Pointwise Motion Image (PMI): A Novel Motion Representation and Its Applications to Abnormality Detection and Behavior Recognition

Qiulei Dong, Yihong Wu, and Zhanyi Hu

***Abstract*—In this paper, we propose a novel motion representation and apply it to abnormality detection and behavior recognition. At first, pointwise correspondences for the foreground in two consecutive video frames are established by performing a salient-region-based pointwise matching algorithm. Then, based on the established pointwise correspondences, a pointwise motion image (PMI) for each frame is built up to represent the motion status of the foreground. The PMI is more suitable for video analysis as it encapsulates a variety of motion information such as pointwise motion speed, pointwise motion orientation, pointwise motion duration, as well as the global shape of the foreground. In addition, it represents all of these pieces of information by a color image in the HSV space, by which many popular techniques in the image processing field can be straightforwardly adopted. By combining the PMI and AdaBoost, a method for abnormality detection and behavior recognition is proposed. The proposed method is shown to possess a high discriminative ability and is capable of dealing with local motion, global motion, and similar motions with different speeds. Experiments including a comparison with two existing methods demonstrate the effectiveness of the proposed representation in abnormality detection and behavior recognition.**

***Index Terms*—Abnormality detection, behavior recognition, motion representation, pointwise motion image (PMI).**

I. Introduction

MOTION representation, which is to extract discriminative information from raw video data for representing the motion status of the foreground, is a central topic in many fields such as visual surveillance and video analysis. Various motion representations have been proposed in recent years such as those reviewed in [1] and [12].

A traditional method of motion representation is motion trajectory [4], [7], [21]. For example, a principal curve is computed to describe gestures in gesture recognition [4]. In [7], the central points of face and hands in a video sequence are fitted into several quadratic curves, and six invariants from each quadratic curve are computed to construct a feature vector. However, the motion trajectory only carries global motion information, i.e., the motion status of the whole foreground, and it does not contain shape or local motion information, i.e., the motion status of each local area (even each pixel) of the foreground, but these two pieces of information are also useful for behavior recognition.

Another traditional method of motion representation is spatio-temporal silhouettes [2], [4], [23], [25]. Bobick and Davis [4] proposed motion-energy image (MEI) and motion-history image (MHI) representations for the recognition of human movements, and these representations contain both shape and global motion information. Blank *et al.* [2] used properties of the solution to the Poisson equation to extract spatio-temporal features for human action recognition. Yilmaz and Shah [25] used spatio-temporal action volumes for action recognition. In [23], an associated sequence of human silhouettes was converted into two types of representations, average motion energy (AME), and mean motion shape (MMS).

In [26], Zhong *et al.* classified spatial histogram feature vectors into prototypes and then detected unusual activities by finding spatially isolated clusters, where some useful local motion information could be lost in their spatial histogram features. Boiman and Irani [5] proposed a probabilistic graphical model using ensembles of spatio-temporal patches to detect irregular behaviors in videos. Xiang and Gong [24] used 7-D feature vectors from a blob of a scene-event to perform behavior profiling and abnormality detection. These 7-D feature vectors can only appropriately describe the gross information of movements. In [6], Cuntoor and Chellappa proposed an epitomic representation for modeling human activities, using kinematics of objects within short-time interval. Optical flow was used as a spatio-temporal descriptor for action recognition in [8].

From the above discussions, we can see that a crucial issue in motion representation is what features should be used to facilitate subsequent operations such as abnormality detection and behavior recognition. To this end, global motion information and shape information seem rather insufficient. In order to discriminate behaviors with local motion differences, local motion information should also be considered. In addition, motion duration, i.e., temporal information, is also quite helpful for discriminating similar behaviors with different speeds.

Hence, we think a good representation should include global motion, local motion, motion duration, and shape information, from which more discriminative features could be extracted for

Manuscript received August 09, 2007; revised November 24, 2007 and March 05, 2008. First published February 13, 2009; current version published April 01, 2009. This work was supported in part by the National Natural Science Foundation of China under Grant 60633070 and Grant 60675020, by the National High Technology R&D Program of China under Grant 2007AA01Z341, and by the National Key Basic Research and Development Program under Grant 2004CB318107. This paper was recommended by Associate Editor S. Takamura.

The authors are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: qldong@nlpr.ia.ac.cn; yhwu@nlpr.ia.ac.cn; huzy@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2009.2013503

further motion analysis. It seems possible to get such a representation by processing pointwise motion status of the foreground pixels, which are extracted by performing pointwise correspondences between consecutive frames. However in most cases, it is difficult to perform the pointwise correspondences due to the fact that nonrigid objects may be deformed and that the existing matching algorithms are mostly suitable for sparse corner points [11] or other keypoints [16]. Although the classical optical flow methods including the Lucas–Kanade method [17] try to find the pointwise correspondences, they usually cannot deal with large motion (a rapid move across frames). In addition, when the pointwise correspondences of the foreground pixels are obtained, if they were used directly as a motion representation, such a representation could not have provided more representative information as shown in the experiments.

Motivated by the above discussions, we propose a novel representation called pointwise motion image (PMI), which is a color image in the HSV color space, where the color components of each pixel represent the pointwise motion speed, pointwise motion orientation and pointwise motion duration respectively.

The PMI representation has the following characteristics.

1) The PMI is under the form of a "color image," which contains abundant motion information and temporal information. Being an image, many popular image and video processing techniques can be conveniently adopted to operate on the PMI, then many discriminative features can be extracted.
2) The PMI representation can be applied to various problems in the field of video analysis and motion analysis. In this study, we show that a method combining the PMI and AdaBoost is quite robust and effective for behavior recognition and abnormality detection, different behaviors and actions, particularly similar actions with different speeds such as running and jogging, can be discriminated effectively.

The remainder of this work is organized as follows. Section II presents the salient-region-based pointwise matching algorithm as well as the PMI construction. A method combining the PMI and AdaBoost is elaborated for abnormality detection and behavior recognition in Section III. Experiments including a comparison with two existing representations and some discussions are reported in Section IV and followed by some concluding remarks in Section V.

