Human Behavior Analysis Based on a New Motion Descriptor

Kaiqi Huang, *Senior Member, IEEE,* Shiquan Wang, Tieniu Tan, *Fellow, IEEE,* and Stephen J. Maybank, *Senior Member, IEEE*

***Abstract*—Human behavior analysis is an important area of research in computer vision and is also driven by a wide spectrum of applications, such as smart video surveillance and human–computer interface. In this paper, we present a novel approach for human behavior analysis. Two research challenges, motion representation and behavior recognition, are addressed. A novel motion descriptor, which is an improved feature based on optical flow, is proposed for motion representation. Optical flow is improved with a motion filter, and feature fusion with the shape and trajectory information. To recognize the behavior, the support vector machine is employed to train the classifier where the concatenation of histograms is formed as the input features. Experimental results on the Weizmann behavior database and the Institute of Automation, Chinese Academy of Science real-world multiview behavior database demonstrate the robustness and effectiveness of our method.**

***Index Terms*—Human behavior, motion analysis, optical flow, surveillance.**

I. Introduction

HUMAN BEHAVIOR analysis is an important area of research in computer vision devoted to detecting, tracking, and understanding people's physical behaviors. This research is driven by a wide spectrum of applications in various areas such as smart video surveillance [1], interactive virtual reality systems [2], advanced and perceptual human–computer interfaces [3], content-based video storage and retrieval [4], sports performances analysis and enhancement [5], clinical studies [6], smart rooms and ambient intelligence systems [7], and so forth. A survey of recent research can be found in [8]. The application area in this paper is video surveillance.

In video surveillance people are tracked and monitored for particular actions. The demand for smart video surveillance systems is due to the large number of security-sensitive areas such as banks, department stores, parking lots, etc. and the vast numbers of images collected from these areas by surveillance cameras. Surveillance camera video streams are often stored in video archives or recorded on tapes. Most of the time, these video streams are only used "after the fact," mainly as an identification tool. There is a need for real-time video analysis, for example to alert security staff if a criminal act is in progress.

The behavior analysis framework is shown in Fig. 1. It consists of feature extraction, basic behavior description and complex behavior description. Complex behaviors are composed of many single behaviors with the temporal relations. Most work of behavior analysis focuses on feature extraction and behavior description, which are connected closely. According to the features used for analysis, the behavior analysis methods can be classified into three kinds: spatial-based (such as shape), motion-based (such as trajectory) and spatial-temporal-based methods. We give an introduction to these three types of feature based methods and propose our method based on a novel motion descriptor.

A. Shape-Based Features

Shape-based features have been commonly used in behavior analysis based on contour or silhouette information. In [10], the authors extracted 3-D shape for recognizing human posture using support vector machines. While for 3-D methods, point correspondences are needed with high accuracy, which costs high computation. 2-D shapes are extracted in [11] for behavior analysis. In [11], the authors use the Canny edge detector to extract shape and some key frames are applied for recognizing behaviors. For more complex activity analysis, different body shape features are employed in many studies [12], [13]. Sato and Aggarwal have developed a hierarchical method for human interaction behaviors, the poses (shape) of body parts are estimated at the low level and the overall poses are recognized at high level [12]. Park *et al*. extract silhouettes to classify more detailed interactions such as "pointing at the other person," "shaking hands," etc. [13].

B. Spatial-Temporal-Based Features

Space-time approaches for behavior analysis have been widely used in recent years. In [14], the empirical distributions of space-time gradients are collected from an entire video clip

Manuscript received August 26, 2008; revised January 5, 2009, February 24, 2009, and April 13, 2009. First version published August 4, 2009; current version published December 1, 2009. This paper was recommended by Associate Editor S. Pankanti. This work is supported by National Basic Research Program of China (Grant No. 2004CB318100), National Natural Science Foundation of China (Grant No. 60736018, 60723005), NLPR 2008NLPRZY-2, National Hi-Tech Research and Development Program of China (2009AA01Z318).

K. Huang, S. Wang, and T. Tan are with the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing 100080, China (e-mail: kqhuang@nlpr.ia.ac.cn; sqwang@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn).

S. J. Maybank is with the Department of Computer Science and Information Systems, Birkbeck College of London University, London WC1E 7HX, U.K. (e-mail: sjmaybank@dcs.bbk.ac.uk).

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.2009.2029024

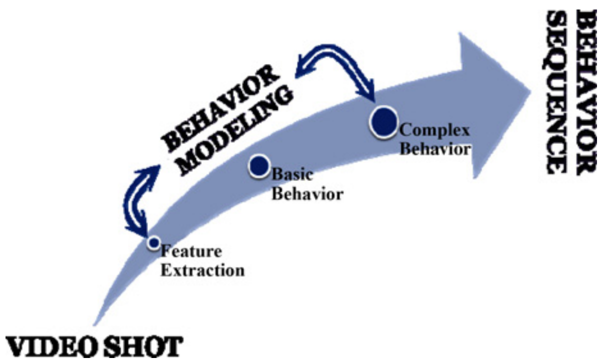

Fig. 1. Framework for behavior analysis.

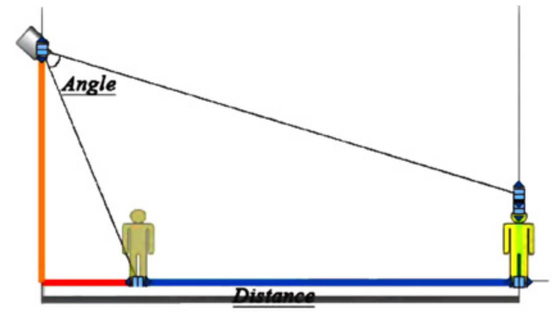

Fig. 2. Two factors between camera and objects: angle and distance.

to recognize a single behavior. However, this method did not capture a detailed geometric description of the behavior. In [15], the authors propose a 3-D space-time video-template correlation for recognizing dynamic action, which needs high computation cost. In [16], Yilmaz and Shah use a two-step graph theoretical approach to generate a spatial temporal volume (STV), which can solve the point correspondence problem between consecutive frames. They then compute the action descriptors by analyzing the differential geometric properties of the STV. Similarly, in [17], Laptev and Lindeberg extend the 2-D Harris detector to 3-D, to find a sparse set of space-time corner points, while maintaining scale invariance. But there are so few such points in a typical motion that the method may be badly affected by occlusions or by misdetections of these corner points. M. Blank uses the Poisson equation to extract space-time features such as corners, local space-time saliency, behavior dynamics and shape orientation, and then integrates these local features into a compact vector of features to represent an action [18], while it costs high computation to solve the Poisson equation. In a word, the spatial-temporal-based features provide more information for behavior analysis, so the discriminative performance is high, while the computation cost is also high.

