Available online at www.sciencedirect.com

SciVerse ScienceDirect

AASRI Procedia 3 (2012) 184 – 190

AASRI Procedia

www.elsevier.com/locate/procedia

2012 AASRI Conference on Modelling, Identification and Control

Human Behavior understanding via Top-View vision

Qing Lin[1], CanCan Zhou[*1,2], Shitong Wang[3], Xiaogang Xu[4]

[1,2,4]*School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang, 212013, China*
[1]*School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, 210094, China*
[3]*School of Information Technology, Southern Yangtze University, Wuxi 214122, China*

Abstract

Aiming to target occlusion problem in complex scenes, human action recognition via a top-view vision is proposed. However, in this view, the human behavior rotated will be mistakenly identified as another behavior. To address this situation, taking into account the rotation invariance of the moments, human static posture are represented by Hu moments, and using the SVM as trainer and classifier. According to the change of coordinate of the binary image centroid, semantic web of dynamic behavior is established. The experimental results show that this method can accurately identify the human dynamic information and has a high recognition rate.

keywords: top-view vision, rotation invariance, Hu moments, SVM, centroid

1. Introduction

Human behavior understanding is very active in the field of computer vision, and with the rapid development of computers and other technology, it has made great progress [1]. In recent years, people have started from theoretical research to practical application scenarios. It is great significance in the real world. Now at many public occasions, the camera is located in the corners or top, so in this paper, based on a positive view would be impractical. At present, most of the behavior studies are based on the corner of the camera, while the top of the camera mostly concentrated on the statistics of the number of people [2,3], and action understanding literature is rare [4].

Typically, vision-based human behavior analysis in general compliances with a few basic processes such as feature extraction, motion characterization, action recognition, high-level behavior understanding and scene understanding [5]. Human action characteristics extracted from image sequences are essential to behavior recognition and behavior understanding. The results based on the positive view and corner view in the complex environment of the room or station will be greatly curtailed by kind of occlusion.

Therefore, a method based on the top-view is proposed. The method has the inherent advantage to solve the occlusion problem. However, in top-view, rotation of people makes the common feature representation is invalid. The translation invariance and rotation invariance of Hu moments are used to represent the gesture feature vector [6]. At the same time more high-level semantic information is drawn according to the change in the coordinates of the centroid.

Corresponding author. Tel.: +86-18796001319.
*E-mail address: youhunzh@163.com，jcs533@qq.com

doi:10.1016/j.aasri.2012.11.031

2. Feature Extraction and Movement Feature Representation

2.1 Feature Extraction

The underlying feature extraction is the first step and is also a crucial step in analysis of human action. Whether to extract high-quality moving target is the key to the success of the entire system. Gaussian mixture model (GMM) is commonly used in motion segmentation [7]. It has a good effect on the leaves jitter and water ripples.

K (3~5) Gaussian distribution is indicated the state of each pixel in the image. The probability of Pixel X at time t is:

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \bar{\mu}_{i,t}, U_{i,t}) \tag{1}$$

Where K represents the number of Gauss, $\bar{\mu}_{i,t}, U_{i,t}, \omega_{i,t}$ denotes the standard deviation, covariance and weights of the i-th Gaussian at time t, and η represents Gaussian density function.

$$\eta(X_t, \mu_{i,t}, U_{i,t}) = \frac{1}{(2\pi)^{\frac{n}{2}} |U_{i,t}|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu_{i,t})^T U_{i,t}^{-1}(X_t-\mu_{i,t})} \tag{2}$$

If $\left|X_t - \bar{\mu}_{i,t-1}\right| \le \lambda\sigma_{i,t-1}$, you need to update the Gaussian model, otherwise update the minimum right of the Gaussian model, and λ is an empirical value, usually take 2.5 to 3.5.

The update process is as follows:

$$\omega_{i,t} = (1-\alpha)\omega_{i,t-1} + \alpha M \tag{3}$$

$$\bar{\mu}_{i,t} = (1-\rho)\bar{\mu}_{i,t-1} + \rho X_t \tag{4}$$

$$\sigma_{i,t}^2 = (1-\rho)\sigma_{i,t-1}^2 + \rho(X_t - \mu_{i,t-1})^T (X_t - \mu_{i,t-1}) \tag{5}$$

$$\rho = \alpha\eta(X_t \mid \bar{\mu}_k, \sigma_k) \tag{6}$$

Among them, α is the learning rate of model, usually take 0.002. ρ is the update rate of model.

