Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (19): 2276-2284, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.2276.2284

Human Action Recognition Based on Multiple Instance Learning

Shaoping Zhu and Dan Song
Department of Information Management,
Hunan University of Finance and Economics, Changsha, 410205, China

Abstract: In this study, a novel method based Multiple Instance Learning is proposed for human action recognition in video image sequences. First of all, HOG and T-HOG model is used for extracting space-time interest points feature, optical flow model is used for extracting motion features which are used to characterize human action. Then we combine spatial-temporal points of interest vector with the optical flow vector to form a hybrid feature vector. Final Multiple Instance Learning algorithm is presented which is used to recognize human actions. Experimental results show the effectiveness of the proposed method in comparison with other related works in the literature and the proposed method can enhance the robustness, also tolerate noise and interference conditions.

Key words: Human action recognition, multiple instance learning, space-time interest points, optical flow model

INTRODUCTION

Recent advances in computer vision have fuelled numerous initiatives which aim to automatically recognize human actions. Recognizing human actions in video sequences has a wide range of potential applications, such as video surveillance system (Zhou *et al*., 2011), interpretation of sport events and human computer interactions and so on. In particular, a human action recognition system may enable the detection of abnormal actions as opposed to the normal action of persons in public places like airports and subway stations. Automated human action recognition may be used for real-time monitoring of the elderly people, patients or babies (Amien and Lin, 2007). Specifically, human action recognition aims at automatically telling the activity of a person, i.e., to identify if someone is walking, dancing or performing other types of actions. However, it remains a challenging problem in computer vision to achieve robust human actions recognition from image sequences due to occlusion, human postures, illumination conditions, changing backgrounds, camera movements, photometric and geometric variances of objects. With moving target, moving cameras, non-stationary background and few vision algorithms could categorize and recognize such human actions well.

In this study, we propose a method for automatically recognizing human action from video sequences. This approach includes two steps: Extracting feature of human action and recognizing human action. In the extracting feature, space-time interest points feature of human action are extracted by using HOG and T-HOG method, motion features are extracted by using optical flow model. Then spatial-temporal point of interest vector is combined with the optical flow vector to form a hybrid feature vector. Final Multiple Instance Learning algorithm is used for human action recognition.

RELATED WORK

A lot of previous work has been presented in recognizing human actions from both still images and video sequences. Various approaches for human action recognition have been proposed in the literature (Ke *et al*., 2005). Human action evolve dynamically over time, Bayesian approaches and Hidden Markov Models (HMM) have been extensively used to detect simple and complex events that occur in the scenarios. e.g., Olivera *et al*. (2004), Bobick and Wilson (1997), Xiang and Gong (2006) who tries to model the full dynamics of videos using sophisticated probabilistic models. The problem with this approach is that those sophisticated models impose too many assumptions and constraints in order to be tractable. Learning those models is also hard since there are usually a large number of parameters which need to be set. This approach may limit

Corresponding Author: Shaoping Zhu, Department of Information Management, Hunan University of Finance and Economics, Changsha, 410205, China

its practical use. Recently, one popular approach is to apply tracked motion trajectories of body parts to recognizing human action (Yilmaz and Shah, 2005; Agarwal and Triggs, 2004). This is done with much human supervision and the robustness of the algorithm is highly dependent on the tracking system. Bobick and Davis (2001) use a representation known as "temporal templates" to capture both motion and shape, represented as evolving silhouettes. Flaherty *et al.* (2005) proposed a LDA method of motion processing for action recognition. Schuldt *et al.* (2004) performed human action recognition by training SVM classifiers. But these approaches ignore the contextual information provided by different frames in a video, the modeling and learning frameworks are rather simple. Another work named "video epitomes" is proposed by Cheung *et al.* (2005). However, all of the above methods demand aligned training examples to learn detectors with high detection accuracy via supervised learning. They model the space-time cubes from a specific video by a generative model. The learned model is a compact representation of the original video, therefore, these approaches are suitable for video super-resolution and video interpolation but not for recognition.

Recognizing human actions in video is both a challenging problem and an interesting research area. Generally, two important questions are involved in action recognition. One is how to extract useful motion information from raw video data and the other is how to model reference movements which effectively deal with variations at spatial and temporal scales within similar motion classes. Space-time gradients and other optical flow-based features are often useful for recognition.

