6

Multiple Image Objects Detection, Tracking, and Classification using Human Articulated Visual Perception Capability

HeungKyu Lee
MarkAny Corporation
Republic of Korea

1. Introduction

This chapter examines the multiple image objects detection, tracking, and classification method using human articulated visual perception capability in consecutive image sequences. The described artificial vision system mimics the characteristics of the human visual perception. It is a well known fact that a human being, first detects and focuses motion energy of a scene, and then analyzes only a detailed color region of that focused region using a storage cell from a human brain.

From this fact, the spatio-temporal mechanism is derived in order to detect and track multiple objects in consecutive image sequences. This mechanism provides an efficient method for more complex analysis using data association in spatially attentive window and predicted temporal location. In addition, occlusion problem between multiple moving objects is considered. When multiple objects are moving or occluded between them in areas of visual field, a simultaneous detection and tracking of multiple objects tend to fail. This is due to the fact that incompletely estimated feature vectors such as location, color, velocity, and acceleration of a target provide ambiguous and missing information. In addition, partial information cannot render the complete information unless temporal consistency is considered when objects are occluded between them or they are hidden in obstacles. To cope with these issues, the spatially and temporally considered mechanism using occlusion activity detection and object association with partial probability model can be considered. Furthermore, the detected moving targets can be tracked simultaneously and reliably using the extended joint probabilistic data association (JPDA) filter. Finally, target classification is performed using the decision fusion method of shape and motion information based on Bayesian framework. For reliable and stable classification of targets, multiple invariant feature vectors to more certainly discriminate between targets are required. To do this, shape and motion information are extracted using Fourier descriptor, gradients, and motion feature variation on spatial and temporal images, and then local decisions are performed respectively. Finally, global decision is done using decision fusion method based on Bayesian framework. The experimental evaluations show the performance and usefulness of introduced algorithms that are applied to real image sequences. Figure 1 shows the system block-diagram of multi-target detection, tracking, and classification.

In section 2, we describe the target detection and feature selection procedure employing occlusion reasoning from detail analysis of spatio-temporal video frame sequences. In section 3, multi-target tracking based on modified joint probabilistic data association filter is described. In section 4, we describe the multi-target classification using local and global decision rules based on Bayesian framework. Finally, concluding remarks are described in section 5.

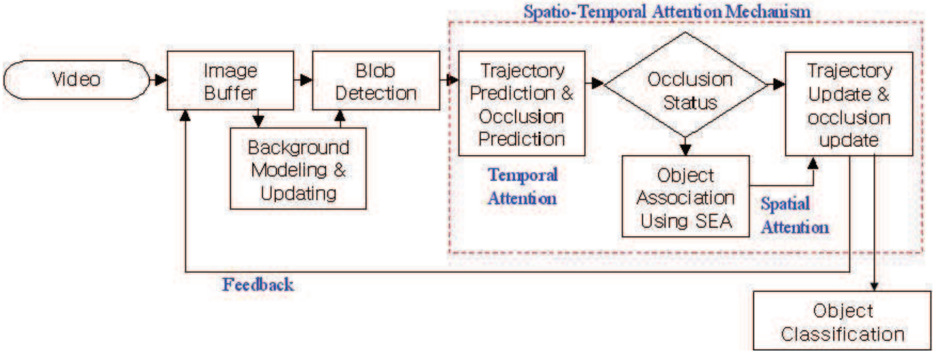

Spatio-Temporal Attention Mechanism
Video
Image Buffer
Blob Detection
Background Modeling & Updating
Trajectory Prediction & Occlusion Prediction
Temporal Attention
Occlusion Status
Object Association Using SEA
Trajectory Update & occlusion update
Spatial Attention
Feedback
Object Classification

Figure 1. System block-diagram for multi-target detection, tracking, and classification

2. Target Detection and Feature Selection

In video frame sequences, extracting moving blobs is very important task to identify the target. Its performance affects the accuracy of the detection, tracking and classification because the detected moving blobs might include the false alarms. To increase the accuracy of moving blob extraction, adaptive background model generation is required. From this model, accurate estimation of moving blob region can be done just by subtracting accurately estimated adaptive background model from original video frame. For doing this, lots of researches have been done (Y.L.Tian et al. 2005, C.Stauffer et al. 1999, A.Elgammal, et al. 2002, K. Kim, et al. 2004). These researches make the time variant background model using temporal information, and then subtract it from the original video frame sequence. In addition, spatial directional information using motion estimation can be applied.

2.1 Moving Blobs Detection

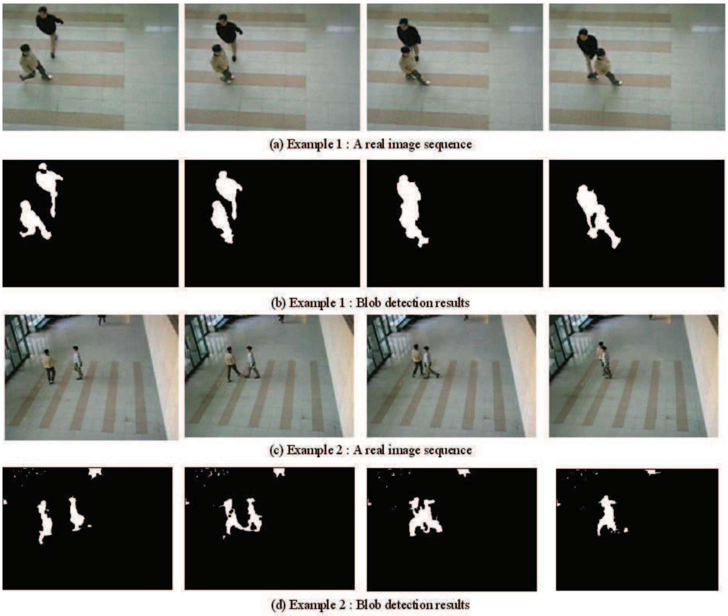
For moving blobs detection, the spatio-temporal information is very important task to accurately estimate the just moving parts from the complex background. The human eye first stimulates motion information such as time difference image to recognize the moving objects, and then focus on the spatial information such as detail color distribution in detected motion information group.

