Human Behavior Recognition Based on Wavelet Moment and Regional Optical Flow

Manyi Wang, Liang Zhang

Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China, Tianjin, 300300, China
my_wang1217@163.com, l-zhang@cauc.edu.cn

Abstract - Motion history image (MHI) and motion energy image(MEI) can recognize simple actions effectively, but they can't represent the velocity information of the actions. Sometimes in the behavior recognition, we must take the speed into account. For example, touching a person, fierce and friendly touches are two totally different movements. The paper combines the wavelet moment of temporal motion descriptors (MHI and MEI) and the speed feature based on optical flow to represent the action. Wavelet moments are rotation, translation and scale invariance. Foreground is obtained from video sequences by background subtraction. Then we use Lucas-Kanade algorithm to get the regional optical flow information, whose direction amplitude can represent the velocity changes in different direction intervals. Experiments based on video sequences outdoor scenarios carried out to verify the effectiveness of the proposed method.

Index Terms - Human behavior recognition, MHI, MEI, Wavelet moment, Optical flow calculating

1. Introduction

Human behavior recognition and understanding is an advanced vision analysis of human motion processing domain and the most challenge research direction of computer vision domain. The ultimate target is to analyze and understand individual behavior and interpersonal behavior. For many missions, automatic recognition of human behavior is very important. It can improve people's ability in sports, provide security monitoring for important places and enhance the ability of human-computer interaction, etc. Usually, human behavior recognition is based on accomplishing human tracking and extracting features in the image sequences successfully, belongs to the higher level vision task.

Till-to-date, there are excellent surveys on human action recognition and analysis [1-5]. These papers have covered many detailed approaches and issues and most of these have referred the Motion History Image (MHI) and Motion Energy Image (MEI) method [6] as one of the important methods for motion representation, which can be employed for recognition and other purposed for human behavior analysis and understanding. The Motion History Image and the Motion Energy Image approach is a view-based temporal template approach, which is simple but robust in representing movements and is widely employed by various research groups for action recognition and motion analysis. The advantage of temporal template-based methods is that silhouette sequence is condensed into gray scale images while dominant motion information is preserved. In 1961, Hu employed the theory of algebraic invariants and derived his seven famous invariants to rotation of 2D objects. Bobick proposed Motion History Image (MHI) and Motion Energy Image(MEI), and calculate their Hu moments as the template. Hu moments can extract global features, but it's sensitive to noise.

This paper proposes a feature description method which combines wavelet moment and regional optical flow to include both contour feature and velocity information. In Section 2, we present detailed approaches of the motion representations by the combination of motion history image and motion energy image. Section 3 illustrates the recognition approach by exploiting these motion templates. Section 4 depicts the results and analysis of this combined cue. Finally, we conclude the paper in Section 5.

2. Temporal Motion Descriptors

To describe the motion-shape and spatial distribution of motion, the Motion Energy Image (MEI) is proposed. In order to describe how the motion is moving in the image sequence, we can form a Motion History Image (MHI). MHI $H_\tau(x,y,t)$ can be computed from an update function $\psi(x,y,t)$:

$$H_\tau(x,y,t)=\begin{cases}\tau & if\ \psi(x,y,t)=1\\ \max(0,H_\tau(x,y,t-1)-\delta) & other\end{cases} \quad (1)$$

Where (x, y) and t show the position and time. The symbol $\psi(x,y,t)$ shows object's presence (or motion) in the current video image, the parameter τ decides the temporal duration of the MHI, and the symbol δ is the decay parameter.

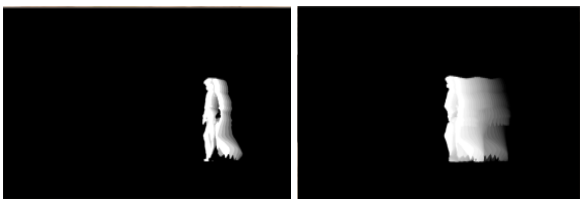

Fig.1: At different time, MHIs are different.

The MEI is the cumulative binary motion image, which can describe where a motion is in the video sequence, computed from the start frame to the final frame. Let

This work was supported by the National Natural Science Foundation of China under Grant 61179045.

$I(x,y,t)$ be an image sequence. Then the binary MEI $E_\tau(x,y,t)$ is defined as follows

$$E_\tau(x,y,t)=\begin{cases}1 & if\ H_\tau(\mathrm{x,y,t})\geq 0\\ 0 & otherwise\end{cases}. \qquad (2)$$

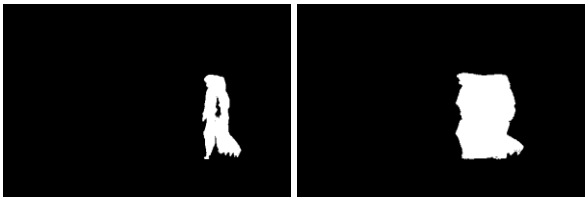

Fig.2: At different time, we can see the MEIs are different.

