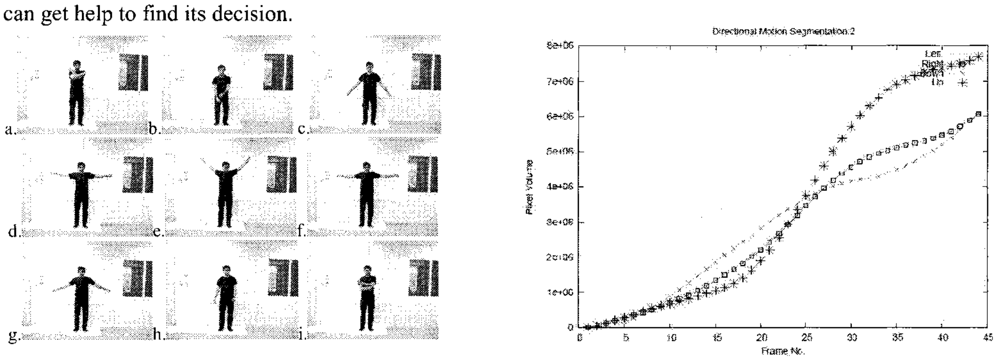

Fig. 3. Directional Motion Segmentation technique: (i) Left side:- Some frames for Rajio-taiso action-2; (ii) Right side:- Graph showing the change of directional motions and gets saturated once there is no motion.

6. Conclusions

In this correspondence, we compared the MHI with our DMHI method. The DMHI can recognize complex human motions with high recognition rate. We have found a satisfactory recognition rate for the DMHI method. We have employed Hu moments to calculate feature vectors. But these have some drawbacks [23], e.g., their dramatic increase in complexity with increasing order; their containment of redundant information about shape. We can try with other moments [24], especially with Zernike moments [20], which might yield higher recognition rates compared to Hu moment features. Moreover, due to multi-dimensional analysis, the total computational cost increases from the original method. To solve this problem, we may employ special hardware to compute the optical flow (e.g., FPGA), or try with different optical flow methods to find the best-fit optical flow method for recognition.

We also explored motion segmentation using the directional motion history images. This process can segment a complex motion sequence into four directions easily and promptly. This segmentation process is quick and can accurately determine the directions, which can lead a robot to better decision instantly. This technique along with the DMHI method can be employed for other application realms to find its suitability for different purposes in future.

Reference:

1. G.D. Borshukov, G. Bozdagi, Y. Altunbasak, A.M. Tekalp, "Motion Segmentation by Multistage Affine Classification", *IEEE Trans. on IP*, vol. 6, no. 11, pp. 1591-1594, Nov. 1997.
2. J. Gao, R.T. Collins, A.G. Hauptmann, H.D. Wactlar, "Articulated Motion Modeling for Activity Analysis", *Proc. Third Int'l Conf. on Image and Video Retrieval (CIVR'04), Workshop on Articulated and Nonrigid Motion (ANM'04)*, 9 pages, July 2004.