II. Motion Representation From Pointwise Correspondences

Here, we give a novel motion representation for moving objects. The representation contains the motion status of every image point on the foreground and thus fully utilizes both local and global motion information. At first, a salient-region-based pointwise matching algorithm is proposed to establish approximate pointwise correspondences for the foregrounds between consecutive frames, which consists of three steps.

Step 1) For every foreground point, a salient region in the current frame and its associated salient region in the preceding frame are constructed.

Step 2) The features of these salient regions and associated salient regions are extracted.

Step 3) A cost function is constructed with the extracted features and minimized to establish approximate pointwise correspondences.

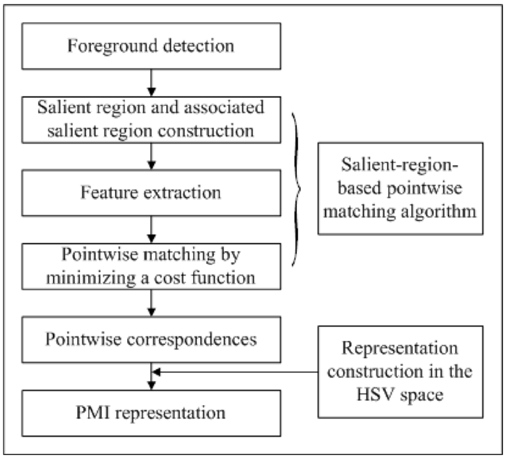

Foreground detection
Salient region and associated salient region construction
Feature extraction
Pointwise matching by minimizing a cost function
Pointwise correspondences
PMI representation
Salient-region-based pointwise matching algorithm
Representation construction in the HSV space

Fig. 1. PMI construction process.

Then, with the obtained pointwise correspondences, a motion representation called PMI, which is under the form of a color image in the HSV space, is proposed to represent pointwise motion status. The whole process is shown in Fig. 1, and the details of these steps are presented in the following subsections.

A. Salient Region and Its Associated Salient Region Construction

Gilles [10] and Kadir and Brady [13] applied an information-theoretic criterion to define salient regions for image registration and object recognition. In their works, salient regions centered at each foreground pixel were identified, and only a small number of them revealed some good discriminative ability, while the majority of them were not quite useful for extracting discriminative features. Here, we define and detect both a new salient region for each foreground pixel in the current video frame and its associated salient region at the same position in the preceding video frame. The considered pixels are only from the foreground, so the motion information is contained in such constructed salient regions and associated regions.

The adaptive Gaussian mixture model [22] is used to detect foreground pixels in this work. Then, the construction of salient regions and associated salient regions is performed as follows.

Let $X^t = (x^t, y^t)$ be a point on the detected foreground in the frame t. A local foreground region within a circular neighborhood of X^t with radius $= s^t_{X^t}$ is denoted as $R^t_{X^t}$, where $s^t_{X^t}$ is called the scale of $R^t_{X^t}$. Similarly, in the frame $(t-1)$, a region $R^{t-1}_{X^t}$ at the same location with scale $s^{t-1}_{X^t}$ is defined.

In order to extract more effective features for point matching, a multiscale approach is generally preferred. In this work, the entropy is used as the saliency measure to determine the appropriate scales for a salient region and its associated salient region.

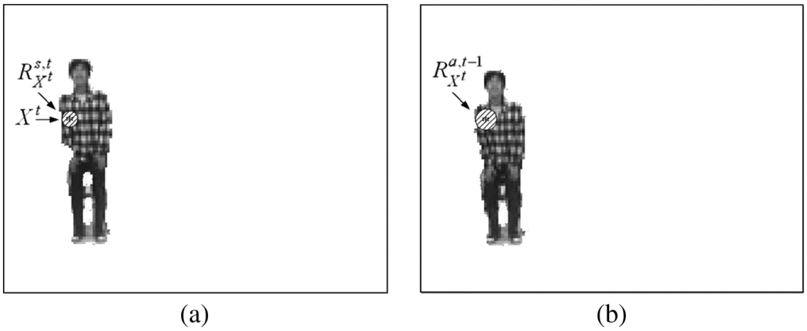


Fig. 2. Salient region pair. (a) In the current frame, the intersection $R^{s,t}_{X^t}$ of the foreground with the shadow circle is the salient region of the star point X^t. (b) In the preceding frame, the intersection $R^{a,t-1}_{X^t}$ of the foreground with the shadow circle is the associated salient region of the point.

More specifically, given a local region R and its intensity descriptor D which takes on values $\{d_1,\ldots,d_l\}$, the entropy of the intensity descriptor is defined as

$$E(R) = -\sum_i P_R(d_i)\log_2 P_R(d_i) \tag{1}$$

where $P_R(d_i)$ is the probability of the descriptor D taking the value d_i in R. Here, the descriptor D ranges from 0 to 255, hence the maximum value of $E(R)$ is 8 and its minimum value is 0. Then our saliency measure is defined as

$$Sr\left(s^t_{X^t}, s^{t-1}_{X^t}\right) = E\left(R^t_{X^t}\right) + \omega E_{\text{joint}}\left(R^t_{X^t} \cup R^{t-1}_{X^t}\right) \tag{2}$$

where $E(R^t_{X^t})$ is the entropy of the intensity histogram of $R^t_{X^t}$, $E_{\text{joint}}(R^t_{X^t} \cup R^{t-1}_{X^t})$ is the joint entropy of the intensity histogram of $R^{t-1}_{X^t}$, and $R^t_{X^t}$, ω is a constant weight. By varying the radii $s^t_{X^t}$ and $s^{t-1}_{X^t}$ of the circular neighborhoods (in this work, the varying range is from 2 to 13 pixels), the measure (2) attains a maximum value at some $(s^t_{X^t}, s^{t-1}_{X^t})$, and this value is called the salient measure of X^t. The corresponding scaled region $R^t_{X^t}$ in the frame t is defined as the salient region of X^t, denoted as $R^{s,t}_{X^t}$, and the corresponding scale is denoted as $s^{s,t}_{X^t}$. Then, the corresponding scaled region $R^{t-1}_{X^t}$ in the frame $(t-1)$ is called the associated salient region of X^t, denoted as $R^{a,t-1}_{X^t}$, and the corresponding scale is denoted as $s^{a,t-1}_{X^t}$. Later on, $R^{s,t}_{X^t}$ and $R^{a,t-1}_{X^t}$ always mean the salient region and the associated salient region of X^t.