FEATURE EXTRACTION

High quality features are essential to improve the recognition accuracy for recognizing human action. A variety of features both in time and frequency domains have been investigated in human action recognition such as variance, FFT coefficients, spectral entropy and correlation etc. However, little work has been done on feature analysis at the primitive level. In this study, we evaluate two feature sets. The first feature set contains space-time interest points for features. However, at primitive level, complex features may not be reliably calculated. Therefore, we only consider statistical features that can be reliably calculated at primitive level. These features are extracted by using HOG and T-HOG method. The second set of features is called motion features by optical flow model which are derived based on the physical parameters of human motion (Zhang and Sawchuk, 2011).

Spatial-temporal interest points feature representation: Spatial-temporal interest points method is one of the most popular approach in human action recognition. There are varieties of methods for space-time interest points detection in still images. Laptev and Lindeberg (2003) proposed an extended version of the space-time interest points detection which was in the spatial domain into space-time domain. Blank *et al.* (2005) extracted space-time features for human action recognition. Schuldt *et al.* (2004) trained a SVM classifier which based on the space-time feature for human action recognition. Nowozin *et al.* (2007) detected local interest points and learnt a set of discriminative subsequences for human action classification, by which used the sequence mining techniques from data mining. However, the interest points detection by using the generalized space-time interest point detector are too sparse to characterize many complex videos in literature (Dollar *et al.*, 2005). Therefore, we detect space-time interest points using HOG and T-HOG method to characterize human action (Laptev and Lindeberg, 2003; Thi *et al.*, 2012). Here we give a brief review of this method.

- Given a stabilized video sequence f(x, y), we detect space-time interest points

f(x, y, t) is the image in pixel (x, y) at time t, $g(x, y, t; \sigma_l^2, \tau_l^2)$ is Gaussian function that time and space parameters can be separated. Linear multi-scale space is defined as:

$$L(x, y, t; \sigma_l^2, \tau_l^2) = g(x, y, t; \sigma_l^2, \tau_l^2) * f(x, y, t) \quad (1)$$

where, * is convolution operator, σ_l^2, τ_l^2 is independent space scale variable and independent time scale variable.

We defined Gaussian function as:

$$g(x,y,t;\sigma_l^2,\tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \times \exp\left(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2\right) \quad (2)$$

Space-time second moment matrix μ is constructed to detect space-time interest points by convolution of multi-scale space L and gaussian weighting function which is expressed as:

$$\mu(x, y, t; \sigma_l^2, \tau_l^2) = g(x, y, t; \sigma_l^2, \tau_l^2) * (\nabla L(x, y, t; \sigma_l^2, \tau_l^2)(\nabla L(x, y, t; \sigma_l^2, \tau_l^2)^T) \quad (3)$$

where, $\sigma^2_i = s\sigma^2_1$ $\tau^2_i = s\tau^2_1$ $\nabla L(x, y, t; \sigma^2_1, \tau^2_1)$ are first derivatives of scale space function L, respectively in the x, y, t direction.

In this study, we use threshold function H to detect space-time interest points. Threshold function T is defined as:

$$T = \det(\mu)\text{-}k\ \text{tracce}^3(\mu) \quad (4)$$

Assuming $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$, λ_1, λ_2, λ_3 are eigen values of the second moment matrix μ. Threshold function T is expressed as:

$$T = \lambda^3_1(\alpha\beta\text{-}k(1+\alpha+\beta)^3) \quad (5)$$

where, $k \leq \alpha\beta/(1+\alpha+\beta)^3$ and a maximum of k is 1/27. The space-time interest points in images f(x, t, y) is the local where, H is the local maxima of time and space. Figure 1 shows examples of the area of space-time interest points for human actions for the KTH dataset.

- Gradient descriptor HOG and T-HOG are used to describe space-time interest point area cube

In STIP area cube of each image block frame f_{t+1}, then we calculate size and direction of each pixel gradient as follows:

$$\rho_{t+i}(x, y) = \sqrt{\begin{matrix}(L(x+1, y) - L(x-1, y))^2 \\ +(L(x, y+1) - L(x, y-1))^2\end{matrix}} \quad (6)$$

$$\theta(x, y) = \arctan\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \quad (7)$$

