Learning Feature Extraction and Classification for Tracking Multiple Objects: A Unified Framework

Xiaotong Yuan, Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science
Beijing, China, 100080

Abstract

A great challenge in tracking multiple objects is how to locate each object when they interact and form a group. We view it as a binary classification problem. It is important to base the classification on the currently most discriminative features. We derive a unified framework for learning feature extraction and classification in appearance-spatial space for multiple object tracking. In this framework, both classifier design and feature evaluation are accomplished by minimizing an criterion which corresponds to an upper-bound of classification error. There, the most discriminative features, as variables, minimize the criterion function, whereas the classifier, as a function, minimizes the criterion functional. The resulting system offers high accuracy for real-time tracking of nearby multiple objects in complex and dynamic scenes.

1. Introduction

An intelligent visual surveillance and monitoring system requires moving objects to be tracked as they move through a scene. Accurate and real-time tracking can provide useful information for applications, such as activity analysis [12] and detection of important or abnormal events [8]. An arsenal of powerful tracking algorithms have been proposed during the last two decades of vision research. A common approach is by detect-then-track where objects are detected and then tracked [13, 2].

Much attention has been paid to multiple objects tracking [13, 9, 10, 14]. A challenge in tracking multiple moving objects is when the objects form groups, interact, or depart from one another, especially in unconstrained cluttered environments. When the objects are spatially nearby, each object is a distractor to another. McKenna et.al. [13] used color distribution models to track groups of people. Han el.al. [9] based their indoor multiple people tracking on state-observation sequence analysis. Probabilistic approaches like Monte Carlo filter [10] or Bayesian correlation [14] is useful in dealing with the problem of background clutter as it allows for the tracking of multiple hypotheses.

In a recent work [3], Collins et. al. argue that the success or failure of tracking depends primarily on how distinguishable an object is from its surroundings. They present an online distractor-resistance feature extraction mechanism to enhance the performance of mean-shift tracker for tracking single object in complex scenes, where distractors include background and nearby objects. The method is simple yet proves powerful for tracking single object under influence of distractors. Similar idea is taken in [1] where tracking is consider as a binary classification problem and an ensemble of weak classifier is trained on-line to distinguish between the object and the background.

Appearance-based features are often used for tracking [4, 5]. Such methods offer flexibility to track deformable or non-rigid objects while being robust to partial occlusion. However, appearance based features could lead to confusion between different objects when the objects and their distractors are very similar in appearance.

In this paper, we address the problem of distractor-resistance tracking for multiple nearby objets. We pose it as a binary pattern classification problem. The key problems are the following: (1) designing an evaluation criterion for effective classification; (2) performing online learning of the best features for the next frame; and (3) exploiting constraints on spatial relations between the tracked region and distractors tracked the region and distractors, to make full use of available information contained in the observation. Solving these problems leads to a unified, theoretically justified, framework , in which both feature evaluation and classifier design are accomplished by minimizing an criterion which corresponds to an upper-bound of classification error. The overall description of the framework is as follows:

(1) Classification. In a tracking system, samples of pixels

from object and its distractor classes are available from previously tracked multiple objects; they form a training data set. We use an upper-bound of error rate as the evaluation criterion for classification. Through functional analysis, the optimal Bayesian classifier is derived by minimizing the evaluation criterion, in the form of log likelihood ratio.
(2) Feature Extraction. We further analyze the upper-bound of classification error rate that can be achieved by the derived Bayesian classifier and derive the most discriminative features. We adopt the linear subspace learning for feature extraction. The optimal projection matrix is learned by minimizing the corresponding Bayessian upper-bound.
(3) Constraints on Spatial Relationship. We augment the appearance features in temporal frames by including spatial relationship between the object regions; and then perform the learning of the classification and feature extraction in the augmented appearance-spatial space.

The work most closely related to ours is that of [3] and [1] that use on-line subspace learning and Adaboost to find the best feature space to work in and tracker ensemble. We extend their work in several aspects. First, we derived a theoretically justified framework, instead of a heuristic one, for on-line tracker design and feature extraction. Second, we introduce the constrains on spatial relationship among the object regions into feature space, and automatically trade off between the appearance feature and spatial feature during the tracking process. Finally we focus on multi-objects tracking which is not involved in [3] and [1] and the resulting system is realtime.

The remainder of this paper is organized as follows. The unified framework and how it is formulated to tackle the three problems are presented in Section 2 . Experiments are provided in Section 3 to demonstrate the performance of the resulting system. Section 4 concludes the paper.

