**Vs-star: A visual interpretation system for visual surveillance**

**Kaiqi Huang**, Tieniu Tan

**National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China**

**ABSTRACT**

In recent years, intelligent visual surveillance has become more and more important for enhanced security. In this paper, we will introduce some recent work in image and video understanding. First, we will give an introduction of the related video surveillance system in recent years, in particular, we will describe algorithms and systems developed in our group for the automatic interpretation of human and vehicle motion in surveillance videos, where automatic interpretation involves object motion detection, object classification and recognition, object tracking and the analysis of object behaviors in order to detect abnormal behaviors. We also give some examples of the real applications of Vs-star system.

© 2010 Published by Elsevier B.V.

**1. Introduction**

Security has become a major world-wide concern since the event of September 11, 2001 in USA and the bomb attacks in London on the July 2005. Video surveillance is a critical component of any effective security system. Most current video surveillance systems are monitored by relatively small teams of human operators. Even though there may be a very large number of cameras. Typically a human watches a set of screens which cycle from one camera to another every few seconds. In addition to problems of fatigue and boredom, the human attention span is limited both spatially and temporally. Studies have shown that the average human can attentively track the movements of up to four independent targets simultaneously and can detect changes to the targets but not any change to nearby parts of the background (Sears and Pylyshyn, 2000). When targets and distractors are too close it becomes difficult to separate out the targets and maintain tracking. Also, if the targets move too fast, the common people are unable to track them accurately (Cavanaugh and Alvarez, 2005). To overcome these limitations the computer vision and pattern recognition field is developing automated systems for the real-time monitoring of humans and vehicles. These systems can interpret the events in the camera and generate an alarm if a suspicious person or an abnormal activity is detected.

In this paper, we present some of our recent work in image and video understanding using the CASIA visual surveillance system for enhanced security. In particular, we will describe algorithms and systems for the automatic interpretation of human and vehicle motion in surveillance videos, where automatic interpretation involves object detection, object classification and recognition, object tracking and the analysis of object behaviors.

**2. Related work**

In recent years, many automated vision based surveillance systems have been proposed, motivated by commercial, military and industrial applications. Here we give a brief review of these systems. Interested readers can be referred to Aggarwal and Cai (1999) and Collins et al. (2000) for detailed surveys.

The VIEWS system (Corrall, 1991) is one of the earliest video surveillance systems for describing scene activity. It is designed to operate in controlled environments where a great deal of information can be specified in advance. This information includes camera models, ground plane representation, salient regions of the environment, 3-D object models and behavior models. The VIEWS system requires significant computational power to achieve real-time performance.

An ambitious system, VSAM (Video Surveillance and Monitoring) (Collins et al., 2000) provided the capability to detect, track, localize and visualize objects within a known environment surveyed by a distributed network of cameras. Moving objects are classified as human, human group or vehicle. The motions of individual humans are classified as running or walking.

Among the earlier automated monitoring systems, Pfinder (Wren, 1997) is perhaps the most well known. The full body of person is tacked in a scene that contains only one unoccluded person in an upright posture. It uses a unimodal background model to locate the moving person.

For military applications, Boul et al. (2001) present a system for monitoring uncooperative and camouflaged targets. They focus on low level analysis such as object detection and tracking.
The W4 system (Haritaoglu et al., 2000) uses dynamic appearance models to track people. Single person and group are distinguished using projection histograms. The heads of the people in a group are tracked individually.

Fujioyshi et al. (2001) developed a system to detect and track multiple people and vehicles in a cluttered scene and monitor activities over a large area and extended periods of time. Their system could also classify objects as a person, group of peoples, vehicles, and so on, using shape and color information.

Recently, Shah et al. (2007) propose an automatic surveillance system named Knight. The Knight system cannot only detect, track and classify objects in a realistic scene surveyed by a camera network, but also can flag significant events and present a summary in terms of key frames and a textual description of observed activities. In contrast with VSAM, Knight is being developed for commercial application.

It is deserved to be mentioned that some work is not introduced by system while they play important role in the automated surveillance system. Stauffer and Grimson (2000) used an adaptive multimodal background subtraction method named Adaptive Mixture Gaussian Model (GMM) for object detection that can deal with slow changes in illumination, repeated motion from background clutter, and long-term scene changes. Their background subtraction method performed well in realistic scenes and was adopted by many systems. They also proposed the detection of unusual activities by statistically learning the common patterns of activities over time. They tracked detected objects using a multiple hypothesis tracker. Ricquebourg and Bouthemy (2000) proposed people tracking algorithm by exploiting spatiotemporal slices. Their detection scheme involves the combination of intensity, temporal differences between three successive images, and a comparison between the current image and a background reference image, which is reconstructed and updated online.