- We calculated time direction changes of the cube of space-time interest points. Scale space gradient between frames is defined as:

$$D(x, y, t) = L(x, y, t+1) - L(x, y, t) \quad (8)$$

Size and direction of scale space gradient difference between frames are calculated as follows:

$$\psi(x, y, t) = \sqrt{D(x, y, t)^2 + D(x, y, t+1)^2} \quad (9)$$

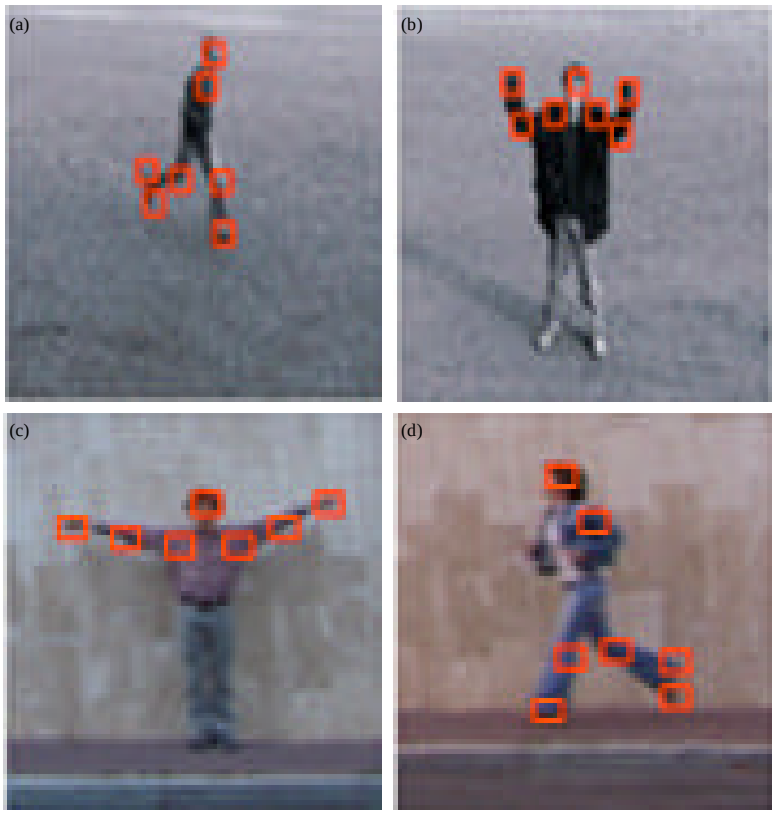

Fig. 1(a-d): Area of space-time interest points for human actions

$$\varphi(x, y, t) = \arctan\frac{D(x, y, t+1)}{D(x, y, t)} \quad (10)$$

- Statistic each frame pixel gradient in the same histogram and the cube space direction change information can be expressed as a feature vector η

Motion information of human representation: The human action is a dynamic event, it must be represented by the motion information of the human. So, we use motion features to characterize human action. The motion features (optical flow vector) are estimated by optical flow model. For example, Little and Boyd (1998) analyzed the periodic structure of gait recognition using optical flow patterns. In Efros *et al*. (2003), the motion descriptor is used to represent the video frames which has been shown to perform reliably with noisy image sequences and has been applied in various tasks such as action classification, motion synthesis, etc.

To calculate the motion descriptor, we must track and stabilize the person in a video sequence. We use an automatic human detection method in our experiments, since the motion descriptor that we use is very robust to jitters introduced by the tracking.

Given a stabilized video sequence, in which the person of interest appears in the center of the field of view, we compute the optical flow at each frame by using the Lucas and Kanade (1981) algorithm.

Optical flow equation is expressed as:

$$I_xF_x+I_yF_y+I_t = 0 \quad (11)$$

Where:

$$I_x = \frac{\partial I}{\partial x},\quad I_y = \frac{\partial I}{\partial y},\quad I_t = \frac{\partial I}{\partial t},\quad F_x = \frac{dx}{dt},\quad F_y = \frac{dy}{dt}$$

where , (x, y, t) is the image in pixel (x, y) at time t, where I(x, y, t) is the intensity at pixel (x, y) and time t, F_x, F_y is the horizontal and vertical motions in pixel (x, y).

We can obtain the optical flow vector field F = (F_x, F_y) by minimizing the objective function:

$$R = \int_D \left[\lambda^2 \|\nabla F\|^2 + (\nabla I \cdot F + I_t)^2\right] dxdy \quad (12)$$

The optical flow vector field F is then split into two scalar fields F_x and F_y which are further half-wave rectified into four none-gative channels F_x^+, F_x^-, F_y^+, F_y^-, so that, F_x= F_x^+-F_x^- and F_x= F_x^+-F_x^-. These four nonnegative channels are then blurred with a Gaussian kernel and normalized to obtain the final four channels Fb_x^+, Fb_x^-, Fb_y^+, Fb_y^-.