2. The Unified Framework

The tracking problem we are dealing with can be illustrated in Fig.1: (a),(b) and(c) show nearby moving objects of similar type. Each object is a distractor to the other before they spit. The object detection results of an object tracker are given in (d),(e) and (f), in which these nearby objects are detected as merged foreground blobs (we use a Gaussian mixture model [15] for moving objects detection and the merge and split events reasoning is similar to the work in [16]). Such cases are common in many tracking applications. We aim to locate each object in blob correctly.

Tracking nearby objects is equivalent to keeping track of the location of each object. This may be done by labeling the pixels inside the blob with 1 or -1, with 1 indicating the pixels belonging the object and -1 those of distractors. This could be viewed as a binary classification problem.

In a tracking system, samples of pixels from object and its distractor classes are available from multiple objects tracked; they form a training data set for each object, denoted as $\{x_i, y_i\}_{i=1}^N$ with x_i a vector valued feature and $y_i = 1$ or -1. We aim to learn from the training data a classification function $f(x) : X \mapsto \{-1, +1\}$ where X is the domain of features x. The training data for tracking in the next frame is collected up to the current frame, and so the learning is done online using temporal information. In the following three subsections, we present solutions for the aforementioned three problems.

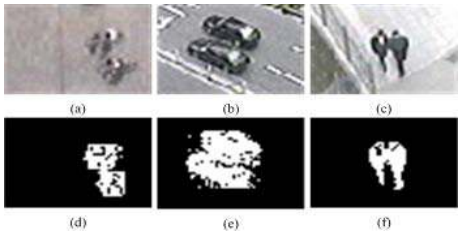

Figure 1. Two moving objects are detected as one foreground blob.

2.1. Upper-Bound of Classification Errors

We adopt the upper-bound estimate used in AdaBoost learning [7], which is viewed to be the most important developments in classification methodology. For the real version of AdaBoost (RealBoost), a stronger classifier $L(x) : X \mapsto (-\infty, +\infty)$ is a real function on features domain X. The class label for a test x is obtained as $f(x) = sign(L(x))$ while the magnitude $|L(x)|$ indicates the confidence.

The following criterion measures the upper-bound on classification error [7][1]:

$$J(L) = \log(E(e^{-yL(x)})) \tag{1}$$

where E stands for the mathematical expectation over the joint (x, y) space. According to Lemma 1 in [7], $J(L)$ is minimized at

$$L^*(x) = \frac{1}{2} \log \frac{P(y = +1|x)}{P(y = -1|x)} \tag{2}$$

Assuming that the two class prior probabilities satisfy $P(y = +1) = P(y = -1)$, and denote $P_+(x) = P(x|y = +1)$,$P_-(x) = P(x|y = -1)$, we get

$$L^*(x) = \frac{1}{2} \log \frac{P_+(x)}{P_-(x)} \tag{3}$$

[1]For the convenience of derivation, we actually use the logarithm of the upper-bound, without changing the minimum

L^* maps object one/two class condition distributions into a positive real number for features distinctive to object one, and a negative one for features distinctive to object two (belonging to distractors). The associated classification rule is $f^*(x) = sign(L^*(x))$, which can be shown as the Bayesian classifier (Lemma 1 in [7]). This log likelihood ratio function has been empirically used as what is called "tuned feature" in [3] for discriminating between the tracked object and its distractors. This gives a solution to problem (1) defined in this paper.

2.2. Extraction of Most Discriminability Features

We adopt linear subspace learning for feature extraction. The transformation can be performed by multiplying the input data by a projection matrix W to produce new feature salient features. After substituting x by extracted features Wx and L by the Bayesian estimate L^*, the upper-bound becomes $J(L^*|W) = \log(E(e^{-yL^*(Wx)}))$. Now, we will find optimal linear transformation W by minimizing $J(L^*|W)$ with respect to W of the evaluation criterion.

It is interesting to notice that the above the Bayesian upper-bound $J(L^*|W)$ is identical to the Bhattacharyya coefficient [11] between the two class conditional distributions (see Appendix A for a proof).

$$J(L^*|W) = \log(\int (P_+(Wx)P_-(Wx))^{\frac{1}{2}} dx) \qquad (4)$$

Generally speaking, the expression of object function J is very complex, and its gradient equation is highly nonlinear with respect to W even when P_+ and P_-. For real time tracking, it is computationally impractical to estimate the best W^* using a gradient-descent method. A compromise is to define a finite set S of seed candidate transformation matrix and choose the most discriminative one from the set to get the sub-optimal feature space [3].