For each foreground point X^t in the frame t, a salient region pair $(R^{s,t}_{X^t}, R^{a,t-1}_{X^t})$ is obtained. Fig. 2 is a salient region pair. The first frame $t=1$ is not considered.

B. Feature Extraction

Usually, the salient regions and the associated salient regions are small and their contours are sensitive to noise, so contour information is not considered at this stage. We are going to only extract gray and texture information from the salient regions and the associated salient regions.

A gray histogram of the union set $R^{s,t}_{X^t} \cup R^{a,t-1}_{X^t}$ with 32 bins is computed as a gray feature vector, where 32 is a tradeoff of accuracy and computational load. This extracted gray feature vector is denoted as $V^t_{gray} = (p_1, p_2 \ldots, p_{32})$.

Gabor features are used as the texture features due to their desirable characteristics of spatial locality and orientation selectivity [15], [18]. A 2-D Gabor wavelet can be expressed as

$$\varphi_{m,n}(X^t) = \frac{\|k_{m,n}\|^2}{\sigma^2} e^{\left(-\|k_{m,n}\|^2\|X^t\|^2/2\sigma^2\right)} \times \left[e^{ik_{m,n}X^t} - e^{-\sigma^2/2}\right] \tag{3}$$

where $X^t = (x^t, y^t)$, $\|X^t\|$ is the norm of X^t, and the wave vector $k_{m,n}$ is defined as

$$k_{m,n} = k_n e^{i\phi_m}$$

where $k_n = k_{\max}/f^n$, $\phi_m = \pi m/8$, $k_{\max}$ is the maximum frequency, and f is the spacing factor [14]. Here, we use eight orientations $m \in \{0,1,2,\ldots,7\}$ and five different scales $n \in \{0,1,2,3,4\}$, and the parameter setting is $k_{\max} = \pi/2$ and $f = \sqrt{2}$.

Given an image $I(X^t)$, its Gabor wavelet transform is

$$G_{mn}(X^t) = I(X^t) * \varphi_{m,n}(X^t) \tag{4}$$

where "$*$" denotes the convolution operator.

For each foreground point X^t in the frame t, the texture feature vector on its salient region $R^{s,t}_{X^t}$ is constructed based on the $G_{mn}(X^t)$, the mean value μ_{mn} and the standard deviation σ_{mn} of the magnitude of the transformed coefficients as

$$V^{s,t}_{\text{gabor}} = [G_{00}, \mu_{00}, \sigma_{00}, G_{01}, \mu_{01}, \sigma_{01}, \ldots, G_{mn}, \mu_{mn}, \sigma_{mn}] \tag{5}$$

where

$$\mu_{mn} = \frac{1}{S_{R^{s,t}_{X^t}}} \sum_{X_s \in R^{s,t}_{X^t}} |G_{mn}(X_s)|$$

$$\sigma_{mn} = \frac{1}{S_{R^{s,t}_{X^t}}} \sqrt{\sum_{X_s \in R^{s,t}_{X^t}} \left(|G_{mn}(X_s)| - \mu_{mn}\right)^2}$$

and $S_{R^{s,t}_{X^t}}$ is the area of the salient region $R^{s,t}_{X^t}$.

Similarly, for each foreground point X^t in the frame t, the texture feature vector $V^{a,t-1}_{\text{gabor}}$ on its associated salient region $R^{a,t-1}_{X^t}$ is obtained.

The texture feature vector at each foreground point X^t in the frame t is a concatenation of $V_{\text{gabor}}^{s,t}$ and $V_{\text{gabor}}^{a,t-1}$ as $V_{\text{texture}}^{t} = (V_{\text{gabor}}^{s,t}, V_{\text{gabor}}^{a,t-1})$.

C. Pointwise Matching by Minimizing a Cost Function

In order to search for the pointwise correspondences of the foreground object between two consecutive frames, the extracted feature vectors in Section II-B are used by minimizing a cost function. The cost function is based on the following two observations:

(ob1) the gray and texture in the local neighborhood at each foreground point does not change much between two consecutive frames;

(ob2) the scales of the salient region and the associated salient region change slightly between two consecutive frames.

Then, the cost function Fun^t in the frame t is defined as

$$Fun^t = \sum_{X^t \in I^t} D(X^t) \tag{6}$$

where I^t is the foreground region in the frame t and

$$\begin{aligned} D(X^t) = & W_1(X^t, X^{t+1}) D_{gabor1}(X^t, X^{t+1}) \\ & + W_2(X^t, X^{t+1}) D_{\text{gabor2}}(X^t, X^{t+1}) \\ & + \frac{W_1(X^t, X^{t+1}) + W_2(X^t, X^{t+1})}{2} \\ & \times D_{\text{gray}}(X^t, X^{t+1}) \end{aligned} \tag{7}$$

and X^{t+1} is a point in the frame $t+1$ within a rectangular neighborhood of X^t as

$$W_1(X^t, X^{t+1}) = \sqrt{1 + \left(s_{X^t}^{s,t} - s_{X^{t+1}}^{s,t+1}\right)^2} \tag{8}$$

$$W_2(X^t, X^{t+1}) = \lambda \sqrt{1 + \left(s_{X^t}^{a,t-1} - s_{X^{t+1}}^{a,t}\right)^2} \tag{9}$$

$$D_{\text{gabor1}}(X^t, X^{t+1}) = \left\| V_{\text{gabor}}^{s,t}(X^t) - V_{\text{gabor}}^{s,t+1}(X^{t+1}) \right\| \tag{10}$$

$$D_{\text{gabor2}}(X^t, X^{t+1}) = \left\| V_{\text{gabor}}^{a,t-1}(X^t) - V_{\text{gabor}}^{a,t}(X^{t+1}) \right\| \tag{11}$$

$$D_{\text{gray}}(X^t, X^{t+1}) = \left\| V_{\text{gray}}^{t}(X^t) - V_{\text{gray}}^{t+1}(X^{t+1}) \right\| \tag{12}$$

where λ is a positive constant.