There are many other systems with specific features adapted to different requirements. These systems include the RoboGuard system (Birk and Kenn, 2001), which focuses on visual interpretation for surveillance, while platform, the IBM Face Cataloger system (PeopleVision: IBM research, xxxx), which records the person tracked by a tracker and identified the person using a frontal view of face obtained by a pan-tile-zoom (PTZ) camera.

In this paper, we will also introduce one automatic visual surveillance system for visual interpretation of human and vehicle motion named Vs-star system, which is for the purpose of research and some components have been used for real applications.

3. Vs-star: Casia visual surveillance system and algorithms

We have developed a visual interpretation system for visual surveillance named visual surveillance star (Vs-star). The purpose of Vs-star is to describe the motions of humans and vehicles in a surveillance scene and give a warning if these motions indicate a breach of security. Fig. 1 shows the framework of Vs-star. Compared with other systems for commercial applications, our framework focuses on core topics for research, also some real applications of techniques will be introduced at the end of this paper. Vs-star integrates low level object detection and tracking, high level activity analysis and semantic interpretation. In the next section, we will introduce the components of Vs-star shown in Fig. 1.

3.1. Moving object detection

Moving object detection is a critical early step in video surveillance. The moving objects provide a focus of attention for later processes such as tracking and object classification. One convenient and effective way for moving object detection is background subtraction. This method constructs a reference background model, and then, takes difference between the current image and background model to detect moving object. Therefore, an accurate and robust background model is prerequisite.

Much work has been done on this field (Wren, 1997; Stauffer and Grimson, 2000; Huang, 2006; Karmann and Von Brandt, 1990; Koller et al., 1994; Elgammal et al., 2002; Radke et al., 2005). Gaussian Mixture Model (GMM) (Stauffer and Grimson, 2000) is one of the widely used method for moving object detection. It models the time series of color observations at a given pixel as a mixture of Gaussians, and a subset of the Gaussians is used to describe the scene background. The parameters for each Gaussian are updated as new observations are obtained, so it can adapt to gradual background changes. In (Elgammal et al., 2002), the authors use the Kernel Density Estimation (KDE) method to represent the color distribution for each pixel. Some other methods are proposed based on these two methods (Wang et al., 2006; Sheikh and Shah, 2005). Wang et al. (2006) propose a dynamic conditional random field model for object and moving cast shadow segmentation. Sheikh and Shah (2005) directly modeled the dependencies between the domain (location) and range (color) by using a non-parametric density estimation method over a joint domain-range representation of image pixels.

Due to the self learning capacity to variations in lighting, periodic dynamic scene, GMM as the background model is considered in many surveillance systems and also in our framework. However, it still has some shortcomings. The number of Gaussians should be decided beforehand. Another limitation is that it does not model the spatial dependencies of neighboring background pixel colors explicitly. Therefore, some false positive pixels will be produced in dynamic scenes where dynamic texture does not repeat exactly. In Vs-star system, we propose an on-line auto-regressive model to capture and predict the behavior of dynamic scenes as in Fig. 2 (Zhou et al., 2008).

The proposed method has three novel contributions:

(1) A probabilistic framework for moving object segmentation in dynamic scene is proposed. Under this framework, we unify background modeling, moving object detection and shadow removing by constructing probability density functions (PDFs) of background, shadow and foreground based on an extended GMM. The method in (Rittscher et al., 2000) also considers background, foreground and shadow based on Hidden Markov Model (HMM). However, parameters of this model are learned off-line and this method fails to distinguish shadow from dark vehicles, which results in poor parameter estimation.

(2) Extended GMM which model spatial dependencies of neighboring pixel colors explicitly is considered in our system. It is inspired by the work of Sheikh and Shah (2005). Moreover, the extended GMM can determine the number of Gaussian components dynamically by Gaussian merging and deleting rules.

(3) We also consider feedback of tracking results to accurate PDF of foreground. The authors (Sheikh and Shah, 2005) also construct PDF of foreground, but tracking information is neglected.

The images as Fig. 3 show the results with Extended GMM on the scene with wind, which caused the clutter background. On this scene, shadow is also distinctive. Ground truth is given manually. It is evident that swaying tree deteriorates the performance of traditional GMM as shown in the forth and last rows of Fig. 3. While with our method, the contour of shadow has been removed and that majority of hollows are filled by our method. The second and the third row show indoor scene, where the nominally moving
camera was simulated by moving the original pictures left and right (motion is about 4 pixels). The shadow here is insignificant. It shows that a slight movement will also cause substantial degradation in performance of GMM, especially on the neighbor of edge. Fig. 4 shows the per-frame moving object detection rates according to precision and recall for the video of Fig. 3. Precision and recall are defined as:

\[
\text{Precision} = \frac{\text{\# of true positive detected}}{\text{total \# of true positives}},
\]

\[
\text{Recall} = \frac{\text{\# of true positive detected}}{\text{total \# of true positives}}.
\]

The results in Fig. 4 are shown for three different learning rates of traditional GMM. The shadow removing method is Cucchiara et al. (2001). These figures demonstrate that our method can deal with dynamic scene and cast shadow effectively considering spatial information and PDF of shadow.