Assuming that the number N of training samples is large enough, we could use the arithmetic average to approximate the expectation E in (1), according to the law of large numbers. This gives the approximate form

$$\tilde{J}(L^*|W) = \frac{1}{N}(\sum_{y_i=-1}(\frac{P_+(Wx_i)}{P_-(Wx_i)})^{\frac{1}{2}} + \sum_{y_i=+1}(\frac{P_-(Wx_i)}{P_+(Wx_i)})^{\frac{1}{2}}) \qquad (5)$$

Given a matrix $W \in S$, the class conditional probability $P_+(Wx)$ and $P_-(Wx)$ in subspace Wx can be determined by using some statistical means such as non-parameter histogram estimation. The best discriminative transformation matrix is then chosen to be:

$$W^* = \arg\min_{W \in S}(\tilde{J}(L^*|W)) \qquad (6)$$

This defines a solution to problem (2) defined in this paper. An exhaustive search will be described in Section 2.4 to find the solution.

2.3. Constraints on Spatial Relationship

In the above section, we have not specified what original features space to work in. In principle, a wide range of features could be chosen for tracking, including color, texture, contour, shape and motion. Collins et. al. [3] chose to represent target appearance using histograms of color filter bank responses applied to the R-G-B pixel values within local image windows. A 11D feature vector that is formed by the combination of local orientation histogram and pixel colors is used in [1] for feature selection.

In this paper, we exploit cues in spatial relations between tracked the region and distractors, and integrate the spatial cues into feature extraction, to make tracking robust to the scene in which appearance itself is indistinctive. This is done by taking into consideration information on spatial locations of both tracked regions so that discrimination between regions can be inferred spatially.

We view the feature vector $x = (u, s)^T$ as a multi-dimension probabilistic variable, where u is the appearance feature vector and s is the the spatial feature vector; and define the two conditional distributions in (3) in the joint appearance-spatial space. In an implementation, the appearance-spatial space is chosen to be a 5D dimensional (3D color plus 2D coordinates) space . The dimensionality reduction may be done via Linear Discriminant Analysis (LDA) [6].

However, it is advantageous to keep the most discriminative appearance features and most discriminative spatial features separately. This means that we want to search for a plane in the 3D color space and a straight line on the 2D image plane, they can best discriminate the samples of two classes. Therefore, we define the transformation matrix in the the following form

$$W = \begin{pmatrix} \omega_1 & \omega_2 & \omega_3 & 0 & 0 \\ 0 & 0 & 0 & \tau_1 & \tau_2 \end{pmatrix}, \omega_i, \tau_i \in R.$$

In an implementation, we define the set of seed candidate transformation matrix as:

$$S = \{ \begin{pmatrix} \omega_1 & \omega_2 & \omega_3 & 0 & 0 \\ 0 & 0 & 0 & \tau_1 & \tau_2 \end{pmatrix} \mid \omega_i, \tau_i \in \{-2, -1, 0, 1, 2\} \}$$

This set of seed matrixes is chosen as such for the following reasons: (1) the matrix computation can be done efficiently when the entries are integer-valued; (2) most commonly used appearance feature combination schemes can be expressed by this form of transformation; for example 2G-R-B used in [3]; (3) the orientations in 2D image plane, such as 0, $\frac{\pi}{4}$, $\frac{\pi}{2}$ and $\frac{3\pi}{4}$ in fig.2, can be expressed by this form.

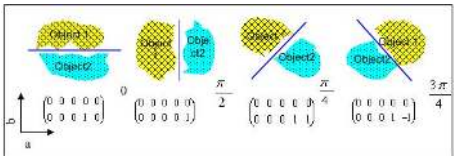

Figure 2. Four typical spatial configurations of nearby objects watched by a top-view camera

