http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (5): 974-982, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.974.982

Human Activities for Classification via Feature Points

Hao Zhang, Zhijing Liu and Haiyong Zhao
School of Computer Science and Technology, Xidian University, P.O. Box 171,
No. 2 South Taibai Road, Xi'an, (710071) China

Abstract: This study presented a new classification method for single person's motion, which is represented by Haar wavelet transform and classified by Hidden Markov Model. We tackle the challenge of detecting the feature points by Haar wavelet transform to improve classification accuracy. We extract binary silhouette and segment them by cycle after creating the background model. Then the low-level features are detected by Haar wavelet transform and principal vectors are determined by Principal Component Analysis. We utilize Hidden Markov Models to train and classify cycle sequences and demonstrate their usability. Compared with others, our approach is simple and effective in feature point detection, strength in scale-invariant and generalized in different motions. Therefore, the video surveillance based on our method is practicable in (but not limited to) many scenarios where the background is known.

Key words: Activity classification, feature points, Haar wavelet transform, Hidden Markov Model (HMM), intelligent video surveillance

INTRODUCTION

Analyzing human actions has always remained a topic of great interest in computer vision. It is seen as a stepping stone for applications such as automatic environment surveillance, assisted living and human computer interaction. The surveys (Aggrawal and Cai, 1999; Turaga *et al*., 2008; Poppe, 2010) provide a broad overview of over three hundred papers and numerous approaches for analyzing human motion in videos, including human motion capture, tracking, segmentation and recognition.

Human activity recognition has been a widely studied topic, but the solutions to this problem that have been submitted to date are very premature and still specific to the dataset at hand. Building a general and effective activity recognition and classification system is a challenging task, due to the various parameters from the environment, objects and activities. Variations in the environment can be caused by cluttered or moving background, camera motion, occlusion, weather and illumination. Variations in the objects are caused by differences in appearance, size or posture of the objects or due to self-motion which is not itself part of the activity. Variations in the activity can make it difficult to recognize semantically equivalent activities.

Our work is motivated an important application of activity recognition in intelligent video surveillance. In this study, we describe a generative method based on silhouette-based shape feature of human motion (Haar wavelet transform) and a generative model (Hidden Markov Model, HMM) in motion classification. Our aim was to offer a discriminative solution to human motion categorization via flexible yet highly descriptive HMM. Its advantage is not only helpful in utilizing internal and external information of images, but also easy to distinguish similar shape sequences. Furthermore, it rarely suffers from such factors, i.e. video alignment, noise and images segmenting. Our contribution is that we summarise the discriminative features of each motion image with Haar wavelet transform and apply them into recognition, rather than highlight many local features excessively. Meanwhile, HMM offsets the disadvantages in template models, i.e. consistency. Therefore, it is beneficial to prompt recognition accuracy and suitable for intelligent surveillance in practicability.

Although detection of spatial points has attracted the interest of many researchers, the spatiotemporal counterpart is less studied. One of the most well known space-time feature point detectors is the extension of the Harris corner detector to 3D (Laptev, 2005). A spatiotemporal corner is defined as a region containing a spatial corner whose velocity vector is changing direction. The resulting points are sparse and roughly correspond to start and stop points of a movement when applied to action classification. Dollar *et al*. (2005) identified the weakness of spatiotemporal corners to represent actions in certain domains (e.g. rodent behavior recognition and facial expressions) and propose a detector based on the response of Gabor filters applied both spatially and temporally. The detector produces a denser set of interest points and proves to be more

Corresponding Author: Hao Zhang, School of Computer Science and Technology, Xidian University, P.O. Box 171, No. 2 South Taibai Road, Xi'an, (710071) China Tel: 0086-(0)-29-88208033

representative of a wider range of actions. According to (Lowe, 2004) sparseness is desirable to an extent, but too few features can be problematic in representing actions efficiently.

Oikonomopoulos *et al*. (2006) used a different measure and propose a spatiotemporal extension of the salient point detector of Kadir and Brady. They relate the entropy of space-time regions to saliency and describe a framework to detect points of interest at their characteristic scale determined by maximizing their entropy. This detector is evaluated on a dataset of aerobic actions and promising results are reported. Wong and Cipolla (2007) reported a more thorough evaluation of the latter and propose their own detector based on global information. Their detector is evaluated against the state-of-the-art in action recognition and outperforms the ones (Laptev, 2005; Dollar *et al*., 2005; Oikonomopoulos *et al*., 2006) on standard datasets highlighting the importance of global information in space-time interest point detection. Recently and Willems *et al*. (2008) proposed a space-time detector based on the determinant of the 3D Hessian matrix, which is computationally efficient (use of integral videos) and is still on par with current methods.