3.2. Object tracking

Object tracking, as the following step of moving object detection, has always been a hot topic in the field of surveillance. The aim of object tracking is to locate the moving objects in the whole

![Diagram](image-url)
video sequences so that the complete trajectory of an object can be obtained. There are mainly two major steps of object tracking: target representation and data association (Gilbert and Bowden, 2006).

Although substantial work has been done on object tracking, there are still many problems deserved further investigation. Firstly, occlusion due to self-occlusion or occlusion between different objects brings uncertainties into the data association. Secondly, tracking objects can be difficult due to loss of information caused by projection of the 3D world onto a 2D image plane. Thirdly, as the need of wide area visual surveillance, the use of multiple cameras also brings challenges into object tracking. Next, we will give a brief introduction to the tracking methods in our group.

According to the number of cameras in surveillance scenarios, we divide object tracking problem into two subset problems: intra-camera tracking (single camera tracking) and inter-camera tracking (multi-camera tracking).

3.2.1. Intra-camera tracking

In intra-camera tracking, the inputs to the tracking module are foreground regions corresponding to the moving objects. Background subtraction algorithm (Cucchiara et al., 2001) is applied to extract the foreground regions. According to the representation of moving objects, tracking methods can be divided into four major categories: region-based tracking, active-contour-based tracking, feature-based tracking, and model-based tracking. It should be
pointed out that this classification is not absolute and the algorithms from different categories can be integrated together. Here, feature is low level information such as color and so on, region and active contour are middle level and model can be high level for tracking, active contour can be looked as one feature and we can also get features from region, so we introduced the tracking algorithms from feature-based and model-based methods.

3.2.1.1. Feature-based tracking. Feature-based tracking algorithms perform recognition and tracking of objects by extracting elements, clustering them into higher level features and then matching the features between images. In our system, we employ “color” as the feature for tracking which is a global feature. To deal with occlusion, we introduce a Bayesian network which involves human 2D ellipse models and an extra hidden process for occlusion relation. In contrast to previous work, the appearance of the target is modeled as a statistical color spatial model other than an appearance template. This color spatial model captures the multi-modal patterns of color of the target. The human models are initialized when people are firstly detected as targets and then updated at each new frame before occlusion. The observation likelihood is then proposed based on the Bhattacharyya distance to people tracking through occlusions. The tracking process is performed using a Condensation algorithm. Fig. 5 illustrates key frames where two people are tracked through occlusions. More results can be found in Min et al. (2004), Lou et al. (2002).

In general, as these algorithms operate on 2-D image planes, feature-based tracking algorithms can work in real-time and track multiple objects which are required in heavy threeway scenes, etc. However, there are several serious deficiencies in feature-based tracking algorithms:

- The recognition rate of objects based on 2-D image features is low, because of the non-linear distortion during perspective projection and the image variations with the viewpoint’s movement.
- These algorithms are generally unable to recover 3-D pose of objects.
- The stability of dealing effectively with occlusion, overlapping and interference of unrelated structures is generally poor.

Model-based tracking in next part provides an effective solution to above problems.

3.2.1.2. Model-based tracking. Model-based tracking algorithms track objects by matching projected 3D object models, which are produced with prior knowledge. The models are usually constructed off-line with manual measurement, such as Computer Aided Design (CAD) tools or computer vision techniques. According to difference of human (non-rigid object) and vehicle (rigid object), we model and track these two types of objects separately (Huazhong et al., 2004; Huazhong et al., 2004). For human, we construct human body models by articulated truncated cones and a sphere as a trade-off between accuracy and complexity as Fig. 6. For simplicity and efficiency, the human body model is represented by a 12-dimensional state vector composed of the global position of the human body and relative positions of different body parts. Fig. 7 shows tracking results in complex real-world outdoor scenes. For vehicle model, 3D wire-frame vehicle models are mainly adapted (Tan and Baker, 2000; Tan et al., 1998) and an improved extended Kalman filter is used for tracking (Jianguang et al., 2005), which outperforms the traditional extended Kalman filter when the vehicle carries out a complicated maneuver. The 3D model-based tracking algorithms can be applied even when objects greatly change their orientations during the motion.

3.2.2. Inter-camera tracking

With the increasing number of the cameras (overlap or non-overlap) in surveillance, multi-camera tracking becomes more and more important. Main issues of multi-camera tracking include camera installation, calibration of multiple cameras, correspondence between multiple cameras, automated camera switching and data fusion, etc. We focus on correspondence between multiple overlapped and non-overlapped cameras. Firstly, multiple cameras are used for improving the tracking of people when people are within the common ground plane of the multiple views, such as in Fig. 9. Secondly, people are tracked continuously from one camera to another. Cameras are widely separated in this case. Experimental setup is shown in Fig. 12.