In (6), Fun^t is the summation of the cost functions of all of the foreground pixels. In (7), $D(X^t)$ is the weighted sum of three terms computed by (10)–(12) derived from the (ob1). $W_1(X^t, X^{t+1})$ and $W_2(X^t, X^{t+1})$ in (8) and (9) derived from the (ob2), are the weights to balance the relative strengths of these three terms. The "1" is added into (8) and (9) to ensure both the two weights nonzero, otherwise two points which are actually not matched but have the same scale of their salient regions or associated regions will minimize $D(X^t)$, which may result in false matching.

It can be seen that for one point in the current frame and its potential corresponding point in the next frame, from the observations (ob1) and (ob2), its $D(X^t)$ should be minimized. Then, for all pairs of potential corresponding points, the function Fun^t should be minimized. Thus, for each foreground point X^t in the frame t, we search a foreground point X^{t+1} in the frame $t+1$. The X^{t+1} that minimizes $D(X^t)$ of (7) is considered as the corresponding point of X^t. We repeat the process for all of the foreground points and finally obtain approximate pointwise correspondences between the foregrounds of the frame t and $t+1$. The whole minimization then is equivalent to minimizing the cost function Fun^t of (6).

Remark: If only the salient region computed from a single frame is used to perform matching, it is hard to seek correct correspondences of all of the foreground points due to the lack of representative features. Here, a region pair in the two consecutive frames is used for setting local scales as well as searching the correspondences of foreground points, as the region pair has more gray and texture information than a single region. In addition, in the proposed matching algorithm, it is not expected that all the computed pointwise correspondences are accurate enough, which is why we term our correspondences as approximate correspondences here.

D. Construction of Pointwise Motion Images

It is desirable to have a motion representation that contains a variety of pieces of useful information such as motion duration, speed, orientation, and the shape of foreground. The obtained pointwise correspondences of foreground points between two consecutive frames in Section II-C just can be used to construct such a motion representation.

For a foreground point X^t at time t, its motion status is represented by

$$\begin{cases} o(X^t) = \text{angle}(X^{t+1} - X^t, L_h) \\ v(X^t) = \frac{d(X^t, X^{t+1})}{\max\{d(X^t, X^{t+1}) | t \in [2, \ldots, T-1], X^t \in I^t\}} \\ tm(X^t) = \frac{t}{T} \end{cases} \tag{13}$$

where X^{t+1} is the corresponding point of X^t in the frame $t+1$, L_h is the image horizontal axis, T is the time duration of a moving foreground, I^t is the foreground region in the frame t, $o(X^t)$ is the orientation defined by the included angle between the vector $(X^{t+1} - X^t)$ and L_h, $v(X^t)$ is the speed computed by the distance from X^t to X^{t+1} normalized by the maximum distance, and $tm(X^t)$ records the relative time of the current frame.

There are different motion statuses from different foreground points. The orientation $o(X^t)$ varies within [0, 360), the speed $v(X^t)$ varies within [0, 1], and the relative time $tm(X^t)$ varies within [0, 1]. It can be noted that the scope [0, 360) of $o(X^t)$, the scope [0, 1] of $v(X^t)$ and the scope [0, 1] of $tm(X^t)$ are respectively coincident with those of the hue, saturation, and value in the HSV color space. It follows that we can assign the values of $o(X^t)$, $v(X^t)$, $tm(X^t)$ to H, S, V in the HSV color space as

$$\begin{cases} H(X^t, t) = o(X^t) \\ S(X^t, t) = v(X^t) \\ V(X^t, t) = tm(X^t) \end{cases} . \tag{14}$$

By such assignments, all of the motion information made up of pointwise motion statuses in a frame is represented by a color image in the HSV space. This color image is called PMI in this study. Different frames give different PMIs. The color information of a point in a PMI describes the motion duration, motion

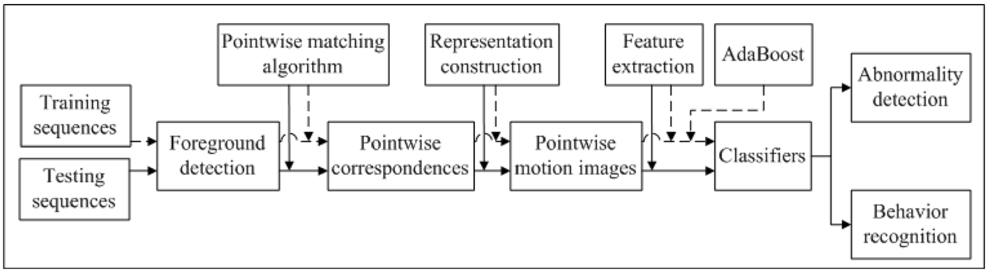
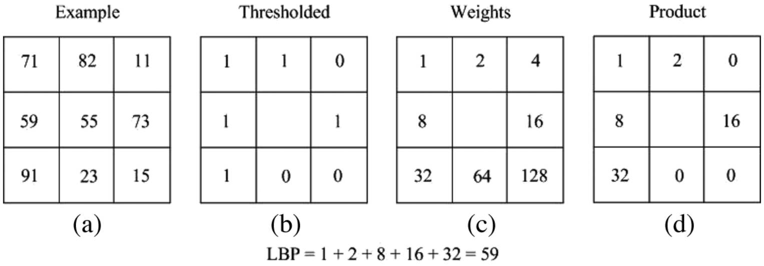


Fig. 3. Flowchart of the proposed method for abnormality detection and behavior recognition based on the PMIs.



