2nd International Symposium on Mechanical Engineering and Material Science (ISMEMS 2017)

Human Action Recognition Based on Global Silhouette and Local Optical Flow

Ning Zhang
Dept. of Information Communication Engineering
Tongmyong University
Busan, Korea
jangneyong0829@hotmail.com

Zeyuan Hu
Dept. of Information Communication Engineering
Tongmyong University
Busan, Korea
dlhzy410@126.com

Suk-Hwan Lee
Dept. of Information Protection Engineering
Tongmyong University
Busan, Korea
skylee@tu.ac.kr

Eung-Joo Lee
Dept. of Information Communication Engineering
Tongmyong University
Busan, Korea
ejlee@tu.ac.kr

Abstract—**Currently, Human Activity Recognition is a research hotspot in the field of machine vision, it involves knowledge of image processing, pattern recognition, artificial intelligence and many other disciplines. Video-based Human Activity Recognition including human area detection, movement and gesture segmentation, objective analysis and behavior understands for activity recognition and so on. 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. In the end, test the model with other samples from the database.**

Keywords— Action recognition; Computer vision; Global silhouette; Local optical

I. 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 a key 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 numerous kinds of methodologies for visual analysis and identification of human motion. Forsyth [8] 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 [9] used 2D and 3D methods to study the behavior of the 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 of the noise, partial occlusion, perspective changes are more sensitive.

Davis [10] 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 [11] and others use r-transform to obtain the silhouette of the human body. Hsuan-Shen [12] extracts the contours of the human body, which uses the star skeleton to describe the angle between the baselines, which are prolonged from the human body's hands, feet, and the like to the contours of the human body. And Wang [13] colleagues used silhouette information and contour information to describe the action, that is, built 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 compare with them. Daniel [14] uses the Euclidean distance to measure its similarity, and then he uses the chamfer distance to measure [15], it eliminates the background subtraction of this preprocessing step.

On the basis of taking into consideration 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 identify 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 hybrid feature.

II. Global Silhouette Feature Representation and Extraction

A. Pretreatment

Image preprocessing can reduce the image area of the study and process the data, reducing the computational complexity.

When extracting the human area of video sequence, we use the Background Subtraction Method. Through the average background model method to structure the background image, after doing margin calculation between the background image and the tested image of video sequence, we use the value of the margin calculation to make absolute value calculation, in order to get the object image.

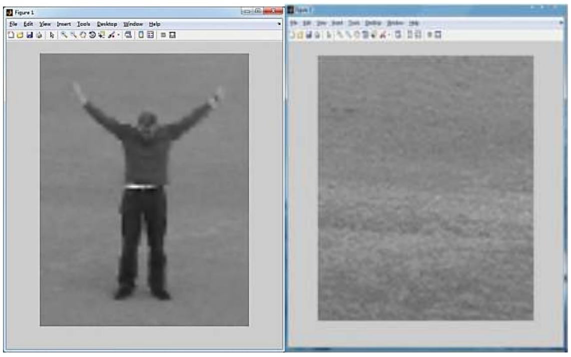
When detecting the human area of present frame, we need to use the pixel of this frame $\mathrm{I}(\mathrm{x},\mathrm{y})$ (in figure 1(a)) to minus the pixel average value $\mathrm{u}(\mathrm{x},\mathrm{y})$ (in figure 1(b)) of the same position in background, and obtain the difference value $\mathrm{d}(\mathrm{x},\mathrm{y})$, using $\mathrm{d}(\mathrm{x},\mathrm{y})$ to compare with a threshold TH, then we will obtain the value of output image:

$$d(\mathrm{x},\mathrm{y}) = \mathrm{I}(\mathrm{x},\mathrm{y}) - u(x,y) \tag{1}$$

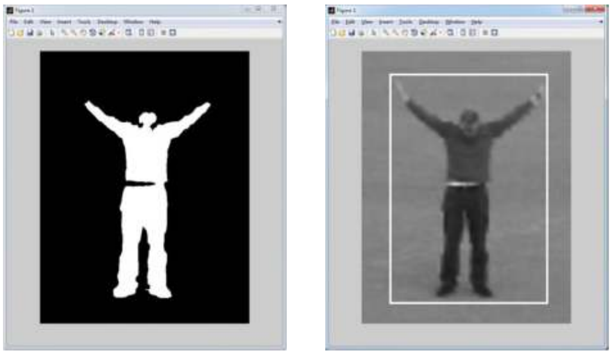
$$output(x,y) = \begin{cases} 1, & |d(x,y)| > TH \\ 0, & otherwise \end{cases} \tag{2}$$

When the action and gesture are segmenting, only need the information such as outline and edge of the activity object, and it doesn't matter whether have color, therefore, after extracting the human area, to make binary processing for difference value image, in order to reduce the complexity of calculate, and improve the real-time of system, figure 1(c) is the binary image, and the threshold is 0.1.