3.2.2.1. Principal axis-based correspondence between multiple cameras for people tracking. For overlapped cameras, the principle axis of each person is proposed to match people across multiple views (Weiming et al., 2006). Estimation of principle axes does not rely on accurate motion detection since the influence of the imperfection of motion detection is counteracted by the symmetrical distributions of the error along the principle axis, which makes principle axes an appealing attribute in real applications. The principle axis of an isolated person is determined by minimizing the median of squared perpendicular distance between the foreground pixels and a vertical axis in Fig. 8. The detection of principle axes of people in a group and under occlusion can be carried out similarly. The correspondence likelihood reflecting the similarity of pairs of principal axes of people is constructed according to the relationship
between “ground-points” of people, which is detected in each camera view, and the intersections of principal axes, which is detected in different camera views and transformed to the same view as illustrated in Fig. 9. The experimental results on several real video sequences from outdoor environments in NLPR database (CASIA gait database, xxxx) and PETS database (PETS database, xxxx) have demonstrated the effectiveness, efficiency, and robustness as in Figs. 10 and 11 respectively.

3.2.2.2. Continuously tracking objects across multiple widely separated cameras.

Another classic multi-camera tracking is handover (Cai et al., 2007; Gilbert and Bowden, 2006). That is to say, we need give a single ID to an object when object travels from one camera to another.

Here, tracking objects across cameras is achieved through computing the probability of correspondence according to appearance and spatio-temporal cues. The probability of two observations in two cameras generated from the same object is dependent on both appearance probability and spatio-temporal probability. In computing the appearance probability, we propose a two-layered histogram representation to incorporate spatial information into common histogram method, which provides more discriminability than computing the histogram of the whole body directly. Each layer of representation under one view is matched against its corresponding layer under another view by diffusion distance (Ling and Okada, 2006) to compensate for illumination changes and camera distortions.

The experiments consist of three cameras with non-overlapping fields of view. The cameras are widely separated, including two outdoor settings and one indoor setting. The layout is shown in Fig. 12a. Some of the experimental results are shown in Fig. 13. More details can be found in (Cai et al., 2007).

3.3. Object classification and recognition

3.3.1. Real-time object classification

Moving object classification is a popular topic in intelligent visual surveillance. With object type information known, more specific and accurate methods can be developed to recognize high
level actions of video objects. Especially in traffic scene videos, classification of moving objects into predefined categories allows the operator to program the monitoring system by specifying events of interest, such as alarming when a pedestrian is coming into a forbidden area or alarming when a vehicle is running in a reverse direction.

Much work has been done in object classification. In (Brown, 2004; Quming and Aggarwal, 2001), foreground objects are detected using motion information and some image features, like area, compactness, speed and bounding box aspect ratio are extracted for training and classification. However, most of these features are based on 2D space, which are often degraded with projective distortion especially in far-field traffic scene videos.

For example, nearby objects in images appear to be larger and move faster than those far away. Therefore, simply using these features for classification is unsuitable and limits the accuracy rate of the results. In addition, most of these algorithms cannot be robust to environment changes. In (Everingham et al., 2006), series of algorithms are described to demonstrate the effectiveness of local features for object detection and classification. However, most of these methods are time consuming and not applicable to low resolution videos. Viola et al. give a good framework for feature selection and object class recognition using boosting (Viola et al., 2003). However, it is a tremendous work to collect large samples of training data in all kinds of conditions and label all of them manually.

In our system framework, three properties are considered in the object classification algorithm (Zhang et al., 2007; Zhang et al., 2008):

- **Practical**: The algorithm should be real-time and achieve high classification accuracy.
- **Robust**: The algorithm should be robust in most kinds of conditions and perform well in different environments.
- **Automatic**: The algorithm should avoid any supervised learning and manual labeling of large samples of training data.

The flowchart for object classification is as Fig. 14. 2D features such as area, velocity, compactness and angle are used. Traditional 2D features on image plane based classification method cannot avoid projective distortion in a far-field scene. In our work, a sub-region based strategy is considered to conquer this disadvantage. With the Gaussian assumption that features in each grid of every category satisfies a multivariate Gaussian distribution, we divide the scene of the far-field video into many equal parts, distortion of 2D features can be ignored in each part because one part just covers a narrow field of view and classification can be realized in each part by using efficient 2D features. The size of sub-regions will affect the classification accuracy. Abundant videos are collected from Vs-star system at different times. The data set is then acquired from these collected videos. As we only focus on moving objects in traffic scene surveillance, motion information is employed here to detect targets of interest in videos. With regions of interest detected, every detected foreground region is fixed to a square according to its mass center, which is then normalized to the same size $20 \times 20$. These foreground targets are then labeled manually. In this way, we collect the whole data set including 58,958 vehicles, 56,942 pedestrians and 3297 other individuals as outliers. Samples of the database are illustrated in Fig. 42. Experiments are conducted with different number of grids and the accuracy is shown in Fig. 15. As we can see, the accuracy increases with smaller size of sub-regions, which demonstrates the advantage of the sub-region strategy. Gaussian assumption is adopted as prior information. Experiments are conducted to illustrate its effectiveness for classification. Experimental
results with Gaussian assumption and that with simply clustering are shown in Tables 1 and 2. It is evident that the approach using Gaussian assumption is much better, which demonstrate the effectiveness of Gaussian Assumption.