Example

71	82	11
59	55	73
91	23	15

(a)

Thresholded

1	1	0
1		1
1	0	0

(b)

Weights

1	2	4
8		16
32	64	128

(c)

Product

1	2	0
8		16
32	0	0

(d)

LBP = 1 + 2 + 8 + 16 + 32 = 59

Fig. 4. Calculating the LBP.

speed, and motion orientation of this point. All the color points in a PMI constitute the same shape as the original foreground. Therefore, the global shape of the foreground, motion information of every foreground point, and temporal information are completely represented in a PMI.

Then, the obtained PMIs from all frames are refined further by image filtering so that some unreliable point correspondences are corrected. Some examples of the PMIs are shown in Fig. 7.

Here are some reasons why we prefer to use the PMI rather than to use pointwise information directly as a motion representation.

1) Although motion information of each point is useful and important for many further operations on the raw video data, it is really hard to extract an effective feature vector directly from the obtained motion information of each point. The PMI represents all these motion information as a color image, therefore, many popular image and video processing techniques, such as image enhancement, image filtering, and regional descriptor, can be applied to improving the PMI representation and facilitating feature extraction, and many effective image-proper features, such as texture and color, can be adopted to describe the motion status of the foreground.
2) The PMI representation, which contains motion information of each point, global shape information, and the relative time of the current frame, is more effective for feature handling later on when we deal with classification problems in abnormality detection and behavior recognition.
3) The PMI gives a more intuitive motion status, and is quite helpful for potential human intervention or human assisted motion analysis. For example, a homogeneous color region indicates a smooth motion, in contrast, if the color distribution is disorderly and unsystematic in a local area, it indicates that some of the pointwise correspondences in this area may be unreliable, and special care should be paid to such areas.

Besides, here are some reasons why the HSV color space is used in this work rather than other color spaces.

1) The scope [0, 360) of $o(X^t)$, the scope [0, 1] of $v(X^t)$, and the scope [0, 1] of $tm(X^t)$ are respectively coincident with those of the hue, the saturation, and the value in the HSV color model.
2) The correlations among the components in the HSV space are smaller than those in other color spaces.
3) Most importantly, as shown in the experiments, we found in the HSV space that the PMI manifests a better discriminative ability.

III. PMI-Based Applications

Being an effective motion representation, the PMI could be used in various fields such as visual surveillance and video analysis. Here, a method combining the PMI and Adaboost is presented for abnormality detection and behavior recognition. The flowchart of the method is shown in Fig. 3.

A. *PMI-Based Feature Extraction and Classifier Training*

As analyzed in Section II, since a PMI is a color image in the HSV space, many image related features can be extracted from it. Here, for each PMI, color, texture, and shape are used to constitute a feature vector F.

The color histogram over the foreground region is computed as the color feature, while the local binary pattern (LBP) histogram [19] is used for the texture feature. An illustrative example is shown in Fig. 4 for the LBP construction. Given a pixel and its 8-pixel neighborhood, at first, the gray level of this point is used as the threshold to binarize its neighborhood as shown in Fig. 4(b). Then, this binary pattern is multiplied in a pixel-wide manner with a predefined binomial weight pattern

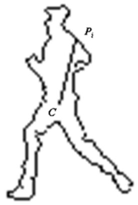


Fig. 5. Shape descriptor of the foreground: C represents the centroid of the foreground, and P_i represents one sample point.

[shown in Fig. 4(c)] to produce a pattern as shown in Fig. 4(d). Finally, the LBP number is obtained by summing up the numbers in Fig. 4(d).

Moreover, the shape descriptor is defined simply as follows. For each frame, n points on the silhouette are evenly sampled, then the distances between the centroid of the foreground and each sample point on the silhouette are computed as shown in Fig. 5. These computed distances are normalized by the maximum distance, and the shape descriptor vector is composed of these normalized n distances. Hence, the shape descriptor is invariant to translation and scale change. In addition, we experimentally found when n changes from 50 to 80, the recognition rate barely changed, hence empirically speaking, the shape descriptor is not sensitive to the sampling rate also. In all our experiments, n is set to 60.

Finally, the feature vector F extracted from a PMI is the concatenation of these three types of features. Both abnormality detection and behavior recognition can be considered as a classification problem. Based on the feature vectors F extracted from the PMIs, a strong classifier is constructed by AdaBoost [9]. More specifically, stumps, i.e., one-level decision trees, are used as the weak classifiers, and the learnt strong classifier $H(F)$ is a linear combination of a series of weak classifiers. Finally, the classification decision is simply the sign of $H(F)$.

B. Abnormality Detection

Abnormality detection is considered as a two-category classification problem.

In the training stage, the PMI sequences representing normal behaviors are selected as positive samples.

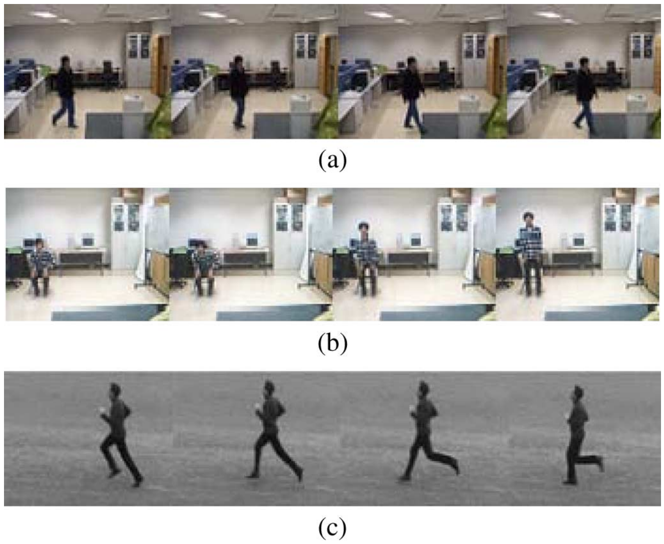
In the testing stage, for an input video sequence, we take different video subsequence V_{pq} from the pth frame to the qth frame ($q - p > \theta_d$, and θ_d is a threshold). Then, the corresponding PMIs and feature vectors F_i $(i = p, \ldots, q)$ of V_{pq} are constructed orderly. With these F_i $(i = p, \ldots, q)$, a score is computed as

$$\mathrm{score}(V_{pq}, H) = \frac{\sum_{i=p}^{q} \mathrm{sgn}\left(H(F_i)\right)}{q - p + 1} \tag{15}$$

where sgn is signum function. If the score is smaller than a preset threshold θ_s, the video sequence V_{pq} is regarded as an abnormality.