3. Feature Vectors Calculating

A. Wavelet moment

Although wavelet moment invariants are rotation invariant naturally, translation invariant and scaling invariant are achieved by using a normalization base on regular moments. The regular moment m_{pq} is defined as follows

$$M_{pq}=\iint x^p y^q f(x,y)dxdy. \qquad (3)$$

Translation invariant is achieved by transforming the original image into a new one whose first order regular moments, m_{01} and m_{10}, are both equal to zero. This is done by transforming the original image to a new one $f(x+\bar{x},y+\bar{y})$. The symbols $\bar{x}$ and $\bar{y}$ are the centroid of the original image:

$$\bar{x}=\frac{m_{10}}{m_{00}},\ \bar{y}=\frac{m_{01}}{m_{00}}. \qquad (4)$$

Scaling invariant can be achieved by transforming the original image $f(x,y)$ to a new function $f(x/a,y/a)$, where $a=\sqrt{s/m_{00}}$, S is the expected size of the image.

Therefore, the original images can be normalized according to the following transformation

$$\begin{pmatrix}x\\ y\end{pmatrix}\rightarrow\begin{pmatrix}(x-\bar{x})/a\\ (y-\bar{y})/a\end{pmatrix}. \qquad (5)$$

After normalization, wavelet moment has not only rotation, shift, scale invariants, but also has multi-resolution characteristic of wavelet. It is not sensitive to noise. It can extract image local features and global features so that it can describe the image comprehensively. In this paper, the wavelet moments of MHI and MEI are used as the templates.

After polar coordinating, the regular moment m_{pq} changes to

$$F_{pq}=\iint f(r,\theta)g_p(r)e^{jq\theta}rdrd\theta. \qquad (6)$$

Where F_{pq} is the $p+q$ order moment, $g_p(r)$ is the kernel function of radial variable r, p and q are integer parameters. $e^{jp\theta}$ is the kernel function of angle variable. Equation (3) can be rewritten as

$$F_{pq}=\int S_q(r)g_p(r)rdr. \qquad (7)$$

where

$$S_q(r)=\int f(r,\theta)e^{jq\theta}d\theta \qquad (8)$$

We use wavelet basis function $\psi^{a,b}=\frac{1}{\sqrt{a}}\psi(\frac{r-b}{a})$ instead of $g_p(r)$, where $a=0.5^m$, $b=n*a$, $m=0,1,2,3$, $n=0,1,2,...,2^m-1$ and $\psi_{m,n}(r)=2^{\frac{m}{2}}\psi(2^m r-n)$. The expression of wavelet moment is

$$\left\|F_{m,n,q}\right\|=\left\|\int S_q(r)\psi_{m,n}(r)rdr\right\|. \qquad (9)$$

The discrete wavelet moment is

$$\left\|F_{m,n,q}\right\|=\sum_{r=0}^{1}\sum_{\theta=0}^{2\pi}f(\mathrm{r},\theta)\psi_{m,n}(\mathrm{r})\,\mathrm{e}^{jq\theta}\,r. \qquad (10)$$

We choose the cubic B-spline wavelet as wavelet basis function

$$\psi(r)=\frac{4\alpha^{n+1}}{\sqrt{2\pi(n+1)}}\sigma_w\cos(2\pi f_0(2\pi-1))*\exp(-\frac{(2r-1)^2}{2\sigma_w^2(n+1)}) \qquad (11)$$

where $n=3$, $a=0.697066$, $f_0=0.409177$, $\sigma_w^2=0.561145$.

B. Mixture of Gaussians

The purpose of foreground extraction is to achieve object's relatively complete foreground region. Background subtraction is one of the most common methods [7]. The principle of the background subtraction is constructing a background to judge which is the foreground through the difference between the video frame and the background. The background subtraction includes the method based on global threshold [8] and the method based on background model [9-10]. The method based on global threshold segments the foreground by supposing there is a best threshold. But it lacks continuity of video frames. The method based on background model usually adopts mathematical methods such as MoG (mixture of Gaussians) [9] and the kernel estimation of parameters [10] etc. to model the changes of pixels in order to get the background model.

This paper uses MoG (Mixture of Gaussians) [9] to model the changes of pixels to get the background model and update the model. When MoG is used for background modelling, the changes of each pixel in the time domain are simulated by K multidimensional Gaussian distributions.

If $\{X_1, X_2, ..., X_t\}$ represents the observation sequence of $P(x, y)$, at the moment t, the probability of the pixel X_t is $P(X_t) = \sum_{i=1}^{K} w_{i,t} N(X_t, \mu_{i,t}, \sum_{i,t})$, where K is the number of the Gaussian distributions, $w_{i,t}$ is the weight of the i th Gaussian distribution, $N(X, \mu, \sum)$ is the Gaussian probability density function, $\mu_{i,t}$ is the mean of the i th Gaussian distribution and $\sum_{i,t}$ is the covariance matrix of the i th Gaussian distribution.