2.4 Implementation Issues

When both appearance and spatial feature are considered jointly, a total of 5^5 candidate features have to be enumerated for finding the optimal matrix W. To alleviate such computational burden for real-time performance, we assume that appearance feature and spatial feature are independent, so that $P(W(u,s)^T|y) = P(wu^T|y)P(\tau s^T|y)$. We then use a greedy strategy to minimize the following two object functions separately:

$$\tilde{J}_1(L^*|\omega) = \frac{1}{N}\sum_{y_i=-1}(\frac{P_+(\omega u_i^T)}{P_-(\omega u_i^T)})^{\frac{1}{2}} + \frac{1}{N}\sum_{y_i=+1}(\frac{P_-(\omega u_i^T)}{P_+(\omega u_i^T)})^{\frac{1}{2}} \quad (7)$$

$$\tilde{J}_2(L^*|\tau) = \frac{1}{N}\sum_{y_i=-1}(\frac{P_+(\tau s_i^T)}{P_-(\tau s_i^T)})^{\frac{1}{2}} + \frac{1}{N}\sum_{y_i=+1}(\frac{P_-(\tau s_i^T)}{P_+(\tau s_i^T)})^{\frac{1}{2}} \quad (8)$$

This obtains a sub-optimum for minimizing (5).

Suppose $\tilde{J}_1(L^*|\omega)$ and $\tilde{J}_2(L^*|\tau)$ are minimized at ω^* and τ^* with value $\tilde{J}_1^*$ and $\tilde{J}_2^*$ separately , then $\begin{pmatrix} \omega^* & 0 \\ 0 & \tau^* \end{pmatrix}$ is chosen as a approximation to the optimal matrix W^*. After normalizing (J_1^*, J_2^*), we set $\lambda_1 = \frac{1}{2}\log\frac{1-\tilde{J}_1^*}{\tilde{J}_1^*}$ and $\lambda_2 = \frac{1}{2}\log\frac{1-\tilde{J}_2^*}{\tilde{J}_2^*}$. Also normalize (λ_1, λ_2), then the tuned feature can be expressed as:

$$L^*(u,s) = \lambda_1 \log\frac{P_+(\omega^* u^T)}{P_-(\omega^* u^T)} + \lambda_2 \log\frac{P_+(\tau^* s^T)}{P_-(\tau^* s^T)} \quad (9)$$

This way, the number of candidate features is reduced to $5^3 + 5^2 = 150$. Using a trick described in [3], this number can be further decreased to 57.

Fig.3 illustrates the advantage of feature extraction in appearance-spatial space over appearance-only space. The two vehicles are nearby and form a group during the tracking. The histograms of the two objects in the two spaces, when the objects merge, are shown in Fig.4 . About 20% of the pixels is "hard " to be classified in appearance space, and this deteriorates the tracking in subsequent frames, as shown in the third row of fig.3. On the other hand, the error rate of the spatial-feature-based classifier is as low as 3.5%. Combining features in the two spaces improves the classification accuracy, as is illustrated in the fourth row of the fig.3.

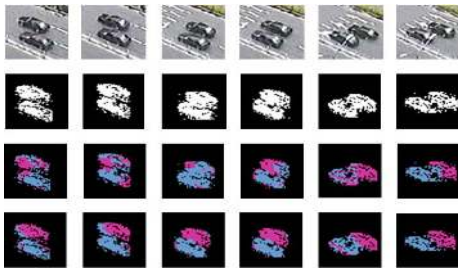

Figure 3. Tracking two nearby moving vehicles which are spatially separable but un-separable by apearance. Row 1:original image. Row 2: Foreground blob. Row 3: Classification results using appearance feature only. Row 4: Classification results using both appearance-spatial feature

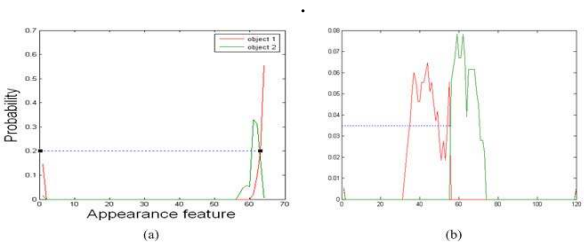

Figure 4. Histograms of the two objects in the RGB space (a) and in the spatial space (b).

The location of the tracked object can be calculated easily based on the labelling result of the pixels through mean-shift procedure[4].

3. Experimental Results

A real-time outdoor tracking system has been implemented based on the methods presented in this paper. The system is running on a standard PC hardware (Pentium IV at 3.0GHz) and works at 10-15 fps. We take the videos captured by a camera system set around our building for test and comparison. The video image size is 320×240(24 bits per pixel).

