

Multi-object Motion Pattern Classification for Visual Surveillance and Sports Video Retrieval

Akira Hayashi, Ryuji Nakashima, Toshihiko Kanbara and Nobuo Suematsu
Faculty of Information Sciences, Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194 Japan
akira@im.hiroshima-cu.ac.jp

Abstract

This paper presents a method to classify scenes based on motion information. While they use object trajectories or optical flow field as motion information in previous work, we use the instantaneous motions of multiple objects in each image. In order to deal with variable number of objects in a scene, we use moment statistics as features. Our approach is based on clustering, a form of unsupervised learning, and needs little human intervention. Furthermore, the probabilistic model based clustering makes it easy to detect scenes with novel patterns.

1 Introduction

Motion information is important for scene activity recognition in visual surveillance [8, 6, 1, 10, 15] and for content based video retrieval [3, 12, 17, 7, 2]. This paper presents a method to classify scenes based on motion information.

In previous work, they use as motion information :

- trajectories of each object [3, 8, 2, 15]
- positional relationship of two objects, and its temporal change [17, 6, 10]
- motion of each pixel (i.e. optical flow field) [12, 7, 1]

We use as motion information the positions and the velocities of multiple objects in each image. Which motion information to use depends on the application. We suppose that the motion information we use is suitable for recognizing the situation in a total scene when there are many moving objects such as people and vehicles in the scene.

Since the number of objects in a scene changes with time, we cannot use object positions and velocities themselves as features for the scene. Instead of them, we use moment statistics for the distribution of the moving objects in position and velocity product space.

The proposed method consists of two phases, the learning phase and the recognition phase. In the learning phase, scenes in the learning data are clustered in the feature space. Suppose the scenes in the learning data correspond to usual situations at the observation site. Then the clustering result is the classification of the scenes for usual situations. In the recognition phase, a newly observed scene is classified according to which cluster the scene belongs. In addition, scenes which do not belong to any of the clusters can be detected as novel scenes.

Generally speaking, (1) which scenes correspond to usual situations, (2) which scenes correspond to novel situations, and (3) how scenes for the usual situations are classified, all of these depend on the application (e.g. the target, the objective, and the method of the surveillance, or the category of the video library, and the objective of the retrieval). But, in our machine learning approach, much of the application dependent knowledge can be learned directly from data through clustering. Hence the proposed method is very much independent from the application. Furthermore, clustering is a form of unsupervised learning and does not need us to label the learning data as usual or novel.

We explain the input data in Sec. 2. We then explain the learning phase and the recognition phase of the proposed method in Sec. 3 and Sec. 4 respectively. In Sec. 5, the result of the experiment using three examples will be discussed. We summarize our presentation in Sec. 6.

2 Input Data

We fix a video camera at an observation site and film continuously. We assume a detection and tracking system for all the objects such as people or vehicles which appear in such video scenes (Interested readers are referred to review papers on tracking [5, 11]. There was a W/S on the subject recently [16].)

At each instance of time t , the tracking system provides

us with a *motion vector set*, $MVS(t)$:

$$MVS(t) = \{(x_i(t), y_i(t), u_i(t), v_i(t)) | i \in ID(t)\}$$

where $ID(t)$ is the index set for the objects which appear in the scene at time t , and $(x_i(t), y_i(t))$ and $(u_i(t), v_i(t))$ are the center position and its velocity of the i -th object at time t in the image coordinate system.

3 Learning Phase

The learning phase proceeds in the next order: (1) feature extraction, (2) principal component analysis (PCA), and (3) clustering.

3.1 Feature Extraction

When we use the scenes during time $0 \leq t \leq T$ for learning, then, input to the learning phase is the set of MVS 's for the period of time: $\{MVS(t) | 0 \leq t \leq T\}$. Since the number of objects in a scene, $|ID(t)|$, changes with time, we cannot use the elements of $MVS(t)$ directly as features. Instead of them, we use a fixed number of moment statistics computed from $MVS(t)$.