The motion descriptors of two different frames are compared by using a version of the normalized correlation. Suppose u_i is the frame i of sequence u and its four channels are u_1, u_2, u_3 and u_4, similarly, v_j is the frame j of sequence v and its four channels are v_1, v_2, v_3 and v_4, then the similarity between frame u_i and frame v_j is as follows:

$$S(u_i, v_j) = \sum_{t\in Z}\sum_{c=1}^{4}\sum_{x,y\in N} u_c^{i+t}(x, y)v_c^{j+t}(x, y) \quad (13)$$

where, Z and N are the temporal and spatial extent of the motion descriptors. T is set as 10 in all of our experiments. The dimensionality of the feature vector S is 4×Z×N.

To improve the accuracy of human action recognition, we combine spatial and temporal points of interest vector to the optical flow vector and form a hybrid feature vector ζ = [η, S].

HUMAN ACTION RECOGNITION BASED ON MULTIPLE INSTANCE LEARNING

AnyBoost algorithm analyzing: The general class of algorithms that named AnyBoost consists of gradient descent algorithms for choosing linear combinations of elements of an inner product space so as to minimize some cost functional. Each component of the linear combination is chosen to maximize a certain inner product. This inner product corresponds to the weighted training error of the base classifier.

Here, we give a brief review of AnyBoost algorithm for classification.

Input: X = {X_1, X_2,…, X_n}, with X_i = (x_i, y_i) as training set. M, the maximum number of classifiers

Output: H(x), a classifier suited for the training set

- Initialize the weights w_i = 1/n, i∈{1,…,n}
- For m = 1 to M
  - Fit a classifier $H_m(x)$ to the training data using weights w_i
  - Get combined classifier H^t_t from H_t, H_{t-1},…, $H_{max(t-r, 1)}$
  - Let:

$$\varepsilon_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq H_m(x_i))}{\sum_{i=1}^{n} w_i} \quad (14)$$

  - Compute:

$$\alpha_m = 0.5\log(\frac{1-\varepsilon_m}{\varepsilon_m}) \quad (15)$$

- Set w_i←$w_i$exp(-α_mI(yi≠H^t_t(x_i))) and renormalize to $\sum_i w_i = 1$

Output:

$$H(x) = \text{sign}(\sum_{m=1}^{M} \alpha_m H_m(x)) \quad (16)$$

MILBoost algorithm analyzing: Multiple Instance Learning (MIL) algorithm is a variant of supervised learning algorithm. The key idea of MIL is to provide a different way in constituting training samples which is instead of using singleton training samples. Samples are organized into positive and negative bags of instances which may contain many instances in each bag. One instance is positive (i.e., object) at least in a positive bag while all instances are negative (i.e., non-object) in an negative bag. To obtain positive training samples, we know that objects are in images, but the exact locations are unknown. Therefore, it is suitable to represent the object by using a bag of multiple instances (non-aligned human images). MIL can learn which instances in the positive bags are positive, along with a binary classifier (Pang *et al*., 2008; Babenko *et al*., 2008). In this study, MIL is employed for human action recognition with non-aligned training samples. We combined MIL with AnyBoost and proposed MILBoost for learning a classifier with nonaligned training samples, so that the recognition efficiency can be increased without compromising accuracy. In MIL, samples come into positive and negative bags of instances. Each instance x_{ij} is indexed with two indices: i for the bag and j for the instance within the bag. All instances in a bag share a bag label y_i. In MILBoost, the probability of x_{ij} being positive is estimated by the logistic function:

$$P_{ij} = \frac{1}{1+\exp(-H(x_{ij}))} \tag{17}$$

Given P_{ij}, the probability of bag being positive is approximated by the Noise-OR model:

$$P_i = 1-\Pi_{j\in i}(1-p_{ij}) \tag{18}$$

Under this model, the cost function is defined as the negative log likelihood:

$$C(H) = -\sum_{i}^{n}(1_{(y_i=1)}\ln p_i + 1_{(y_i=0)}\ln(1-(p_i)) \tag{19}$$

where, 1(z) is the indicator function that equals 1 when is true and 0 otherwise. According to AnyBoost, the weight of each instance is set as the negative derivative of the cost function with respect to the score of each instance:

$$w_{ij} = -\frac{\partial C}{\partial H_{ij}} = \begin{cases} \frac{p_{ij}(1-p_i)}{p_i} & \text{if} \quad y_i = 1 \\ -p_{ij} & \text{if} \quad y_i = 0 \end{cases} \tag{20}$$