C. Behavior Recognition

Behavior recognition is considered as a multicategory classification problem.


(a)

(b)

(c)

Fig. 6. (a) Example of a walking behavior pattern in the walking database. (b) Example of an action pattern in the action database; (c) Example of an action pattern in the KTH database.

Assume that there are M normal behavior patterns. In the training stage, each kind of these normal behavior patterns is selected as a kind of positive samples. From each kind of the positive samples, we obtain a strong classifier H_k $(k = 1, \ldots, M)$.

In the testing stage for an input video sequence V, we construct its PMIs, then its feature vectors. With the feature vectors and the learnt classifiers H_k, we compute the corresponding $\mathrm{score}(V, H_k)$ for V by (15). By taking

$$k_0 = \arg\max_k \{\mathrm{score}(V, H_k), k = 1, \ldots, M\} \tag{16}$$

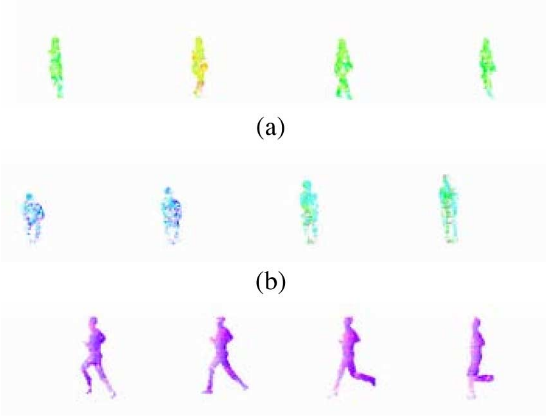
i.e., by taking the index of the strong classifier H_{k_0} whose $\mathrm{score}(V, H_{k_0})$ is the maximum value of $\{\mathrm{score}(V, H_k), k = 1, \ldots, M\}$, the input V is regarded as the k_0th behavior pattern.

IV. Experiments

We use three databases, the walking behavior database, the action database and the KTH database [20], to test the proposed method. The walking behavior database and the action database are recorded by us, where the foreground moves entirely for some sequences but partially for other sequences. Fig. 6 shows some examples of these three databases.

A. With the Walking Behavior Database

The database contains six kinds of walking behaviors in an office scene (walking from the side-door to the computer, from the computer to the side-door, from the computer to the printer, from the printer to the computer, from the side-door to the printer, and from the printer to the side-door), each one of which is performed by four different people 25 different times. These video sequences are captured by a digital camcorder and then are converted to 300 × 240 BMP files. In total, we have 600 video sequences. For each run, the videos of three people are used for training, and the remaining videos are used for the testing of behavior recognition. The reported result for behavior recognition is the average of four runs. In addition, a long video sequence

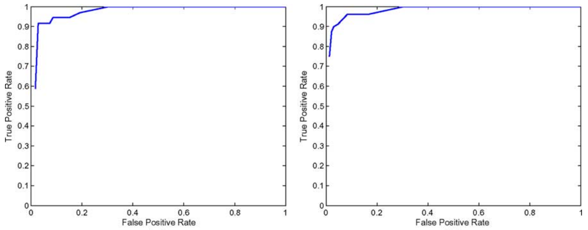


(a)

(b)

(c)

Fig. 7. (a) Filtered PMI sequence of Fig. 6(a). (b) Filtered PMI sequence of Fig. 6(b). (c) Filtered PMI sequence of Fig. 6(c).



(a)

(b)

Fig. 8. (a) ROC curve of the walking behavior sequence. (b) ROC curve of the action sequence.

of about one and a half hours is recorded for the testing of abnormality detection. This long video contains not only 620 defined normal behaviors but also 204 abnormal behaviors such as jumping, running, and other behaviors.

For the video sequences in the database, the PMIs are constructed by pointwise matching. Then alpha-trimmed mean filter is performed on the PMIs to improve the correspondence reliability. The filtered PMIs of the behavior sequence in Fig. 6(a) are shown in Fig. 7(a). Since a PMI is a color image, the feature vector F consisting of three popular types of image features, color histogram, the LBP descriptor and the shape descriptor, is extracted easily from a PMI.

Abnormality Detection: Based on the extracted feature vector F, the AdaBoost algorithm is used to construct a strong classifier. Then, the method of Section III-B is applied to the input long video, and the true positive rate and the false positive rate are computed. The true positive rates and the false positive rates by varying the threshold θ_s (for θ_s definition, see Section III-B) are recorded by a receiver operating characteristic (ROC) curve as shown in Fig. 8(a). It is seen that the curve increases fast at first, and the true positive rate attains 0.9 when the false positive rate is at a small value of about 0.027, which shows that our method can achieve high abnormality detection rate and low false positive rate simultaneously.

Behavior Recognition: The behavior recognition is considered as a multicategory classification problem. From the training data, we obtain the corresponding strong classifiers. Then, the method of Section III-C is applied to the test data. The average recognition rate is 89.3% as shown in the last row of Table I. In addition, two existing methods, the AME-based method [23] and the MHI-based method [3], are also tested with the same data. AME is an image where the pixel-wise intensity is the mean intensity of the pixels at the same location in a sequence of binary foreground silhouettes, and MHI is a scalar-valued image where pixel intensity is a function of the temporal history of motion. In our implementation, the parameter setting in MHI is: $\tau_{\max} = 21$, $\tau_{\min} = 9$, and $n = 10$ (with this setting, the best result is obtained for the MHI-based method in our experiment). The recognition rates of the two methods are 78.7% and 83.3% as shown in Table I.