3.3.2. Gait based human recognition

Object classification just gives the coarse result of the object type such as pedestrian and vehicle, while some time, object recognition is urgently needed especial for human identification. Gait is an attractive biometric feature for human identification at a distance, and recently has gained much interest from computer vision researchers. Gait is a particular way or manner of moving on foot (Shakhnarovich et al., 2001). Compared with those traditional biometric features, such as face, iris, palm print and finger print, gait has many unique advantages such as non-contact, non-invasive and perceivable at a distance.

Many gait recognition algorithms have been developed in recent years (Human ID Challenge Problem at USF, xxxx; Sudeep et al., 2005). Some of them are model-based approaches (Wang et al., 2004; Wang et al., 2003), some are appearance-based ones (Yu et al., 2004; Yu et al., 2007; Wang et al., 2003). Model-based approaches aim to explicitly model the human body or motion, and they usually perform model matching in each frame of a walking sequence in order that the parameters such as trajectories, limb lengths and angular speeds are measured on the model. The majority of current approaches are appearance-based. There are many gait databases for this research such as (Shutlter et al., 2002; Automatic Gait Recognition for Human ID at a Distance at Soton, xxxx; Center for Biometrics and Chinese Academy of Science Security Research, xxxx), while no enough variations have been considered in these databases.

We have created a large gait database, the CASIA Gait Database (CASIA gait database, xxxx) including three datasets: Dataset A (small set, shown in Fig. 16), Dataset B (multi-view set shown in Fig. 17) and Dataset C (infrared set shown in Fig. 18). Based on these databases, we make deep research into gait recognition (Wang et al., 2004; Wang et al., 2003; Yu et al., 2004; Yu et al., 2007; Wang et al., 2003; Yu et al., 2009), gait recognition based on Procrustes shape analysis is model-free (Wang et al., 2004), which achieve good results in CASIA Database (Dataset A). Considering the body appearance and the dynamics of human walking motion information of gait (Wang et al., 2003), a silhouette analysis based gait recognition algorithm using Principal Component Analysis (PCA) is proposed (Yu et al., 2004). The algorithm implicitly captures the structural and transitional characteristics of gait, especially the shape cues of body biometrics. Although it is very simple in essence, the experimental results are surprisingly promising in CASIA Database (Dataset A). Due to the difficulties of automatic parameter recovery from video, few methods except (Cunado et al., 2003; D’iaz de Le’on and Sucar, 2000) used Prior information, we propose a gait recognition method with combination of model-based and appearance-based approaches (Wang et al., 2003). It not only analyzes spatiotemporal motion pattern of gait dynamics but also derives a compact statistical appearance description of gait as a continuum.

3.4. Activity recognition and abnormality analysis

In visual surveillance for security, motion detection, tracking and suspicious object classification and recognition can make sense in some degree, while understanding human physical behavior for surveillance is more significant, which is also driven by a wide spectrum of other applications in various areas such as

![Fig. 10. Tracking and correspondence of multiple people with three cameras: From the top to bottom, the frame numbers are 33, 54, and 133, respectively.](image-url)
interactive virtual reality systems, advanced and perceptual human–computer interfaces (HCI), content-based video storage and retrieval sports, and so forth (Haritaoglu et al., 1998). Here, according to the view of surveillance, we focus on shape based near-field activity analysis and trajectory based far-field activity analysis.

3.4.1. Trajectory based activity analysis

Trajectory that records the object’s position from entering to exiting a scene is one of the most useful information for behavior of moving objects. Much work on trajectory analysis for visual surveillance focused on learning several statistical motion routes most commonly taken by motion objects in a scene. Johnson and Hogg (1996) learn probability density functions of object trajectories generated from image sequences. The movement of an object is described in terms of a sequence of flow vectors, where each vector consists of four elements representing the position and velocity of the object in the image plane. The patterns of object trajectories are formed with two competitive learning networks which are connected by leaky neurons. Stauffer and Grimson (2000) learn motion

Fig. 11. Tracking and correspondence of a group of people through occlusion: From the top to bottom, the frame numbers are 878, 908, and 940, respectively.