Fig.5 shows results of tracking cars on road . Tracking methods based on color, shape or appearance information are likely to fail in this situation. The first row shows results obtained using the method of [3]: At frame 53, the two vehicles merge into a group, where a white rectangle is used to indicate merge into a single blob; and it is unable to recover two different objects when they split at frame 65. In contrast, the use of the spatial constraint in the present methods can give correct results after the objects merge or

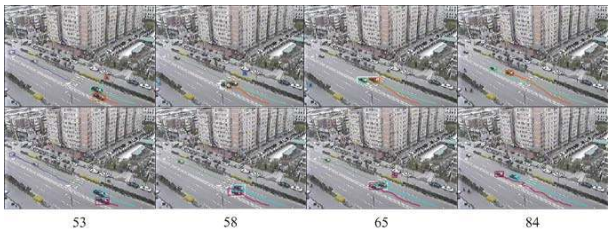

Figure 5. Tracking cars on road. Appearance-only based methods fail (row 1). The appearance-spatial feature based method works well (row 2).

split (the second row). Table 1 provides some feature extraction results.

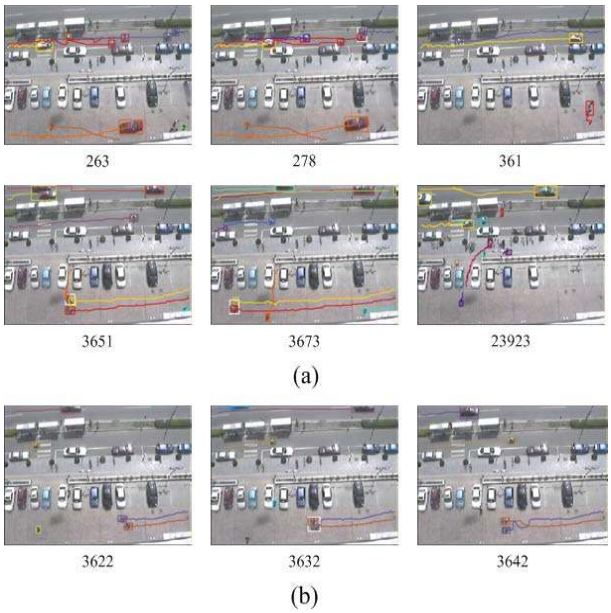

Figure 6. Tracking people walking and cycling in a parking lot. (a) Our method well handled in this challenge sequence for departure from on another (263∼361), merge and split (3651∼3673). (b) Tracking fails by appearance only method.

Fig.6 shows results of tracking people walking and cycling in a parking lot. A parking lot surveillance video about 30 minutes is used for this test. It is a challenging dynamic scene for multiple objects tracking. Difficulties due to partial occlusion, merge, split and departure from one another are frequently encountered during the period. Fig.6(a) shows results of the present method. In frame 263, 278 and 361, at the bottom side, a person and a red car depart from each other. In the upper part of the video images, six vehicles or bicycles are moving, five from left to the right and the other in the other direction. Our results made

Table 1. Feature extraction results for Fig. 5 and Fig.6

Frame ID	λ_1	λ_2	W^*
(Fig.5) 58	0.44	0.56	$\begin{pmatrix} 1 & 1 & -2 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 \end{pmatrix}$
(Fig.5) 65	0.44	0.56	$\begin{pmatrix} 1 & 1 & -2 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 \end{pmatrix}$
(Fig.6(a)) 278	0.13	0.87	$\begin{pmatrix} -2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$
(Fig.6(a)) 3651	0.07	0.93	$\begin{pmatrix} -2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$
(Fig.6(a)) 3673	0.06	0.94	$\begin{pmatrix} -2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$

no mal-tracking after merge and split happened five times. In frames 3651 and 3673, at the lower part, three persons interacted to each other and the trajectories were still accurately computed for each target. Frame 23923 is the last frame of the sequence. Our method worked well throughout the sequence. Some feature extraction results are shown in table 1. Note that the tracker pick the spatial coordinates as discriminative feature over time. Fig.6(b) shows a failure results made by the method of [3] for frames 3622 to 3642, in which only appearance space is used for feature extraction..

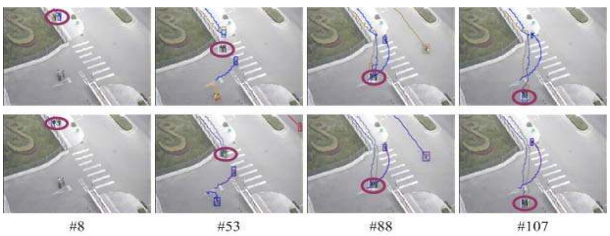

Figure 7. Pedestrian tracking. Two pedestrians (highlighted by red ellipse) merge into a group and walk nearby until they leave the scene. Row 1: results obtained by appearance only tracker. Row (2) results obtained by present tracker.