Moments are widely used to characterize probability distributions in statistics as well as to characterize mass distributions in mechanics. An image, a mapping from 2D coordinates to grey scales, can be considered as a density distribution in 2D. Therefore, moments are used also as image features [13]. The popularity of the moments can be explained by the *uniqueness theorem* in statistics: the moments all together uniquely determine the density distribution [14].

In our method, we use moments to characterize distributions in 4D position and velocity product space, of the moving objects within a scene. As features for a scene at time t as a whole, we use

1. the number of objects in the scene, $N(t) = |ID(t)|$.
2. the sample means of the positions and velocities (the first order moments), $(\bar{x}(t), \bar{y}(t), \bar{u}(t), \bar{v}(t))$.
3. n -th (the second and higher) order *central* moments defined below.

$$\hat{M}_{n_x n_y n_u n_v} = \frac{1}{N(t)} \cdot \sum_{i \in ID(t)} (x_i - \bar{x})^{n_x} (y_i - \bar{y})^{n_y} (u_i - \bar{u})^{n_u} (v_i - \bar{v})^{n_v}$$

where n_x, n_y, n_u, n_v are non negative integers, and $n_x + n_y + n_u + n_v = n$ holds.

The covariances are the second order central moments. For example, \hat{M}_{2000} , which is defined as $\frac{1}{N(t)} \sum_i (x_i(t) - \bar{x}(t))^2$, is the variance $\sigma_{xx}^2(t)$. \hat{M}_{0011} , defined as $\frac{1}{N(t)} \sum_i (u_i(t) - \bar{u}(t))(v_i(t) - \bar{v}(t))$, is the covariance $\sigma_{uv}^2(t)$.

Which of the higher order moments are necessary as features depends on the application. In this paper, we use up to the second order central moments (i.e. 4 variances and 6 covariances) as features¹. Hence, a scene at time t is expressed with a 15 dimensional feature vector: $\mathbf{f}(t) = (N(t), \bar{x}(t), \bar{y}(t), \bar{u}(t), \bar{v}(t), \sigma_{xx}^2(t), \dots, \sigma_{uv}^2(t))$.

We then normalize each component of the feature vector, $\mathbf{f}(t)$, so that its average and the variance over all the learning data will become 0 and 1 respectively. We denote the normalized feature vector as $\hat{\mathbf{f}}(t)$.

3.2 Principal Component Analysis

It is well known that clustering in high dimensional space is difficult. We apply principal component analysis (PCA), a well established method in data analysis, for dimensionality reduction. PCA is applied to the set S of 15 dimensional feature vectors: $S = \{\hat{\mathbf{f}}(t) | 0 \leq t \leq T\}$.

Then, the 15 dimensional feature vectors are expressed as a linear combination of *the principal components*, the eigenvectors of the covariance matrix for S . Only the major principal components will be retained such that the sum of their contribution rates exceeds some threshold. We denote the coefficient vector, which is a new feature vector with reduced dimensionality as $\mathbf{x}(t)$, and call it as a principal component vector.

3.3 Clustering

The principal component vectors for all the scenes, $\{\mathbf{x}(t) | 1 \leq t \leq T\}$, are then clustered using MCLUST [4], a model based clustering package. Using EM algorithm, MCLUST estimates the mixture model of Gaussian probability distributions (1), each of which corresponds to a cluster.

$$p(\mathbf{x}) = \sum_{k=1}^K \tau_k \cdot \phi_k(\mathbf{x} | \mu_k, \Sigma_k) \quad (1)$$

$$\phi_k(\mathbf{x} | \mu_k, \Sigma_k) = (2\pi)^{-\frac{d}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right]$$

¹In the experiment of simulated road crossing, which will be discussed, we tested both of the 15 dimensional feature vector and 35 dimensional feature vector which contains all of the third order moments. The results were not very much different.

where τ_k is the prior probability for \mathbf{x} to be contained by the k -th cluster, K is the total number of clusters, and d is the length of the feature vector.