Fig. 12. (a) The layout of the camera system and (b) three views from three widely separated cameras.
patterns using real-time tracking. Their method develops a codebook of representations using an online vector quantization on the entire set of representations acquired by the tracker. Then joint co-occurrence statistics are accumulated over the codebook by treating the set of representations in each trajectory as an equivalency multi-set. Finally, a hierarchical classification is performed using only the accumulated co-occurrence data. Porikli and Haga (2004) proposed a Hidden Markov Model (HMM) based distance to measure the similarity between two trajectories, and a kind of spectral clustering algorithm is adopted to acquire several clusters of single object. Compared with the research for trajectory, the spatial geography and temporal information are considered in trajectory based activity analysis here.

3.4.1.1. Spatial geography based analysis on single object’s trajectory. In early work, spatial geography information is more considered in trajectory analysis. Authors present a simple framework ranging from motion pattern learning to natural language description (Jianguang et al., 2002). First the spatial information is used to cluster trajectories into several coarse clusters. Here, a Hausdorff like distance is used to measure the spatial similarity between trajectories, where the distance can be considered as global information of trajectories. And then a simple k-means like clustering method is adopted to obtain several clusters of single object. In the second layer clustering, an identical clustering algorithm is used to acquire several clusters representing different spatial motion patterns of single moving object have been extracted. Fig. 21 presents the learned results in a model scene and a real traffic scene. From these figures, we can see the motion patterns well reflect the distributions of the sample trajectories.

For each trajectory cluster, a chain of probability distributions is built to recognize the event type of new trajectory. Every probability distribution in the chain is assumed to be Gaussian distribution, which corresponds to successive point feature vectors in a sample trajectory. The mean and the covariance matrix can be estimated according to maximum likelihood evaluation. Finally, the probability of a trajectory under a certain motion pattern is estimated as an exponential distribution. The details can be referred to Zhang et al. (2007).

Based on the learned motion patterns, some anomaly detection and behavior prediction applications can be realized as Figs. 22–25.

3.4.1.2. Temporal structure analysis on trajectory series. As presented in the above part, most of work on trajectory analysis focuses on spatial geography trajectory analysis. A set of spatial motion patterns of single moving objects are obtained by trajectory clustering, so that the simple single agent event (“Agent A goes from... to...”, etc.) could be described by these motion routes, and some anomaly detection and behavior prediction are also validated. However it is also very interesting to learn the temporal structure between these learned spatial motion patterns in a trajectory series. These work have been presented in Zhang et al. (2006), Zhang et al. (2007). Some major processing steps are shown in Fig. 26.

We adopt a minimum description length (MDL) principle based grammar induction framework to analysis the temporal structure in the trajectory series (primitive series), grammar induction can refer to Stolcke and Omohundro (1994), Allen (1994) and structure discovery (Cook and Holder, 1994). Primitive series are modeled by a Hidden Markov Model (HMM) (Rabiner, 1989) and Principal Component Analysis (PCA) with Euclidean distance is adopted to measure the trajectory similarity due to its advantage in outdoor surveillance scene (Grunwald, 1996). The clustering results in a traffic crossroad are shown in Fig. 27. Here, the output is a set of temporal rule which is represented as a form of grammar production with temporal relation matrix. An example of temporal rule
and the corresponding instance is shown in Fig. 28. The effectiveness is valid by traffic light rules in a crossroad scene illustrated by Figs. 29 and 30.

3.4.2. Shape based behavior analysis

As for near-field behavior analysis, shape-based features are commonly considered. Two types of shape-based features are used, i.e. silhouette and contour. The silhouette method takes into account all the pixels within a shape, and the contour method only extracts the boundary of a shape. Feature description is a key bridge between low level image feature and high level activity understanding (Filiberto et al., 2005; Hongeng et al., 2004). General contour-based descriptors include wavelets, Fourier descriptors and Hough transform (Chuang and Kuo, 1996; Zhang and Lu, 2002; Leavers, 1992; Zhang and Lu, 2005). Since contour descriptors are based on the boundary of a shape, they cannot capture the internal structure information. Consequently, they are limited to certain applications. Common silhouette-based shape descriptors include invariant moment, Zernike moment and wavelet moment (Teh and Chin, 1988; Prokop and Reeves, 1992). The moments are computationally intensive and sensitive to disjoint shapes or shapes with noise where the silhouette information is not correct. In a real scene, accurate shapes is difficult to be extracted because of complex background, the size of moving object also varies with its distance to camera. Therefore we need features invariant to geometry transformation and robust to noise, which is in accordance with the performance of R transform (Tabbone et al., 2006; Wang et al., 2007).