To mimic the human eye, motion information is first estimated. For doing this, moving blob detection is achieved by adaptively estimating the fixed background model and then by subtracting the background model from the original video frame sequences. For estimating background model in this chapter, the extended adaptive change detection algorithm (Huwer, et al. 2000) that improves change detection accuracy by combining both the temporal difference and the spatial difference using weighted accumulation is applied. The function that accumulates consecutive video frame sequences is given by $\phi(\hat{f}_i, f_i, \tau)$ representing a measure for the number of past values.

$$\begin{aligned}\hat{f}_{i+1}(x, y) &= \phi(\hat{f}_i(x, y), f_i(x, y), \tau) \\ &= f_i(x, y)\left(1 - e^{-1/\tau}\right) + \hat{f}_i(x, y)e^{-1/\tau}\end{aligned} \tag{1}$$

where f_i is the video frame sequences and τ is the length of the video frame accumulation. Using this accumulated video frames, mean background image μ_i is computed. Then, the detection of the background change region, B_i is then done by thresholding the absolute difference between the current video frame f_i and the mean background video frame, μ_i with the background standard deviations, σ_i. Figure 2 shows the moving blob detection example.

$$B_i(x, y) = \{(x, y) \in I / |\mu_i(x, y) - f_i(x, y)| > \sigma_i(x, y)\} \tag{2}$$


(a) Example 1 : A real image sequence

(b) Example 1 : Blob detection results

(c) Example 2 : A real image sequence

(d) Example 2 : Blob detection results

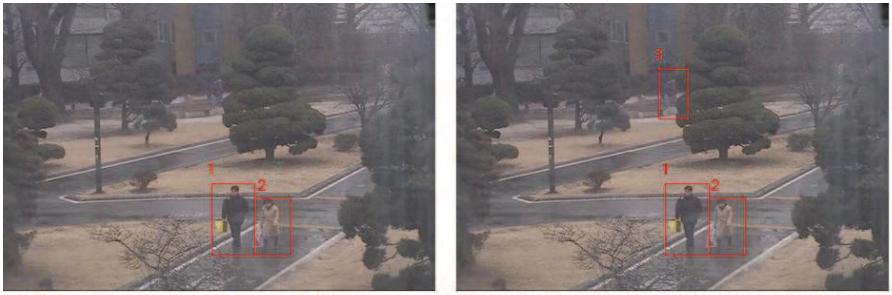
Figure 2. Moving blobs detection using time difference of current video frame and adaptive background model

In addition, the optical flow estimation (S.S. Beauchemin, et al. 1995) is done between the previous video frame sequence and current video frame sequence. The optical flow estimation result, B_{opt} is combined for the detection of the background change region as given

$$B_i(x, y) = \max(B_i(x, y), B_{opt}(x, y)) \tag{3}$$

Background adaptation procedure is recursively performed to deal with changes in illumination. To reliably detect moving blobs as shown in Figure 3, the time difference method using shape and motion information is applied as follows:

$$B_{t,i}(x,y) = \left|B_i(x,y) - f_i(x,y)\right| \tag{4}$$



(a) Object detection using shape information (b) Object detection using shape and motion information

Figure 3. Moving blobs detection example

On the segmented image, $B_{t,i}(x,y)$, a connected components analysis is then applied in order to fill holes in probable regions of interest. It is due to the fact that initial segmentation region is usually noisy. So, the low-pass filter and morphological operations are required. Next time, the segmented foreground region is labeled. At this time, a blob map of current video frame is computed. The blob map, $b_i(t)$ is represented by

$$b_i(t) = \bigcup_x \left|d_x(t) > \Gamma\right| \tag{5}$$

where $d_x(t)$ is a segmented foreground region, and Γ is a threshold to rule out small region. The blob map, $b_i(t)$ is recomputed for obtaining the color distribution (M. J. Swain, et al. 1991) of a blob as follows.

$$MB_{i,j}(x,y) = \begin{cases} f_i(x,y) & \text{if } \mathrm{b}_{\mathrm{i}}(x,y) == 1 \\ 0 & else \end{cases} \tag{6}$$

where $MB_{i,j}(x,y)$ is a moving bob having color model, i is a moving blob index, and j is a frame index. This moving blob color model can be used in order to associate a specific blob of occluded region with a real target when the occlusion status is enabled. Thus, this model is saved during short-time period. Meanwhile, it is not stored in queue during the occlusion status is enabled. We then compute the centroids (center points) of labeled blobs as feature vectors by calculating the geometric moment of moving blobs by using

$$M'_{p,q} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^p f(x,y)\,dx\,,dy \tag{7}$$

where f(x,y) is a moving blob to be analyzed and (p_x, p_y) is a centroid. The center point is stored at the trajectory variable, and it computes the width, $MV_w(i)$ and height, $MV_h(i)$ to represent the bounding region as a minimum bounding rectangle (MBR) (Rasmussen, et al. 1998). The respective centroid points in video frame sequences can give the object's kinematic status information such as walk, running, turn over and so on. Thus, we would be able to utilize them for analyzing objects behavior pattern.

