Human Behavior Analysis with Optical Flow and Median-filtered Temporal Motion Segmentation Method

Md. Atiqur Rahman Ahad, J.K. tan, H. Kim and S. Ishikawa

Department of Control Engineering, Kyushu Institute of Technology,
Kitakyushu, Tobata, Sensui-cho 1-1, T-804-8550, Japan,
(Tel : 81-93-884-3183; Fax : 81-93-884-3183)
(Email: atiqahad@yahoo.com)

*Abstract***:** We focus on human activity analysis so that an intelligent system (e.g., a robot) can easily understand some important activities and help thereby. Hence we present an improved method for activity analysis, called Median-filtered Temporal Motion Segmentation (MfTMS) method, which can segment and understand motion temporally from the video sequence. It is based on the computation of optical flow and thereafter split it into four different channels. Later median filtering is applied and we produce four motion-history templates based on the directional motion vectors. Based on the total pixel volumes on these history templates and their related variations, various directions of the action primitives are segmented temporally. We conduct experimentations both indoor and outdoor environments and achieved sound performance. This segmentation method can assist an intelligent system or a robot to understand activities and take decisions afterwards. It is a simple but robust approach.

Keywords: Motion segmentation, behavior understanding, DMHI, optical flow, MfTMS**.**

I. INTRODUCTION

Motion understanding and behavior analysis are important research areas in computer vision arena with various important applications [1-2]. In this paper, we attempt to understand and segment human behavior based on the concept of our Directional Motion History Image (DMHI) method [3], which is based on four channels of optical flow. There are various action segmentation approaches. For temporal motion segmentation approach, Kahol et al [4] focused on developing a gesture segmentation system based on HMM states. Yuan et al [5] 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. Ref. [6] proposed a method to divide all the motion data into segmental motions based on breaking points. After segmentation, all the data are classified into clusters, called basic motion. Ihara et al. [7] proposed a gesture-description model based on synthesizing fundamental gestures. They developed a gesture semantics database based on 'fundamental gestures', which are defined on some set of meaningful actions. Inspired by the natural language processing, an approach for automatically segmenting sequences of natural activities into atomic sections and clustering them is proposed [8]. Ref. [9] segmented video based on the connections between certain low level and computationally simple features; and high level, semantically meaningful features.

Ref. [10] proposed methods for motion patterns of humanoid robots observed as a continuous flow using pattern correlations and associative memory. Ref. [11] proposed approaches that have a more similar connotation with our approach: to segment behaviors into distinct behaviors. Ref. [12] proposed a scheme based on the basic movement transition graph for extracting motion beats from given rhythmic motion data. Similarly, [13] developed a method that automatically detect the musical rhythm and segment the original motion to classify to the primitive motions.

Section II presents the method. Next in Section III, experimental results and analyses are presented. Finally, we conclude the paper in Section IV.

II. THE MfTMS METHOD

Based on the directional motion history evaluation, we can temporally segment a motion sequence into its action primitives [3,14]. Fig. 1 shows the system flow diagram of the MfTMS method.

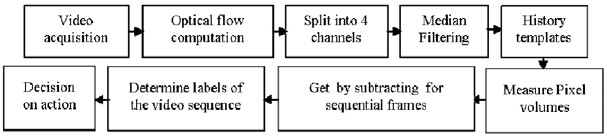


Fig. 1 System flow diagram of MfTMS method.

We calculate the directional motion *history* templates based on the DMHI [3] in order to develop the Median-filtered Temporal Motion Segmentation (MfTMS) method. The MfTMS is the technique for the intermediate interpretation of complex motion into four directions, namely, right, left, up and down. In this method, extracted features are created from pixel-wise optical flow method. We compute the optical flow vector directly from consecutive frames. Then it is split into two scalar fields corresponding to the horizontal and vertical components of the flow. Later, these components are half-wave rectified into four separate channels. We exploit the median filter to filter out the noise, so that the noise remains on the boundary of the object. Based on the four directions, four separate optical flow motion history templates (i.e., summation of optical flow on a sliding temporal window for each pixel) are created after deriving the four optical flow components. After having the motion templates for a complex action or activity, we calculate the volume of pixel values ($\nu_t \in \{\nu_t^{x+}, \nu_t^{x-}, \nu_t^{y+}, \nu_t^{y-}\}$) after summing up the motion templates' brightness levels. For consecutive frames, it is defined by,