Each of the aforementioned detectors is based on a different measure related to cornerness, entropy-based saliency, global texture or periodicity. Study of the published results confirms the trade-off, highlighted by Lowe (2004) between sparsity and discriminative power of the points. Inspired by methods related to visual attention and saliency modeling we study the incorporation of more features apart from intensity and the interaction of local and global visual information in a single measure. We derive some feature points consisting of motional information, which describes human actions in each video sequence. The main contribution of our work is the integral representation of motional information with Haar wavelet transform to obtain minimal feature points and then training and classification human action with HMMs. Henceforth, we refer that the training sequences are termed galleries and the testing ones are probes.

MATERIALS AND METHODS

In our work, we create background model at the beginning of each video sequence, then extract human contour (Zhang *et al*., 2010) from binary silhouette. The image with feature points is obtained after their implementation of Haar wavelet transform. We can select some principal vectors and combine them into one. These vectors in each sequence are employed to training or classification in HMM. The flowchart is shown in Fig. 1.

Haar wavelet transform: Alfred Haar proposed Haar wavelet function (Alfred, 1910) consisting of piecewise-constant function. Its defined field is [0, 1) and in certain

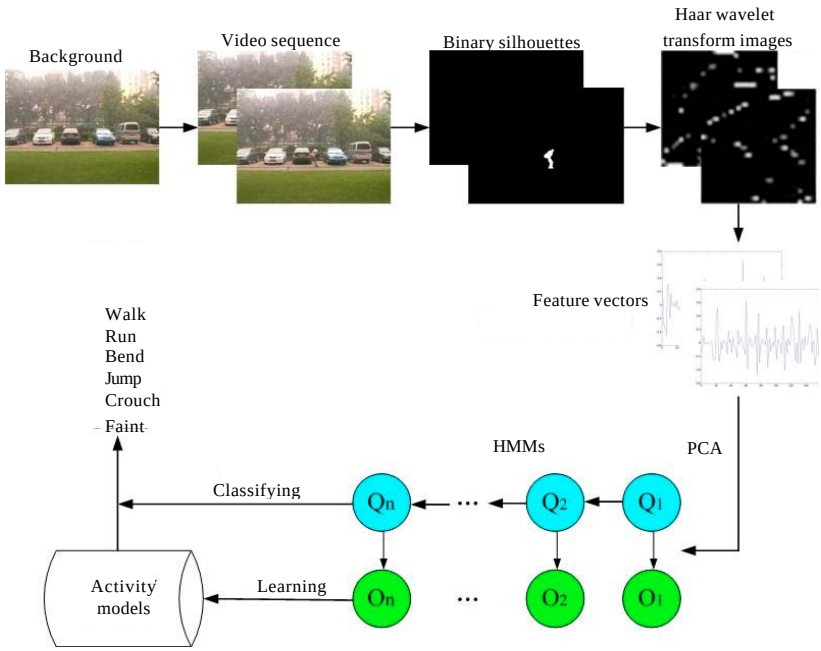

Fig. 1: The flowchart of activities classification

field every piecewise-constant function is 1 while in other field it is 0. Wavelet function is usually described as $\psi_i^j(x)$. Compared with frame function, it is called Haar wavelet function and defined as:

$$\psi(x)=\begin{cases}1 & 0\le x<\frac{1}{2}\\ -1 & \frac{1}{2}\le x<1\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

The measuring function of Haar wavelet is defined as:

$$\psi_i^j(x)=\psi(2^j x-i),\quad x=0,1,\cdots,2^j-1 \tag{2}$$

The vector space represented in W^j consisting of wavelet function is described by:

$$W^j=sp\{\psi_i^j(x)\},\quad i=0,1,\cdots,2^j-1 \tag{3}$$

where, sp represents line production. The measuring factor is j, which varies with the size of graph function. The shift parameter is i, which can bring the function to shift along x-axis.

It hypothesizes that $\{c^L_{n,m}\}$ (n, m = 0, 1, …, N-1) is the signal of 2-D and discrete image, in which $c^L_{n,m}$ his between 0 and 255. The distance between pixels is N-1, in which N is 2L and L is (j+1). It is supposed that $\tilde{h},\tilde{g}$, h and g is a filter set of biorthogonal Haar wavelet transform, whose algorithm of Mallat in convolution form is defined as:

$$\begin{aligned}c^j_{k,m}&=\sum_{l,n}\tilde{h}_{l-2k}\tilde{h}_{n-2m}c^{j+1}_{i,n}\\ d^{j,1}_{k,m}&=\sum_{l,n}\tilde{h}_{l-2k}\tilde{g}_{n-2m}c^{j+1}_{i,n}\\ d^{j,2}_{k,m}&=\sum_{l,n}\tilde{g}_{l-2k}\tilde{h}_{n-2m}c^{j+1}_{i,n}\\ d^{j,3}_{k,m}&=\sum_{l,n}\tilde{g}_{l-2k}\tilde{g}_{n-2m}c^{j+1}_{i,n}\end{aligned} \tag{4}$$