Stauffer et al. point out that in the Mixture of Gaussians, K Gaussian distributions are sorted by $w_{i,t} / \left|\sum_{i,t}\right|^{1/2}$. Then we take the first B Gaussian distributions as the background models. If the difference between the current pixel and the background model of the Gaussian distribution is less than a threshold, we determine it is the background. It is represented by a binary variable B_t.

$$B_t = \begin{cases} 1, & \left\|X_t - \mu_t\right\| \le \beta \left\|\Sigma\right\|^{1/2}, \beta = 2.5 \sim 3.0 \\ 0, & other \end{cases} \quad (12)$$

The weight and background model are updated by

$$w_{i,t} = (1-\lambda) w_{i,t-1} + \lambda B_t \quad (13)$$

$$\mu_{i,t} = (1-\alpha)\mu_{i,t-1} + \alpha X_{i,t} \quad (14)$$

$$\sum_{i,t} = (1-\alpha)\sum_{i,t-1} + \alpha (X_{i,t} - \mu_{i,t})(X_{i,t} - \mu_{i,t})^T \quad (15)$$

where λ is the learning rate, $\alpha = \lambda / w_{i,t}$.

Fig.3 A frame of video and after background subtraction we can get the foreground.

C. Optical flow

Although the wavelet moment of MHI and MEI can extract the shape features of motion process preferably, it can't describe the action comprehensively without reflecting the speed characteristics. Optical flow features can solve this problem. Optical flow is the good spatial and temporal characteristic and contains pixel's instantaneous velocity vector information. Its disadvantages are the large amount of calculation and easily affected by environment. The model of mixture of Gaussions is proposed to indicate the variation of background pixels in order to speed up the calculation of optical flow and against the interference of environment.

The optical flow is calculated by Lucas-Kanade algorithm [11].We let the optical flow of $m \times m$ (we take $m = 5$) feature window be (u, v). At the condition of meeting the optical flow constraints $I_x u + I_y v + I_t = 0$, we can get the equation as follows

$$\begin{vmatrix} I_{x_1} & I_{y_1} \\ I_{x_2} & I_{y_2} \\ \vdots & \vdots \\ I_{x_n} & I_{y_n} \end{vmatrix} \begin{vmatrix} u \\ v \end{vmatrix} = - \begin{vmatrix} I_{t_1} \\ I_{t_2} \\ \vdots \\ I_{t_n} \end{vmatrix} \quad (16)$$

where n is the number of pixels in the feature windows, I_x and I_y are spatial gradient of image, I_t is the temporal gradient of image. By solving (8), we can get

$$\begin{vmatrix} u \\ v \end{vmatrix} = \begin{vmatrix} \sum I_{x_i}^2 & \sum I_{x_i} I_{y_i} \\ \sum I_{x_i} I_{y_i} & \sum I_{y_i}^2 \end{vmatrix}^{-1} \begin{vmatrix} -\sum I_{x_i} I_{t_i} \\ -\sum I_{y_i} I_{t_i} \end{vmatrix} \quad (17)$$

The computational complexity of regional optical flow is $O(\eta N / m^2)$, η is the ratio that foreground occupies the entire image, N is the number of the pixels and m^2 is the size of the feature windows.

Generally, foreground area is less than 40 percent of the current video frame size. This method reduces the amount of computation greatly. Calculation of optical flow based on motion region can effectively remove the disturbance of environment.

We extract optical flow features $V = \{F_1, F_2, ..., F_n\}$ of motion foreground region, where n is the number of regions. Regional optical flow feature is $F = (x_c, y_c, w, h, a, d, r)$, where x_c and y_c are region's centre coordinates, w and h are region's width and height, a and d are region's amplitude and direction of optical flow and r is region labelling.

The speed and the direction of each kind of actions has its own characteristics. We use the method of direction and amplitude weighted to describe the regional optical flow feature. Firstly, the direction is divided into several angle intervals. Secondly, vectors are classified to each interval. Then we sum the vector amplitude of each interval to form a n-dimension feature vector. The larger the amplitude of the direction of the action is, the higher the speed of the it is. Regional optical flow feature vector is $H(R) = \{h_j(R)\}_{j=1,2,...,m}$, m is the number of the intervals. In this paper, we make $m = 4$ and the size of the interval is 90 °.

$$h_j(R) = C_{norm} \sum_{i=1}^{B} w_{F_i} \delta(a(F_i) - j) \quad (18)$$

where B is the number of optical flow features in region R, w_{F_i} is the weight of the i th optical flow feature, $a(F_i)$

represents F_i 's direction interval, j represents the direction interval, δ is the Kronecker delta function, C_{norm} is the normalized parameter and w_{F_i} is the weight in this direction.