C. Trajectory-Based Features

Trajectory-based approaches have often been proposed for outdoor behavior analysis. In [20], Stauffer *et al.* acquired a set of concept prototypes by using online vector quantization of trajectories. Then they used hierarchical clustering to obtain several motion routes in an outdoor surveillance scene. Based on these motion routes, a single person event in the scene ("one person goes from entrance A to exit B," etc.) can be recognized.

A trajectory on its own does not provide enough detailed information about behavior. Local motion descriptors are required. In [21], Ribeiro *et al.* evaluated the performance of two large sets of features for recognizing five categories of human activity such as walking, running, fighting, etc. The first set consists of trajectory-based features, such as velocity. The second set is based on estimates of the optical flow or instantaneous pixel motion inside a bounding box. Then the authors investigated a hierarchical classifier with different combinations of features. In [22], Robertson *et al.* proposed a combined approach to recognize single person actions that are described by trajectory information (position and velocity), and a set of local motion descriptors (coarse optic flow).

Trajectories are also very useful for recognizing multiperson behaviors. In [23], Galata *et al.* proposed an automatic approach to learn several qualitative spatial relations of primitive object interactions. They first extract primitive units from the trajectories of single person, then the variable length Markov model is used to infer the temporal structure of typical interactive behavior. In [24], Oliver *et al.* extracted the relative distance, the derivative of the relative distance, the degree of alignment of the moving directions and the magnitudes of their velocities of two pedestrians to make a feature vector describing their activity. Then coupled hidden Markov models are used for modeling the evolving relative spatial relationships.

In Table I, we compare three types of features which are often used for behavior analysis. Shape features and spatial-temporal features are often used for single person behavior analysis, motion features can be used for interactive person behavior. The favored camera view is also different for the three types of features. For shape and spatial-temporal features, the object should be close to the cameras and the width of the field of view should be limited to about 30°. Under these conditions, the shape features can be extracted better. Trajectory-based motion features are used if the camera is some distance away from the object, while there is no obvious distance limitation on the use of optical flow. Motion-based features have lower discriminative performance compared with shape and spatial temporal-based features, but they can be computed more quickly, making motion-based features more useful in real-time applications.

According to the above analysis, we find that optical flow features are more robust than other features in different views and cost lower computation, while the discriminative performance of optical flow-based feature cannot satisfy the requirements of view invariant behavior analysis. We improve the optical flow feature and find a novel motion descriptor that uses both shape and trajectory information. We test our method not only on a public behavior database (Weizmann behavior database) but also on the Institute of Automation, Chinese Academy of Science (CASIA) multiview real world behavior database. Experimental results show that

TABLE I
COMPARISON OF THREE TYPES OF FEATURES USED FOR BEHAVIOR ANALYSIS

Feature	Application	Favored Angle View[a]	Discriminative Performance	Computational Cost
Shape (contour, silhouette)	Single person behavior	Near distance and horizontal view (0–30°)	High	Medium
Motion (trajectory)	Simple single person behavior and interactive behavior	Far distance view (30–90°)	Low	Low
Optical flow		No limitation		
Spatial-temporal (ST, STV, ...)	Single person behavior	Near distance and horizontal view (0–30°)	High	High

[a]The angle and distance between objects and camera referred to Fig. 2

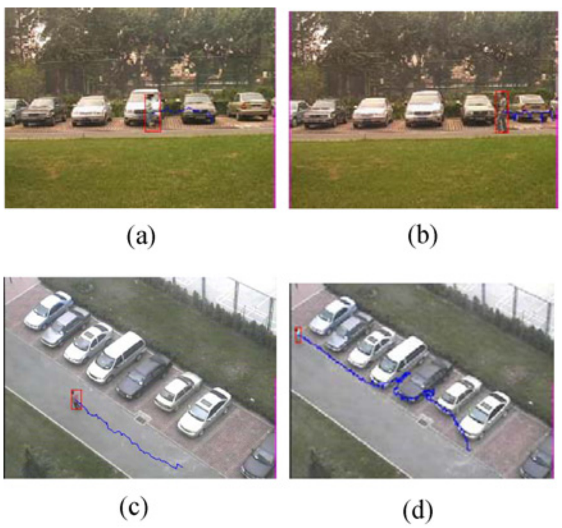

Fig. 3. Object detection and tracking results. (a) Running in horizontal view. (b) Walking in horizontal view. (c) Running in vertical view. (d) Wandering in vertical view.

our method is robust and effective, and that the computational cost is relatively low.

This paper is organized as follows. In Section II, we give the description of CASIA behavior database. The detection and tracking methods are given in Section III. In Section IV, the improved optical flow-based method is presented in detail and a lot of experimental results and evaluation based on our challenging database (CASIA behavior database) and Weizmann behavior database are given in Section V. Conclusion will be given in Section VI.

II. CASIA BEHAVIOR DATABASE

The CASIA behavior database contains image sequences of 11 actions, each performed three times by 24 actors (13 males/ 11 females). Every action is captured by three cameras at the same time in different views: horizontal view (HV: 0–30°), vertical view (VV: 60–90°) and bird's eye view (BV: 30–60°). Every view includes seven types of single-person behaviors (walking, running, jumping, bending, crouching, lying, wandering) and four types multiperson interactive behaviors (meeting, robbing, following, fighting) as shown in Table II. Each sequence is about 8–10 s.

III. OBJECT DETECTION AND TRACKING

Currently, we use Gaussian mixture functions to model the probabilistic distributions of image pixel values, and we update the parameters of all stochastic models following Stauffer and Grimson [20]. Then we employ the point (center of mass) representation to describe each detected object and make use of the nearest neighbor criterion to track moving objects. Meanwhile, we exploit Kalman filtering to predict the position and size of tracked objects. When object occlusion happens, the predicted values are used to replace object states at last time instant. One limitation of the current tracking algorithm lies in trajectory crossings and object merging. Hence we employ two strategies to solve this problem. One is scale-invariant feature transform descriptor [6] based appearance matching. The other is to combine the nearest neighbor criterion with particle filtering-based probabilistic inference. Fig. 3 illustrates some detection and tracking results.