It consists of 4 parts after first rank transform of Haar wavelet, as shown below:

$$\begin{bmatrix}c^j_{k,m} & d^{j,1}_{k,m}\\ d^{j,2}_{k,m} & d^{j,3}_{k,m}\end{bmatrix}$$

where, each subimage is one quarter of mother image in size, whose structure is shown in Fig. 2. It can be seen that the regions of cA, cH, cV and cD correspond to $c^j_{k,m}$, $d^{j,1}_{k,m}$, $d^{j,2}_{k,m}$ and $d^{j,3}_{k,m}$, respectively. Each of them, which is a half in length and a quarter in area, displays the subimage after Haar wavelet.

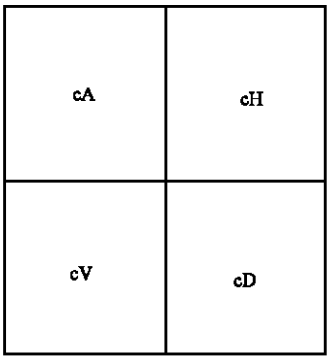

Fig. 2: The region of Haar wavelet transform

After the filter of Haar wavelet transform is replaced and simplified, the equations are as follow:

$$\begin{aligned}c^j_{k,m}&=\frac{1}{4}(c^{j+1}_{2m,2k}+c^{j+1}_{2m,2k+1}+c^{j+1}_{2m+1,2k}+c^{j+1}_{2m+1,2k+1})\\ d^{j,1}_{k,m}&=\frac{1}{4}(c^{j+1}_{2m,2k}+c^{j+1}_{2m,2k+1}-c^{j+1}_{2m+1,2k}-c^{j+1}_{2m+1,2k+1})\\ d^{j,2}_{k,m}&=\frac{1}{4}(c^{j+1}_{2m,2k}-c^{j+1}_{2m,2k+1}+c^{j+1}_{2m+1,2k}-c^{j+1}_{2m+1,2k+1})\\ d^{j,3}_{k,m}&=\frac{1}{4}(c^{j+1}_{2m,2k}-c^{j+1}_{2m,2k+1}-c^{j+1}_{2m+1,2k}+c^{j+1}_{2m+1,2k+1})\end{aligned} \tag{5}$$

Feature extraction: We adopt action videos of Dataset A in CASIA database as gallery sequences. Gaussian Mixture Modeling (GMM) is used as background modeling and motive object detecting. Binary images of human silhouette are extracted by background subtraction. The connective contour within bounding rectangle is obtained by pre-process of binary images.

After obtaining the contour, we change its size to l×l for Haar wavelet transform. Then we implement Haar wavelet transform as well as Zhang and Liu (2009) and obtain 4 subimages, as shown in Fig. 3a and b. It concludes that the subimage in cH region:

$$\frac{1}{2}\times\frac{1}{2}$$

highlights motional information, so they are thought as feature points. Principal Component Analysis (PCA) is employed to select vectors, before they are formed another vector.

We assume that C is one subimage and defined by:

$$C_{\frac{1}{2}\times\frac{1}{2}}=\begin{pmatrix}c_{11} & c_{12} & \cdots & c_{1\times\frac{1}{2}}\\ c_{21} & c_{22} & \cdots & c_{2\times\frac{1}{2}}\\ \vdots & \vdots & \ddots & \vdots\\ c_{\frac{1}{2}\times 1} & c_{\frac{1}{2}\times 2} & \cdots & c_{\frac{1}{2}\times\frac{1}{2}}\end{pmatrix}=(C'_1\quad C'_2\quad\cdots\quad C'_{\frac{1}{2}}) \tag{6}$$

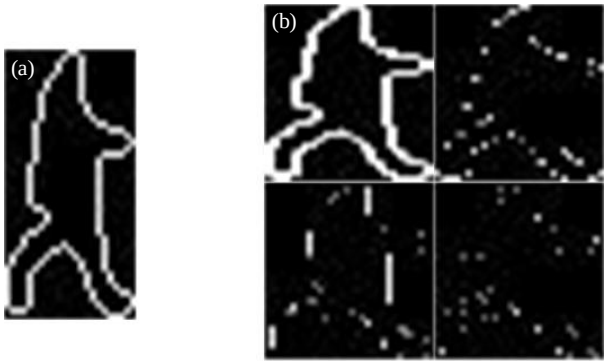

Fig. 3: (a) Source image before Haar wavelet transform and (b) Four subimages after Haar wavelet transform

We reduce the dimensionalities of C with PCA and obtain the principal components as:

$$C'' = (C_1'' \quad C_2'' \quad \cdots \quad C_{\frac{1}{2}}'') \tag{7}$$

So we select m components and form a new vector:

$$D = (C_1'' \quad C_2'' \quad \cdots \quad C_m'')^T, 1 < m < \frac{1}{2}$$

to train and recognize in HMMs.