Note that the weights w_{ij} are signed and the sign interprets the label of the instance x_{ij}. A positive instance x_{ij} is assigned with a high weight if it has a high p_{ij} (i.e., close to the target) or low p_i (i.e., far away from the target). High p_{ij} depicts that x_{ij} is likely to be a true positive. Low p_i indicates that the bag does not have a good prediction yet and so the algorithm gives high weights to all instances in the bag. As for negative instances, if p_{ij} is predicted incorrectly (i.e., p_{ij} approaches), a high negative weight is assigned. In selecting the weak learner, MILBoost will pay much attention to important instances that have high absolute weights, $|w_{ij}|$. The h_t is selected to satisfy $h_t(x_{ij}) = w_{ij}$ for all i, j. So, as to most reduce the cost over training examples, h_t can be rewritten as:

$$h_t = \arg\max \sum_{i,j} w_{ij} h(x_{ij}) \tag{21}$$

Human action recognition: Keeler *et al*. (1990) proposed originally the idea for the Multiple Instance Learning for handwritten digit recognition in 1990. It was called Integrated Segmentation and Recognition (ISR) and it is the key idea to provide a different way in constituting training samples. Training samples are not singletons, at the same time they are in "bags", where all of the samples in a bag share a label (Dietterich *et al*., 1997). Samples are organized into positive bags of instances and negative bags of instances which each bag may contain a number of instances (Marson and Lozano-Perez, 1998). At least one instance is positive (i.e., object) in a positive bag while all instances are negative (i.e., non-object) in a negative bag. In MILBoost, learning must simultaneously learn which samples in the positive bags are positive along with the parameters of the classifier. MILBoost can learn which instances in the positive bags are positive, along with a binary classifier. In this study, MILBoost is employed for human action recognition with non-aligned training samples. In human actions recognition, the sample is represented by a hybrid feature vector and as input of classifier. Assuming the observed data be independent of each other, the MILBoost-based human action recognition proceeds as follows:

Input: Given dataset $\{X_i, y_i\}^n_{i=1}$, n is the number of all weak classifiers, $X_i = \{x_{i1}, x_{i2},\ldots, x_{ij},\ldots, x_{in},\}$ is training bags and $y_i \in \{0,1\}$ is the score of the sample

Pick out K weak classifiers and consist of strong classifier.

Update all weak classifiers with data $\{x_{ij}, y_i\}$. Initialize all strong classifier: $H_{ij} = 0$ for all i, j.

for k = 1 to K do

for m = 1 to n do

We calculate the probability that the j-th sample is positive in the i-th bag as follow:

$$P^m_{ij} = \sigma(H_{ij}+h_m(x_{ij})) \tag{22}$$

Where:

$$P_{ij}^{m} = p(y_i|x_{ij}) = \frac{1}{1+\exp(-y_{ij})}$$

We calculate the probability that the bag is positive as follow:

$$P^{m}_{i} = 1-\Pi_j(1-p^{m}_{ij}) \quad (23)$$

where, $p^{m}_{i} = p(y_i|X_i)$.

The likelihood assigned to a set of training bags is:

$$C^{m} = \sum_i (y_i \log(p_i^m) + (1 - y_i)\log(1 - (p_i^m)) \quad (24)$$

End for

Finding the maximum m* from n as the current optimal weak classifier as follow:

$$m^* = \arg\min_m C^m \quad (25)$$

The m* come into the strong classifier:

$$h_k(x) \leftarrow h_{m^*}(x) \quad (26)$$

$$H_{ij} = H_{ij} + h_k(x) \quad (27)$$

End for

Output: Strong classifier which consist of K weak classifiers as follow:

$$H(x) = \sum_k h_k(x) \quad (28)$$

where, h_k is a weak classifier and can make binary predictions using $\text{sign}(H_k(x))$

EXPERIMENTAL RESULTS AND ANALYSIS

The effectiveness of the proposed algorithm was verified by using C++ and Matlab hybrid implementation on a PC with Pentium 3.2 GHz processor and 3G RAM. We test our algorithm using KTH human motion dataset and Weizmann human action data set.

KTH data set: KTH data set is the largest available video sequence dataset of human actions, whose sample images are shown in Fig. 2.

In this database, there are six groups of images (e.g., "walking", "jogging", "running", "boxing", "hand waving" and "hand clapping") by 25 subjects in different scenarios of outdoor and indoor environment with scale change.