2.2 Occlusion

In feature based multiple target tracking, occlusion issue is challenging one to be considered. Combining feature points derives the tracking failure on the tracking filter. Thus, the separation procedure should be done. To perform the separation procedure, detail analysis should be done in combined (or occluded) region between moving objects. For doing this, temporal information having time difference energy and motion can be utilized. If the modeling of object movement is applied, we can predict the object movement from the LTM(Long Term Memory). Thus, we can utilize the predicted motion information when the multiple objects are occluded between them or hidden back to obstacles even if it is an inaccurate estimation. For doing this, occlusion activity detection algorithm can be applied (H. K. Lee, et al. 2006). This method predicts the occlusion status of next step by employing a kinematics model of moving objects as shown in Figure 4, and notifies it for next complex analysis. Thus, this describes the temporal attention model. Then, the occlusion status is updated in current time of captured image after comparing the MBR of each object in attention window. Proposed occlusion activity detection algorithm has two-stage strategies as follows.

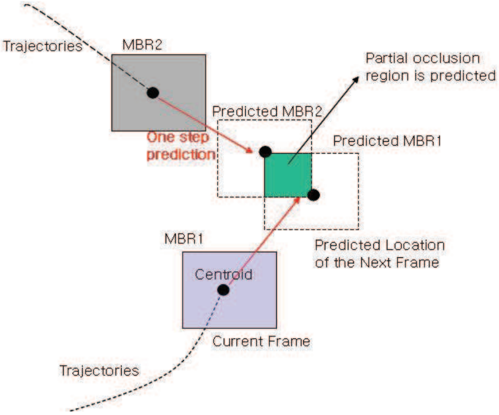


Figure 4. Occlusion reasoning and prediction using Kalman prediction

- STEP 1: Occlusion Prediction Stage

This step predicts the next center points of blobs by employing the Kalman prediction (Y. Bar-Shalom, et al. 1995) as follows:

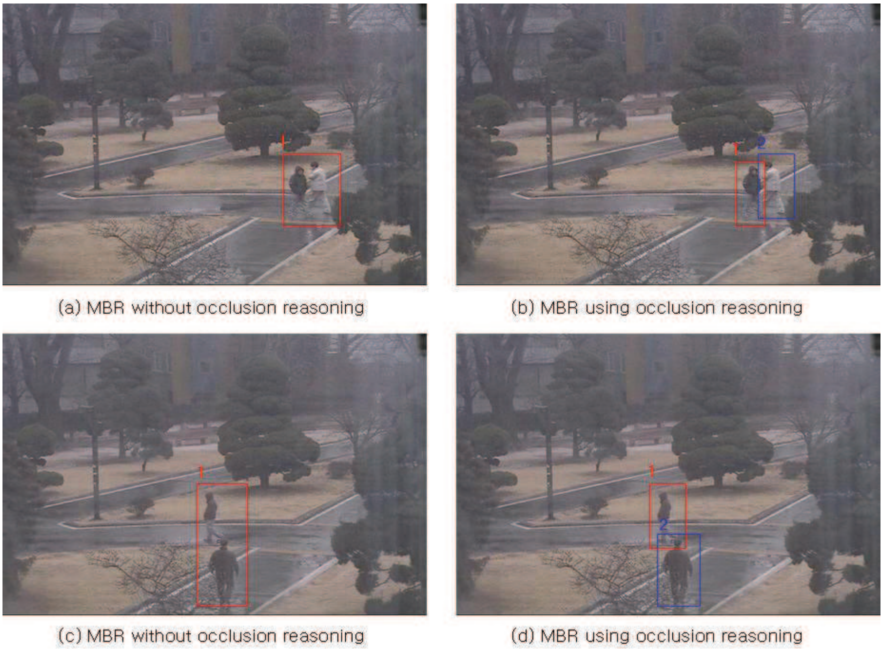
$$\hat{S}(k+1/k) = F(k)\hat{S}(k/k) + u(k) \tag{8}$$

$$\hat{Z}(k+1/k) = H(k+1)\hat{S}(k+1/k) \tag{9}$$

where S(k+1/k) is the state vector at time k+1 given cumulative measurements to time k, F(k) is a transition matrix, and u(k) is a sequence of zero-mean, white Gaussian process noise. Using the predicted center points, we can determine the redundancy of objects using the intersection measure in attention window. The occlusion activity is computed by comparing if or not there is an overlapping region between MBR_i of each object in the predicted center points as follows.

$$Fg = \begin{cases} 1 & if \ \left(MBR_i \cap MBR_j\right) \neq \phi \\ 0 & \quad otherwise \end{cases} \tag{10}$$

where the variable, i, j=1,…,m, the variable, Fg is an occlusion alarm flag, the subscript i and j are the index of the detected target at the previous frame, and m is a number of a target. If a redundant region has occurred at the predicted position, the probability of occlusion occurrence in the next step will be increased. Therefore, the occlusion activity status is notified for next complex analysis.



(a) MBR without occlusion reasoning

(b) MBR using occlusion reasoning

(c) MBR without occlusion reasoning

(d) MBR using occlusion reasoning

Figure 5. Minimum bounding rectangle for representing a validation region using occlusion reasoning

- STEP 2: Update Stage of Occlusion Status

The occlusion activity status can be updated in the current frame. The first, the size of the labeled blobs is verified whether they are contained within the validation region or not. If the shape of labeled blobs is contained within the validation region, the occlusion status flag is disabled. Otherwise, we conclude that the occlusion has occurred at the region, and the occlusion status is enabled. At this time, we apply the predicted center points of the previous step to the system model and the predicted MBR is recomputed as in Figure 5. Then, the Kalman gain is computed and the measurement equation is updated.

2.3 Feature Selection

Each feature sets describing multiple objects is integrated into a set of feature map. This feature map is used for visual search process to associate each blob with a real target. In this paper, color, location, velocity, and acceleration are used to describe object shape and model the kinematics of moving objects (Y. Bar-Shalom, et al. 1995).