Activity training and classification: Hidden Markov Models (HMMs) are natural tools for representing the time-varying correlations in stochastic processes. HMMs have been successfully used in the areas of speech recognition (Rabiner, 1989) motion recognition (Yamato *et al*., 1992) gesture recognition (Wilson and Bobick, 1999; Hyeon-Kyu and Kim, 1999) and general image analysis (Aas *et al*., 1999). In all cases, suitably constructed models have shown great promise with the recognition of time-varying signals and patterns using appropriately defined observation sequences. In this paper, we extend the application of HMMs to the analysis of motion classification.

Given a sequence of image features for activity j, $X^j = \{x^j(1), x^j(2), ..., x^j(T)\}$, we wish to build a model for the activity and use it to recognize them from the different in the database.

A reasonable way to build a structural representation for a person is to pick N exemplars (or stances) $E = \{e_1, e_2, ..., e_N\}$ from the pool of images that will minimize the error in representation of all the images of that activity. The specifics of choice of exemplars may differ for different approaches. Given the image sequence for an unknown activity $Y = \{y(1), y(2), ..., y(T)\}$ these exemplars can be directly used for recognition as:

$$S = \arg\min_j \sum_{i=1}^{T} \min_{n \in \{1,...,N\}} \{d(y(t), e_n^j)\}$$

where, y(t) represents the image of an unknown activity at the tth time instant, while e_n^j represents the nth exemplar of the jth activity.

Because the transitions are systematic, it is possible to model this probabilistic dependence by a Markov matrix, as follows:

$$A = [P(e_i(t) \mid e_j(t-1))] \tag{8}$$

for $i, j \in \{1, 2, ..., N\}$. The matrix A encodes the dynamics in terms of state duration densities and transition probabilities.

Activity representation: In this approach, we use the feature vector in its entirety to estimate the HMM $\lambda = (A, B, \pi)$ for each activity. One of the important issues in training is learning the observation probability B. In general, if the underlying distribution of the data is non-Gaussian, it can be characterized by a mixture of Gaussians. The reliability of the estimated B depends on the number of training samples available and the dimension of the feature vector. In order to deal with the high dimensionality of the feature vector, we propose an alternative representation for B.

During an activity cycle, it is possible to identify certain distinct phases or stances. We build a structural representation for a activity by picking N exemplars (or stances) from the training sequence, $X = \{x(1), x(2), ..., x(T)\}$. We now define B in terms of the distance of this vector from the exemplars as follows:

$$b_n = (x(t)) = P(x(t) \mid e_n) = \beta e^{-\alpha D(x(t), e_n)} \tag{9}$$

The probability $P(x(t) \mid e_n)$ is defined as a function of $D(x(t), e_n)$, the distance of the feature vector x (t) from the

nth exemplar, e_n. The motivation behind using an exemplar-based model in the above manner is that the recognition can be based on the distance measure between the observed feature vector and the exemplars. During the training phase, a model is built for all the subjects in the gallery. Note that B is completely defined by E if α and β are fixed beforehand. An initial estimates of E and λ is formed from X and these estimates are refined iteratively using expectation maximization (Dempster *et al*., 1977). We can iteratively refine the estimates of A and π by using the Baum-Welch algorithm (Rabiner, 1989) with E fixed. The algorithm to refine estimates of, while keeping A and π fixed, is determined by the choice of the distance metric. We describe in the following sections the methods used to obtain initial estimates of the HMM parameters, the training algorithm and, finally, identification from a probe sequence.

Initial estimate of HMM parameters: In order to obtain a good estimate of the exemplars and the transition matrix, we first obtain an initial estimate of an ordered set of exemplars from the sequence and the transition matrix and then iteratively refine the estimate. We observe that the gait sequence is quasiperiodic and we use this fact to obtain the initial estimate $E^{(0)}$. We first divide the sequence into cycles. We can further divide each cycle into N temporally adjacent clusters of approximately equal size. We visualize the frames of the nth cluster of all cycles to be generated from the nth state. Therefore, we can get an initial estimate of e_n from the feature vectors belonging to the nth cluster of all cycles. In order to get reliable initial estimates of the exemplars, we need to robustly estimate the cycle boundaries (Sundaresan *et al*., 2003). A corresponding initial estimate of the transition matrix is also obtained. The initial probabilities are set to be equal to 1/N.