We use MCLUST for two reasons:

- MCLUST selects the best model ² and determines the most appropriate number of clusters on the basis of BIC, Bayesian Information Criterion.
- We can use the estimated probabilistic model for novelty detection in the recognition phase, which will be explained shortly.

4 Recognition Phase

In the recognition phase, for a newly observed scene, (1) its 15 dimensional feature vector, $\mathbf{f}(t)$, is computed and normalized, and then (2) is projected onto the subspace obtained by PCA. This gives us the principal component vector, $\mathbf{x}(t)$, to be used in (3) novelty detection and (4) classification. (5) Temporal smoothing will then be applied to the detection and classification results.

4.1 Novelty Detection

Our novelty detection makes full use of the probabilistic model based clustering method, and is based on the following lemma [9].

Lemma 1 *Given a d dimensional random vector \mathbf{x} from a Gaussian distribution with mean μ and covariance Σ , the probability that \mathbf{x} satisfies the inequality (2) is $1 - \alpha$.*

$$(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \leq \chi_d^2(\alpha) \quad (2)$$

where $\chi_d^2(\alpha)$ is the 100 α % point of χ^2 distribution with d DOFs.

The set of \mathbf{x} 's which satisfies the inequality (2) is called as the 100(1 - α)% *certainty ellipsoid*.

Since we have estimated a mixture of Gaussian probability distribution using MCLUST in clustering, we can compute the certainty ellipsoid for each cluster. Given $\mathbf{x}(t)$, the feature vector of a scene at time t , we check for each of the clusters whether $\mathbf{x}(t)$ is contained in its certainty ellipsoid using (2). If $\mathbf{x}(t)$ is not contained in any of the ellipsoids, the scene is determined as novel.

²MCLUST provides various models according to the volume, the shape, and the orientation of the density contours for each cluster. They can be allowed to vary between clusters, or constrained to be the same for all clusters.

4.2 Classification

If a scene with a feature vector $\mathbf{x}(t)$ has not been detected as novel, it is classified by finding the cluster to which it belongs. The cluster, k^* , is the one which maximizes $p(k|\mathbf{x})$ for $1 \leq k \leq K$. This can be computed using the next formula derived from Eq.(1) and Bayes's theorem.

$$k^* = \operatorname{argmax}_k p(k|\mathbf{x}) = \operatorname{argmax}_k \tau_k \cdot \phi_k(\mathbf{x}|\mu_k, \Sigma_k) \quad (3)$$

4.3 Temporal Smoothing

We have treated each scene independently so far in clustering and classification, without considering its temporal continuity. As a simple method to enforce temporal continuity, we compute, as the cluster ID of a scene at time t , the ID of the cluster which appears most frequently in the window of scenes between time $t - f$ and $t + f$.

The temporal smoothing is applied to novelty detection as well. We use the majority of the outputs (i.e. novel or not) for the scenes in the window.

5 Experiment

In order to show the validity of our method, three experiments have been carried out with the following settings.

- In PCA, only the major principal components are retained such that the sum of their contribution rates exceeds 80%.
- 95% certainty ellipsoids ($\alpha = 0.05$) are used for novelty detection ³.
- In temporal smoothing, we set $f = 5$.

5.1 Simulated Road Crossing

In this example, we simulated a road crossing. See Fig. 1. The motion patterns of the people who walk across the roads were learned and then classified. There are 4 crosswalks, two of them are in the vertical direction, and two in the horizontal direction. In accordance with the traffic signals which change in every 50th scene, pedestrians cross the roads in either of the two directions.

Learning Phase

We used 1000 scenes for learning. The number of objects (people) per scene is 8 at the minimum, and 44 at the maximum. In PCA, the first 3 principal components were retained (i.e. $d = 3$). Fig. 2 shows the clustering result up to

³Generally speaking, novel scenes are more reliably detected as we increase α , but more detection errors are made as to usual scenes.