4. Classification Results and Analysis

To test the effectiveness of the algorithm, we use the Weizmann Database that contains ten different actions.

Fig.4 Ten different actions-each row represent two actions, which are later mentioned in the Tables.

We show some sequential frames for each action in Fig.4. The first two images present the action of "bending". In the similar fashion, the following actions are jumping jack; jumping; vertical jumping (pjump in the Table I); running; sideways jumping (side); skipping; walking; waving one arm; waving both arms. The dress, height, size and age of the subjects were different for each person.

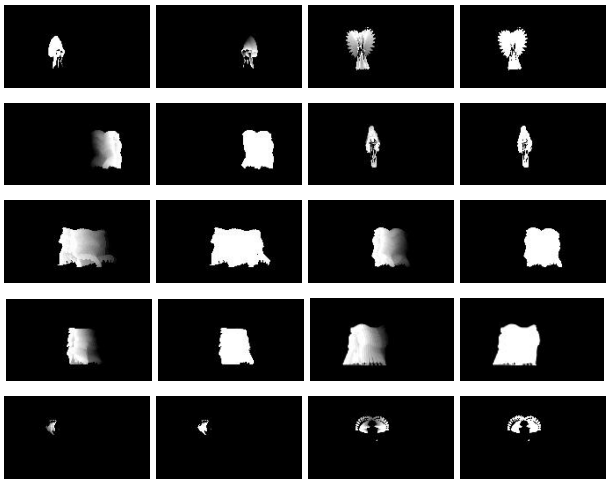

Fig.5 MHI and MEI of ten kinds of actions

Support Vector Machine (SVM) is utilized to classify the actions. The leave-one-out method is used as the evaluation method. Table I shows the corresponding recognition results for the dataset. The various actions demonstrate 85.5% recognition results. Four kinds of actions in the database, the recognition rate is 100%. Extracting the features by HU moment, we can see the average recognition rate is 71.5%.The feature extraction method combined wavelet moment and regional optical flow can better describe the action.

TABLE I. the Recognition Rate Comparison

Action	Wavelet Moment (100%)	HU Moment(100%)
Bend	88.8	66.6
Jack	100	100
Jump	77.7	66.6
Pjump	100	88.8
Run	77.7	66.6
Side	66.6	44.4
Skip	55.5	33.3
Walk	100	77.7
Wave1	100	88.8
Wave2	88.8	77.7
Average	85.5	71.5

5. Conclusions

This paper presents a feature extraction method combined wavelet moment and regional optical flow for the human behavior recognition. We describe the shape feature of the motion process by calculating the wavelet moments of the temporal descriptors. At the same time, by calculating the regional optical flow's direction amplitude of the motion, we can get the speed characteristic in different angle interval. The method is robust. The experiment shows that this feature extraction method can recognize human behavior effectively.

References

[1] M. A. R. Ahad, J. K. Tan, H. Kim and S. Ishikawa. "Human activity recognition: various paradigms", *Int'l Conf Control Automation and Systems*, pp.1896-1901.2008.

[2] A. Jaimes and N. Sebe. "Multimodal human-computer interaction: a survey", *Computer Vision and Image Understanding*, 108(1-2): pp.116-134, 2007.

[3] R. Poppe. "Vision-based human motion analysis: an overview", *Computer Vision and Image Understanding*, 108(1-2):pp. 4-18,2007.

[4] M. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. "Human activity analysis: concentration on Motion History Image and its variants", *SICE-ICASE Joint Annual Conf*, Japan, Aug.2009.

[5] T. TB. Moeslund, A. Hilton, and V. Kruger. "A survey of advances in vision-based human motion capture and analysis", *Computer Vision and Image Understanding*, pp.104: 90-126,2006.

[6] A. Bobick and J. Davis. "The recognition of human movement using temporal templates", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 23(3).pp. 257-267, 2001.

[7] M. Piccardi. "Background subtraction techniques: a review", *IEEE International Conference on Systems*, Man and Cybernetics. pp. 3099-3104 2004.

[8] N. Otsu. "A threshold selection method from gray-level histogram". *IEEE Trans on System Man Cybernetics*,pp.62-66, 1979.

[9] C. Stauffer and W. E. L. Grimson. "Adaptive background mixture models for real-time tracking", *Computer Vision and Pattern Recognition*. pp. 246-252, 1999.

[10] A. Elagammal, D. Hanvood, L.S.Davis. "Nonparametric model for background subtraction". *Proceedings of the 6th European Conference on Computer Vision-Part II*, pp.751-767, 2000.

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", *Imaging Understanding Workshop.* pp. 121-130, 1981.