In addition, in order to examine the effects of the extracted features from the PMIs, single features and their different combinations (single shape descriptor, single HSV histogram, single LBP, HSV histogram + Shape descriptor, LBP + Shape descriptor, LBP + HSV histogram) are tested for behavior recognition, the corresponding recognition rates are shown from the fourth row to the ninth row in Table I.

Moreover, in order to illustrate the rationality of choosing the HSV color space rather than other color spaces to construct the PMIs, we also convert the motion information of each point into a color image in other three color spaces, the RGB space, the YIQ space and the YCbCr space, respectively, then extract the corresponding features for recognizing different behaviors. The recognition rates on the walking database are shown in the second column of Table II, and the results show that the HSV space performs the best.

B. With the Action Database

The data used in this experiment contain eight different kinds of actions (standing-sitting, waving left hand, waving two hands, tilting head, kicking left foot, kicking right foot, clapping, and boxing), and each action is performed by eight different people 15 different times. We have 960 action sequences in total and then they are converted into 160 × 120 BMP files. For each run, we arbitrarily select the sequences of four people for training, and the remaining sequences are used for the testing of action recognition. The reported result for action recognition is the average of ten runs, and the constructed PMIs of the action sequences in Fig. 6(b) are shown in Fig. 7(b). In addition, a long video sequence containing 80 abnormal actions and 240 normal actions is recorded for the testing of abnormality detection.

We apply the method of Section III-B to the long video. The ROC curve is shown in Fig. 8(b). It is seen that the curve increases at first, and the true positive rates attains 0.9 when the false positive rates is at a small value of about 0.029, which also shows that our method can achieve high abnormality detection rate and low false positive rate simultaneously.

By applying the method of Section III-C to the testing data of action recognition, the average recognition rate of the proposed PMI-based method is 90.6%, while the recognition rate of AME and MHI is 81.0% and 80.8% respectively, and the results with different combinations of the extracted features from the constructed PMIs are shown in the fourth column of Table I. In addition, the recognition rates corresponding to different color spaces are shown in the third column of Table II.

TABLE I
RECOGNITION RATE COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS ON THE THREE DATABASES

Methods	Used features	Recognition rate		
		Walking database	Action database	KTH database
AME-based	Gray values	78.7 %	81.0 %	69.4 %
MHI-based	Hu moments	83.3 %	80.8 %	75.8 %
PMI-based	Shape descriptor	59.3 %	73.5 %	55.8 %
PMI-based	HSV histogram	79.3 %	71.5 %	66.8 %
PMI-based	LBP	81.3 %	72.7 %	68.4 %
PMI-based	HSV histogram + Shape descriptor	80.7 %	87.5 %	77.4 %
PMI-based	LBP + Shape descriptor	82.7 %	88.3 %	78.4 %
PMI-based	LBP + HSV histogram	87.3 %	83.5 %	81.9 %
PMI-based	LBP + HSV histogram + Shape descriptor	89.3 %	90.6 %	84.8 %

TABLE II
RECOGNITION RATE COMPARISON OF THE HSV COLOR SPACE WITH THE OTHER THREE COLOR SPACES ON THE THREE DATABASES

Color spaces	Recognition rate		
	Walking database	Action database	KTH database
YIQ	83.3 %	88.5 %	80.7 %
YCbCr	85.3 %	85.6 %	76.1 %
RGB	84.7 %	87.9 %	78.7 %
HSV (PMI)	89.3 %	90.6 %	84.8 %

The confusion matrix of action recognition by the proposed method is computed and shown in Table III. The confusion values are small, indicating that the proposed method can classify actions effectively.

C. With the KTH Database

The KTH database [20] contains six kinds of actions (walking, jogging, running, hand waving, clapping, and boxing) performed several times by 25 people. It contains 598 sequences, and the spatial resolution is 160 × 120 pixels. In this database, walking, jogging and running are three kinds of similar actions with different speeds. For each run, we select the sequences of 12 people for training, and the rest sequences are for action recognition testing. The reported result is the average of 10 runs. The constructed PMIs of the action sequences in Fig. 6(c) are shown in Fig. 7(c).

The average recognition rate of the proposed PMI-based method is 84.8%, while the recognition rate of the AME-based method and the MHI-based method is 69.4% and 75.8% respectively, and the results with different combinations of the extracted features from the constructed PMIs are shown in the fifth column of Table I.

It is seen that when only the shape information is used, the method has a poor performance on classifying the actions in this database. A single color or single texture feature cannot classify these actions effectively either. When two of these extracted features are combined, the corresponding recognition rate is increased. The PMI-based method combining the three kinds of features outperforms all other combinations, and can discriminate similar actions with different speeds more effectively.

In addition, the recognition rates corresponding to different color spaces on the KTH database are shown in the fourth column of Table II, and the results show that the HSV space performs the best.

D. Observations and Discussions

Here are some points from our experiments.

1) The experiments on the three databases show that the proposed PMI-based method has a higher ability to handle global motion, local motion, and similar motions, and can be used effectively for abnormality detection and behavior recognition. This is chiefly due to the fact that the PMI representation contains different pieces of motion information, including pointwise motion time, pointwise motion speed, pointwise motion orientation, and the global shape of the foreground. Moreover, the PMI is a color image where popular image processing techniques and image features can be adopted and extracted conveniently, therefore it can also be applied to other applications, for example, human-assisted motion analysis and human assisted behavior modeling.
2) From Table I, it can be seen that the performance using only a single feature, such as color (i.e., pure motion information of each point), shape or texture feature, is poor, which indicates that a single feature is inadequate for behavior and action recognition. In contrast, feature combinations perform better. In addition, the proposed PMI-based method outperforms the AME-based method and the MHI-based method. The possible reasons could be that the AME-based method lacks temporal information and is hard to discriminate similar behaviors with different performing speeds like running and jogging, and the MHI-based method used seven Hu moments, and those global features lack effective local discriminability. For the proposed PMI representation, unreliable pointwise correspondences are rectified by alpha-trimmed mean filtering on the PMI, and more effective features such as LBP texture and color histogram are used, so it gives a better recognition rate.
3) The PMIs in the HSV space performs better than those in other color spaces. The possible reasons could be that the correlations among the three components H, S, V in the HSV space are smaller. However, the correlations among the components in other color spaces are usually bigger, which bring adverse effects.