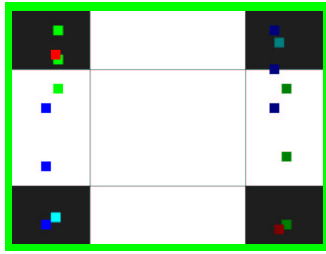


Figure 1: Exp.1: Simulated road crossing

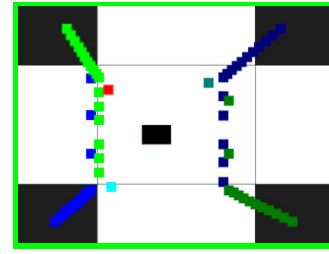


Figure 3: Exp.1: Scene with a novel pattern

scene 300. There are 4 clusters. In the figure, we show in a xy graph how the cluster ID (i.e. the classification) changes as the scene (i.e. the time) changes. Shown is the graph after the temporal smoothing has been applied.

Since the traffic signals change with the period of 50 scenes, we can tell from Fig. 2 that cluster #2, for the scenes in the middle of intervals 0-49, 100-149, 200-249, corresponds to the steady state where people are walking in the vertical direction. Similarly, cluster #4, for the scenes in the middle of intervals 50-99, 150-199, 250-299, corresponds to the steady state where people are walking in the horizontal direction.

The rest of the clusters correspond to transient states. Cluster #1 corresponds to the state when pedestrians are beginning/finishing to cross in either of the two directions. Cluster #3 corresponds to the state when all of the pedestrians are standing still, i.e. pedestrians in the vertical (horizontal) direction have finished crossing, and pedestrians in the horizontal (vertical) direction are waiting for the signal to change.

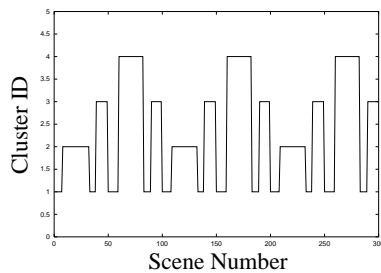


Figure 2: Exp.1: Clustering result

Recognition Phase

300 scenes are provided for the recognition phase. We set up so that a still object suddenly appears in the center and all pedestrians gather around the object in scenes 110-160. See Fig. 3.

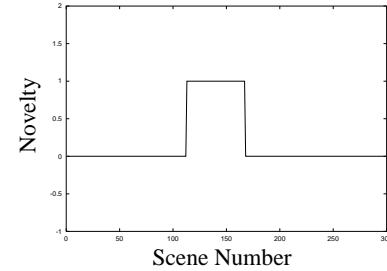


Figure 4: Exp.1: Novelty detection

Fig. 4 shows the result of the novelty detection. Scenes 113-167 are detected as novel, because the motion patterns in these scenes are not contained in the learned scenes. Fig. 5 shows the classification result. We can see the classification in the figure is approximately equal to that in Fig.2 except for the novel scenes.

5.2 Road Crossing

In the second example, we used 3 min. 20 sec. long observation data at a real road crossing (Fig. 6). In order to improve the computational efficiency and to increase the accuracy of position and velocity data, 10 consecutive frames (1/3 second) are combined to a scene, for which the average position

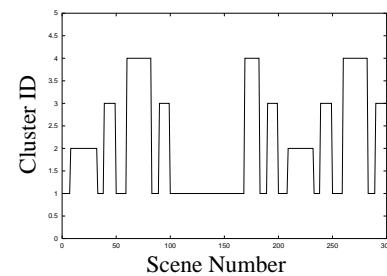


Figure 5: Exp.1: Classification



Figure 6: Exp.2: Scene of a real road crossing

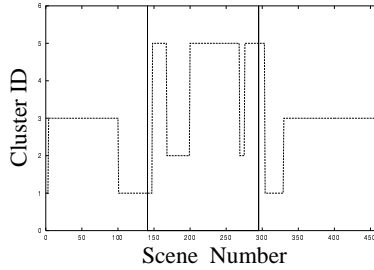


Figure 7: Exp.2: Clustering Result and Changes noticed by human subjects

and velocity are computed and used. The number of the objects (vehicles and people) per scene is 4 at the minimum, and 25 at the maximum.