IV. IMPROVED OPTICAL FLOW-BASED BEHAVIOR ANALYSIS

A. Problems in Optical Flow-Based Behavior Analysis

Shapes, trajectories and optical flow are often used for behavior analysis. The use of shape features is better in near distances and horizontal view, while trajectory features are suitable for distant objects and vertical views. In comparison with shape and trajectory features, optical flow is better in view and distance invariance, so optical flow is an intuitive choice for behavior analysis [25]–[29]. Simple global velocity and a global orientation are often extracted from optical flow to analyze the behaviors [26]–[28]. In [29], the authors combine optical flow with other features to improve its performance. However, optical flow has not been so far successful for behavior analysis because of two problems, the first problem is the effect of noise on the computation of optical flow. The second problem is the low discriminative performance.

1) *Noises and Errors:* Errors and noises will lead to inaccurate features during the optical flow computation, e.g., direction, speed, and so on [30]. We compute optical flow using the Horn–Schunck algorithm [34]. As shown in Fig. 4, the left image gives the original bending person behavior and the middle image gives the flow field with

TABLE II
CASIA BEHAVIOR DATABASE [33]

	Walking	Running	Jumping	Bending	Crouching	Wandering	Lying
HV							
VV							
BV							

(a) Single-person behaviors

	Meeting	Robbing	Following	Fighting
HV				
VV				
BV				

(b) Multiperson interaction behaviors

the noises and errors. Compared with the right image (correctly optical flow), the flow field in the middle image will cause low accuracy for behavior analysis.

2) *Low Discriminative Performance:* Optical flow is often extracted as global feature, which does not contain enough information to discriminate different behaviors. As shown in Fig. 5, the bending behavior and the falling behavior have the same optical flow, which causes the failure to discriminate these two behaviors.

We define a motion descriptor based on improved optical flow feature. Two aspects are considered to be robust and discriminative motion descriptor: 1) making full use of optical flow feature considering local patches information, and 2) an improved descriptor is formed fusing the trajectory, shape and optical flow features, which is effective in different views and cost low computation. With the computer of P4 3.0 GHz CPU and 1.5 GB RAM, the processing time for one single frame of 320×240 is about 6 ms, which is appropriate for real-time applications.

B. Improved Motion Descriptor with Optical Flow Features

In this part, we put forward the new motion representation based on optical flow for behavior analysis, which also combines trajectory and shape cues. The framework is shown in Fig. 6 and the details of feature extraction are given as following.

1) *Motion Filtering:* As we have mentioned above, noises and errors in optical flow computation will affect the discriminative power of extracted features for behavior analysis. Noises and errors are mainly affected by camera quality, video transmission or camera vibration and so on as shown in Fig. 7.

Here, instead of considering how noise and error affect optical flow, which is very difficult to analyze due to complexity of the calculation of optical flow, we analyze the image quality statistical information about the effects of noise and errors on the optical flow. As we can see in Fig. 8, the speed distribution of noise, optical flow and error can be modeled as a Gaussian distribution, respectively. Obviously, the noise and error are in the different parts of the distribution, the noise is in the

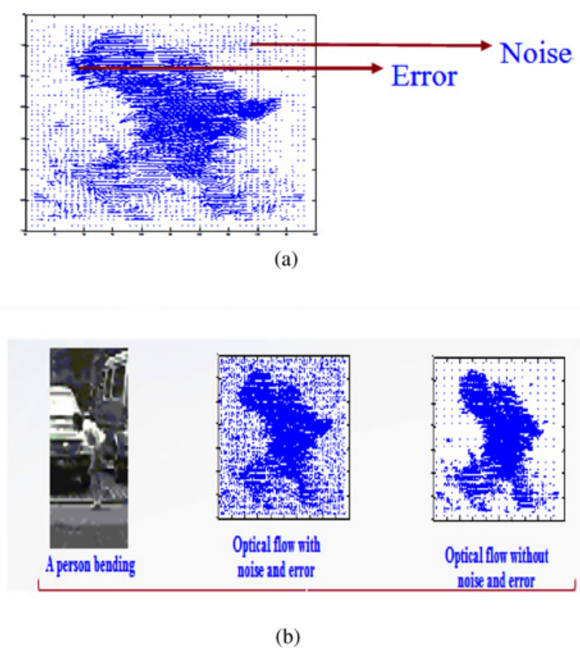

Fig. 4. Noises and errors in the optical flow field. We compute optical flow using the Horn–Schunck algorithm [34]. (a) Noise and error examples. (b) Optical flow example of bending behavior.

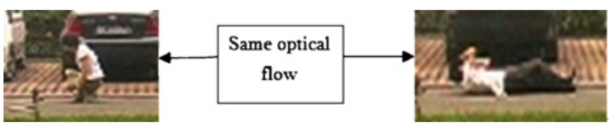

Fig. 5. Two behaviors: bending and falling.

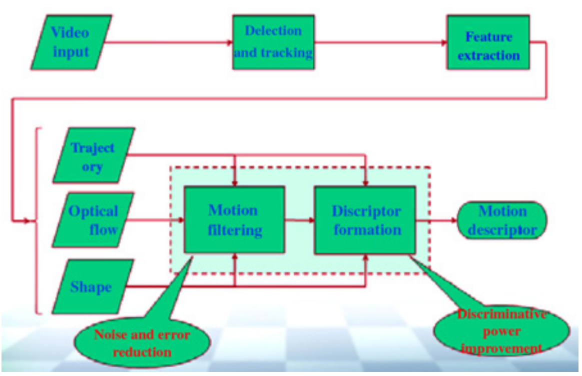

Fig. 6. Flowchart of our motion descriptor.

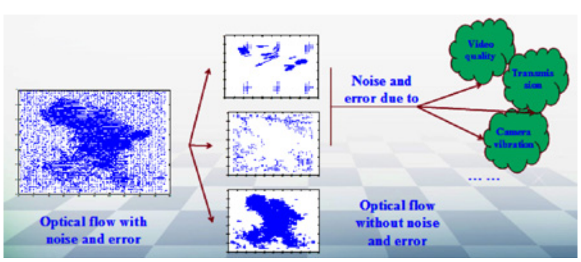

Fig. 7. Noise and error for optical flow.

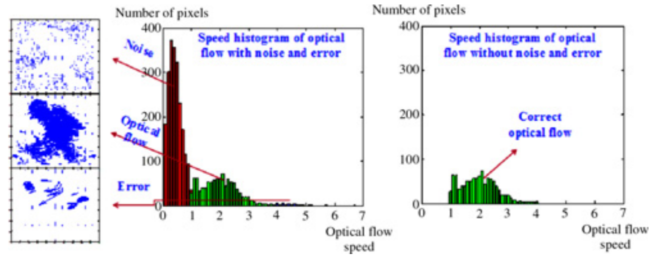

Fig. 8. Optical flow distribution of bending behavior. We compute optical flow using the Horn–Schunck algorithm [34]. Noises, optical flow, and errors can be modeled as a Gaussian distribution.