In the experiments we found that the scene transient probability is relatively small so model update rate should be taken a smaller one. We sort the weight of GMM, and take the 65% of sum of weights as a useful model. Moving target can be obtained quickly by the above method. However, because of noise, light, and many other factors, the image will appear hollow noise points, shadows, etc. so image post-processing is required. In this paper, the color model is used to calculate the shaded area [8], and the model is described as follows.

Assume that the shadow of the target pixel Shd in the YCbCr space vector represent as Shd(Cx,Cy,Cz). The background pixels 'Back' in the YCbCr space vector is expressed as Back(Bx,By,Bz). Among them, x is the Cb components, y is the Y components and z is the Cr components. The model of YCbCr space is shown in Figure 1.

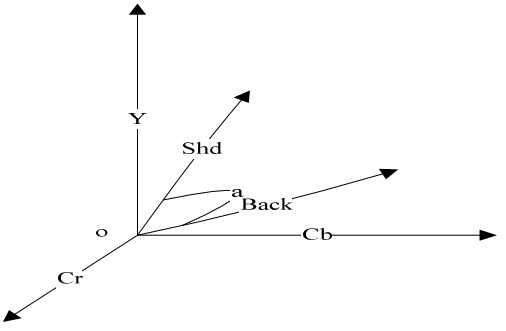

Figure 1 YCbCr space

Figure 2 original picture and processed

Angle a represents the value of the current pixel value relative to the background. The current image and background image pixels of YCbCr known, a is easy to calculate as follows:

$$\cos a = \frac{Shd \bullet Back}{|Shd| \bullet |Back|} \tag{7}$$

The shadow of the role played by the three components is different. The brightness information Cx of the shadow has not change when compared to the background information Bx, so as to avoid the interference of the brightness information, it need to reduce the luminance Y component weight and increase the weight of the Cb and Cr. $\cos a$ obtained from the above equation:

$$a = a\cos\frac{\beta^2 Bx \cdot Cx + \lambda^2 By \cdot Cy + \beta^2 Bz \cdot Cz}{\sqrt{\beta^2 Bx^2 + \lambda^2 By^2 + \beta^2 Bz^2}\sqrt{\beta^2 Cx^2 + \lambda^2 Cy^2 + \beta^2 Cz^2}} \tag{8}$$

$$\beta = 1 - \frac{\lambda}{2} \tag{9}$$

And the value of λ is 0.3.

Next, apply the threshold T (T is 4) to determine the eigenvalue a, if a >T Binary is target, otherwise is background.

$$Binary = \begin{cases} shadow \ \ if(a < T) \\ foreground \ \ if(a > T) \end{cases} \tag{10}$$

After these steps and through the expansion and corrosion, the effect of morphological processing is shown in Figure 2.This article only research three basic static postures (standing, squat, lying).

2.2 Movement Feature Representation

The shape features need to be same after rotation in the top-view. So, the invariance of the geometric transformation is employed to represent the behavioral characteristics. Certain moments of the image area have the same characteristics such as translation, rotation, scale and other geometric transformations, so it is widely used in pattern recognition.

In recent years, Hu moments and its improved algorithm are widespread concerned by researchers. (p+q) order moment m_{pq} of Two-dimensional f(x, y) and center moment μ_{pq} are represented as follow：

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \tag{11}$$

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{12}$$

Otherwise, p and q are in range of [0, n]. $(\bar{x}, \bar{y})$ is the centroid coordinate of the two-dimensional image.

According to the center distance of second-order and third-order nonlinear combination, Hu constitute the seven invariant moments, and its expression as follows：

$$\begin{aligned}
M_1 &= \eta_{02} + \eta_{20} \\
M_2 &= (\eta_{02} - \eta_{20})^2 + 4\eta_{11}^2 \\
M_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
M_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
M_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
M_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
M_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{13}$$

What's more, $\eta_{pq} = \mu_{pq} / \mu_{00}^r$, r=(p+q+2)/2

Every moment of the gesture feature vector K_t represents as follows:

$K_t = [M_1, M_2, M_3, M_4, M_5, M_6, M_7]$

3. Behavior Recognition

3.1 Static Posture Recognition

Static posture recognition is the basis of high-level behavior understanding. It plays a crucial role in the accurate classification of static posture recognition rate for high-level behavior. This article will take the following steps to identify the static posture.

(1) Use the training data to train SVM. SVM is a learning method based on structural risk minimization criteria and takes training error as the constraints of the optimization problem [9]. SVM classifier with kernel is shown in formula and we can get a classification.

$$f(x)=\sum_{i}^{N}\alpha_i k(x_i,x)+b \tag{14}$$

$$f(x)=\begin{cases}\geq 0 & positive\ \ class\\ <0 & negative\ \ class\end{cases} \tag{15}$$

α_i is the weight of SVM. N is the number of training data. $k(x_i,x)$ represents a kernel.