Let o = [o_1, o_2, …, o_M] denote the set of objects to track, φ denotes the movement directions for object o_i and x=[x_i , y_i]^T denote the vector of points of center corresponding to o_i , with v= [$\dot{x}_i, \dot{y}_i$]^T ,where $\dot{x}_i$ and $\dot{y}_i$ denote the derivative of x_i and y_i with respect to t, respectively. First, center points of moving objects are computed, and then movement directions are computed using motion vectors extracted by the optical flow method (Kollnig, et al. 1994). To obtain the movement directions of objects, we compute the direction of motion vector for each pixel. The direction, φ of the vector is defined and computed using the Lucas-Kanade tracking equation (Tomasi, C. et al.) as follows:

$$\begin{aligned}\varphi(rad) &= angle(\frac{v_y}{v_x}) \qquad 0 \le \varphi < 2\pi \\ &= \{\varphi / \sin\varphi = v_y /\|v\|\} \cap \{\varphi / \cos\varphi = v_x /\|v\|\} \cap \{\varphi / \tan\varphi = v_y / v_x\}\end{aligned} \tag{11}$$

where v_x and v_y are motion vectors for x and y direction respectively, and $\|v\| = \sqrt{v_x^2 + v_y^2}$. From Equation (11), we know $\dot{x} = \|v\|\cos\varphi$ and $\dot{y} = \|v\|\sin\varphi$. The equations are differentiated with respect to t as follows.

$$\frac{d}{dt}\varphi = -\frac{1}{\|v\|\sin\varphi}\ddot{x} = \frac{1}{\|v\|\cos\varphi}\ddot{y} = \frac{1}{2\|v\|}\left(\frac{1}{\cos\varphi}\ddot{y} - \frac{1}{\sin\varphi}\ddot{x}\right) \tag{12}$$

Using equation (11) and (12), the proposed system model is given by

$$\dot{s} = \Psi s + \Pi u^e + v \qquad v \sim M(0, Q) \tag{13}$$

$$\Psi = \begin{bmatrix} O_{2\times2} & I_2 & O_{2\times2} & O_{2\times1} \\ O_{2\times2} & -G^{-1}\Sigma & O_{2\times2} & O_{2\times1} \\ O_{2\times2} & O_{2\times2} & O_{2\times2} & O_{2\times1} \\ O_{1\times2} & O_{1\times2} & \frac{1}{2\|v\|}[-\csc\varphi \quad \sec\varphi] & 0 \end{bmatrix} \tag{14}$$

$$\Pi = \begin{bmatrix} O_{2\times2} \\ -G^{-1}I_2 \\ O_{2\times2} \\ O_{1\times2} \end{bmatrix} \tag{15}$$

where $O_{m\times n}$ is an m×n zero matrix, I_m is an m×m identity matrix and s = [x^T , v^T , a^T, φ]^T denote the system state, which is composed of center points, velocity, acceleration and direction of moving object. In the proposed method, the acceleration component in state vector is included to cope with maneuvering of object. The model assumes random acceleration with covariance Q, which accounts for changes in image velocity. As the eigen-values of Q become larger, old measurements are given relatively low weight in the adjustment of state. This allows the system to adapt to changes in the object velocity. Since time interval Δt between one frame and next is very small, it is assumed that F is constant over the (t_k , t_{k+1}) interval of interest. The state transition matrix is simply given by

$$F_k = e^{\Psi\Delta t} = \begin{bmatrix} I_2 & I_2\Delta t & \frac{\Delta t^2}{2} I_2 & O_{2\times1} \\ O_{2\times2} & I_2 - G^{-1}\Sigma\Delta t & O_{2\times2} & O_{2\times1} \\ O_{2\times2} & O_{2\times2} & I_2 & O_{2\times1} \\ O_{1\times2} & O_{1\times2} & \frac{\Delta t}{2\|v\|}[-\csc\varphi \quad \sec\varphi] & 1 \end{bmatrix} \tag{16}$$

Let z = [z_1, z_2, …, z_M] and z_i denote the measurement vector for object o_i . In the proposed model, center points and movement directions for each object are treated as system measurements. The measurement vector satisfies:

$$z_i = Hs + w \qquad w \sim N(0, R) \tag{17}$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{18}$$

where matrix H connects the relationship between z_i and s. After all, the object kinematics model is determined by setting the appropriate parameters.

3. Multi-Target Tracking using Data Association

For multi-target tracking, the joint probabilistic data association filter (JPDA) (Y. Bar-Shalom, et al. 1995 , Samuel Blackman, et al. 1999, Rasmussen, et al. 1998) is applied. Similarly to the PDA algorithm (Y. Bar-Shalom, et al. 1995 , Samuel Blackman, et al. 1999), the JPDA computes the probabilities of association of only the latest set of measurements Z(k) to the various targets. The key to the JPDA algorithm is the evaluation of the conditional probabilities of the following joint association events pertaining to the current time k. First, it computes the probabilities of association of only the latest set of moving blob Z(k) to the targets to pursue multiple people simultaneously. Next, steps depict the process for calculating the association probability between multiple people.

Step 1: Construction of validation matrix

First, it defines the validation matrix for the evaluation of the conditional probabilities of the following joint association events pertaining to the current image frame, k.

$$\theta = \bigcap_{j=1}^{m_k} \theta_{jt_j} \tag{19}$$

where θ_{jt} is the moving blob, j originated from person, t, j=1,…,m_k, t=0,…,T. A joint association event, θ can be represented by the matrix;

$$\hat{\Omega}(\theta) = \left[\hat{\omega}_{jt}(\theta)\right] \tag{20}$$

consisting of the units in Ω corresponding to the association in θ, ie.,

$$\hat{\omega}_{jt}(\theta) = \begin{cases} 1 & \text{if } \theta_{jt} \subset \theta \\ 0 & otherwise \end{cases} \tag{21}$$

At this point, a moving blob can have only one source, and no more than one moving blob can originate from one person. This is a necessary condition for validation matrix. On the contrary, if an occlusion is occurred, such a condition is not satisfied.