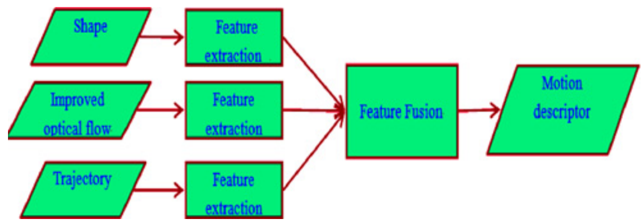

Fig. 9. Framework of motion descriptor.

left and the error is in the right of the distribution. We can remove the noise and error by thresholding the magnitude of the optical flow to get a robust estimation of optical flow [30].

a) Noise due to random perturbations of short optical flow vectors. Magnitude $< VL$.
b) Error due to a random optical flow vector with a large magnitude. Magnitude $> VH$.

The VL and VH stand for the threshold of low and high magnitude. We estimate the thresholds according to the size and speed of normalized blob. So we can reduce noise and error of optical flow with the help of blob (coarse shape) and trajectory.

First, optical flow is normalized in both size and time scale as (1), where $\rho n_k(i, j)$ indicates the normalized speed of optical flow for pixel $\rho_k(i, j)$ in the kth frame. F is the video frame rate according to the compression rate, 25 frames/s in our video data

$$\rho n_k(i, j) = \frac{F \times \rho_k(i, j)}{\sqrt{Hu_k^2 + Wu_k^2}} \tag{1}$$

where Hu_k and Wu_k are the height and width of the union of human blobs in the $(k-1)$th and the kth frames. Then the first derivative of blob position (∇_x, ∇_y) and trajectory (X_{n_k}, Y_{n_k}) in the kth frame are normalized. Third, the threshold $[\alpha\ \beta]$ is determined as

$$[\alpha\ \beta] = \begin{cases} [\alpha 1\ \beta 1], & \sqrt{Xn_k^2 + Yn_k^2} < tt \\ [\alpha 2\ \beta 2], & \sqrt{Xn_k^2 + Yn_k^2} > tt \end{cases} \tag{2}$$

where tt is the criteria of normalized trajectory speed between low and high speed. Here tt is 1.5 by choosing walking and running as reference motion for low speed motion and high speed motion, respectively.

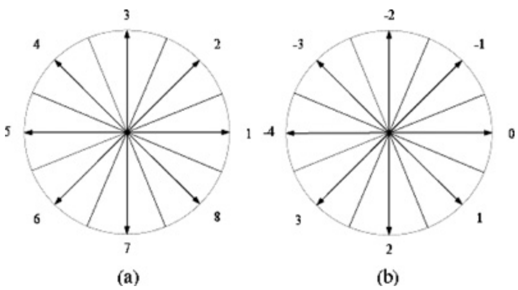

Fig. 10. Direction assignment. (a) Direction histogram bin. (b) Relative majority direction.

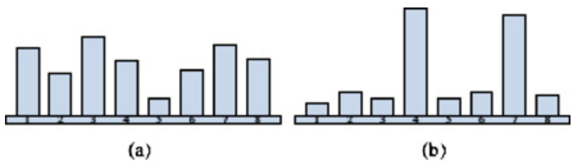

Fig. 11. VDD value. (a) Small VDD implies complex motion pattern. (b) Large VDD implies simple motion pattern.

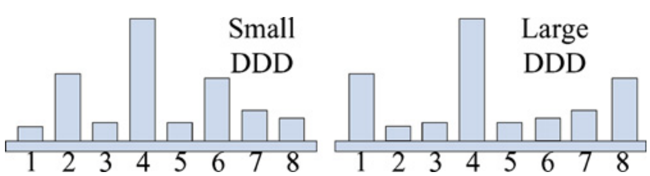

Fig. 12. Divergence of direction distribution. Motion pattern with same MDP may have different DDD.

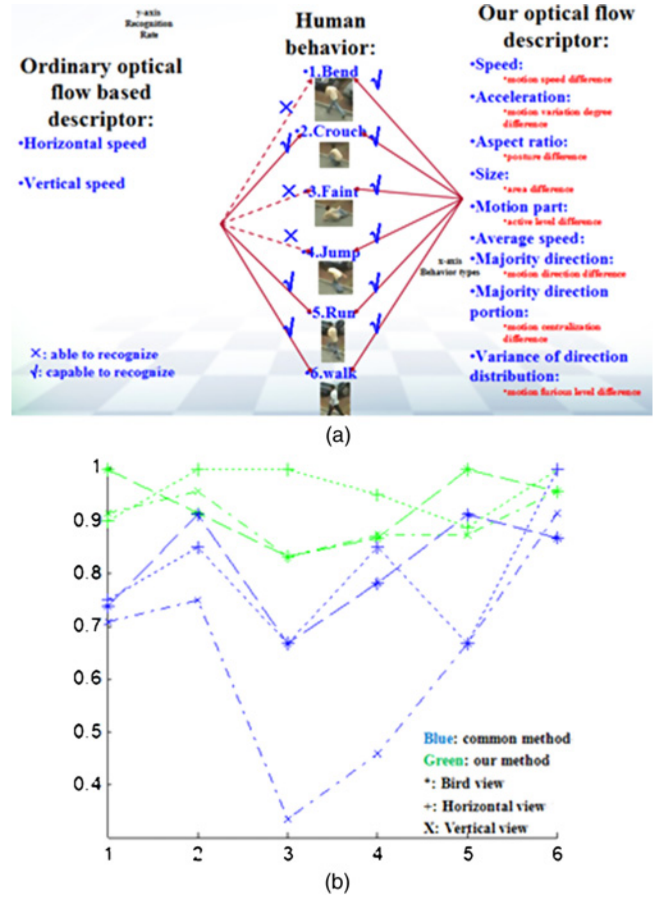

Fig. 13. Comparison of our method with common optical flow feature in three different views. (a) Deference between the original and our optical flow descriptor. (b) Results of two descriptors in three views.

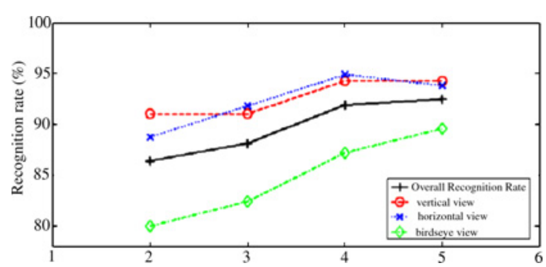

Fig. 14. Test results on our database. Four sets of test results are given for $N = 2, 3, 4, 5$, respectively. The test results are given by sequence.