We run an automatic preprocessing step to track and stabilize the video sequences, so that all the figures appear in the center of the field of view. In this experiment, we study recognition accuracy of six kinds of human actions from KTH human motion dataset. On the KTH data set, the confusion matrix for human action recognition based on MILBoost method is shown in Fig. 3, where action 1-6 indicate "walking", "jogging", "running", "boxing", "hand waving" and "hand clapping", respectively. We can see that the algorithm can correctly recognize most human actions. Most of the mistakes are confusions between "running" actions and "jogging" actions, since "running" and "jogging" actions are similar actions with each other.

Weizmann data set: The Weizmann human action data set contains 83 video sequences which show nine different people and each perform nine different actions (e.g., "running", "walking", "jumping-jack",

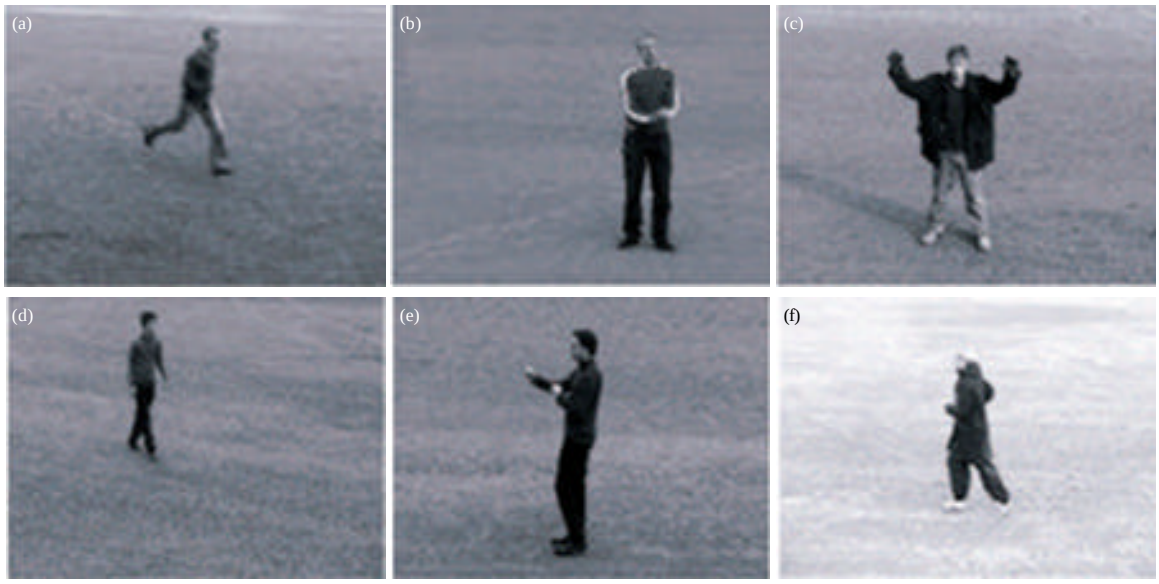

Fig. 2(a-f): Key frames for KTH human action data set

"jumping-forward-on-two-legs", "jumping-in-place-on-two-legs", "galloping-sideways", "waving-two-hands","waving-one-hand" and "bending"). We track and stabilize the figures by preprocessing images which come with this data set and leave the videos of one person as test data each time. Some sample frames are shown in Fig. 4.

In this experiment, we study recognition accuracy of nine kinds of human actions from Weizmann human motion data set. Recognition results using MILBoost method are presented in the confusion matrices as shown in Fig. 5. Where action 1-9 indicate "running", "walking", "jumping-jack", "jumping-forward-on-two-legs", "jumping-in-place-on-two-legs", "galloping-sideways", "waving-two-hands", "waving-one-hand" and "bending" respectively. Each cell in the confusion matrix is the average result of every human action, respectively.

Action							
1	0.97	0.01	0.02	0	0	0
2	0.01	0.89	0.10	0	0	0
3	0.02	0.11	0.87	0	0	0
4	0	0	0	0.00	0	0
5	0	0	0	0	0.99	0.01
6	0	0	0	0	0	0.98
	1	2	3	4	5	6

Action

Fig. 3: Confusion matrix for human action recognition on the KTH data set

As Fig. 5 shows the algorithm correctly classifies most actions. Most of the mistakes that the algorithm make are confusions between "jumping-forward-on-two-legs" and "jumping-in-place-on-two-legs" actions, are confusions between "running" and "jumping-jack" actions. These are intuitively reasonable since "jumping-forward-on-two-legs" and "jumping-in-place-on-two-legs" are similar actions with each other, "running" and "jumping-jack" are similar actions with each other.