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

Here, ℓ is a label (which can be one of the four directions based on $\ell \in \{up, down, left, right\}$) of the segmented motion after some threshold values (to determine the starting point for a motion above Θ_α and to determine the stop of motion at Θ_β) as shown:

$$\Delta_t^\ell = \nu_t^\ell - \nu_{t-k}^\ell \tag{2}$$

$$\ell = \begin{cases} \ell & if \quad \Delta_t^\ell > \Theta_\alpha \\ \Phi & if \quad \Delta_t^\ell < \Theta_\beta \end{cases} \tag{3}$$

Here, Δ_t^ℓ is the difference between two *volume* of pixel values (ν_τ) for two frames. Variable k is 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. These four different labels are identified. When the difference Δ_{t+k}^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.

So the total pixel volumes obtained by the summation of the motion histories over the pixel neighborhood, decides the direction of the action. Therefore, based on this mechanism from the motion history templates, we can easily segment a complex motion sequence into four directions. Note that in this case, based on the relationships of total pixel volumes for four directions and individual pixel volume in one direction, the system sets the ν_τ to zero, considering no change in that direction. In this way, the system can perform consistently. Figures below illustrate this concept as we find that when an action has finished in one direction, the variation of pixel volumes sets back to zero, and in some cases, increases instantly from zero to some values to denote the action.

III. EXPERIMENTAL RESULTS

We demonstrated two experiments. Fig. 2 shows 'sit-down' and 'stand-up' sequences in indoor. One person sits down, then stands up and repeats the actions in indoor with various speeds. If we employ consecutive frames for analysis, *k*=1 as in Fig. 3 and Fig. 5(a). The ground truth is presented with the segmentation result.

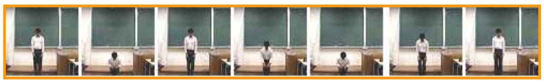


Fig. 2. Sequences for sit-down and stand-up actions

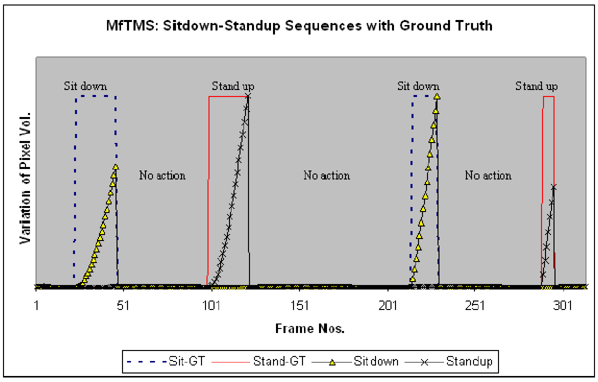


Fig. 3. Results of segmentation for actions of Fig. 2

The output is shown with the ground-truth data (GT as unit step-pulse-shaped representations) for better evaluation, where the abscissa represents 'Frames numbers' and the ordinate stands for 'Variation of pixel volumes'. From Fig. 3, we note that the action primitive extraction and identification are done accurately. After first sitting down and stand-up, we find some small spikes which can be considered as noise. We manually checked the original action sequences frame by frame and the corresponding experimented output. We also consider an experiment in outdoor and cluttered environment having textured road-marks and static but complex background. In this experiment, one person walks on the zebra-crossing from one side of the road to other side. He walks to left and back to right.