3. N. Gheissari, A. Bab-Hadiashar, “Motion Analysis: Model Selection and Motion Segmentation”, *Proc. 12th Int'l Conf. on Image Analysis and Processing (ICIAP'03)*, pp. 442-447, Sept. 2003.
4. Griesbeck, “Introduction to Labanotation”, http://user.uni-frankfurt.de/~griesbec/LABANE.HTML and http://en.wikipedia.org/wiki/Labanotation
5. K. Kahol, P. Tripathi, S. Panchanathan, “Documenting Motion Sequences with a Personalized Annotation System”, *IEEE Journal of Multimedia*, vol. 13, issue 1, pp. 37-45, Jan-March 2006.
6. K. Kahol, P. Tripathi, S. Panchanathan, “Automated Gesture Segmentation from Dance Sequences”, *Proc. 6th IEEE Int'l Conf. on Automated Face and Gesture Recognition*, pp. 883-888, 2004.
7. C. Yuan, G. Medioni, J. Kang, I. Cohen, “Detecting Motion Regions in the Presence of a Strong Parallax from a Moving Camera by Multiview Geometric Constraints”, *IEEE Trans. on PAMI*, vol. 29, issue 9, pp. 1627-1641, Sept. 2007.
8. J.K. Aggarwal and Q. Cai, ‘Human Motion Analysis: A Review’, *Proc. IEEE Nonrigid and Articulated Motion Workshop*, pp. 90-102, June 1997.
9. A.F. Bobick, J.W. Davis, “The Recognition of Human Movement using Temporal Templates”, *IEEE Trans. on PAMI*, vol. 23, no. 3, pp. 257-267, March 2001.
10. D. Son, T. Dinh, V. Nam, T. Hanh, H. Lam, “Detection and Localization of Road Area in Traffic Video Sequences Using Motion Information and Fuzzy-Shadowed Sets”, *Proc. 7th IEEE Int'l Symposium on Multimedia*, pp. 725-732, 12-14 Dec. 2005.
11. J. Davis, G. Bradski, “Real-time Motion Template Gradients using Intel CVLib”, *Proc. IEEE ICCV Workshop on Frame-rate Vision*, pp.1-20, Sept. 1999.
12. A.F. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, A. Wilson, “The Kidsroom: A Perceptually-Based Interactive and Immersive Story Environment", *Presence: Teleoperators and Virtual Environments*, vol. 8, no. 4, pp. 367-391, 1999.
13. W.C. Yau, D.K. Kumar, S.P. Arjunan, S. Kumar, “Visual Speech Recognition Using Image Moments and Multiresolution Wavelet”, *Proc. of Int'l Conf. on Computer Graphics, Imaging and Visualization*, pp. 194-199, 26-28 July 2006.
14. W.C. Yau, D.K. Kumar, S.P. Arjunan, “Voiceless Speech Recognition Using Dynamic Visual Speech Features”, *HCSNet Workshop on the Use of Vision in HCI*, Canberra, Australia, 2006.
15. M. Valstar, M. Pantic, I. Patras, “Motion History for Facial Action Detection in Video”, *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, vol. 1, pp. 635-640, 10-13 Oct. 2004.
16. Z. Yin, R. Collins, “Moving Object Localization in Thermal Imagery by Forward-backward MHI”, *Proc. Conf. on Computer Vision and Pattern Recognition Workshop*, pp. 133-140, 17-22 June 2006.
17. S. Kumar, D. Kumar, A. Sharma, N. McLachlan, “Classification of Hand Movements Using Motion Templates and Geometrical Based Moments”, *Proc. of ICISIP Int'l Conf. on Intelligent Sensing and Information Processing,* Chennai India, pp. 299-304, 2003,.
18. M.K. Hu, “Pattern Recognition by Moment Invariants”, *Proc. IRE.* vol. 49, p. 1218, 1961.
19. M.K. Hu, “Visual Pattern Recognition by Moment Invariants”, *IRE Trans. on Information Theory*, pp. 179-187, 1962.
20. A. Khotanzad, Y.H. Hong, “Invariant Image Recognition by Zernike Moments”, *IEEE Trans. on PAMI*, vol. 12, no. 5, pp. 489-497, May 1990.
21. Rajio Taiso, accessed at July 2007, http://chinmusicpress.com/books/kuhaku/literature/glossary/entries/rajio_taiso.html
22. Radio Exercise (Japanese page), accessed at July 2007, http://www.kampo.japanpost.jp/kenkou/radio/taisou.html
23. D. Shen, H.S.H. Ip, “Discriminative Wavelet Shape Descriptors for Recognition of 2-D Patterns”, *Pattern Recognition*, vol. 32, pp. 151-165, 1999.
24. D. Zhang, G. Lu, “Review of Shape Representation and Description Techniques”, *Pattern Recognition*, vol. 37, pp. 1-19, 2004.
25. A.A. Efros, A.C. Berg, G. Mori, J. Malik, “Recognizing Action at a Distance”, *Proc. Int'l Conf. on Computer Vision (ICCV)*, pp. 726-733, 2003.