Training the HMM parameters: The iterative refinement of the estimates is performed in two steps. In the first step, a Viterbi evaluation (Rabiner, 1989) of the sequence is performed using the current values for the exemplars and the transition matrix. We can, thus, cluster feature vectors according to the most likely state they originated from. Using the current values of the exemplars $E^{(i)}$ and the transition matrix $A^{(i)}$, Viterbi decoding on the sequence X yields the most probable path $Q = \{q^{(i)}(1), q^{(i)}(2), ..., q^{(i)}(T)\}$ where, $q^{(i)}(t)$ is the estimated state at time and iteration. Thus, the set of observation indices, whose corresponding observation is estimated to have been generated from state n is given by $T_n^{(i)} = \{t : q^{(i)}(t) = n\}$. We now have a set of frames for each state and we would like to select the exemplars so as to maximize the probability in Eq. 10. If we use the definition of Eq. 9 in Eq. 11 follows:

$$e_n^{(i+1)} = \arg\max_{e} \prod_{t \in T_n^{(i)}} P(x(t) \mid e) \tag{10}$$

$$e_n^{(i+1)} = \arg\max_{e} \prod_{t \in T_n^{(i)}} D(x(t), e) \tag{11}$$

The actual method for minimizing the distance in Eq. 11, however, depends on the distance metric used. We use the inner product (IP) distance Eq. 12. We have experimented with other distance measures, namely the Euclidean (EUCLID) distance and the sum of absolute difference (SAD) distance (Sundaresan *et al*., 2003):

$$D_{IP}(x,e) = 1 - \frac{x^T e}{\sqrt{x^T x e^T e}} \tag{12}$$

Note that x though and are 2-D images, they are represented as vectors of dimension D×1 for ease of notation. $1_{D\times 1}$ is a vector of D ones. The equation for updating the exemplars is given by Eq. 13. $\tilde{x}$ denotes the normalized vector:

$$e_n^{(i+1)} = \sum_{t \in T_n^{(i)}} \tilde{x}(t) \tag{13}$$

Given $E^{(i+1)}$ and $A^{(i)}$, we can calculate $A^{(i+1)}$ using the Baum-Welch algorithm (Rabiner, 1989). We set $\pi_n^{(i+1)} = 1/N$ at each iteration. Thus, we can iteratively refine our estimates of the HMM parameters. It usually takes only a few iterations to obtain an acceptable estimate.

Identifying from a test sequence: Given the sequence of the unknown activity Y and the exemplars and HMM model parameters for the different activities in the database, we wish to recognize the unknown activity. As before, the given image sequence of the unknown activity is subjected to the same image processing operations as the training image sequence to extract the relevant image features. As explained before, the likelihood that the observation sequence was produced by the jth activity in the database is computed using the forward algorithm as:

$$P_j = \log (P (Y \mid \lambda_j)) \tag{14}$$

Note that λ_j implicitly includes the exemplar set corresponding to activity j.

RESULTS AND DISCUSSION

Activity dataset and data selection: The videos used are taken from activity database shot by Institute of Automation, Chinese Academy of Sciences (CASIA). In this database, there are two sets of activities, including single person and two interactive persons. Each of them

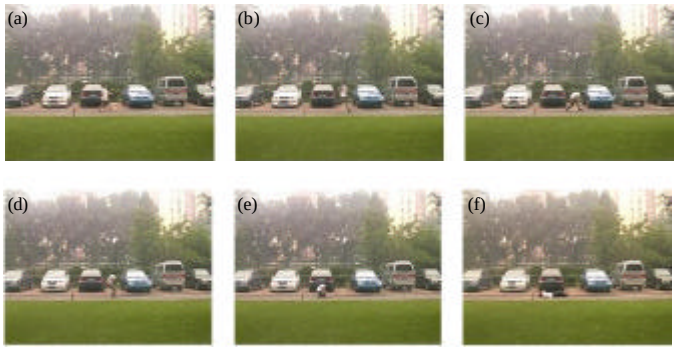

Fig. 4: Samples of activities in CASIA dataset (a) Walking; (b) Running; (c) Bending; (d) Jumping; (e) Crouching and (f) Fainting

Table 1: The number of cycles in six motions

Type of motion	Walking	Running	Bending	Jumping	Crouching	Fainting
Total cycle	125	85	126	32	8	8
Gallery cycle	63	38	56	15	4	4
Probe cycle	62	47	70	17	4	4
Frame number per cycle	12	9	10	18	50	50

is screened in 3 visual angles, i.e. horizontal, vertical and angle. The dataset of single person contains 8 categories per angle, including 11 or 16 flips (320×240, 25fps), respectively. All flips are shot outdoors by stationary camera, including 16 persons in 8 types of activities. In our experiment, the video flips of single person in horizontal view, i.e. walk, run, bend, jump, crouch and faint, are used, as shown in Fig. 4a-f.

