Journal of Korea Multimedia Society Vol. 21, No. 12, December 2018(pp. 1473-1480)
https://doi.org/10.9717/kmms.2018.21.12.1473

Human Action Recognition Based on An Improved Combined Feature Representation

Ning Zhang†, Eung-Joo Lee††

ABSTRACT

The extraction and recognition of human motion characteristics need to combine biometrics to determine and judge human behavior in the movement and distinguish individual identities. The so-called biometric technology, the specific operation is the use of the body's inherent biological characteristics of individual identity authentication, the most noteworthy feature is the invariance and uniqueness. In the past, the behavior recognition technology based on the single characteristic was too restrictive, in this paper, we proposed a mixed feature which combined global silhouette feature and local optical flow feature, and this combined representation was used for human action recognition. And we will use the KTH database to train and test the recognition system. Experiments have been very desirable results.

Key words: Computer Vision; Action Recognition; Global Silhouette; Local Optical

1. INTRODUCTION

The study of behavioral identification analysis can be traced back to an experiment in Johansson in 1975 [1]. The authors propose a 12-point human model, which is an important guiding role in the behavioral description algorithm based on human structure. Since then, the research history of behavior recognition can be divided into three stages, the first stage is the preliminary research stage of behavior analysis in the 1970s; the second stage is the gradual development stage of love analysis in the 1990s; The three stages are the rapid development of behavior analysis in recent years. From the literature [2-7] these six more well-known behavior recognition review papers can be seen, the number of research behavior recognition is increasing, the number of papers is also increasing, and produced a number of important algorithms and ideas.

There are many kinds of methodologies for visual analysis and identification of human motion. Hu [8] sue the improved Dense Trajectories (iDT) algorithm to optimize and extract the optical flow features in the movement of human action, then they combined with Support Vector Machin methods to identify human behavior. Forsyth [9] and others focus on the action from the video sequence of human posture and movement information recovery, which belongs to a regression problem, and human behavior identification is a classification problem, these two problems have many similarities, such as its The features are extracted and described in many ways. Turaga [5] and others to human behavior identification is divided into three parts, namely, mobile identification, motion recognition and behavior recognition, these three categories were in the lower visual, middle vision, high-level visual corresponding. Gavrila [10] used 2D and 3D methods to study the behavior of the

※ Corresponding Author : Eung-Joo Lee, Address: 428, Sinseon-ro, Nam-gu, Busan, Korea, TEL : +82-51-629-1143, E-mail : kkkkkk@knu.ac.kr
Receipt date : Oct. 8, 2018, Revision date : Oct. 23, 2018
Approval date : Nov. 1, 2018

† Dept. of Information Communication Engineering, Tongmyong University
(E-mail : jangneyong0829@hotmail.com)
†† Dept. of Information Communication Engineering, Tongmyong University

human body.

The global feature is to describe the entire human body of interest, usually through the background subtraction or tracking method to get, usually using the human body edge, silhouette contour, optical flow and other information. And these features on the noise, partial occlusion, perspective changes are more sensitive.

Davis [11] and others first use the contour to describe the movement information of the human body, which uses MEI and MHI two templates to save the corresponding action information, and then use the Markov distance classifier to identify. MEI is the movement energy map, used to indicate where the movement occurred; MHI is the movement history map, in addition to the movement of the space position, but also reflects the movement of the time sequence.

In order to advance silhouette information, Wang [12] and others use r-transform to obtain the silhouette of the human body. Hsuan-Shen [13] extracts the contours of the human body, which uses the star skeleton to describe the angle between the baselines, which are extended from the human body's hands, feet, and the like to the contours of the human body. And Wang [14] colleagues used silhouette information and contour information to describe the action, that is, based on the contour of the average motion shape (MMS) and based on the movement of the average energy (AME) two templates to describe. When the contours and silhouettes are preserved, the newly extracted features are compared with them. Daniel [15] uses the Euclidean distance to measure its similarity, and then he uses the chamfer distance to measure [16], it eliminates the background subtraction of this preprocessing step.