1. Occlusion Case:

Using the recalculated moving blobs, the proposed system satisfies above condition. It employs the state transition model to handle various occlusion scenarios according to the state transition mode (occlusion mode and non-occlusion mode) within the JPDA filter. The transition of the current state that can be altered according to occlusion prediction and detection rules is just conditionally processed. The occlusion process that consists of the procedure of occlusion prediction and detection, and a splitting of coupled objects according to state transition mode, is performed.

Figure 6 shows a state transition diagram with two states. Each state is used to reflect the states of occlusion at each image frames. Under the occlusion state, a recalculating procedure of the occluded people is performed and then the tracking flow is continued. Seven transition modes are applied as follows. (1) A specific target enters into the scene. (2) Multiple targets enter into the scene. (3) A specific target is moving and forms a group with other targets, or just moves beside other targets or obstacles. (4) A specific target within the group leaves a group. (5) A specific target continues to move alone, or stops moving and then starts to move again. (6) Multiple targets in a group continue to move and interact between them, or stop interacting and then start to move again. (7) (8) A specific target or a group leaves a scene. The events of (1), (4), (5), and (7) can be tracked using general Kalman tracking. In addition, the events of (2), (3), (6) and (8) can be tracked reliably using predictive estimation method.

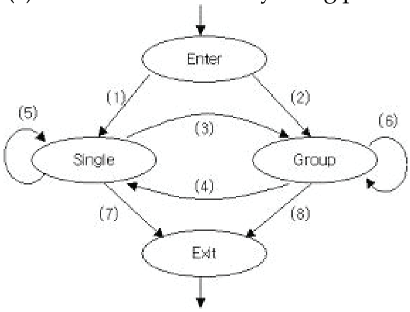


Figure 6. State transition diagram using occlusion reasoning

2. Non-Occlusion Case:

Under the non-occlusion state, a normal JPDA tracking filtering is performed.

Step 2: Compute Joint Association Probability

The purpose at this step is to compute the marginal association probability, β_{jt} that is the probability to be associated between j-th moving blob and person t at current frame k using image sequences. Then, in order to estimate the state and for the purpose of deriving the joint probabilities, it defines the person detection indicator $\delta_t(\theta)$, the moving blob association indicator $\tau_j(\theta)$ and the number of false alarm blobs $\phi(\theta)$ as in Equation (4-22), (4-23), and (4-24).

$$\delta_t(\theta) \equiv \sum_{j=1}^{m_k} \hat{\omega}_{jt}(\theta) \leq 1, \qquad t = 1,\ldots,T \tag{22}$$

$$\tau_j(\theta) \equiv \sum_{t=1}^{T} \hat{\omega}_{jt}(\theta), \qquad j = 1,\ldots,m_k \tag{23}$$

$$\phi(\theta) = \sum_{j=1}^{m_k} \left[1 - \tau_j(\theta)\right] \tag{24}$$

1. Conditional Probability:

The conditional probability of the joint association event, $\theta(k)$ given the set Z^k of validated moving blobs at current image frame k using Bayes' rule is as follows:

$$\begin{aligned} P\{\theta(k)/Z^k\} &= P\{\theta(k)/Z(k), Z^{k-1}\} \\ &= \frac{1}{c} p\left[Z(k)/\theta(k), Z^{k-1}\right] P\{\theta(k)/Z^{k-1}\} \\ &= \frac{1}{c} p\left[Z(k)/\theta(k), Z^{k-1}\right] P\{\theta(k)\} \end{aligned} \tag{25}$$

where c is the normalization constant.

2. Likelihood Function

The PDF on the right-hand side in equation (25) is

$$p\left[Z(k)/\theta(k), Z^{k-1}\right] = \prod_{j=1}^{m_k} p\left[z_j(k)/\theta_{jt_j}(k), Z^{k-1}\right] \tag{26}$$

The conditional PDF of a moving blob given its origin is assumed to be

$$p\left[z_j(k)/\theta_{jt_j}(k), Z^{k-1}\right] = \begin{cases} N_{t_j}\left[z_j(k)\right] & \text{if } \tau_j\left[\theta(k)\right] = 1 \\ V^{-1} & \textit{otherwise} \end{cases} \tag{27}$$

where a moving blob associated with person t_j has Gaussian PDF. Moving blobs not associated with any person are assumed uniformly distributed in the field of view of volume V. Using Equation (27), the PDF (26) can be written as follows:

$$p\left[Z(k)/\theta(k), Z^{k-1}\right] = V^{-\phi(\theta)} \prod_{j=1}^{M_e} \left[N\left(\hat{x}_j; x_i, \Sigma_i\right)\right]^{\tau_j(\theta)} \tag{28}$$

3. Prior Probability:

The prior probability of a joint association event $\theta(k)$ combining equations (30) and (31) in equation (29) yields the equation (32).

$$\begin{aligned} P\{\theta(k)\} &= P\{\theta(k), \delta(\theta), \phi(\theta)\} \\ &= P\{\theta(k)/\delta(\theta), \phi(\theta)\} \cdot P\{\delta(\theta), \phi(\theta)\} \end{aligned} \tag{29}$$

Assuming each event a priori equally likely, first factor in equation (27) has

$$P\{\theta(k)/\delta(\theta), \phi(\theta)\} = \left(P^{m_k}_{m_k - \phi(\theta)}\right)^{-1} = \left(\frac{m_k!}{\phi!}\right)^{-1} = \frac{\phi!}{m_k!} \tag{30}$$

and the last factor is

$$P\{\delta(\theta),\phi(\theta)\}=\prod_{t=1}^{T}\left(P_D^t\right)^{\delta_t}\left(1-P_D^i\right)^{1-\delta_t}\mu_F(\phi) \tag{31}$$

where P_D^t is the detection probability of person t and $\mu_F(\phi)$ is the prior PMF of the number of false moving blobs.