(2) Take the SVM for classification. Because SVM is a single classifier, so we should build M kinds of classifiers.

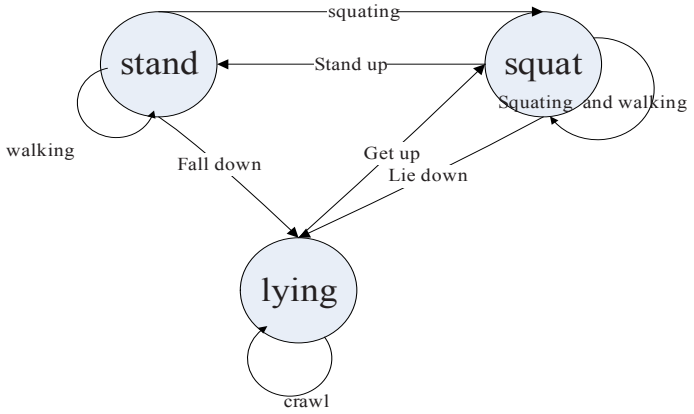

Figure 3 dynamic behavior of the Semantic Web

3.2 High-Level Behavior Recognition

Static posture recognition is far from meeting the needs of people. Accurately our aim is to identify the dynamic information. In recognition of the static posture, you can employ the coordinates of the centroid changes to identify the higher level of semantic information.

In the experiment, a static posture is recorded every five frame. If the frame dissatisfies all the above set of values, it's beginning to find eligible value and record it every three frames. Then you get all the records according to the current record and previous records to identify the dynamic behavior information of Figure 3.

4. Experiment Result and Analysis

In the experiment, the camera fixed at the top of hall and the lens direction is vertically downward. The distance between ground and hall is about 6m.The experimental environment is VS2010 and Opencv2.2. And experimental set with 30 frames /s video data.

4.1 Static Posture Recognition

200 samples are taken to train each posture model. Then the static behavior at a frame is detected by the SVM. And the behavioral sequence such as walking, crawling and squatting is showing in Fig 4.

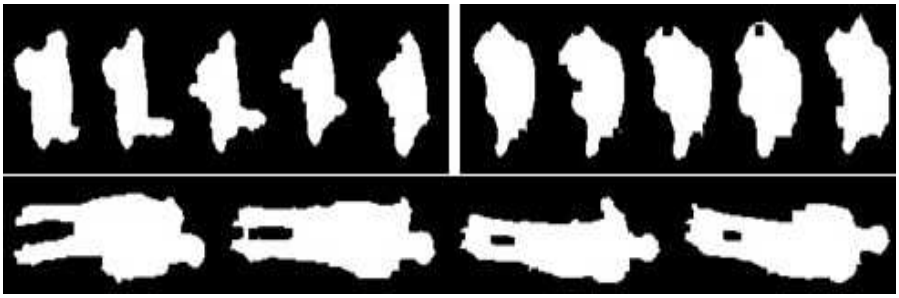

Figure 4 some test set sequence

It is only taken a sequence of 100 frames to test each gesture. The first moment of Hu moments is shown in Figure 5. Each gesture contains a rotating action and other factors. And we all know the meaning of first value of Hu moment is area.