Learning Phase

We used 450 scenes in the preceding 2.5 minutes for the learning phase. The first 5 principal components were retained in PCA ($d = 5$). The Broken lines in Fig. 7 show the clustering result. There are 5 clusters ⁴.

Changes noticed by Human subjects

In order to evaluate the clustering result of the proposed method, a comparison was made with that by 9 human subjects. Since it is difficult for humans to classify scenes, they are asked to tell us when the situation changes. The subjects were told beforehand that the situation change can be either (1) a change of the traffic signals, or (2) a congestion and its dissolution in the lanes, or (3) others.

Solid lines in Fig. 7 show when the situation changes according to the human subjects. Among the scene changes noticed by any of the subjects, only the changes which are close within ± 8 scenes and were noticed by more than 3

⁴Cluster #4 disappears from the figure because of the temporal smoothing.

people were collected. Then, for each collection (i.e. each situation change), its average scene number was computed and used as the result. As explained, broken lines in the same figure are the clustering result of our method.

In order to evaluate our method quantitatively, the *precision* and the *recall* of the method were computed, assuming that the changes noticed by the human subjects are true (correct).

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

where

TP True Positive. Total changes found both by the subjects and by our method. Small differences in time are neglected ⁵.

FP False Positive. Total changes not found by subjects, but by our method.

FN False Negative. Total changes found by the subjects, but not by our method.

The precision and the recall of our method were 14.3%, and 100%, respectively. The low precision can be explained like this. While changes from cluster #1 to #5, and those from #5 to #1 correspond to traffic signal changes and are considered to be important by the subjects, changes from cluster #3 to #1, and those from #5 to #2 do not correspond to signal changes and are not considered important ⁶. Considering the high recall, our method is good at finding changes important for the subjects. In summary, the clustering result is considered to be redundant and thorough for the humans in this experiment.

Recognition Phase

150 scenes in the last 50 seconds were passed to the novelty detection, and scenes from 55 to 66 were detected as novel (Fig. 8). In the detected scenes, only one of the two vertical lanes is crowded, and in the opposite lane, vehicles within the oval are resting (Fig. 9). Scenes with such a pattern are not contained in the learning data.

⁵The maximal difference in time was 10 scenes, and the average difference was 8 scenes (2.7 sec.).

⁶In both cluster #3 and #1, many of the objects are moving vertically. In both cluster #5 and #2, many of the objects are moving horizontally. While there are as many objects in the right lane as in the left lane in cluster #3, there are more objects in the right lane in cluster #1. While there are objects in the near road as well as in the far road in cluster #5, there are objects only in the near road in cluster #2.

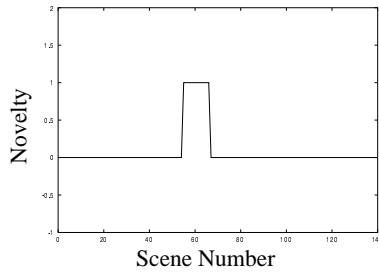


Figure 8: Exp.2: Novelty Detection



Figure 9: Exp.2: Scene found by the novelty detection

Tracking Errors

As stated before, the proposed method assumes that it is possible to detect and track all the moving objects in the scenes. Here, we try to assess the effects, when the assumption is violated and 10% of the moving objects in each scene are not detected/tracked⁷. We consider three cases: (1) no error in the learning data, and 10% error in the recognition data, (2) 10% error in the learning data, and no error in the recognition data, (3) 10% error in the learning data, and 10% error in the recognition data.