R transform, a new feature representation, has low computational cost and is effective to recognize similar activity even in the case of disjoint silhouette, silhouette with holes or frame loss.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Classification result: with Gaussian assumption.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pedestrians (%)</td>
</tr>
<tr>
<td>Pedestrians</td>
<td>98.2</td>
</tr>
<tr>
<td>Bicycles</td>
<td>3.4</td>
</tr>
<tr>
<td>Vehicles</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Classification result: with simply clustering.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pedestrians (%)</td>
</tr>
<tr>
<td>Pedestrians</td>
<td>87.2</td>
</tr>
<tr>
<td>Bicycles</td>
<td>11.3</td>
</tr>
<tr>
<td>Vehicles</td>
<td>8.4</td>
</tr>
</tbody>
</table>

Fig. 15. Different accuracy with different extent of division.
data. Moreover, many experiments prove that it outperforms common shape descriptors in activity sequence recognition. Some features for R transform:

- Translation in the plane does not change the R transform.
- Scaling the original image would not change the shape of the R transform, but the amplitude would change accordingly.

Fig. 16. Sample images from Dataset A.

Fig. 17. Sample images from Dataset B.

Fig. 18. Sample images from Dataset C.

Fig. 19. Clusters obtained by multi-layers clustering & action learning.
Therefore, R transform is invariant under translation and scaling if we resize the image into a normalized scale, which is feasible in activity recognition. For each frame in the video, a 180-dimensional feature vector, instead of the 2D image matrix, is extracted to represent the motion of the human body. The information that the initial silhouette sequences carry is transformed in a more compact way and invariant to geometry transformation.

According to the comparison experiments with common shape descriptors as shown in Fig. 31, the method based on R transform has several advantages. First, it does not require video alignment and is applicable in many scenarios where the background is known, because R transform is invariant to scale and translation. Second, R transform gets the high recognition rate for similar but actually different shape sequences, and even in the case of frame loss data and incomplete data. Third, our shape descriptor captures both boundary and internal content of the shape. For this reason, it is more robust to noise, internal holes and separated shape. Finally, the computation of shape descriptor is linear, so the computation cost of 2D R transform is low. Moreover, this approach can also be applied to other tasks such as gait recognition, content-based image retrieval and face animation.

- A rotation of $\theta_0$ in the original image leads to the phase shift of $\theta_0$ in R transform. This rarely happens in human activities.
3.5. Semantic Interpretation and event analysis

Achieving semantic understanding and interpretation of what behaviors or events happened about moving objects in video is the final task for computer vision. Compared with the lower level processing, the higher level phase involves spatio-temporal relationship mining, reasoning under uncertainty, semantic representation, and so on (Kojima, 2002).

For the goal of semantic analysis of events, there comes out a central problem of the definition of semantics. In various research areas, the proper meanings of semantics are quite different. In semiotics, semantic has a more narrow meaning that indicates the relationship between signs and objects. As a part structure of language, semantic refers to the meaning and relationship of words. Here we follow the view that semantic is the result of mapping and integration between concepts (Bobick, 1997).

But there is a gap between measurable features and semantic meanings. According to the ability and the procedure of human in perception and understanding for the world, event can be looked as the semantically meaningful changes happened in the scenes. The basic elements of the cognition for event analysis and understanding are various concepts. Each concept denotes a special semantic meaning. And all these concepts are grouped into different clusters according to their semantic functions. For the purpose of semantic analysis and understanding of events happened in dynamic scenes, all related concepts should be obtained firstly, and all these concepts should be organized in a well-defined structure.

In recent years, more and more researchers tend to use probabilistic frameworks to express and analyze events, such as Bayes-Networks, Hidden Markov Models, etc. All these models have...
a common peculiarity that stochastic parameters can be acquired automatically without any assumptions of prior knowledge under uncertainty. Considering idiographic demands of different circumstances, some variations have emerged. Galata et al. (2001) mention a method to present human behavior by variable length Markov models (VLMM). The algorithm of coupled hidden Markov models (CHMM) to model two-handed interactions is presented in Brand et al. (1997). At the same time, the superiority of these methods mentioned above brings obvious shortages. The computation of parameters for the given structure of a model is time-costly. To fit another problem, the structure of the model must be changed, and the learning for variable structures is more difficult. The obvious shortcomings of these methods are costly computation of parameters for given model structures and difficulties for fitting different demands of variable model structure.

To simulate the ability and procedure of human perception to the world, we present an ontology-based event representation and modeling method in the dynamic scenes (Xin and Tan, 2005). One ontology example for motion pattern is showed in Fig. 32b intuitively. All these motion patterns can be linked to relevant concepts directly by the method, and these motion related concepts can be clustered in a tree in Fig. 32a. This taxonomic tree can be regarded as the ontology for movement information of moving objects. Elliptical nodes denote concepts and rectangular nodes denote conditions for furcating. Another ontology structure is about concepts for interaction of objects or between objects and the environment. These concepts can also be represented in a taxonomic tree.

Based on hierarchical ontology, we propose a framework for semantic event analysis as shown in Fig. 34 (Xin and Tan, 2006). We integrate all ontologies into three levels. From bottom to top,
there are elementary level, interaction level and semantic level respectively. The elementary level supplies all related information of basic concepts for scene modeling and the movement of moving objects. And the interaction level produces types of all interactions in the scene for occupation or transition of moving objects in different regions and interaction between objects. The task of the semantic level is to create semantic representations of events in the scene.