TABLE III
CONFUSION MATRIX OF ACTION RECOGNITION BY THE PROPOSED METHOD. EACH ROW REPRESENTS THE PROBABILITIES OF THAT ACTION BEING CONFUSED WITH ALL THE OTHER ACTIONS

Stand-sit	0.92	0.05	0.00	0.00	0.00	0.03	0.00	0.00
Waving-l	0.07	0.85	0.00	0.00	0.08	0.00	0.00	0.00
Waving-two	0.00	0.00	0.90	0.00	0.00	0.00	0.10	0.00
Tilting head	0.00	0.02	0.00	0.90	0.03	0.00	0.00	0.05
Kicking-l	0.00	0.03	0.00	0.02	0.95	0.00	0.00	0.00
Kicking-r	0.00	0.00	0.00	0.05	0.00	0.95	0.00	0.00
Clapping	0.00	0.00	0.15	0.00	0.00	0.00	0.83	0.02
Boxing	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.95
	Stand-sit	Waving-l	Waving-two	Tilting head	Kicking-l	Kicking-r	Clapping	Boxing

V. CONCLUSION

A novel motion representation is proposed and applied to abnormality detection and behavior recognition. Our major contributions include the following.

1) A salient-region-based pointwise matching algorithm is proposed to establish approximate pointwise correspondences of foregrounds.
2) A PMI is introduced for motion representation, which includes a variety of motion information of each point such as motion speed, orientation and duration, and also preserves the shape of motion foreground simultaneously. Hence, the PMI representation contains not only the local motion information of every foreground point but also the global motion information and shape information of the foreground. In addition, since the PMI representation is under the form of a color image, many popular image and video processing techniques can be adopted to extracted more effective features.
3) With the PMI representation, a method for abnormality detection and behavior recognition under the AdaBoost paradigm is presented. The method is shown to be able to achieve high detection and recognition rates, and is capable of dealing with global motion, local motion, and similar motions with different speeds successfully.

Finally, we would also point out that, though the PMI representation processes the above listed properties, it is view-dependent and cannot handle video sequences with varying frame rate at its present form. In the future, how to extend the PMI representation to varying frame rate and how to use the PMI representation to segment videos will be investigated.

REFERENCES

[1] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE Int. Conf. Comput. Vision*, Beijing, China, 2005, pp. 1395–1402.

[3] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[4] A. Bobick and A. Wilson, "A state-based approach to the representation and recognition of gesture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 12, pp. 1325–1337, Dec. 1997.

[5] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *Proc. IEEE Int. Conf. Comput. Vision*, Beijing, China, 2005, pp. 462–469.

[6] N. Cuntoor and R. Chellappa, "Epitomic representation of human activities," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2007, pp. 846–853.

[7] Q. Dong, Y. Wu, and Z. Hu, "Gesture recognition using quadratic curves," in *Proc. 7th Asian Conf. Comput. Vision*, Hyderabad, India, 2006, pp. 817–825.

[8] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Comput. Vision*, Nice, France, 2003, pp. 726–733.

[9] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. ECCLT*, 1995, pp. 23–37.

[10] S. Gilles, "Robust description and matching of images," Ph.D. dissertation, Dept. of Eng. Sci., Univ. of Oxford, Oxford, U.K., 1998.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, Manchester, U.K., 1988, pp. 147–151.

[12] W. Hu, T. Tan, and L. Wang, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334–352, Mar. 2004.

[13] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[14] M. Lades, J. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.*, vol. 42, pp. 300–311, 1993.

[15] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[18] B. Manjunath and W. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[19] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recogn.*, vol. 29, no. 1, pp. 51–59, 1996.

[20] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. Pattern Recogn.*, 2004, pp. III:32–III:36.

[21] M. Shin, L. Tsap, and D. Goldgof, "Gesture recognition using Bezier curves for visualization navigation from registered 3-D data," *Pattern Recogn.*, vol. 37, no. 5, pp. 1011–1024, 2004.

[22] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[23] L. Wang and D. Suter, "Informative shape representations for human action recognition," in *Proc. Int. Conf. Pattern Recogn.*, 2006, pp. 1266–1269.

[24] T. Xiang and S. Gong, "Video behaviour profiling and abnormality detection without manual labelling," in *Proc. IEEE Int. Conf. Comput. Vision*, Beijing, China, 2005, pp. 1238–1245.

[25] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2005, pp. I:984–I:989.

[26] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2004, pp. 819–826.

Qiulei Dong received the B.S. degree in automation from Northeastern University, Shenyang, China, in 2003, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2008.

He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. His research interest covers visual surveillance, motion analysis, 3-D localization, and visual metrology.

Yihong Wu received the Ph.D. degree from the Institute of Systems Science, Chinese Academy of Sciences, Beijing, in 2001.

She is currently a Professor with the Institute of Automation, Chinese Academy of Sciences. Since then, she has been with the National Laboratory of Pattern Recognition of China. She was a Visiting Scholar with the MEEM Department, City University of Hong Kong, in 2005, 2006, and 2007. Her research interests include polynomial elimination application, geometric invariant application, camera calibration, camera pose determination, image matching, and stereovision.

Zhanyi Hu received the B.S. degree in automation from the North China University of Technology, Beijing, in 1985 and the Ph.D. degree in computer science from the University of Liege, Liege, Belgium, in 1993.

Since 1993, he has been with the Institute of Automation at the Chinese Academy of Sciences. From May 1997 to May 1998, he was a Visiting Scholar with the Chinese University of Hong Kong. He now is a Research Professor of computer vision, an Associate Editor for the *Journal of Computer Science and Technology*, and an Associate Editor-in-Chief for the *Chinese Journal of CAD and CG*. His current research interests are in robot vision, which include camera calibration, 3-D reconstruction, active vision, geometric primitive extraction, and vision guided robot navigation.