To examine the accuracy of our proposed human action recognition approach, we compare our method to two state-of-the-art approaches for human action recognition using the same data and the same experimental settings. The first method is SVM (Support Vector Machine) (Schuldt *et al*., 2004). The second method is LDA (Flaherty *et al*., 2005). The 200 different human action images are used for this experiment. The average recognition accuracy was observed which is displayed in Fig. 6.

We can see that our method improves the recognition accuracies. It achieves 95.5% average recognition rate, whereas "SVM" obtain a result of 83.2%, "LDA" gets a result of 87.4%. The reason is that we improve the recognition accuracy in the two stages of human action feature extraction and human action recognition. In the stage of human action feature extraction, we use a hybrid feature vector which combines space-time interest points

Fig. 4(a-j): Key frames for Weizmann human action data set

Action	1	2	3	4	5	6	7	8	9
1	0.92	0.02	0.05	0.01	0	0	0	0	0
2	0.02	0.98	0	0	0	0	0	0	0
3	0.04	0.01	0.93	0.01	0.01	0	0	0	0
4	0.01	0	0.01	0.95	0.03	0	0	0	0
5	0.04	0	0.01	0.01	0.94	0	0	0	0
6	0	0	0	0	0.02	0.96	0	0	0
7	0	0	0	0	0	0	0.97	0.02	0
8	0	0	0	0	0	0	0.01	0.99	0
9	0	0	0	0	0	0	0	0	1.00

Action

Fig. 5: Confusion matrix for human action recognition on the Weizmann data set

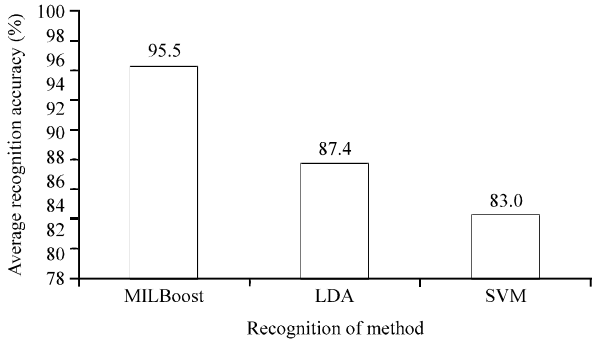

Fig. 6: Comparison of recognition accuracy for three methods

features and motion features that are reliably with noisy image sequences and describe human action effectively. In the stage of human action recognition, we use MILBoost algorithm to classify human action images. Our method performs significantly better.

CONCLUSION

Human action recognition can provide significant advantage in video surveillance. In this study, we present a novel method to recognize human action in video sequences. The main contribution can be concluded as follows.

In feature extraction and representation, we extracted space-time interest points for feature using HOG and T-HOG method and Optical flow model is used for extracting motion features. Then we combine them and form a hybrid feature vector.

In action modeling and recognition, we combined MIL with AnyBoost and proposed MILBoost for human action recognition, so that the recognition efficiency can be increased without compromising accuracy.

Experiments were performed on KTH human motion data set, Weizmann human action data set and evaluate the proposed method. Experimental results reveal that our proposed method performs better than previous ones. Our algorithm can also recognize multiple actions in complex motion sequences containing multiple actions.

ACKNOWLEDGMENTS

This study was supported by Research Foundation for Science and Technology Office of Hunan Province under Grant (No. 2014FJ3057), by Hunan Provincial Education Science and "Twelve Five" planning issues (No. XJK012CGD022), by the Teaching Reform Foundation of Hunan Province Ordinary College under Grant (No. 2012401544) and by the Foundation for Key Constructive Discipline of Hunan Province.

REFERENCES

Agarwal, A. and B. Triggs, 2004. Learning to track 3d human motion from silhouettes. Proceedings of the 21st International Conference on Machine Learning, July 2004, Banff, Canada, pp: 9-16.

Amien, M.B.M. and J. Lin, 2007. Dual-mode continuous arrhythmias telemonitoring system. J. Applied Sci., 7: 965-971.

Babenko, B., P. Dollar, Z. Tu and S. Belongie, 2008. Simultaneous learning and alignment: Multi-instance and multi-pose learning. Proceedings of the Workshop on Faces in Real-Life Images: Detection, Alignment and Recognition, October 17, 2008, ECCV.

Blank, M., L. Gorelick, E. Shechtman, M. Irani and R. Basri, 2005. Actions as space-time shapes. Proceedings of the 10th International Conference on Computer Vision, Volume 2, October 17-21, 2005, Beijing, pp: 1395-1402.