On the basis of taking into account the advantages and disadvantages of the different features and the scope of application, this paper presents a hybrid feature that combines the static features described by the global shape and the dynamic characteristics of the local optical flow description. Firstly, the background subtraction method is used to determine the approximate area of the motion, and the silhouette of the human body is obtained, and the whole information of the human body's appearance is expressed by the silhouette contour vector. Then, the optical flow is extracted in the moving area and the local optical flow information So as to improve the anti-noise ability of the optical flow. Finally, the global silhouette feature and the local optical flow feature are combined as the mixed feature.

2. GLOBAL SILHOUETTE FEATURE REPRESENTATION AND EXTRACTION

2.1 Pre-treatment

Image preprocessing can reduce the image area of the study and process the data, reducing the computational complexity. The background area and the human body silhouette of the movement are usually determined by the background subtraction method, as showed in Fig. 1(b). Assuming that all actions are performed before the static

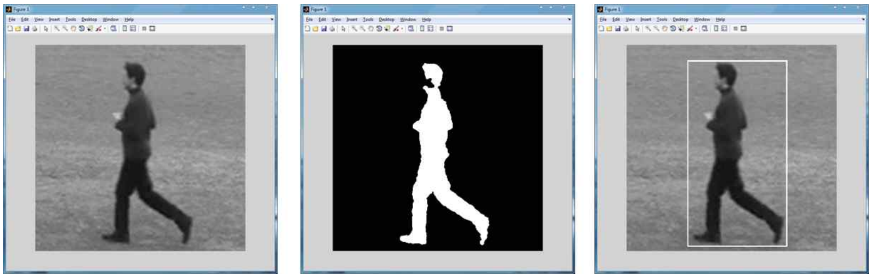


Fig. 1. Background subtraction method.

background, the interest area can be determined based on the silhouette information, as showed in Fig. 1(c), for the area of interest in the white rectangle. After the region of interest is determined, only the information in the area of interest can be processed.

2.2 Contour Vector Extraction

The silhouette of a human body in a single frame image can be used to describe the overall shape change information of human motion. Selection of silhouette features have the following advantages: a) silhouette features can be simple and intuitive description of the shape of the movement of human information; b) silhouette features easy to extract; c) the binary silhouette is not sensitive to the texture and color of the foreground image. This step is intended to convert the motion information of the original video into a sequence of morphological features associated with it, which reflect changes in the motion process.

There is a T frame image I in a motion video V, $V=[I_1,I_2,...,I_\gamma]$, the corresponding motion silhouette sequence is $S=[s_1,s_2,...,s_\gamma]$, s has been obtained during image preprocessing. For the sake of simplicity, the contours vector is used to describe the overall shape information of human silhouettes. The specific process is as follows:

a) Using the Canny operator to obtain the edge contour of each frame is showed in Fig. 2(a), and the coordinate representation of the edge contour is obtained, as showed in Fig. 2(b). So that the body contour can be used for n_t points that, $[(x_1,y_1),(x_2,y_2),...,(x_{n_t},y_{n_t})], t=1,2,3,...,T.$

b) The centroid of the contours of the human body is obtained by the formula (1).

$$(x_c,y_c)=\left(\frac{1}{n_t}\sum_{i=1}^{n_t}x_i,\frac{1}{n_t}\sum_{i=1}^{n_t}y_i\right) \tag{1}$$

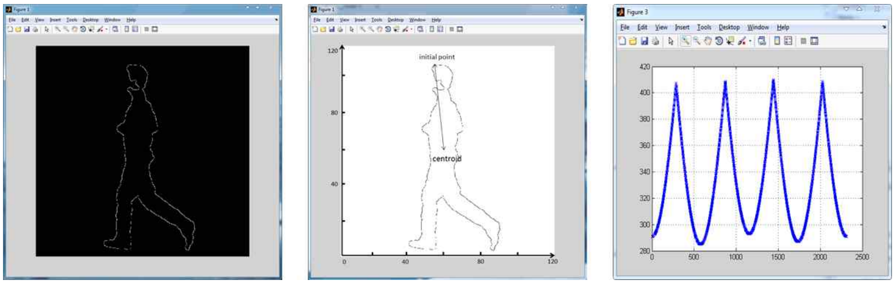
Where (x_c,y_c) is the center of mass, (x_i,y_i) is the edge of the contour, and n_t is the number of edge points in the t-th image.