2) *Motion Descriptor by Fusing Three Cues:* In the above step, we compute optical flow using the Horn–Schunck algorithm [34], then the noises and errors are removed to obtain a robust optical flow feature. However, it is not enough to discriminate many categories behaviors. To improve the discriminative performance, on one hand, we will extract more features from optical flow by considering local patches. On the other hand, we will combine the optical flow with other features such as shape and trajectory as shown in Fig. 9.

a) *Feature extraction from local optical flow computation:* After thresholding, we separate the optical flow of one object into $N \times N$ blocks. Each block is numbered from the 1st to the $(N \times N)$th, the optical flow of the whole blob is labeled as the zeroth block. We then compute the direction histogram of each block with the eight bins as shown in Fig. 10(a) and normalized to get $ND_{ij}, i = 0, ..., N \times N, j = 1, ..., 8$ and $\sum_{j=1}^{8} NDi_j = 1$.

For the optical flow of whole blob and each block, the following statistical features are extracted.

Valid Pixel Portion (VPP): Valid pixels are defined after thresholding. The valid pixel portion is calculated as

$$VPP_i = \frac{\dim\{PV_i\}}{\dim\{PA_i\}} \tag{3}$$

where PV_i is the set of valid pixels in the ith block and PA_i is the set of all pixels in the ith block, dim{} is the operator used to compute the number of pixels.

Average Speed (AS): Instead of computing average optical flow, we make use of average speed because the average optical flow is not accurate to represent the motion speed of a block. A special case is that average optical flow doesn't correctly represent the motion speed level in its region. The average speed is calculated as

$$\mathrm{AS_i} = \frac{1}{\dim\{PV_i\}} \sum_{P(u,v)\in PV_i} p(u, v). \tag{4}$$

Relative Majority Direction (RMD): Majority direction is the direction of valid pixels and is computed as

$$MD_i = \arg\max_{j=1,...,8} \{NDi_j\} \tag{5}$$

TABLE III

RECOGNITION RESULTS FOR N = 5 IN THREE VIEWS. WE ACHIEVE AN AVERAGE RECOGNITION RATE OF 92.49% BY SEQUENCE

Vertical View	*Bend*	*Crouch*	*Faint*	*Jump*	*Run*	*Walk*	Recognition Rate
Bend	22	0	0	0	0	1	95.65%
Crouch	0	24	0	0	0	0	100.00%
Faint	0	2	4	0	0	0	66.67%
Jump	0	1	0	20	1	1	86.96%
Run	0	0	0	0	24	0	100.00%
Walk	0	0	0	1	0	22	95.65%
Horizontal View	*Bend*	*Crouch*	*Faint*	*Jump*	*Run*	*Walk*	Recognition Rate
Bend	20	0	0	0	0	0	100.00%
Crouch	0	19	1	0	0	0	95.00%
Faint	0	0	6	0	0	0	100.00%
Jump	1	0	0	17	1	1	85.00%
Run	0	0	0	2	16	0	88.89%
Walk	0	0	0	0	0	14	100.00%
Birdeye View	*Bend*	*Crouch*	*Faint*	*Jump*	*Run*	*Walk*	Recognition Rate
Bend	22	0	0	2	0	0	91.67%
Crouch	0	22	2	0	0	0	91.67%
Faint	0	1	5	0	0	0	83.33%
Jump	0	0	0	19	1	4	79.17%
Run	2	0	0	1	20	0	86.96%
Walk	0	0	0	0	0	24	100.00%

TABLE IV

ROBUSTNESS TEST AGAINST IRREGULAR ACTIVITIES ON CASIA DATABASE: TABLE SHOWS THE PERCENT OF FRAMES THAT ARE CORRECTLY CLASSIFIED, IN CASES N = 2 AND N = 5

Test Sequences	N = 2				N = 5			
	First Best		Second Best		First Best		Second Best	
Normal walk	Walk	77.55%	Jump	12.24%	Walk	53.06%	Run	30.61%
Walking in a skirt	Walk	100.00%	NA	NA	Walk	89.74%	Bend	5.13%
Carrying briefcase	Walk	100.00%	NA	NA	Walk	98.96%	Jump	1.04%
Limping man	Walk	100.00%	NA	NA	Walk	94.38%	Bend	3.37%
Occluded legs	Walk	100.00%	NA	NA	Walk	94.55%	Run	3.64%
Knees up	Walk	100.00%	NA	NA	Walk	87.50%	Jump	10.71%
Walking with a dog	Walk	93.75%	Run	6.25%	Walk	93.75%	Run	6.25%
Sleep walking	Walk	83.33%	Jump	10.42%	Walk	87.50%	Run	10.42%
Swinging a bag	Walk	100.00%	NA	NA	Walk	100.00%	NA	NA
Occluded by a "pole"	Walk	85.11%	Run	14.89%	Walk	89.36%	Run	10.64

TABLE V

ROBUSTNESS TEST AGAINST IRREGULAR ACTIVITIES ON WEIZMANN DATABASE: THE TABLE SHOWS THE PERCENT OF FRAMES THAT ARE CORRECTLY CLASSIFIED, IN CASES N = 2 AND N = 5

Test Sequences	N = 2				N = 5			
	First Best		Second Best		First Best		Second Best	
Normal walk	Walk	56.25%	Jack	20.83%	Walk	75.00%	Jack	10.42%
Walking in a skirt	Walk	68.09%	Side	25.53%	Walk	95.74%	Skip	4.26%
Carrying briefcase	Walk	51.28%	Side	43.59%	Walk	76.92%	Side	23.08%
Limping man	Walk	92.73%	Bend	3.63%	Walk	89.09%	Wave1	9.09%
Occluded legs	Walk	66.67%	Side	16.67%	Walk	93.75%	Skip	4.17%
Knees up	Walk	91.01%	Wave1	6.74%	Walk	86.52%	Wave1	10.11%
Walking with a dog	Walk	94.79%	Jack	3.13%	Walk	88.54%	Jack	8.33%
Sleep walking	Walk	58.93%	Side	25.00%	Walk	76.79%	Jump	16.07%
Swinging a bag	Walk	79.59%	Side	12.24%	Walk	91.84%	Jack	2.04%
Occluded by a "pole"	Walk	92.86%	Side	7.14%	Walk	100.00%	NA	NA

where MD_i is the majority direction of the ith block. Relative majority direction of each block is assigned relative to the whole optical flow majority direction as shown in Fig. 10(b). This feature represents the motion direction of the block relative to the whole body motion direction and is calculated as

$$RMD_i = \text{mod}(MD_i - MD_0) - 8 \times (\text{mod}(MD_i - MD_0, 8) \geq 4) \quad (6)$$

where MD_0 is the majority direction and $\text{mod}(MD_i - MD_0)$ can get value from 0 to 7, the latter part can get 0 or 8, then the RMD can be obtained from −4 to 3 as shown in Fig. 10(b).