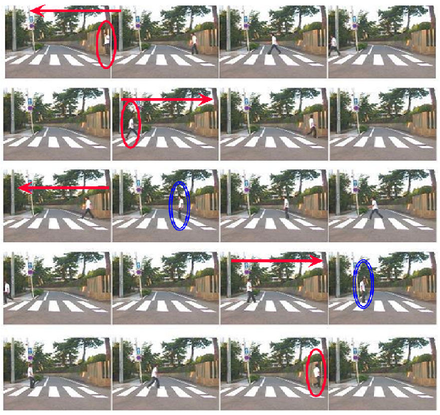
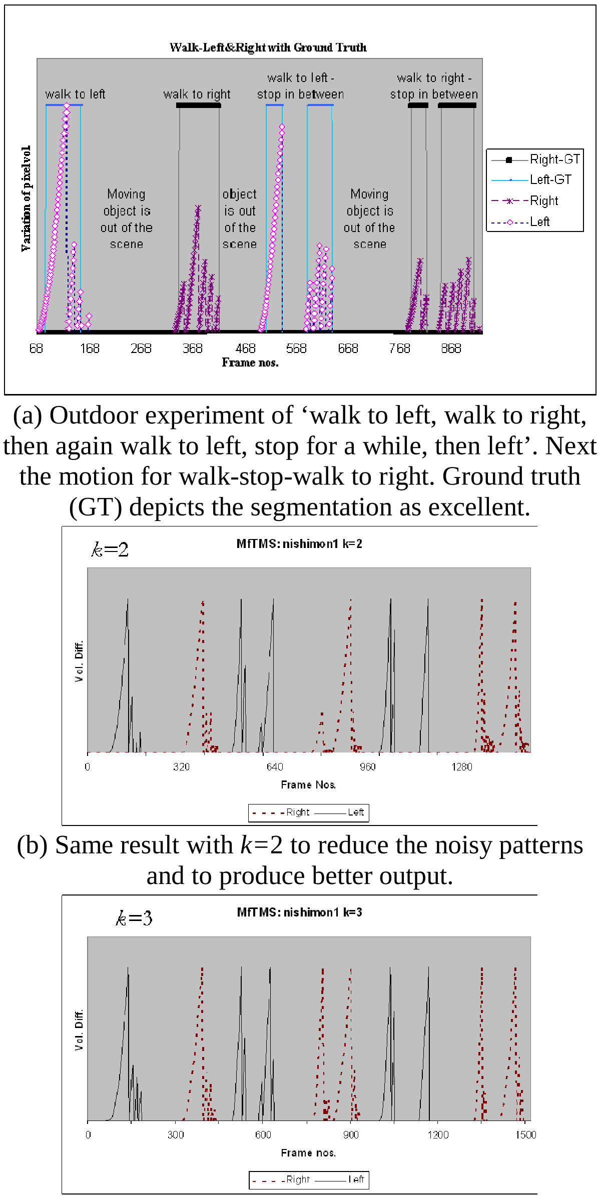


Fig. 4. Sequences of outdoor experiment



(a) Outdoor experiment of 'walk to left, walk to right, then again walk to left, stop for a while, then left'. Next the motion for walk-stop-walk to right. Ground truth (GT) depicts the segmentation as excellent.

(b) Same result with *k*=2 to reduce the noisy patterns and to produce better output.

(c) Same result with *k*=3.

Fig. 5. Experimental results for outdoor experiment

Then again come to the scene as walking to left with sudden stop for a while and does the same in right direction. So we tried to identify the action primitives using the MfTMS method. Fig. 4 shows some sequences where a person is in motion – to left, right or out of the scene. Fig. 5 demonstrates the result of segmentation and identification. The output shows satisfactory resemblance for this long video sequence. Around frame no. 568, we notice a sudden stop ('no motion') for a while and then again the start of motion. Similarly, after 820 frames, we can observe this 'no motion' part (i.e., sudden stop while walking towards right). In this graph, we notice more than one steep curves under one action (e.g., for 'walk to left' action, we see three major steeps). As explained above, the system reduces the volume of pixel values (V_τ) to zero based on the relationship of total pixel volumes of four directions. This is not tricky as in between there is no other motion directions and these curves are just adjacent to another to cover one block of action based on that direction. However, another finding depicts (Fig. 5(b)) that introduction of different values of *k* can reduce the noisy patterns and provide better output. From Fig. 5(b)~(c), it is evident that by the introduction of *k* as 2 and 3, the noisy spikes are less.

However, as the frame nos. in between are a few and no other actions, we consider the entire block as one single action primitive. For example, from frames around frame no. 70 till 'walk to left' part (Fig. 5(a)), we note two other almost consecutive step-pulses to denote 'walk to left'. Similarly, we can get the block of one action for walk to right and so on. These clearly separate the actions in different directions and this information can easily aid an intelligence system or robot to identify a sequential or continuous activities. We manually evaluated the video data to analyze the reasons for this nature of output in this case. We manually checked the possible *start* frame and *stop* frame for each action primitives from original video for both activities, and the experimental results. Except in one cases for both, we find that the start and stop for each action primitive is the same or differed by less than five frames on average. In the second 'Sitting' action, the ending lingered for more frames in the experiment, and for the first 'Walk to Left' action, the experimental results show early stop of walking. We think that noise in optical flow seems responsible for this disarray after the evaluation of frames and output. Moreover, due to the outdoor environment, change in illumination, variable speed of walking, cluttered background, textured road-marks, the computed optical flow has varied and hence the final output has some variations like setting to zero values back and forth.