c) The distance from the center of mass at the edge can be obtained by equation (2).

$$d_i=\sqrt[2]{(x_i-x_c)^2+(y_i-y_c)^2},\ i=1,2,...,n_t \tag{2}$$

d_i is the distance from the centroid to the edge of the i-th edge. Calculated from the top leftmost point of the contour chart, clockwise in order to calculate, so that from the single-frame image I_t two-dimensional contour map to obtain the corresponding one-dimensional contour vector D_i.

d) In order to eliminate the influence of spatial scale and distance length, the 2-norm is used to normalize the contour vector. Since the edge point of each frame image is constant, the normalized contour vector is re-sampled at equal intervals to obtain a fixed number of points N. In this paper, we experiment with different values of N, we can get that N=200 is relatively small, and can express the motion information completely, and the single feature and mixed feature can achieve the highest recognition rate. When the sampling point N=200,



(a) (b) (c)

Fig. 2. Contour vector results.

the contour vector results shown in Fig. 2(c) below.

3. LOCAL OPTICAL FLOW FEATURE REPRESENTATION AND EXTRACTION

3.1 The extraction and presentation of light flow

Inaccurate extraction of the silhouette image may cause the contour vector feature information to not express the action characteristic accurately. At this time, the optical flow feature can effectively and accurately represent the action information in the video sequence. In the motion area to extract the optical flow, and the use of sub-regional local optical flow information to represent the local characteristics of human motion, in order to improve the anti-noise ability of optical flow. The extraction and presentation of light flow are as follows:

a) Determine the location of the region of interest corresponding to the current frame image, and cut out the gray scale image area corresponding to the current frame and the previous frame image, as shown in Fig. 3(a) and Fig. 3(b). And then use Lucas-Kanade method in the current frame and the previous frame of interest area for optical flow detection, the resulting optical flow field as shown in Fig. 3(c), and the optical flow is divided into two vertical and horizontal components, As shown in Fig. 3(d) and Fig. 3(e).

b) In order to reduce the dimension of the optical flow information, find the data representation with the ability to identify the sub-regional radial histogram method to calculate the optical flow characteristics. First, according to the long side of the premise of scaling, will get the region of interest optical image is normalized to 120×120 dimensional uniform size optical flow diagram, as shown in Fig. 3(c) and Fig. 3(d) below. The normalized optical flow map is divided into 2×2 sub-borders S_1, S_2, S_3, S_4, and the sub-frame is 60×60, the center points are C_1, C_2, C_3, C_4, as showed in Fig. 4(a) below. Then, the sub-frame is divided into 18 sub-regions centered on the center point of the sub-frame. $S_{i,1}, S_{i,2}, \ldots, S_{i,18}$, $(i=1,2,3,4)$, each center angle is 20°, thus forming 72 sub-regions, as showed in Fig. 4(b).

c) In the sub-region $S_{x,v}$ there are k longitudinal optical flows (or lateral optical flows), and all the longitudinal optical flows (or lateral optical flows) are summed to obtain the sum of the longitudinal