$$P\{\theta(k)\}=\frac{\phi(\theta)!}{\varepsilon\cdot m_k!}\prod_{t=1}^{T}\left(P_D^t\right)^{\delta_t}\left(1-P_D^t\right)^{1-\delta_t} \tag{32}$$

The joint association probabilities with Poisson prior are

$$P\{\theta(k)/Z^k\}=\frac{\lambda^{\phi}}{c'}\prod_{j=1}^{m_k}\left[N_{t_j}\left(z_j(k)\right)\right]^{\tau_j}\prod_{t=1}^{T}\left(P_D^t\right)^{\delta_t}\left(1-P_D^t\right)^{1-\delta_t} \tag{33}$$

where c′ is the new normalization constant and λ is the special density of false moving blobs.

4. Association Probability:

Thus the marginal association probability β_{jt} is calculated as

$$\beta_{jt}\equiv P\{\theta_{jt}/Z^k\}=\sum_{\theta}P\{\theta/Z^k\}\widehat{\omega}_{jt}(\theta) \tag{34}$$

where j=1,…,m$_k$, t=0,…,T because a probabilistic inference can be made on the number of moving blobs in the validation region from the density of false alarms or clutter as well as on their location.

Step 3: State Estimation

Finally, the state estimation equation for each person is computed. The state is assumed to be normally Gaussian distributed according to the latest estimate and covariance matrix. The state update equation is processed as

$$\hat{x}(k/k)=\hat{x}(k/k-1)+W(k)\nu(k) \tag{35}$$

where

$$\nu(k)\equiv\sum_{i=1}^{m_k}\beta_i(k)\nu_i(k) \tag{36}$$

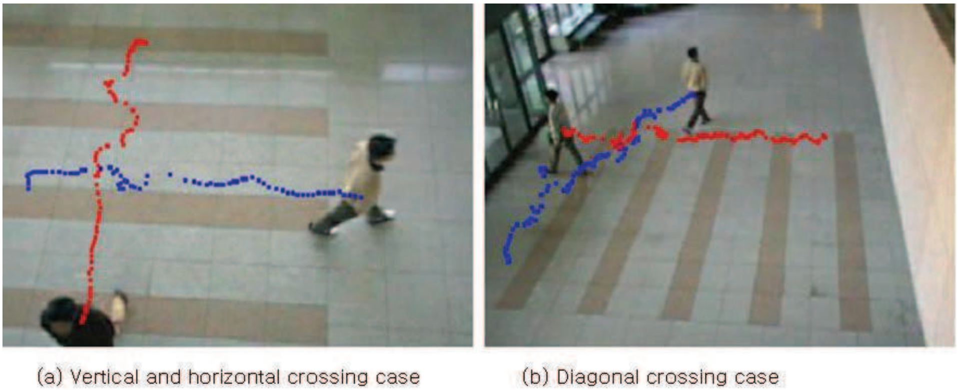
It is highly nonlinear due to the probabilities $\beta_i(k)$ that depend on the innovations. Unlike the standard Kalman filter, the covariance equation is independent of the moving blobs and the estimation accuracy of the error covariance

$$P(k/k)=\beta_0(k)P(k/k-1)+[1-\beta_0]P^c(k/k)+\widetilde{P}(k) \tag{37}$$

Associated with the update state estimate depends upon the data that are actually encountered. Prediction of the state and measurement to image frame k+l is done as in the standard Kalman filter. This JPDA filter extended for resolving occlusion problem in image based tracking is recursively processed. If the step 3 is finished, the step 1 is started again repeatedly in image sequences.

For experimental evaluation, obtained video files were sampled at video rate: example 1 (Figure 7, (b)) (total 640 frames, 15 frames per seconds, and its size is 240×320) and example

2 (Figure 7, (a)) (total 570 frames, 15 frames per seconds, and its size is 240×320) which is processed in a gray level image. In the initial value of the JPDA algorithm to track multi-targets in Figure 7, the process noise variance = 10 and the measurement noise variance = 25 are used. An occlusion state is maintained for 34, 24 frames respectively. We assumed that we know the size of a target to track within field of view. Assumed size of target is set with the following parameters: validation region is (100 pixel, 60~150 pixel) in example 1. In example 2, validation region is (100~120 pixel, 60~170 pixel).



(a) Vertical and horizontal crossing case

(b) Diagonal crossing case

Figure 7. Multi-target tracking result and its trajectories

Robustness has been evaluated mainly in terms of location accuracy and error rate of feature extraction and capability to track under occlusion in complex load scenes. The table 1 is an error rate that extracted blobs are not targets within field of view.

Error Rate(ε)	Error rate of feature extraction		
	Method 1(H. K. Lee, et al. 2005)	Method 2(H. K. Lee, et al. 2005)	Spatio-temporal attention Scheme (H. K. Lee, et al. 2006)
Example 2	0.786	0.796	0.561
Example 1	0.424	0.341	0.336

Table 1. Simulation result of test video sequences

4. Target Classification using Decision Fusion

In this section, the decision problem is considered as binary hypothesis testing to classify the given features into human, vehicle, and animal (H. K. Lee, et al. 2006). From multivariate feature vectors, respective local decisions are made. To extract multivariate feature vectors, shape and motion information are computed using Fourier descriptor, gradients, and motion feature variation (A. J. Lipton, et al. 1998, Y. Kuno, et al. 1996) derived from equation (6) on spatial and temporal images. And then, we apply the global fusion rule based on Bayesian framework to combine local decisions ui, i=1,2,3 based on some optimization criterion for global decision u0. This method provides effective method for the combined decision of respective local feature analysis obtained from shape and motion information.