Table 1 shows the detection and classification error rates for the three cases, assuming that the detections and the classifications are correct when there are no tracking errors. The detection (classification) error rate is the percentage of the scenes which were detected (classified) differently when there are tracking errors in either (or both) of the learning and recognition phases. Both error rates were low, considering 10% tracking errors. Our method is considered to be robust to tracking errors in this experiment.

5.3 Basketball Game

In the third example, we experimented with 9.5 minute long observation data of a basketball game (Fig. 10). 15 consec-

⁷94.6% success rate of tracking vehicles in occlusion situations has been reported [10].



Figure 10: Exp.3: Scene from a basketball game

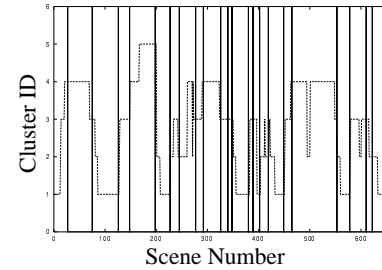


Figure 11: Exp.3: Clustering Result and Changes noticed by human subjects

utive frames (1/2 second) are combined and averaged to a scene. The players were tracked, but the judges, the audiences, and the ball were not tracked. Therefore, the number of objects in each scene is fixed and is equal to the number of players (10).

Learning Phase

We used about 650 scenes during 5.5 minute period in the learning phase. In PCA, the first 10 principal components were retained ($d = 10$). The Broken lines in Fig. 11 show the clustering result. There are 5 clusters.

Changes noticed by Human subjects

The clustering result was then compared with that by 9 human subjects. Since it is difficult for humans to classify scenes, they are asked to tell us when the situation changes,

Table 1: Exp.2: Error rates for detection and classification

tracking error		recognition error	
learning	recognition	detection	classification
none	10%	5.0%	6.4%
10%	none	8.6%	7.1%
10%	10%	2.1%	6.4%

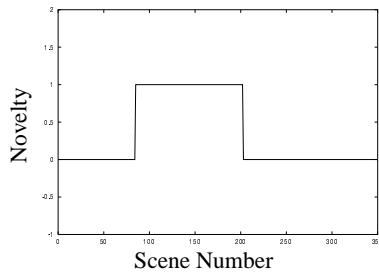


Figure 12: Exp.3: Novelty Detection



Figure 13: Exp.3: Scene from a timeout

where the situation change can be either (1) a change in defense and offense, or (2) a suspension or a resumption because of fouls, or (3) others.

Solid lines in Fig. 11 show when the situation changes according to the subjects. As explained earlier, the broken lines are the clustering result of our method. Assuming that the changes noticed by the subjects are true, the precision and the recall of our method were computed. They are 57.6%, and 82.6%, respectively ⁸. Since both of the precision and the recall are high, we consider that the clustering result of our method is reasonable for the humans in this experiment.

Recognition Phase

350 scenes (2 min. 55. sec long) which had not been used in the learning phase were passed to the novelty detection. Scenes from 85 to 202 were detected as novel (Fig. 12). They are the scenes for a time out (Fig. 13). Such scenes were not contained in the learning data. When we check with the video tape, the time out was from scene 76 to 217, and has a good correspondence with the detection result.

⁸For the changes found by both the subjects and our method, the maximal difference in time was 6 scenes, and the average difference was 2.7 scenes (1.3 sec.).

Tracking Errors

As in the second experiment, we assess the effects of tracking errors. Table 2 shows the detection and classification error rates for the three cases, assuming that the detections and the classifications are correct when there are no tracking errors. The detection error rates were low considering 10% tracking errors. But the classification error rates were high when there are errors in the learning data. When then learning data has errors, the probability model and the number of clusters can become different ⁹, and this explains why the error rates are high. The proposed method may not be robust in this experiment.