Assuming that all actions are performed before the static background, the interest area can be determined based on the silhouette information, as showed in figure 1(d), 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.



(a) (b)



(c) (d)

Fig. 1. Background subtraction method

B. Global Silhouette Feature

The silhouette of a human body in a single frame image can be used to describe the overall shape change information of human motion. The 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, $\mathrm{V} = [I_1, I_2, \dots, I_Y]$, the corresponding motion silhouette sequence is $\mathrm{S} = [s_1, s_2, \dots, s_Y]$, 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 figure 2(a), and the coordinate representation of the edge contour is obtained, as showed in figure 2(b). So that the body contour can be used for n_t points that, $[(x_1,y_1),(x_2,y_2),\dots,(x_{n_t},y_{n_t})]$, $t = 1,2,3,\dots,T$.

b) The centroid of the contours of the human body is obtained by the formula (3).

$$(x_c, y_c) = \left(\frac{1}{n_i}\sum_{i=1}^{n_t} x_i, \frac{1}{n_t}\sum_{i=1}^{n_t} y_i\right) \tag{3}$$

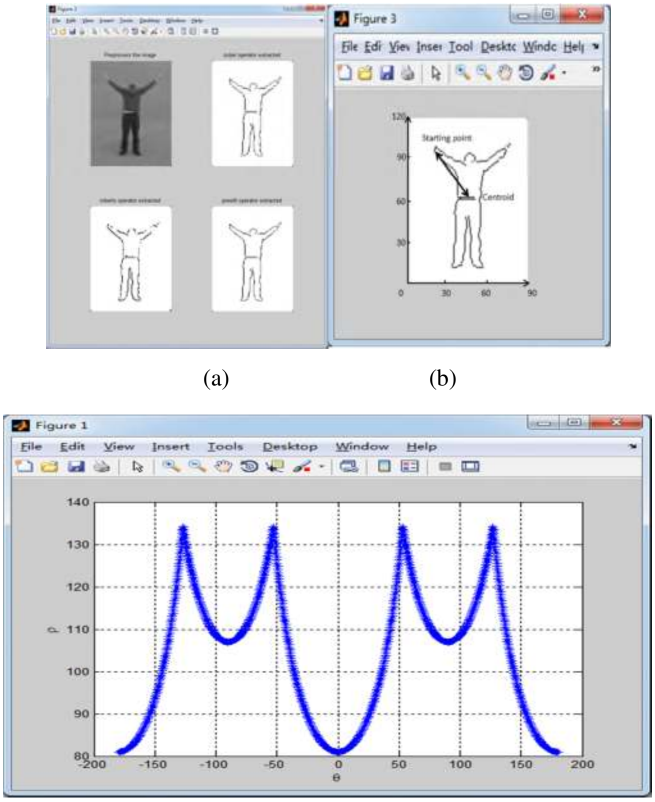
Where (x_c, y_c) is the center of mass, (x_i, y_i) is the edge of the contour, and n_t is the number of edge points in the t-th image.

c) The distance from the center of mass at the edge can be obtained by equation (4).

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2},\ i = 1,2,\dots,n_t \tag{4}$$

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 long, the 2-norm is used to normalize the contour vector. Since the edge point n_t 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, the contour vector results shown in figure 2(c) below.



(a) (b)

(c)