We have tried with few other activities (e.g., walking and fall-down on the floor, walking towards the camera or off the camera along its optical axis) and have found satisfactory and promising results. In this manner, a robot can annotate a complex motion and take the appropriate decision as per the situation. For example, if a robot finds one person walking and suddenly stops and then sitting down on the floor or ground in a park or rehabilitation center and thereafter no motion, then it can sense that the person is sick or need some help immediately. Based on this, the robot can help the person and send a signal or message to the appropriate operator to initiate assistance. It is clear from these Figures that changes in activity can be detected, and the system is able to determine what kind of motion is performed. It produces the action information based on the action itself, hence timing issue is not a problem. However, four optical flow vectors is a concern, as sometimes, it incorporates some noise that presumes some motion components though there is originally no motion in that specific direction.

IV. CONCLUSION

We presented a simple but robust Median-filtered Temporal Motion Segmentation (MfTMS) method, based on the directional motion *history* templates. This process can separate a complex motion sequence into four directions promptly. In this method, we do not need to calculate feature vectors or we do not require any classification method or recognition scheme. Therefore, this segmentation process is simple and thus it is fast, too. As shown in the experimental results, the MfTMS can accurately determine the directions, which can lead a robot to better decision. This method can be employed for other application areas too.

One of the key issues concerning to this method is related to the proper correlation between previous state and the current movement. Employing Hidden Markov Model may improve the system with increasing complexity. Moreover, the presence of noise due to optical flow errors (e.g., low-textured people) limits the performance, which is inherently related to the computation of optical flow. Proper choice of optical flow or a new robust mechanism to deal in calculating the optical flow vectors are very crucial to improve the system. In future, one may consider magnitude and orientation of optical flow for a patch (e.g., patch of 5x5 pixels) and then run RANSAC to ignore outliers and consider the dominant parts of optical flow. We hope that this way can minimize noises and hence can improve the four-directed motion history templates.

REFERENCES

[1] Ahad MAR, Tan J, Kim S, Ishikawa S (2008), Human activity recognition: various paradigms, Int. Conf. Control, Automation and Systems 1896-1901

[2] Ahad MAR, Tan J, Kim S, Ishikawa S (2009), Human activity analysis: concentrating on motion history image and its variants, Int. Conf. ICASS-SICE

[3] Ahad MAR, Tan J, Kim S, Ishikawa S (2009), Temporal motion recognition and segmentation approach, Int. J. of Imaging Systems & Technology 19:91-99

[4] Kahol K, Tripathi P, Panchanathan S (2006), Documenting motion sequences with a personalized annotation system, J. of Multimedia 13:37-45

[5] Yuan C, Medioni G, Kang J, Cohen I (2007), Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints, IEEE Trans. on PAMI 29:1627-1641

[6] Osaki R, Shimada M, Uehara K (1999), Extraction of primitive motion for human motion recognition, Int. Conf. on Discovery Science 1721:12 pages

[7] Ihara M, Watanabe N, Nishimura K (1999), A gesture description model based on synthesizing fundamental gestures, IEEE SouthEastCon 47-52

[8] Wang T, Shum H, Xu Y, Zheng N (2001), Unsupervised analysis of human gestures, IEEE Pacific Rim Conf. on Multimedia 174-181

[9] Peker K, Alatan A, Akansu A, Low-level motion activity features for semantic characterization of video, IEEE Conf. on Multimedia and Expo 801-804

[10] Kadone H, Nakamura Y (2006), Segmentation, memorization, recognition and abstraction of humanoid motions based on correlations and associative memory, IEEE-RAS Int. Conf. on Humanoid Robots 1-6

[11] Barbic J, Safonova A, Pan J, Faloutsos C, Hodgins J, Pollard N (2004), Segmenting motion capture data into distinct behaviors, Graphics Interface 185-194

[12] Kim T, Park S, Shin S (2003), Rhythmic-motion synthesis based on motion-beat analysis, ACM Trans. Graphics 22:392-401

[13] Shiratori T, Nakazawa A, Ikeuchi K (2004), Detecting dance motion structure through music analysis, IEEE Conf. Automatic Face and Gesture Recognition 857-862

[14] Ahad MAR, Uemura H, Tan J, Kim S, Ishikawa S (2008), A simple real-time approach for action separation into action primitives, Int. Workshop on Tracking Humans for the Evaluation of Their Motion in Image Sequences, 69-78