6 Summary and Future Work

We have presented a method to classify scenes based on motion information. While they use object trajectories or optical flow field as motion information in previous work, we use the instantaneous motions of multiple objects in each image. In order to deal with variable number of objects in a scene, we have proposed to use moment statistics as features. The proposed method consists of two phases. In the learning phase, scenes in the learning data are clustered. In the recognition phase, a newly observed scene is classified. In addition, scenes with novel motion patterns are detected. We have carried out three experiments and showed the validity of our method.

As future work, we plan to (1) learn and recognize the temporal dynamics of the scenes, and (2) experiment with observations with much longer period of time.

Table 2: Exp.3: Error rates for detection and classification

tracking error		recognition error	
learning	recognition	detection	classification
none	10%	0.0%	4.4%
10%	none	13.1%	33.1%
10%	10%	15.1%	32.6%

⁹We manually converted the cluster IDs for each of the 10% error cases, so that we can compare them with the IDs in the error free case.

References

- [1] M. Brand and V. Kettner, "Discovery and Segmentation of Activities in Video", *IEEE transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 844-851, 2000.
- [2] S. Dagtas et al., "Models for Motion-Based Video Indexing and Retrieval", *IEEE Trans. on Image Processing*, Vol. 9, No. 1, pp. 88-101, 2000.
- [3] N. Dimitrova and F. Golshani, "Motion Recovery for Video Content Classification", *ACM Trans. on Information Systems*, Vol. 13, No. 4, pp. 408-439, 1995.
- [4] C. Fraley and A. E. Raftery, "MCLUST: Software for Model-Based Cluster and Discriminant Analysis", *Technical Report no.342*, Department of Statistics, University of Washington, 1998.
- [5] D. Gavrilu, "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding*, Vol. 73, No. 1, pp. 82-98, 1999.
- [6] Y. Ivanov, C. Stauffer, A. Bobick, and W.E.L. Grimson, "Video Surveillance of Interactions", *Proc. Second IEEE Int. W/S on Visual Surveillance*, pp. 82-89, 1999.
- [7] A. Jain, A. Vailaya, and X. Wei, "Query by video clip", *Multimedia Systems*, Vol. 7, pp. 369-384, 1999.
- [8] N. Johnson and D. C. Hogg, "Learning the distribution of object trajectories for event recognition", *Image and Vision Computing*, Vol. 14, pp. 609-615, 1996.
- [9] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, Prentice Hall, 1992.
- [10] S. Kamijo and Y. Matsushita and K. Ikeuchi and M. Sakauchi, "Traffic Monitoring and Accident Detection at Intersections", *IEEE transactions on Intelligent transportation systems*, Vol. 1, No. 2, pp.108-118, 2000.
- [11] T. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture", *Computer Vision and Image Understanding*, 2001 (to appear).
- [12] K. Otsuka, T. Horikoshi, S. Suzuki, and M. Fujii, "Feature Extraction of Temporal Texture Based on Spatiotemporal Motion Trajectory", *Proc. Int. Conf. on Pattern Recognition*, pp. 1047-1051, 1998.
- [13] R. Prolop and A. Reeves, "A Survey of Moment-Based Techniques for Unoccluded Object Representation and Recognition", *CVGIP: Graphical Models and Image Processing*, Vol. 54, No. 5, pp. 438-460, 1992.
- [14] A. Shiryaev, *Probability*, Springer, 1995.
- [15] C. Stauffer and W. E. L. Grimson, "Learning Patterns of Activity Using Real-Time Tracking", *IEEE transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 747-757, 2000.
- [16] *Proc. IEEE Workshop on Multi-Object Tracking*, IEEE Computer Society, 2001.
- [17] A. Yoshitaka et al., "Content-based retrieval of Video data based on Spatiotemporal Correlation of Objects", *Proc. IEEE Multimedia Computing and Systems*, pp. 208-213, 1998.