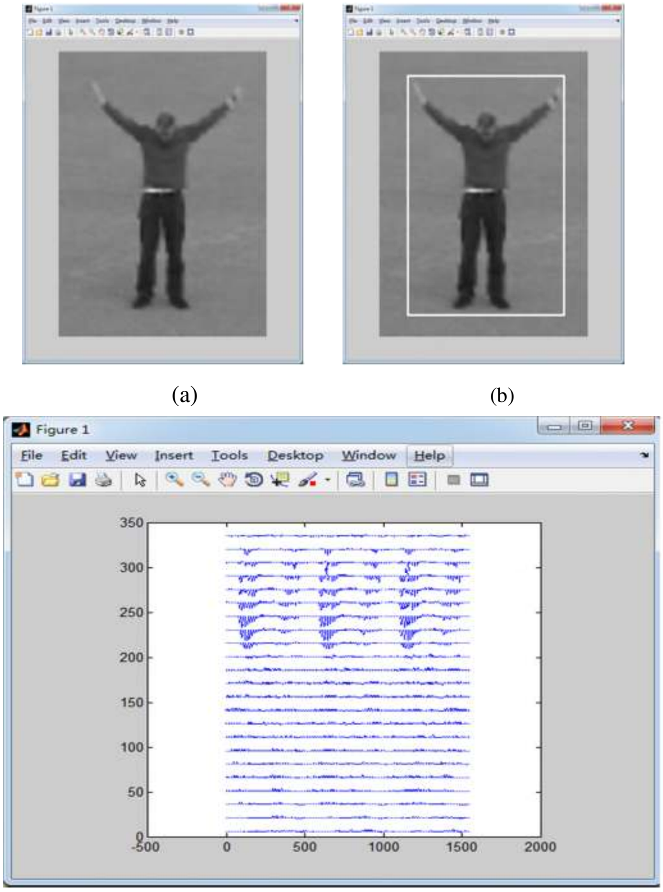
Fig. 2. Contour vector results

III. Local Optical Flow Feature Representation and Extraction

A. 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, in order 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:

1) 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 showed in figure 3 (a) and (b).

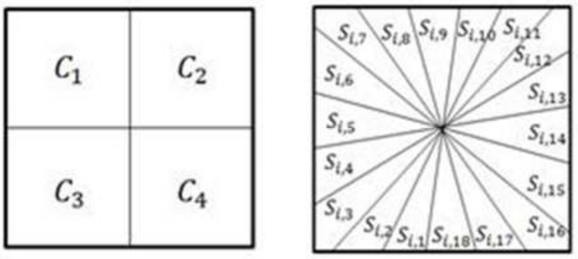


(a) (b)

(c)

Fig. 3. Light flow diagram

2) In order to decrease 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 showed in figure 3 (c) 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 figure 4 (a) below. Then, the sub-frame is split into 18 sub-regions centered on the center point of the sub-frame. $S_{i,1}$, $S_{i,2}$, ... $S_{i,18}$, (i = 1,2,3,4), each center angle is 20°, thus forming 72 sub-regions, as showed in figure 4 (b).



(a) (b)

Fig. 4. Optical flow feature extraction

3) In the sub-region $S_{x,y}$ 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 optical flows $O_{I_{x,y}}$ of the sub-region $S_{i,j}$ (the sum of lateral light flows $O_{H_{i,j}}$). Equation (5) and (6) calculates the sum of the sum of the longitudinal optical flows and the lateral optical flow, respectively.

$$O_{I_{i,j}} = \sum_{m=1}^{k} O_{Lm}\ , \qquad \left(x_{O_{Lm}}, y_{O_{Lm}}\right) \in S_{i,j} \tag{5}$$

$$O_{H_{i,j}} = \sum_{m=1}^{k} O_{Hm}\ , \qquad \left(x_{O_{Hm}}, y_{O_{Hm}}\right) \in S_{i,j} \tag{6}$$

4) 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 (7) to (9). Optical flow characteristics of the extraction, the parameters set reference [16].

$$O_L = \left[O_{L_{1,1}}, \dots, O_{L_{1,18}}, \dots, O_{L_{4,1}}, \dots, O_{L_{4,18}}\right] \tag{7}$$

$$O_H = \left[O_{H_{1,1}}, \dots, O_{H_{1,18}}, \dots, O_{H_{4,1}}, \dots, O_{H_{4,18}}\right] \tag{8}$$

$$O_t = [O_L, O_H] \tag{9}$$

5) 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 showed in (10).

$$F_t = [O_t, D_t] \tag{10}$$

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.

B) 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:

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

$$s_q = \min\left\|M_Q^t - M_T^n\right\|, \qquad n = 1,2,\dots,N \tag{11}$$

2) 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.

3) 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.

IV. 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 MATLAB2010b 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, and each original image selected is a 514*670 pixel screenshot; Background is relatively static, in addition to the lens closer / pull away, the camera's movement is relatively slight, as shown in figure 5.