Once the foreground region is extracted, proposed system consists of three feature extraction procedures: First, Fourier descriptor and gradients representing the shape that is invariant to several change such as translation, rotation, scale, and starting point, is computed, and then classification task for local decision is performed using neural network. Second, classification task for local decision using temporal information is performed using motion information to be obtained from rigidity condition analysis of moving objects. For doing this, skeletonization of the motion region is done, and then motion analysis is done to compute motion feature variation using selected feature points (R. Cutler, et al. 2000, H. Fujiyosi, et al. 2004). Finally, we can classify moving objects through decision fusion method based on Bayesian framework using locally obtained results from shape and motion analysis (M. M. Kokar, et al. 2001, Li. X. R., et al. 2003).

Then, we derive the optimum fusion rules that minimize the average cost in a Bayesian framework (M. M. Kokar, et al. 2001, Li. X. R., et al. 2003). This rule is given by the following likelihood ratio test:

$$\frac{P(u_1,u_2,u_3/H_1)}{P(u_1,u_2,u_3/H_0)} \overset{u_0=1}{\underset{u_0=0}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})} \cong \eta \tag{38}$$

where Cij is the cost of global decision. The left-hand side can be rewritten as given in equation (39) because the local decisions have characteristic of independence.

$$\begin{aligned}\frac{P(u_1,u_2,u_3/H_1)}{P(u_1,u_2,u_3/H_0)} &= \prod_{i=1}^{3}\frac{P(u_i/H_1)}{P(u_i/H_0)} \\ &= \prod_{S_1}\frac{P(u_i=1/H_1)}{P(u_i=1/H_0)}\prod_{S_0}\frac{P(u_i=0/H_1)}{P(u_i=0/H_0)}\end{aligned} \tag{39}$$

where Sj is the set of all those local decisions that are equal to j, j=0,1. In addition, equation (39) can be rewritten in terms of the probabilities of false alarm and miss in detector i. That is why each input to the fusion center is a binary random variable characterized by the associated probabilities of false alarm and miss.

$$\prod_{S_1}\frac{P(u_i=1/H_1)}{P(u_i=1/H_0)}\prod_{S_0}\frac{P(u_i=0/H_1)}{P(u_i=0/H_0)} = \prod_{s_1}\frac{1-P_{M_i}}{P_{F_i}}\prod_{s_0}\frac{P_{M_i}}{1-P_{F_i}} \tag{40}$$

where PFi=P(ui=1/H0) and PMi=P(ui=0/H1). We substitute equation (40) in equation (38) and take the logarithm of both sides as follows:

$$\sum_{s_1}\log\frac{1-P_{M_i}}{P_{F_i}} + \sum_{s_0}\log\frac{P_{M_i}}{1-P_{F_i}} \overset{u_0=1}{\underset{u_0=0}{\gtrless}} \log\left(\frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})}\right) \cong \log\eta \tag{41}$$

The equation (41) can be expressed as shown in equations (42) and (43).

$$\sum_{i=1}^{3}\left[u_i\log\frac{1-P_{M_i}}{P_{F_i}} + (1-u_i)\log\frac{P_{M_i}}{1-P_{F_i}}\right] \overset{u_0=1}{\underset{u_0=0}{\gtrless}} \log\eta \tag{42}$$

$$\sum_{i=1}^{3}\left[\log\frac{(1-P_{M_i})(1-P_{F_i})}{P_{M_i}P_{F_i}}\right]u_i \overset{u_0=1}{\underset{u_0=0}{\gtrless}} \log\left[\eta\prod_{i=1}^{3}\left(\frac{1-P_{F_i}}{P_{M_i}}\right)\right] \tag{43}$$

Thus, the optimum fusion rule is applied by forming a weighted sum of the incoming local decisions, and then comparing it with a threshold. At this time, the threshold depends on the prior probabilities and the costs.

For experimental evaluation, training images are obtained in real image sequences. Each class includes three kinds of image view having front, side, and inclined image. Each kinds of image are composed of total 1200 files respectively for training, which is extracted from respective image sequences. Test images were also obtained from image sequences (image size is 480×320) about three targets (human, car, and animal): human (total 400 frames), car (total 400 frames), and animal (total 400 frames) which is a gray level image.

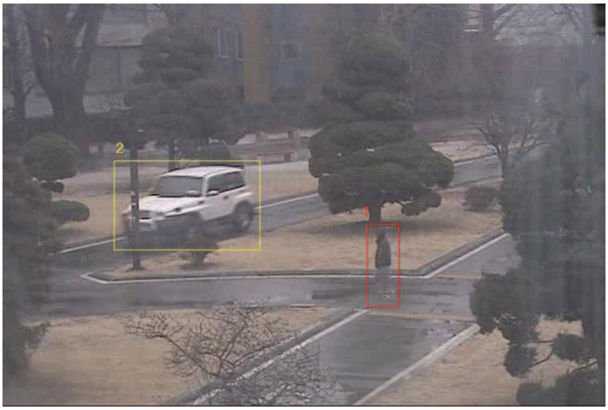


Figure 8. Object detection and classification

Figure 8 shows the detection and classification result of moving objects. To show the robustness of proposed method, we evaluated the proposed method which is compared to other methods such as neural net, and some fusion methods (Y. Bar-Shalom, et al. 1995, Samuel Blackman, et al. 1999, R. R. Brooks, et al. 1998). Majority voting and weighted average score method are fusion methods that are compared to the proposed fusion method. Table 2 shows the experimental results of three methods respectively with respect to the FAR(False Acceptance Rate) and the FRR(False Rejection Rate). Local decision method using neural network showed the lowest classification rate. Each local decision did not show satisfactory results. Thus, some fusion methods are secondly compared using respective local decisions. From the results, majority voting and weight average score methods showed low performance compared to the proposed method considering spatio-temporal information. Thus, we can know that the simple fusion method of final decision does not bring the good performance. In addition, we can know that the fusion method of redundant and complementary information is good choice in feature level and decision level fusion.