The example for semantic event analysis and representation is shown in Figs. 35 and 36. Events at different time of moving objects in the dynamic scene are semantically represented with related concepts. And all these semantic representations are triggered by values of MSVs semantic similarities which exceed a threshold. Noticeable word is “Along”. It means one type of object motions that is an unremitting movement in the same region without obvious direction change.

4. Vs-star system and application

Based on the above work in our group, we integrate some algorithms into our real time surveillance system platform—Vs-star system. This system consists of 21 cameras, which are around the building of CASIA for real time security. Three of them are Pan-Tile-Zoom (PTZ) cameras. The system can monitor the scene including the indoor and outdoor in 24 h and most of the experimental data for the work of our group are captured from this platform. The interface of Vs-star is as Fig. 37. Some algorithms results such as object detection, tracking, classification and abnormal behavior alarming have been shown as Fig. 38. (a) is the detection and tracking for traffic scene, (b) is the human detection for indoor scene, (c) is the classification result for traffic scene, three categories are given: vehicle, pedestrian and biking person. Fig. 39 is the abnormal alarming results. (a) is alarming for climbing the fencing and (b) is the alarming for anti-dromic vehicle. Both indoor and outdoor environments, the object detection algorithms are used, while according to the different environments, the tracking, classification and behavior analysis algorithms are used. The model based methods are often used in far distance and feature

![Fig. 30. A traffic signal cycle recovered by our algorithm.](image)

![Fig. 31. The recognition rates of moments based methods and R transform based method.](image)

![Fig. 32. Ontology for movement in a taxonomic tree.](image)
(appearance, shape) based methods are often used in near distance camera. Especially for the behavior analysis, trajectory feature is often used in bird-view camera in outdoor environment. Our system has been used for some public security fields such as public transportation. The Fig. 40 gives the real alarming example for suspicious person in cable anti-stolen application, and (a) is camera position along the subway (b) is monitor room for security person. In this application, the main algorithms include object detection, tracking and classification. The system really works well in 24 h with considering the prior knowledge of the environment, while there are also many problems when vision system is used in real
application such as the video quality and some unexpected disturbances. The system has been also used in the Command Center in Olympic Park (CCOP) for the 2008 Beijing Olympic game. Fig. 41a is a Map of the Olympic Park. Six regions labeled A, B, C, D, E, F (including bus station, entrance1, entrance 2, ...) are analyzed, the predicted results of crowd density and the number of people in plaza are shown in Fig. 41b. The top half shows the predicted number of people and the lower half shows the real-time curves of crowd density. The deviation of the predicted number of people does not exceed 10% of the actual value in the final statistics (Li et al., 2008).
Fig. 40. Real application for cable anti-stolen along subway. (a) Camera position, (b) security room and (c) alarming for suspicious person.

Fig. 41. Real application for 2008 Beijing Olympic Games. (a) and (b) Map of the Olympic Park and some sketch maps.

Fig. 42. Samples in the database.
5. Conclusion

Automatic interpretation of human and vehicle motion in surveillance videos is very important for enhanced security. In this paper, we give a review about the recent work about the algorithm and system of surveillance based on some work of our group. The algorithms include object detection, tracking, classification and gait based recognition, activity analysis and abnormal alarming and then the Vs-star system is introduced for the purpose of research and real application. Although lots of works have been done in this field including the work introduced here, there are still huge challenges especially for real application such as object detection and tracking robustly for atrocious weather and clutter background, view-invariance object classification and recognition, more robust feature and clustering algorithm for activity analysis and semantic interpretation according with human cognition. The algorithms’ breakthrough will push the real application for enhanced security and the requirement from real application will bring more new challenges for the researchers.

Acknowledgments

This work is supported by National Basic Research Program of China (Grant No. 2004CB318100), National Natural Science Foundation of China (Grant No. 60736018, 60723005, 60605014, 60875021), NLPWRNLPRZY-2, National Hi-Tech Research and Development Program of China (2005A01Z318). The authors also thank the anonymous reviewers and the editor for their valuable comments.

References

CASA gait database: <http://www.cbr.sr.ia.ac.cn/China/CAS120Databases%20CH.asp>.
Huang, K. et al., 2006. Detecting and tracking distant objects at night based on human visual system. ACCV 2, 822–831.
Human ID Challenge Problem at USF. <http://figment.csee.usf.edu/GaitBaseline>.
Min Li, Xiaoqiang Zhang, Huang Kaiqi, Tieniu Tan, Estimating the Number of People in Crowded Scenes by MID Basedforeground Segmentation and Head-shoulder Detection,ICPR, 2008.
PETS database: <http://www.cvg.rdg.ac.uk/VSPETS/>.