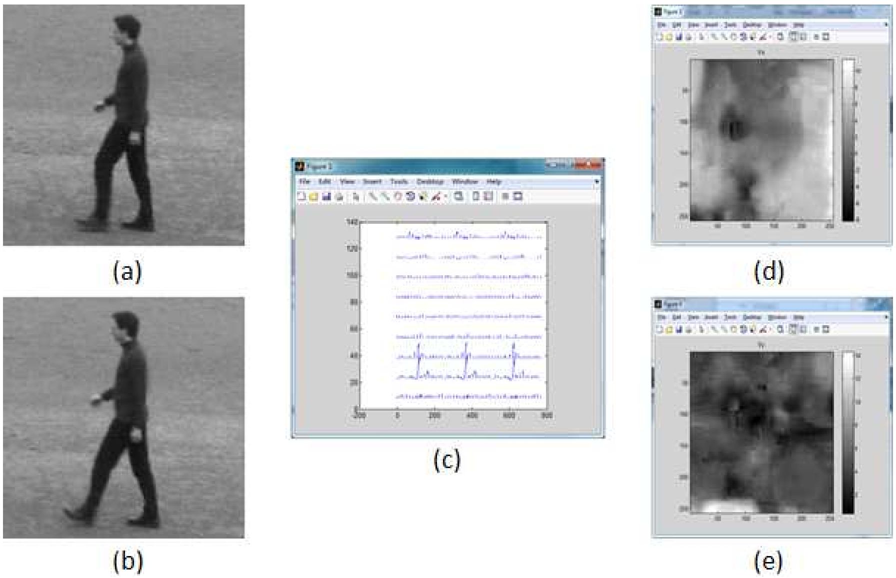


Fig. 3. Light flow diagram.

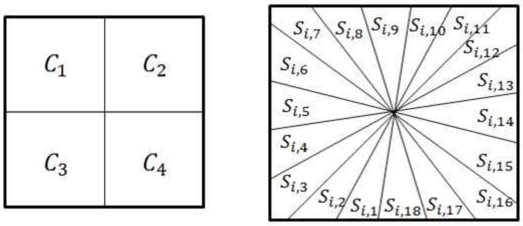


Fig. 4. Optical flow feature extraction.

optical flows O_I of the sub-region $S_{i,j}$ (the sum of lateral light flows O_H). Equation (3) and (4) calculates the sum of the sum of the longitudinal optical flows and the lateral optical flow, respectively.

$$O_{I_{i,j}} = \sum_{m=1}^{k} O_{Lm},\ (x_{O_{Lm}}, y_{O_{Lm}}) \in s_{i,j} \tag{3}$$

$$O_{H_{i,j}} = \sum_{m=1}^{k} O_{Hm},\ (x_{O_{Hm}}, y_{O_{Hm}}) \in s_{i,j} \tag{4}$$

d) the entire frame image the optical flow information can be expressed by the sum of the longitudinal optical flows of the 72 sub-regions , the sum of the horizontal optical flows , as showed in equations (5) to (7). Optical flow characteristics of the extraction, the parameters set reference [16].

$$O_L = [O_{L_{1,1}}, \ldots, O_{L_{1,18}}, \ldots, O_{L_{4,1}}, \ldots, O_{L_{4,18}}] \tag{5}$$

$$O_H = [O_{H_{1,1}}, \ldots, O_{H_{1,18}}, \ldots, O_{H_{4,1}}, \ldots, O_{H_{4,18}}] \tag{6}$$

$$O_t = [O_L, O_H] \tag{7}$$

e) The normalized histogram representation of the local optical flow vector of the current frame image I_t is obtained using the 2-norm for O_t normalization.

In order to improve the accuracy of motion recognition, the contour vector and local optical flow vector are combined to form a mixed feature vector, as shown in equation (8).

$$F_t = [O_t, D_t] \tag{8}$$

Where: F_t, O_t, D_t are the mixed feature vectors, local optical flow vectors, and contour vectors of single frame image I respectively.

3.1 The specific algorithm

This article mainly tests the ability to identify features, so here the most simple nearest neighbor classifier. The specific algorithm is as follows:

a) Find the nearest neighbor of each frame of the test sequence.

$$s_q = \min\|M_Q^t - M_T^n\|, n = 1, 2, \ldots, N \tag{9}$$

b) Assign the label of the action of the nearest neighbor training frame to the current text frame so that each test frame of the test sequence will be labeled with an action.

c) The action sequence of each frame of the test sequence is counted, and the test sequence category corresponds to the action corresponding to the label with the largest number of votes.