Fig.7 shows results of pedestrian tracking, where the motion of objects involves direction changes. The two pedestrians (high lighted by red ellipse) moved into the scene and then merged into a group after frame 8. In frame 53 they turn right together at the parking lot entrance. Velocity-prediction based trackers will fail in such cases. So is appearance-only based method (row 1). In contrast, our tracker tracked every person correctly (row 2) until they leave the scene.

4. Conclusion

In this paper, we have developed a unified, theoretically justified, framework for classification and feature extraction for tracking multiple objects in group. The novel contributions are: (1) The two problems are solved by minimizing an error bound function; this is demonstrated by theoretical analysis derivation. (2) Spatial constraint, in addition to appearance features space, is used to achieve better tracking performance than using either appearance or spatial features, as has been demonstrated by experimental results.

5 Acknowledgement

This work was supported by the following funding: Chinese National Natural Science Foundation Project 60518002, and Chinese National 863 Program Projects 2004AA1Z2290 & 2004AA119050.

Appendix A

Proof of the identity between Bayesian upper-bound and the Bhattacharyya coefficient.

$$\begin{aligned} J(L^*) &= \log(E(e^{-yL^*(x)})) \\ &= \log(E(e^{-L^*(x)}|y=+1)P(y=+1) \\ &\quad +E(e^{-L^*(x)}|y=-1))P(y=-1) \\ &= \log(E((\frac{P_-(x)}{P_+(x)})^{\frac{1}{2}}|y=+1)P(y=+1) \\ &\quad +E((\frac{P_+(x)}{P_-(x}})^{\frac{1}{2}}|y=-1)P(y=-1)) \\ &= \log(P(y=+1)\int(\frac{P_-(x)}{P_+(x)})^{\frac{1}{2}}P_+(x)dx \\ &\quad +P(y=-1)\int(\frac{P_+(x)}{P_-(x)})^{\frac{1}{2}}P_-(x)dx) \\ &= log(\int(P_-(x)P_+(x))^{\frac{1}{2}}dx) \end{aligned} \quad (10)$$

The last form of expression is defined as Bhattacharyya coefficient between two distributions in [11]

References

[1] S. Avidan. Ensemble tracking. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 1:970–975, 2005.

[2] Y. Bar-Shalom and T. Fortmann. Tracking and data association. *Academic Press.*, 1998.

[3] R. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking fea-tures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 2005.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objets using mean shift. *In Proceeding of IEEE Conference on Computer Vision and Pattern Recongition.*, 2:142–149, 2000.

[5] A. Elgammal, R. Duraiswami, and L. Davis. Probabilistic tracking in joint feature-spatial spaces. *Proceedings IEEE Conference of Computer Vision and Pattern Recognition, Wisconsin, Madison*, 1:781–788, 2003.

[6] R. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics.*, 8:376–386, 1938.

[7] J. Friedman, T. Hastie, and Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 38, 2000.

[8] W. E. L. Grimson, C. Stauffer, R. Romano, L. Lee, P. Viola, and O. Faugeras. Forest of sensors: Using adaptive tracking to classify and monitor activities in a site. *in DARPA Im-age Understanding Workshop.*, 1998.

[9] M. Han, W. Xu, H. Tao, and Y. Gong. An algorithm for multiple object trajectory tracking. *Proceedings of Computer Vision and Pattern Recognition*, 1:864–871, 2004.

[10] M. Isard and A. Blake. Condensationcconditional density propagation for visual tracking. *International Journal of Computer Vision*, 29, 1998.

[11] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Tech.*, 15:52–60, 1967.

[12] F. Lv, J. Kang, R. Nevatia, I. Cohen, and G. Medioni. Automatic tracking and labeling of human activities in a video sequence. *eProceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS04), Prague, Czech Republic*, 2004.

[13] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80, 2000.

[14] M.Isard and J. MacCormick. Bramble: a bayesian multiple-blob tracker. *Proceedings of Eighth IEEE International Conference on Computer Vision*, 2:34–41, 2001.

[15] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time track-ing. *In IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[16] T. Yang, Q. Pan, J. Li, and S. Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2:494–501, 2005.