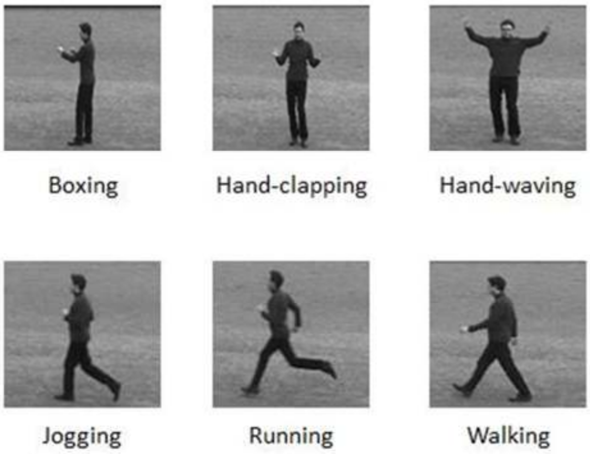


Fig. 5. Kth database six action diagram

In this paper, we extract the contours of human motion contours, optical flow characteristics and mixed features to characterize the action. In order to get 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 verified as a test sample, and statistical identification results. And in order to do classification and identification we use the Nearest Neighbor Method. Optical flow characteristics, contour distance characteristics and the combination of the two characteristics of the hybrid feature recognition as showed in Table 1.

TABLE I. SELECT THE RECOGNITION RATE CORRESPONDING TO THE DIFFERENT CHARACTERISTICS

Database	***Contour***	***Optical Flow***	***Mixed***
KTH	83.33%	93.33%	95.00%

This method and the recent related methods based on KTH database identification performance comparison shown in Table 2. It can show in 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 not difficult to extract and characterize, have high reliability, and avoid the complex operation of methods based on feature extraction of the human body model.

TABLE 2. THE DIFFERENT FEATURES COMBINE THE CORRESPONDING RECOGNITION RATES

Method	***Feature***	***Rate***
Ahmad[17]	Optical Flow +Shape Flow	88.29%
Sawant[18]	Silhouette +Local Optical Flow	90.80%
Tran[16]	Local Silhouette +Local Optical Flow	91.70%
Improved Combined Feature	Global Silhouette +Local Optical Flow	95.00%

V. CONCLUSION

The recognition of human action recognition is one of the research subjects. It has significant 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 being mounted 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 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, detection data are too monolithic. Only one database is used; and there are no experiments on a variety of complex conditions. So we will go ahead with study, perfect our algorithm.

ACKNOWLEDGMENT

This research was supported by the Busan Metropolitan City, Korea, under the 2017 BB21 Program grants.

REFERENCES

[1] Johansson, G. Visual motion perception. Scientific American. 1975

[2] Aggarwal, J. K. and A. Cai. Human motion analysis: A review. IEEE.1997

[3] Moeslund, T.B. and E. Granum. A survey of computer vision-based human motion capture. Computer vision and image understanding 81(3):231-268. 2001

[4] Moeslund, T.B., A. Hilton, et al. A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding 104(2):90-126.2006

[5] Turaga, P., R. Chellappa, et al. Machine recognition of human activities: A survey. Circuits and Systems for Video Technology, IEEE Transactions on 18(11):1473-1488. 2008

[6] Poppe, R. A survey on vision-based human action recognition. Image and Vision Computing. 28(6):976-990. 2010

[7] Aggarwal, J. and M. S. Ryoo. Human activity analysis: A review. ACM Computing Surveys (CSUR) 43(3):16. 2011

[8] Forsyth, D. A., O. Arikan, et al. Computational studies of human motion: Tracking and motion synthesis, Now Pub. 2006

[9] Gavrila, D. M. The visual analysis of human movement: A survey. Computer vision and image understanding 73(1):82-98. 1999

[10] Bobick, A. F. and J. W. Davis. The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence. IEEE Transactions on 23(3):257-267. 2001

[11] Wang Y., K. Huang, et al. Human activity recognition based on r transform. IEEE. 2007

[12] Chen H. S., H. T. Chen, et al. Human action recognition using star skeleton, ACM Computing Surveys.2006

[13] Wang L. and D. Suter. Informative shape representations for human action recognition. IEEE. 2006

[14] Weinland D.,E. Boyer, et al. Action recognition from arbitrary views using 3D exemplars. IEEE. 2007

[15] Weinland D. and E. Boyer. Action recognition using exemplar-based embedding. IEEE. 2008

[16] Tran D, Sorokin A. Activity recognition with metric learning [C]. IEEE Press,2008:61-66.

[17] Ahmad M, Lee S. Human action recognition using shape and CLG-motion flow from multi-view image sequences [J]. Pattern Recognition, 2008,41(7):2237-2252

[18] Sawant N, Biswas K. Human action recognition based on spatiotemporal feature [C]. Berlin: Springer-Ver-lag, 2009:357-362

.