4. EXPERIMENT ANALYSIS

In order to verify the effectiveness of this algorithm, a large number of comparative experiments were made on the published KTH database.

This experiment is run in MATLAB 2010b implementation. KTH database has six kinds of actions, respectively, boxing, hand-clapping, hand-waving, jogging, running, walking, each action by 25 different people in four scenes to complete a total of 599 video; Background is relatively static, in addition to the lens closer / pull away, the camera's movement is relatively slight, as shown in Fig. 5.

In this paper, we extract the contours of human motion contours, optical flow characteristics and mixed features to characterize the action. In order to obtain unbiased estimation accuracy, leave one out to verify the experimental results, that is, each experiment to select a database of all the action for the test sample set, and the remaining as a training sample set. And then cycle, each person's actions will be tested as a test sample, and statistical identification results. Using the nearest neighbor method for classification and identification. Using the nearest neighbor method for

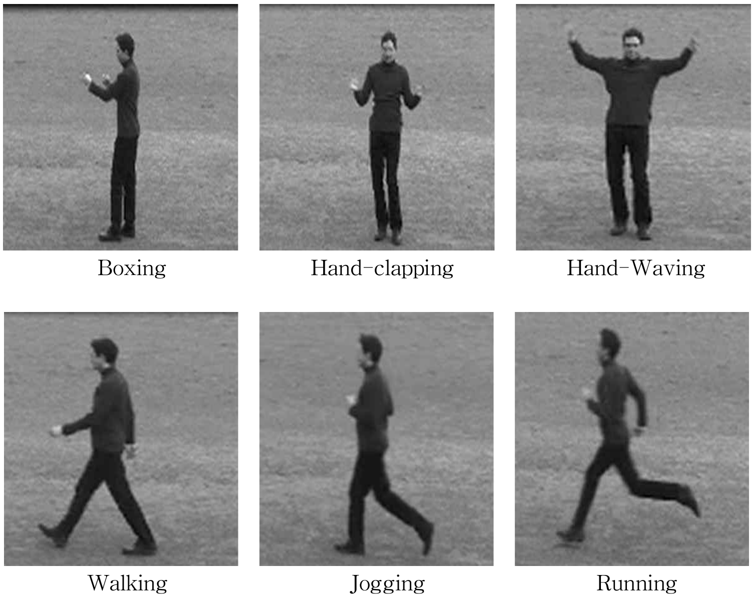


Fig. 5. KTH database six action diagram.

classification and identification. Optical flow characteristics, contour distance characteristics and the combination of the two characteristics of the mixed feature recognition as showed in Table 1.

This method and the recent related methods based on KTH database identification performance comparison shown in Table 2. It can be seen from Table 2 that although the selected features are similar. The characteristics of this paper and its algorithm are better than other algorithms. In addition, the proposed features are easy to extract and characterize, have high reliability, and avoid the complex operation of methods based on feature extraction of the human body model.

Table 1. Select the recognition rate corresponding to the different characteristics

Database	Contour[%]	Optical Flow[%]	Mixed[%]
KTH	83.33	93.33	95.00

5. CONCLUSION

The recognition of human action recognition is one of the research subjects It has important research significance and broad application prospect. Johansson first began studying human action recognition in psychology from the end of last century, and then, the research method based on computer vision becomes the hotspot of research, and it's going up every year. Especially in the late 20th century with the emergence of smart environment and computer, the need for human action recognition is more urgent, more rigorous and precise

Table 2. The different features combine the corresponding recognition rates

Method	Feature	Rate[%]
Ahmad[18]	Optical Flow+Shape Flow	88.29%
Sawant[19]	Silhouette+Local Optical Flow	90.80%
Tran[17]	Local Silhouette+Local Optical Flow	91.70%
Combined Feature	Global Silhouette+Local Optical Flow	95.00%

identification of human behavior in various environments becomes an inevitable trend. On the basis of in-depth analysis of domestic and foreign research status, in this paper, Character extraction of human action recognition from global silhouette and local feature. In short, the results of this paper can be realized based on the simple action recognition of video, but there is still some dissatisfactions. For example, the detection data is too monolithic, only one database is used; and there are no experiments on a variety of complex conditions. So we will continue our study, perfect our algorithm.