In experiments, the number of cycles and average frames per cycle in 6 types of activities is shown in Table 1. Each category is divided into 2 parts, i.e. gallery and probe frame. Note that we select 50 frames with equal interval according to different length of sequences, i.e. crouch and faint, because of their length and only one cycle in each sequence. From this table, we can see that the number of galleries and probes is almost equal, so that we obtain the height in the recognition accuracy compared with other methods. Actually, the number of frames in different motions is unequal because of its length in different motion sequences, i.e. walking 12 frames, running 9, bending 10, jumping 18, crouching 50 and fainting 50.

In our approach, it is very important to denote the scale of Haar wavelet l and the number of valid vector m in the feature extraction. Because of Haar wavelet transform, the size of image is transformed into 2n in length and that is l as 2n. In this experiment, we denote n as 4, 5, 6, 7 and 8, in other word, l as 16, 32, 64, 128 and 256. Figure 5 shows that a walking image and its transformations in Haar wavelet transforms vary with the value of l. Figure 5a and b give the binary image and contour, respectively. From Fig. 5c to g, we can conclude that the key points play an important role in the feature description. That is, more points are useful in the feature representation. The experimental results indicated that it not only represented the feature, but also maintained low computational complexity, when l was 64.

Our experimental result indicated that the eigenvalues after Haar transform are variant in scale. After analyzing, we define the valid vector whose eigenvalue is more than 10^{-4}. Therefore, Fig. 6 indicates the number of valid vectors in different values of l. We can conclude that the number of valid vectors is proportional to l after it is equal to 64. Therefore, we denote l as 64 in our experiment. In Fig. 7, it shows the images of 6 activities after Haar transform in Eq. 5.

To balance the computational complexity and recognition accuracy, we perform our experiments to obtain optimal value of m according to the number of gallery cycle in Table 1. The result is shown in Fig. 8. We can conclude that the recognition accuracy improves as m is from 3 to 8. When m is 5, the recognition accuracy of crouching and fainting is 100% and the other activities are more than 85%. Moreover, when m is larger than 5, i.e. 6, 7 and 8, the recognition accuracy improves less apparently, while crouching and fainting are still 100%. So we denote m as 5 in our experiment.

After deciding l and m, we finish the feature extraction with Eq. 6 and 7. The galleries are employed in activity representation, initial and training parameters with Eq. 9 and 13 in HMMs. Then the probes are used to identify for classification with Eq. 14 and the results are shown in Fig. 9. At first, we conclude that crouching and fainting can distinguish from the others easily because their shapes are more discriminative. Additionally, bending and jumping obtain more scores. Finally, as the integrity of walking and running is similar, especially fast walking and slow running, both of them are more confusable.

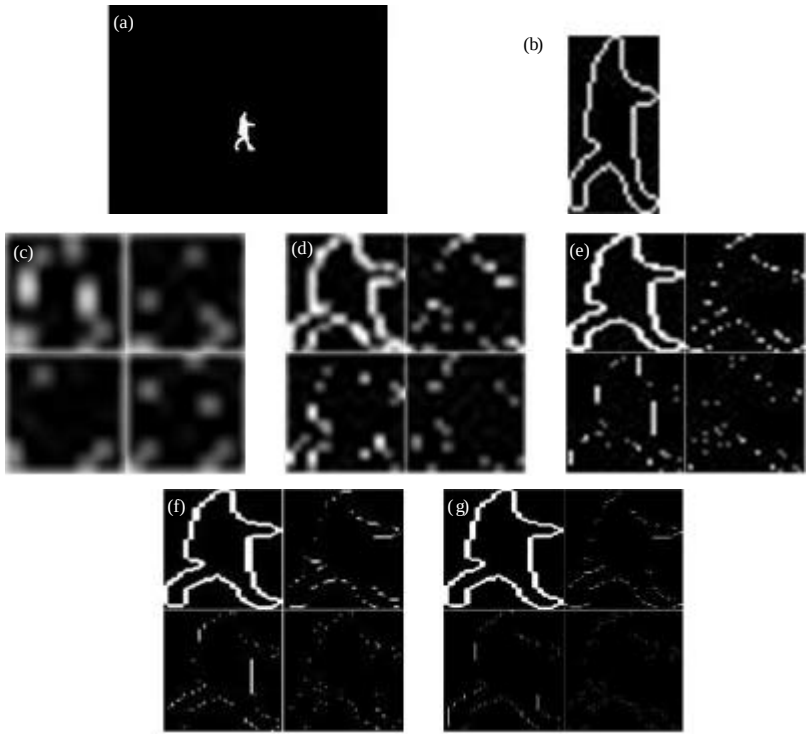

Fig. 5: Images after Haar wavelet transform vary with l; (a) Binary image; (b) Contour image; (c) l = 16; (d) l = 32; (e) l = 64; (f) l = 128 and (g) l = 256