Bobick, A.F. and A.D. Wilson, 1997. A state-based approach to the representation and recognition of gesture. Trans. Pattern Anal. Mach. Intell., 12: 1325-1337.

Bobick, A.F. and J.W. Davis, 2001. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intel., 23: 257-267.

Cheung, V., B.J. Frey and N. Jojic, 2005. Video epitomes. Proceedings of the International Coference on Computer Vision and Pattern Recognition, Volume 1, June 20-26, 2005, San Diego, California, pp: 42-49.

Dietterich, T.G., R.H. Lathrop and T. Lozano-Perez, 1997. Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell., 89: 31-71.

Dollar, P., V. Rabaud, G. Cottrell and S. Belongie, 2005. Behavior recognition via sparse spatio-temporal features. Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, October 15-16, 2005, Beijing, China, pp: 65-72.

Efros, A.A., A.C. Berg, G. Mori and J. Malik, 2003. Recognizing action at a distance. Proceedings of the 9th International Conference on Computer Vision, Volume 2, October 13-16, 2003, Nice, France, pp: 726-733.

Flaherty, P., G. Giaever, J. Kumm, M.I. Jordan and A.P. Arkin, 2005. A latent variable model for chemogenomic profiling. Bioinformatics, 15: 3286-3293.

Ke, Y., R. Sukthankar and M. Hebert, 2005. Efficient visual event detection using volumetric features. Proceedings 10th International Conference on Computer Vision, Volume 1, October 17-21, 2005, Canada, pp: 166-173.

Keeler, J.D., D.E. Rumelhart and W.K. Leow, 1990. Integrated segmentation and recognition of hand-printed numerals. Proceedings of the IEEE Conference on Neural Information Processing Systems, November 26-29, 1990, Denver, Colorado, USA., pp: 557-563.

Laptev, I. and T. Lindeberg, 2003. Space-time interest points. Proceedings of the 9th International Conference on Computer Vision, Volume I, October 13-16, 2003, Nice, France, pp: 432-439.

Little, J. and J.E. Boyd, 1998. Recognizing people by their gait: The shape of motion. Videre: J. Comput. Vision Res., 2: 1-32.

Lucas, B.D. and T. Kanade, 1981. An iterative image registration technique with an application to stereo vision. Proceedings of the 7th International Joint Conference on Artificial Intelligence, Volume 2, Vancouver, BC, Canada, August 24-28, 1981, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA., pp: 674-679.

Marson, O. and T. Lozano-Perez, 1998. A framework for multiple-instance learning. Proceedings of the IEEE Conference on Neural Information Processing Systems, November 30-December 5, 1998, Denver, Colorado, USA., pp: 570-576.

Nowozin, S., G. Bakir and K. Tsuda, 2007. Discriminative subsequence mining for action classification. Proceedings of the 11th International Conference on Computer Vision, October 14-21, 2007, Rio de Janeiro pp: 1-8.

Olivera, N., A. Garg and E. Horvitz, 2004. Layered representations for learning and inferring office activity from multiple sensory channels. Comput. Vision Image Understanding, 2: 163-180.

Pang, J., Q. Huang, S. Jiang and W. Gao, 2008. Pedestrian detection via logistic multiple instance boosting. Proceedings of the 15th International Conference on Image Processing, October, 12-15, 2008, San Diego, CA, pp: 1464-1467.

Schuldt, C., I. Laptev and B. Caputo, 2004. Recognizing human actions: A local SVM approach. Proceedings of the 17th International Conference on Pattern Recognition, Volume 3, August 23-26, 2004, Cambridge, UK., pp: 32-36.

Thi, T.H., L. Cheng, J. Zhang, L. Wang and S. Satoh, 2012. Structured learning of local features for human action classification and localization. Image Vision Comput., 30: 1-14.

Xiang, T. and S. Gong, 2006. Beyond tracking: Modelling activity and understanding behaviour. Int. J. Comput. Vision, 1: 21-51.

Yilmaz, A. and M. Shah, 2005. Recognizing human actions in videos acquired by uncalibrated moving cameras. Proceedings of the 10th International Conference on Computer Vision, Volume 1, October 17-21, 2005, Beijing, China, pp: 150-157.

Zhang, M. and A.A. Sawchuk, 2011. A feature selection-based framework for human activity recognition using wearable multimodal sensors. Proceedings of the 6th International Conference on Body Area Networks, November 7-10, 2011, Beijing, China, pp: 92-98.

Zhou, R., F. Ding and L. Lu, 2011. Development and Implementation of intelligent video surveillance alarm system. Inform. Technol. J., 10: 1295-1304.