REFERENCE

[1] G. Johansson, "Visual Motion Perception," *Scientific American*, Vol. 232, No. 6, pp. 76-89, 1975.

[2] J.K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *Computer Vision and Image Understanding*, Vol. 73, No. 3, pp. 428-440, 1999.

[3] T.B. Moeslund and E. Granum, "A Survey of Computer Vision-based Human Motion Capture," *Computer Vision and Image Understanding*, Vol. 81, Issue 3, pp. 231-268, 2001.

[4] T.B. Moeslund, A. Hilton, and V. Kruger, "A Survey of Advances in Vision-based Human Motion Capture and Analysis," *Computer Vision and Image Understanding*, Vol. 104, No. 2-3, pp. 90-126, 2006.

[5] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 11, pp. 1473-1488, 2008.

[6] R. Poppe, "A Survey on Vision-Based Human Action Recognition," *Image and Vision Computing*, Vol. 28, No. 6, pp. 976-990, 2010.

[7] J.K. Aggarwal and M.S. Ryoo, "Human Activity Analysis: A Review," *Association for Computing Machinery Computing Surveys*, Vol. 43, No. 3, 2011.

[8] Z.Y. Hu, S.K. Lee, and E.J. Lee, "Improved DT Algorithm Based Human Action Features Detection," *Journal of Korea Multimedia Society*, Vol. 21, No. 4, pp. 478-484, 2018.

[9] D.A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational Studies of Human Motion: Tracking and Motion Synthesis," *Foundations and Trends in Computer Graphics and Vision*, Vol. 1, No. 2-3, pp. 77-254, 2006.

[10] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, Vol. 73, No. 1, pp. 82-98, 1999.

[11] 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.

[12] Y. Wang, K.Q. Huang, and T.N. Tan, "Human Activity Recognition Based on R Transform," *Proceeding of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[13] H.S. Chen, H.T. Chen, Y.W. Chen, and S.Y. Lee, "Human Action Recognition Using Star Skeleton," *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks*, pp. 171-178, 2006.

[14] L. Wang and D. Suter, "Informative Shape Representations for Human Action Recognition," *Proceeding of 18th International Conference on Pattern Recognition*, pp. 1266-1269, 2006.

[15] D. Weinland, E. Boyer, and R. Ronfard, "Action Recognition from Arbitrary Views Using 3D Exemplars," *Proceeding of 2007 IEEE 11th International Conference on Computer Vision*, pp. 1-7, 2007.

[16] D. Weinland and E. Boyer, "Action Recogni-

tion Using Exemplar-Based Embedding," *Proceeding of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-7, 2008.

[17] D. Tran and A. Sorokin, "Human Activity Recognition with Metric Learning," *Proceeding of European Conference on Computer Vision 2008*, pp. 548-561, 2008.

[18] M. Ahmad and S.W. Lee, "Human Action Recognition Using Shape and CLG-motion Flow from Multi-view Image Sequences," *Pattern Recognition*, Vol. 41, No. 7, pp. 2237-2252, 2008.

[19] N. Sawant and K. Biswas, "Human Action Recognition Based on Spatio-temporal Features," *Pattern Recognition and Machine Intelligence*, pp. 357-362, 2009.

Ning Zhang

received his B. S. at Dalian Polytechnic University in China (2008-2012), and received his M. S. at Busan Tongmyong University in Korea (2014-2016). Currently, he is studying in Department of Information and Communications Engineering Tongmyong University, Korea for Ph. D. His main research areas are image processing and pattern recognition.

Eung-Joo Lee

received his B. S., M. S. and Ph. D. in Electronic Engineering form Kyungpook National University, Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997 he has been with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a professor. From 2000 to July 2002, he was a president of Digital Net Bank Inc. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, China. His main research interests include biometrics, image processing, and computer vision.