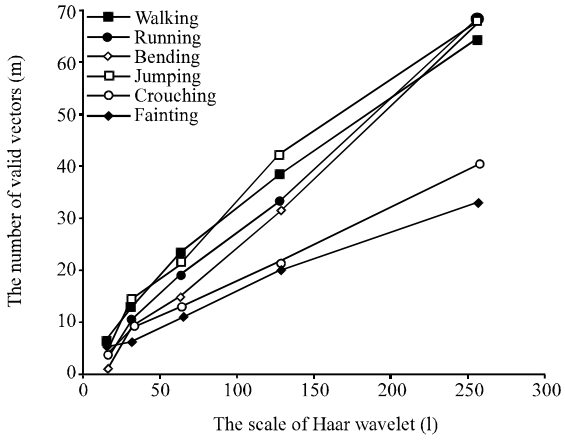

Fig. 6: The number of valid vectors

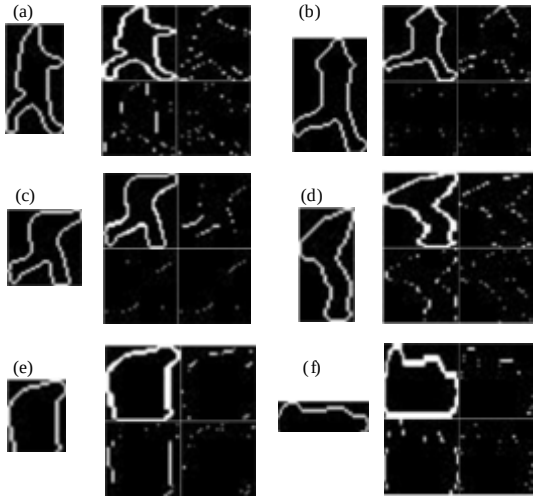

Fig. 7: Samples of six motions after Haar wavelet transform as l is 64; (a) Walking; (b) Running; (c) Bending; (d) Jumping; (e) Crouching and (f) Fainting

As our motivation focuses on application in video surveillance, it generally outperforms the state-of-the-arts in representation and recognition accuracy, while the experimental results are obtained from similar datasets. Though the interest points are always detected before action recognition, there are several distinct differences between ours and state-of-the-arts. Comparison with the state-of-the-arts, it automatically detects the keypoints of actions with Haar wavelet transform rather than builds the bag of words or dictionary or visual vocabulary (Dollar *et al*., 2005; Messing *et al*., 2009; Niebles *et al*.,

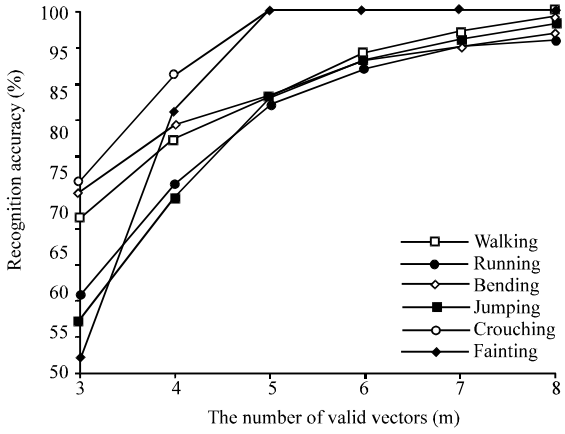

Fig. 8: The curve of classification accuracy with different value of m

	Walk	Run	Bend	Jump	Crouch	Faint
Walk	88.7	8.5	1.4	5.8		
Run	8.1	87.2	1.4	5.8		
Bend			90			
Jump	3.2	4.2		88.2		
Crouch			7.1		100	
Faint						100

Fig. 9: The confusion matrix of six motions

2008; Willems *et al*., 2008; Laptev *et al*., 2008) all of which are not similar to ours in feature detection naturally, before action recognition. Therefore, its limitation lies in supervision and camera motion Niebles *et al*. (2008) but it largely reduces computational complexity and is suitable for application in video surveillance. We utilize Haar wavelet and HMMs instead of probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA) (Niebles *et al*., 2008) to solve the overfitting issues in pLSA. Additionally, there are fewer parameters to be given in our method than Niebles *et al*. (2008) and smarter in temporal scale invariance. In Laptev *et al*. (2008) they employed histogram of oriented gradient (HoG) and optical flow (HoF) to extract feature points in order to build spatio-temporal grid, anther type of bag of words. This leads to a fusion problem of different features, which is unnecessary for application in our approach. Dollar *et al*. (2005) also build cuboids to form a dictionary and each of them is assigned to a type of action. It is not beneficial to obtain the recognition accuracy efficiently, so it is weaker than ours. In (Messing *et al*., 2009) the velocity history of tracked keypoints and generative mixture model are presented for activity recognition, which is quite different to our approach in feature detection. It requires substantive videos in recognizing activities, so it rarely meets the real-time in video surveillance.