Majority Direction Portion (MDP): This feature describes the motion direction the motion in the corresponding block is. It is calculated as

$$MDP_i = \max_{j=1,\ldots,8}\{NDi_j\}. \quad (7)$$

Variance of Direction Distribution (VDD): The direction histogram describes the direction distribution and this feature represents how complex the motion pattern is in the corresponding block as shown in Fig. 11. It is calculated as

$$VDD_i = \frac{1}{8}\sum_{j=1}^{8}(NDi_j - \overline{NDi})^2. \quad (8)$$

Divergence of Direction Distribution (DDD): The divergence of direction distribution is an auxiliary feature for MDP as shown in Fig. 12 and is calculated as

$$DDD_i = \sum_{j=1}^{8} NDi_j \times RMD\{(j - \arg\max_{l=1,\ldots 8}\{NDi_l\})\}^2 \quad (9)$$

where the $RMD\{\}$ indicates the mapping method as mentioned in calculating RMD.

b) *Features fusion:* Besides the above local features from optical flow, we additionally employ some assistant features from shape and trajectory—blob size as $H_k \times W_k$, blob ratio W_k/H_k, acceleration of trajectory in the vertical direction as $\nabla^2 Y_k$. In the fusion stage, for simplification, we just consider the linear combination mode [32], and the weight of every feature is the same. Other complex fusion methods can also be considered, while it is not the task here. Then we have the final motion representation of (5 + *6N2* + 3) dimensions, which is much smaller than the dimensions of original optical flow. Compared with common optical flow-based descriptors, our motion descriptor is discriminative and effective in different three views on our behavior database as shown in Fig. 13.

c) *Classifiers:* Many supervised learning algorithms can be employed to train a behavior pattern recognizer. Support vector machine (SVM) [35] is used in our approach. SVM has been successfully applied to a wide range of pattern recognition and classification problems because it is fast and deterministic. The concatenation of features obtained above is fed as a feature vector into support vector machine. The radial basis function $k(x, y) = \exp(-\lambda \|x - y\|)$ is utilized to map training vectors into a high dimensional feature space for classification.

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
jack	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
jump	.00	.00	.78	.00	.00	.00	.22	.00	.00	.00
pjump	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
run	.00	.00	.00	.00	.89	.00	.11	.00	.00	.00
side	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00
skip	.00	.00	.22	.00	.11	.00	.67	.00	.00	.00
walk	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00
wave1	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00
wave2	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0

Fig. 15. Sequence classification on Weizmann database. All results are from nine runs in a leave-one-out procedure with $N = 4$. The method correctly classifies 93.33% of all testing sequences.

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	.85	.04	.00	.03	.00	.00	.00	.00	.05	.03
jack	.02	.92	.00	.04	.00	.00	.00	.00	.02	.01
jump	.00	.00	.68	.01	.02	.05	.17	.06	.00	.00
pjump	.05	.04	.00	.90	.00	.00	.00	.00	.01	.00
run	.00	.00	.02	.00	.75	.04	.19	.00	.00	.00
side	.00	.00	.08	.00	.05	.76	.05	.05	.00	.00
skip	.00	.00	.22	.01	.21	.07	.47	.02	.00	.00
walk	.01	.01	.01	.00	.00	.01	.01	.95	.00	.00
wave1	.09	.02	.00	.00	.00	.00	.00	.00	.84	.05
wave2	.03	.01	.00	.00	.00	.00	.00	.00	.08	.88

Fig. 16. Frame classification on Weizmann database. All results are from nine runs in a leave-one-out procedure with $N = 4$. The method correctly classifies 82.37% of all testing frames.

Fig. 17. Irregular walking sequences, from left to right and top to bottom: swinging a bag, walking with a briefcase, walking with the knees up, walking with a limp, sleepwalking, occluded feet, normal walking, occluded by a "pole," walking in a skirt, walking with a dog.

TABLE VI
ROBUSTNESS TEST AGAINST HORIZONTAL VIEWPOINT CHANGE ON CASIA DATABASE

Test Sequence	First Best		Second Best	
Walking in 0°	Walk	98.53%	Run	1.47%
Walking in 5°	Walk	100.00%	NA	NA
Walking in 10°	Walk	100.00%	NA	NA
Walking in 15°	Walk	100.00%	NA	NA
Walking in 20°	Walk	100.00%	NA	NA
Walking in 25°	Walk	98.80%	Crouch	1.20%
Walking in 30°	Walk	98.53%	Crouch	1.47%
Walking in 40°	Walk	62.73%	Crouch	24.55%
Walking in 45°	Walk	79.55%	Crouch	12.50%

V. EXPERIMENTAL RESULTS AND ANALYSIS

Our approach has been evaluated on two databases, one is CASIA multiview behavior database and another is a publicly available database (Weizmann action database [18]). The experimental results show the improvement of our approach over the baseline method.

A. Results on CASIA Behavior Database [33]

We test our method on this database with N = 2, 3, 4, 5. The test results are shown in Fig. 14. It is apparent that the results for the vertical and horizontal viewpoints are better than those obtained from the bird's eye view. It is reasonable given that people are smaller and many behavior types look similar from the bird's eye view. The recognition rate increases as N increases, but for larger N the improvement of recognition rate is small. From Table III, our method performs well in different views (overall recognition rate: horizontal view 95%, vertical view 90%, bird's eyes view 89%), which verified that the proposed method is robust to viewpoint variation. It is reasonable that the results in horizontal view achieve best because shape information can be extracted better. The results from vertical view and bird's eye view are similar as the shape information is not helpful for these views, while optical flow and trajectory information are both extracted better. To improve the discriminative ability of optical flow, we improve it from two aspects: 1) optical flow computation, especially the noise removal, and 2) the more information is considered as trajectory and shape information, which could provide the complementary information from different viewpoints.

B. Results on Weizmann Database [18]

For comparisons, we follow the leave-one-out strategy: video clips of one subject are kept as testing data and other video clips are training data. We evaluate the performance of our method in frame-by-frame classification as well as video sequence classification. The confusion tables are shown in Figs. 15 and 16 when N = 4. Compared with Niebles *et al.*'s result [31] of 72.8% by sequence and 55.0% by frame, our method obtained better performance of 93.3% by sequence and 82.37% by frame[1]. It is noticed that confusions are mostly among *jump*, *run* and *skip* behaviors, which is reasonable because they are very similar to each other. The method by Black *et al.* [18] performs very well (97.5%) on this database, while this method considers the space-time shape as feature, which costs high computation in solving the Poisson equation and extracting features (110 × 70 × 50 videos need about 30 s on a P 4.3 GHz computer with MATLAB language). Our method can process a frame of size 320 × 240 about 6 ms with computer of P4 3.0 GHz CPU and 1.5 GB RAM. This processing time is sufficiently low to allow real-time applications.