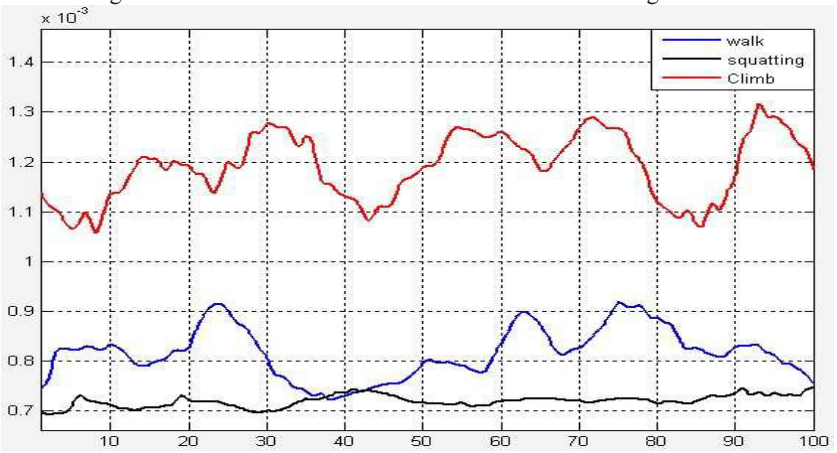

Figure 5 first value of Hu moment

The recognition result is shown in Table 1 by literature [4]. It can be seen from the experiment that the recognition rate of lying is 100% using this article method, and seven frames of standing will be detected as squatting and three frames of squatting will be detected as a standing. According to the analysis of the test video sequences, squatting, arms and some stretch will be detected as a standing, and the relatively larger stretch even can be detected as lying. At the same time, leaning forward will be identified as the standing when squatting. If standing with body tilt down, it may be detected as a squatting. The recognition rate of this article has significantly improved than literature 4.

Table 1 recognition rate of static behavior

behavior	function	number of samples	correct number	wrong number
stand	this article	100	97	3
	literature 4	100	90	10
squat	this article	100	93	7
	literature 4	100	88	12
lying	this article	100	100	0
	literature 4	100	85	15

4.2 High-Level Behavior

According to the eight high-level behaviors in Figure 3, each dynamic behavior takes 10 samples for testing. And the recognition result is shown in Table 2.

Table 2 recognition rate of high-level behavior

high-level behavior	total sample	Correct number	wrong number	rate（%）
walking	10	8	2	80
squatting	10	9	1	90
stand up	10	9	1	90
get up	10	8	2	80
crawl	10	10	0	100
lie down	10	7	3	70
Squatting and walking	10	8	2	80
fall down	10	9	1	90

5. Conclusion

A top-view behavior recognition method is presented in the paper. It is able to identify three static postures and eight high-level dynamic behaviors. The experimental results show that the identification of this article is significantly higher than literature [4]. However, arm lift or not is very serious to behavior recognition, and how to overcome the impact of the arm is the focus of future research. This article is the basis of the behavior recognition of this view and we will focus on the other behavior, people interaction and tracking research.

Acknowledgements

This work was financially supported by National Natural Science Foundation (61272211) and Jiangsu Provincial Natural Science Foundation (BK20111156).

References

[1] Turaga P, Chellappa R, Subrahmanian V S, Udrea O. Marchine recognition of human activities: A survey. IEEE Transactions onCircuits and Systems for Video Technology, 2008,18(11):1473-1488
[2] X.Liu, P.H.Tu, J.Rittscher, A.Perera,,and N. Krahnstoever, “Detecting and counting people in surveillance applications” , in Proc. IEEE Conf.Adv. Video Signal Based Surveillance, 2005, pp. 306–311.
[3] I. Cohen, A. Garg, and T. S. Huang, “Vision-based overhead view person recognition,” in Proc. Int. Conf. Pattern Recog.,2000, vol. 1,pp. 1119–1124.
[4] Skulkittiyut Weerachai, Prof. Makoto MIZUKAWA,"Human Behavior Recognition via Top-View Vision for Intelligent Space",IEEE international conference on Control,pp.1687 - 1690 ,2010.
[5] ZhiGang Lin, ChunHui Zhao,Yan Liang, Quan Pan, Yan Wang. Vision-based Human Behavior Understanding Summary [J]. Application Research of Computers, 2008,25(9)
[6] HU MING-KUEI.Visual pattern recognition by moment invariants [J].IRE Transactions on Information Theory, 1962, IT-8(2) : 179－187.
[7] J.H Piater and J.L. Crowley. Multi-modal tracking of interacting targets using gaussian approximations. In International Workshop on Performance Evaluation of Tracking and Surveillance, 2001.
[8] Kobus Barnard, Graham Finlayson. Shadow Identification Using Colour Ratios [C].In: IS&T/SID 8th Color Imaging Conference: Color Science, Systems and Applycation. 2000: 97- 101
[9] Meyer, David, Leisch, Friedrich, and Hornik, Kurt. The Support Vector Machine Under Test, Neurocomputing 55(1–2): 169–186, 2003
[10] A. Mokhber, C. Achard, M. Milgram, "Recognition of human behavior by space-time silhouette characterization," Pattern Recognition Letters, vol. 29, pp. 81-89, 2008.

[11] S. Belongie, J. Malik, J. puzicha, "Shape matching and object recognition using shape contexts," IEEE Transactions on Pattern Analysis and Machine Intelligence, vo1.24, pp.509-522, 2002.
[12] Thomas B. Moeslund, Adrian Hilton and Volker Kruger, “A survey of advances in vision-based human motion capture and analysis”, Computer Vision and Image Understanding 104 (2006) 90–126.