CONCLUSIONS

We have presented a novel method of spatiotemporal feature point, which is based on Haar wavelet transform and HMM. We employ the contour images with Haar wavelet transform which are extracted from video sequences and obtain the combined vectors from each image. Moreover, we obtain and denote the optimal value of the parameters. Then these vectors are divided into two parts for training and classification in HMMs. In this phase, we perform activity representation, initial estimate HMM parameters, training the HMM parameters and identifying from a test sequence. The performance achieves more than 92% in comparison with the state-of-the-arts. Our motion descriptor has several advantages as follow. First, it is simple and effective, since we extract activity feature points from human contour in the whole image and form a temporal model. Further, it has more strength in scale-invariant. It keeps the scale-invariant with Haar wavelet. Finally, our method is general for different motion recognition in surveillance. As our experiments show, the method is robust to significant changes in scale. We show that, by effective classification of such model, reliable human activity recognition is possible. We demonstrate the effectiveness of our method over the state-of-the-art dataset from CASIA in motion recognition literature. Our results are intuitively comparable and even superior to the reported results. Based on our experiments, the proposed approach can be used in many applications with appreciable performance.

Because of simple activities in our study, the details of tracking problem are avoided to some extent. As the activity becomes more complex and the number of subjects increases, we have to focus on the understanding between the subjects and scenarios and obtain implications. Moreover, we can implement this method to intelligent video surveillance so that the surveillance becomes smarter.

ACKNOWLEDGMENTS

The authors would like to thank CASIA to provide activity database and the anonymous reviewers for their constructive comments.

REFERENCES

Aas, K., L. Eikvil and R.B. Huseby, 1999. Applications of hidden markov chains in image analysis. Pattern Recognit., 32: 703-713.

Aggrawal, J. and Q. Cai, 1999. Human motion analysis: A review. Comput. Vision Image Understand., 3: 428-440.

Alfred, H., 1910. Zur theorie der orthogonalen Funktionensysteme. Math. Ann., 71: 38-53.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., 39: 1-38.

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, Oct. 15-16, IEEE Computer Society, Beijing, China, pp: 65-72.

Hyeon-Kyu, L. and J.H. Kim, 1999. An HMM-based threshold model approach for gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 21: 961-973.

Laptev, I., 2005. On space-time interest points. Int. J. Comput. Vision, 64: 107-123.

Laptev, I., M. Marszalek, C. Schmid and B. Rozenfeld, 2008. Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, IEEE Computer Society, Anchorage, pp: 1-8.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60: 91-110.

Messing, R., C. Pal and H. Kautz, 2009. Activity recognition using the velocity histories of tracked keypoints. Proceedings of the IEEE International Conference on Computer Vision, Sept. 29-Oct. 2, IEEE Computer Society, Kyoto, Japan, pp: 104-111.

Niebles, J.C., H. Wang and L. Fei-Fei, 2008. Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vision, 79: 299-318.

Oikonomopoulos, A., I. Patras and M. Pantic, 2006. Spatiotemporal salient points for visual recognition of human actions. IEEE Trans. Syst. Man Cybern. B, 36: 710-719.

Poppe, R., 2010. A survey on vision-based human action recognition. Image Vision Comput., 28: 976-990.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77: 257-286.

Sundaresan, A., A. RoyChowdhury and R. Chellappa, 2003. A hidden Markov model based framework for recognition of humans from gait sequences. Proceedings of the International Conference on Image Processing, Sept. 14-17, IEEE Computer Society, Barcelona, Spain, pp: 93-96.

Turaga, P., R. Chellappa, V.S. Subrahmanian and O. Udrea, 2008. Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol., 18: 1473-1488.

Willems, G., T. Tuytelaars and V.G. Gool, 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. Lecture Notes Comput. Sci., 5303: 650-653.

Wilson, A.D. and A.F. Bobick, 1999. Parametric hidden markov models for gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 21: 884-900.

Wong, S.F. and R. Cipolla, 2007. Extracting spatiotemporal interest points using global information. Proceedings of the IEEE 11th International Conference on Computer Vision, Oct. 14-21, IEEE Inc., Rio de Janeiro, Brazil, pp: 1-8.

Yamato, J., J. Ohya and K. Ishii, 1992. Recognizing human action in time-sequential images using hidden Markov model. Proceedings of the Conference on Computer Vision and Pattern Recognition, June 15-18, IEEE Computer Society, Champaign, pp: 379-385.

Zhang, H. and Z. Liu, 2009. Gait representation and recognition using haar wavelet and radon transform. Proceedings of the WASE International Conference on Information Engineering, July 10-11, IEEE Computer Society, Taiyuan, Shanxi, China, pp: 83-86.

Zhang, H., Z. Liu and H. Zhao, 2010. Contour extraction of human with single-pixel width. Applied Mechan. Mater., 20-23: 376-381.