[1]It should be mentioned that Niebles *et al.* [31] used the unsupervised learning method. The video sequence is represented by spatial-temporal patches and this method does not use background subtraction.

C. Robustness Evaluation

To evaluate the robustness of our method, we consider both irregular activities and the change of horizontal view in behavior recognition.

D. Robustness Test Via Irregular Activities

The database of irregular activities [18] shown in Fig. 17, which includes nine irregular walking sequences under different conditions and one normal walking sequence.

We test the robustness on the irregular activities database. Tables IV and V give results when N = 2 and N = 5 on CASIA database and Weizmann database. From Table IV, the overall recognition rates are 94% and 89% when N = 2 and 5 on CASIA database. From Table V, the overall recognition rates are 75% and 88% when N = 2 and N = 5 on Weizmann database. The "walking" behavior can be recognized correctly both on CASIA database and Weizmann database, which shows that our method is robust against irregular activities of various types.

E. Robustness Test Via Horizontal Viewpoint Changes

The database of horizontal viewpoint changes contains walking sequences with horizontal viewpoint change of {0°, 5°, 10°, 15°, 20°, 25°, 30°, 40° and 45°}, respectively. Fig. 18 shows some example frames for each viewpoint.

Tables 6 and 7 show the classification results when N = 3. Other values of N lead to similar results. As we can see, the recognition rates of "walk" behavior are 91% and 80% on CASIA database and Weizmann database, which shows that our method is robust against horizontal view changes. On both databases, the results of the last two test sequences are worse

TABLE VII
ROBUSTNESS TEST AGAINST HORIZONTAL VIEWPOINT CHANGE ON WEIZMANN DATABASE

Test Sequence	First Best		Second Best	
Walking in 0°	Walk	89.71%	Side	8.82%
Walking in 5°	Walk	95.31%	Jump	1.56%
Walking in 10°	Walk	98.46%	Side	1.54%
Walking in 15°	Walk	96.05%	Side	2.63%
Walking in 20°	Walk	94.81%	Jump	2.60%
Walking in 25°	Walk	87.95%	Jump	6.02%
Walking in 30°	Walk	73.53%	Bend	11.76%
Walking in 40°	Walk	33.64%	Jump	21.82%
Walking in 45°	Jump	48.86%	Jack	17.05%

Fig. 18. Horizontal viewpoint change data. From top to bottom and left to right are horizontal viewpoint changes of {0°, 5°, 10°, 15°, 20°, 25°, 30°, 40°, and 45°}, respectively.

than other results because the shape information from these two viewpoints (40° and 45°) cannot be helpful.

VI. CONCLUSION

Behavior analysis is important for many applications such as visual surveillance and human computer interaction. View invariant behavior analysis has become a hot topic in recent years. In this paper, based on the analysis of features for behavior analysis, we proposed a novel motion descriptor based on improved optical flow for view invariant behavior analysis. We improve the optical flow by first removing noises and errors removal and then fusing the optical flow information with trajectory and shape information.

To evaluate our method, we have tested on CASIA database (ranging over 11 behaviors, with three views for each type of behavior) and Weizmann behavior database. The experimental results show the advantages of our method: It is real-time with good classifying performance; it is effective from three different viewpoints and robust against horizontal viewpoint change; it is also robust against irregular activities under varying conditions. In the future work, we will continue to investigate how to evaluate the effectiveness of different features in the fusion step and improve the fusion algorithms.

ACKNOWLEDGMENT

The authors would like to thank their anonymous reviewers, and Associate Editor S. Pankanti, for their valuable feedback.

REFERENCES

[1] I. Haritaoglu, D. Harwood, and L. Davis, “W4: Who? When? Where? What? A real-time system for detecting and tracking people,” in *Proc. 3rd Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 222–227.

[2] C. R. Wren, A. Azarbayejani, T. J. Darrell, and A. P. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.

[3] *Project: The European Taskforce Creating Human–Machine Interfaces Similar to Human–Human Communication* [Online]. Available: http://www.similar.cc

[4] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias, “Efficient content-based retrieval of humans from video databases,” in *Proc. 2nd Int. Workshop Recognit. Anal. Tracking Faces Gestures Real-Time Syst.*, 1999, pp. 89–95.

[5] N. Gehrig, V. Lepetit, and P. Fua, “Golf club visual tracking for enhanced swing analysis,” in *Proc. British Machine Vision Conf.*, 2003, pp. 1–10.

[6] M. Köhle, D. Merkl, and J. Kastner, “Clinical gait analysis by neural networks: Issues and experiences,” in *Proc. 10th IEEE Symp. Comput.-Based Med. Syst.*, 1997, pp. 138–143.

[7] P. Maes, T. J. Darrell, B. Blumberg, and A. P. Pentland, “The ALIVE system: Wireless, full-body interaction with autonomous agents,” *ACM Multimedia Syst.*, vol. 5, no. 2, pp. 105–112, Mar. 1997.

[8] L. Wang, W. M. Hu, and T. N. Tan, “Recent developments in human motion analysis,” *Pattern Recognit.*, vol. 36, no. 3, pp. 585–601, Jan. 2003.

[9] V. Parameswaran and R. Chellappa, “View invariance for human action recognition,” Int. J. comput. Vision, vol. 66, no. 1, pp. 83–101, Jan. 2006.

[10] I. Cohen and H. Li, “Inference of human postures by classification of 3-D human body shape,” in *Proc. IEEE Int. Workshop Automatic Face Gesture Recognit.*, 2003, pp. 74–81.

[11] S. Carlsson and J. Sullivan, “Action recognition by shape matching to key frames,” in *Proc. IEEE Comput. Society Workshop Models Versus Exemplars Comput. Vision*, 2002, pp. 263–270.

[12] K. Sato and J. K. Aggarwal, “Recognizing two-person interactions in outdoor image sequences,” in *Proc. IEEE Workshop Multiobject Tracking*, 2001, pp. 87–94.

[13] S. Park and K. Aggarwal, “A hierarchical Bayesian network for event recognition of human actions and interactions,” *ACM J. Multimedia Syst.*, vol. 10, no. 2, pp. 164–179, Aug. 2004.

[14] L. Zelnik-Manor and M. Irani, “Event-based analysis of video,” in *Proc. IEEE Comput. Vision Pattern Recognit.*, vol. 2. Dec. 2001, pp. 123–130.

[15] E. Shechtman and M. Irani, “Space-time behavior-based correlation,” in *Proc. Comput. Vision Pattern Recognit.*, vol. 1. 2005, pp. 405–412.

[16] A. Yilmaz and M. Shah, “Action sketch: A new method to recognition,” in *Proc. Comput. Vision Pattern Recognit.*, 2005, pp. 984–989.

[17] I. Laptev and T. Lindeberg, “Space-time interest points,” in *Proc. Int. Conf. Comput. Vision*, vol. 1. 2003, pp. 432–440.

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

[19] A. R. Madabhushi and J. K. Aggarwal, "Using movement to recognize human activity," in *Proc. Int. Conf. Image Process.*, 2000, pp. 698–701.

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

[21] P. C. Ribeiro and J. S. Victor, "Human activity recognition from video: Modeling, feature selection and classification architecture," in *Proc. Workshop Human Activity Recognit. Modeling*, 2005, pp. 61–70.

[22] N. Robertson and I. Reid, "Behavior understanding in video: A combined method," in *Proc. Int. Conf. Comput. Vision*, 2005, pp. 808–815.

[23] A. Galata, A. Cohn, D. Magee, and D. Hogg, "Modeling interaction using learnt qualitative spatio-temporal relations and variable length Markov models," in *Proc. Eur. Conf. Artif. Intell.*, 2002, pp. 741–745.

[24] N. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[25] A. Verri and T. Poggio, "Motion field and optical flow: Qualitative properties," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 5, pp. 490–498, May 1989.

[26] T. Nakata, "Recognizing human activities in video by multiresolutional optical flows," in *Proc. IEEE (RSJ) Int. Conf. Intell. Robots Syst.*, 2006, pp. 1793–1798.

[27] A. Efros, G. Berg, A. C. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. Comput. Vision Pattern Recognit.*, 2003, pp. 726–733.

[28] F. Niu and M. Abdel-Mottaleb, "View-invariant human activity recognition based on shape and motion features," in *Proc. IEEE 6th Int. Symp. Multimedia Softw. Eng.*, Dec. 2004, pp. 546–556.

[29] N. Robertson and I. Reid. "A general method for human activity recognition in video," *Comput. Vision Image Understanding*, vol. 104, no. 2, pp. 232–248, Nov. 2006.

[30] L. Hongche, T.-H. Hong, M. Herman, and R. Chellappa, "Accuracy versus efficiency trade-offs in optical flow algorithms," in *Proc. Eur. Conf. Comput. Vision*, vol. 2. Apr. 1996, pp. 174–183.

[31] J. C. Niebles, H. C. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, Sep. 2008.

[32] C. V. Christopher and W. C. Garrison, "Fusion via a linear combination of scores," *Information Retrieval*, vol. 1, no. 3, pp. 151–173, Oct. 1999.

[33] *CASIA Behavior Database* [Online]. Available: http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp

[34] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, 1981.

[35] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

Kaiqi Huang (SM'09) received the B.S and M.S. degrees, both in electrical engineering, from the Nanjing University of Science and Technology, Nanjing, China, in 1998 and 2001, and the Ph.D. degree in signal and information processing from Southeast University, Nanjing, China, in 2004.

After receiving the Ph.D. degree, he became a Postdoctoral Researcher with the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, where he is currently an Associate Professor. He is the author of more than 40 conference papers. His research interests include visual surveillance, image and video analysis, human vision and cognition, and computer vision, among others.

Dr. Huang is a Program Committee Member of more than 30 international conferences and workshops, and is a Board Member of the IEEE Systems, Man, and Cybernetics Technical Committee on Cognitive Computing. He is the Deputy Secretary-General of the IEEE Beijing Section.

Shiquan Wang received the B.S. degree in electrical engineering in 2006, in video processing and multimedia communication from Chengdou University, Chengdou, China. He is currently a Ph.D. student in pattern recognition and intelligent systems, the National Laboratory of Pattern Recognition at the Chinese Academy of Sciences, Beijing, China.

His current research interests include computer vision, pattern recognition, and human behavior analysis, among others.

Tieniu Tan (F'03) received the B.S. degree in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. and Ph.D. degrees in electronic engineering from the Imperial College of Science, Technology, and Medicine, London, U.K., in 1986 and 1989, respectively.

In October 1989, he joined the Computational Vision Group, Department of Computer Science, University of Reading, Reading, U.K., where he was a Research Fellow, Senior Research Fellow, and Lecturer. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Sciences, Beijing, China, where he is currently a Professor and the Director of the NLPR. He is the author of more than 200 research papers in refereed journals and conferences in the areas of image processing, computer vision, and pattern recognition. His current research interests include image processing, machine and computer vision, pattern recognition, multimedia, and robotics.

Dr. Tan was a Guest Editor of the International Journal of Computer Vision (June 2000), and is an Associate Editor or Member of the editorial board of eight international journals, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, AND PATTERN RECOGNITION. He serves as a Referee or Program Committee Member and Chair for many major national and international journals and conferences. He is the Chair of the International Association for the Pattern Recognition Technical Committee on Signal Processing for Machine Intelligence, and the Chair of the Fellow Committee of the IEEE Beijing Section. He currently serves as the Acting President of the Chinese Society of Image and Graphics, and the Deputy President of the China Computer Federation and the Chinese Automation Association.

Stephen J. Maybank (SM'06) received the B.A. degree in mathematics from King's College, Cambridge, U.K., in 1976, and the Ph.D. degree in computer science from Birkbeck College of London University, London, U.K. in 1988.

He joined the Pattern Recognition Group, Marconi Command and Control Systems, Frimley, U.K., in 1980 and moved to the GEC Hirst Research Center, Wembley, U.K., in 1989. From 1993 to 1995, he was a Royal Society and Engineering and Physical Sciences Research Council Industrial Fellow with the Department of Engineering Science, University of Oxford, Oxford, U.K. In 1995, he joined the University of Reading, Reading, U.K. as a Lecturer at the Department of Computer Science. In 2004, he joined the School of Computer Science and Information Systems at the Birkbeck College of London University, where he is currently a Professor. His research interests include the geometry of multiple images, camera calibration, visual surveillance, information geometry and the applications of statistics to computer vision.

Dr. Maybank is a Fellow of the Royal Statistical Society, a Fellow of the Institute of Mathematics and its Applications, a Member of the British Machine Vision Association, and a Member of the Société Mathématique de France.