Method	FRR(%)	FAR(%)	Recognition Rate(%)
Neural Net	4.0	3.0	96
Majority Voting	2.7	2.5	97.3
Weight Average Score	2.5	2.0	97.5
Decision fusion	1.5	1.3	98.5

Table 2. Experimental evaluations compared to some methods

5. Discussions and Concluding Remarks

The objects are identified consciously within the attentional aperture from human visual system. The particular region of interest rendering motion sensation is focused, and then the complex analysis can be applied. By using this concept, both temporal attention and spatial attention can be considered because temporal attention provides the predictable motion model, and spatial attention provides the detailed local feature analysis. From this fact, the spatio-temporal mechanism is derived in order to detect and track multiple objects in consecutive video frame sequences. This mechanism provided an efficient method for more complex analysis using data association in spatially attentive window and predicted temporal location.

The challenging issue is when multiple objects are moving or occluded between them in areas of visual field. At this time, a simultaneous detection and tracking of multiple objects tend to fail. This is due to the fact that incompletely estimated feature vectors such as location, color, velocity, and acceleration of a target provide ambiguous and missing information. In addition, partial information cannot render the complete information unless temporal consistency is considered when objects are occluded between them or they are hidden in obstacles. Thus, the spatially and temporally considered mechanism using occlusion activity detection and object association with partial probability model should be considered.

Besides, multi-target tracking task has lots of challenging issues under complex situations such as environmental weather conditions; snow, rain, fog, night, and so on. Thus, preprocessing stage is also seriously considered before moving blob detection process. Accurate moving blob detection would derive a high performance of target tracking and recognition. Thus, preprocessing techniques under natural environments would be studied using sensor fusion scheme such as CCDs and IR sensors.

6. References

Y.L.Tian and A.Hampapur, (2005). Robust Salient Motion Detection with Complex Background for Real-time Video Surveillance, in *Proc. Of IEEE Computer Society Workshop on Motion and Video Computing*, January.

C.Stauffer and W.E.L.Grimson, (1999). Adaptive background mixture models for real tracking. *Int. Conf. Computer Vision and Pattern Recognition*, Vol.2, pp.246-252.

A.Elgammal, R.Duraiswami, D.Harwood and L.Davis, (2002). Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance, *Proceeding of the IEEE*, Vol.90, No.7, July.

K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis, (2004). Background Modeling and Subtraction by Codebook Construction, *IEEE International Conference on Image Processing (ICIP).*

Huwer, S.and Niemann, H., (2000). Adaptive change detection for real-time surveillance applications, Visual Surveillance, *IEEE International Workshop* on, pp 37 -46, July.

S.S. Beauchemin and J.L.Barron, (1995). The Computation of Optical flow, *ACM Computing Surveys*, Vol.27, pp.433 - 466.

M. J. Swain and D. H. Ballard, (1991). Colour indexing, *International journal of Computer Vision*, 7(1):11-32.

Rasmussen, C, Hager, G.D, (1998). Joint probabilistic techniques for tracking multi-part objects, Computer Vision and Pattern Recognition, *Proceedings. IEEE Computer Society Conference* on, pp 16 -21, June.

H. K. Lee, June Kim, and Hanseok Ko (2006). Prediction Based Occluded Multi-Target Tracking Using Spatio-Temporal Attention, *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) Special Issue on Brain, Vision, and Artificial Intelligence*, Vol.20, No. 6, pp.1-14, World Scientific Press, Sept..

Y. Bar-Shalom and X. R. Li, (1995). Multitarget-multisensor tracking: principles and techniques, YBS Press.

Kollnig, Nagel, Otte, (1994). Association of Motion Verbs with Vehicle Movements Extracted from Dense Optical Flow Fields, *proc. of ECCV94*, pp. 338-350.

Tomasi, C. and Kanade, T., Detection and tracking of point features, *Tech. Rept.* CMUCS-91132, Pittsburgh:Carnegie Mellon University, School of Computer Science.

Samuel Blackman, Robert Popoli, (1999). Design and Analysis of Modern Tracking Systems, *Artech House*.

M. M. Kokar, and J. A. Tomasik, (2001). Data vs. decision fusion in the category theory framework, *Proc. 2nd Int. Conf. on Information Fusion*.

Li. X. R., Zhu, Y., Wang, J., Han, C., (2003). Optimal linear estimation fusion-Part I: Unified fusion rules, *IEEE Trans. Information Theory*. Vol. 49, No. 9, Sep.

A. J. Lipton, H. Fujiyosi, and R. S. Patil, (1998). Moving target classification and tracking from real-time video, *Proc of IEEE Workshop. on Applications of Computer Vision*, pp.8~14, 1998.

Y. Kuno, T. Watanabe, Y. Shimosakoda and S. Nakagawa, (1996). Automated detection of human for visual surveillance system. *Proc. of Int. Conf. on Pattern Recognition*, pp.865~869.

R. Cutler and L. S. Davis, (2000). Robust real-time periodic motion detection, analysis, and applications, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol.22 pp.781~796, Aug.

H. Fujiyosi and J. A. Tomasik, (2004). Real-time human motion analysis by image skeletonization, *IEICE Trans, on Info & Systems*, Vol. E87-D, No. 1, Jan.

M. M. Kokar, and J. A. Tomasik, (2001). Data vs. decision fusion in the category theory framework, *Proc. 2nd Int. Conf. on Information Fusion*.

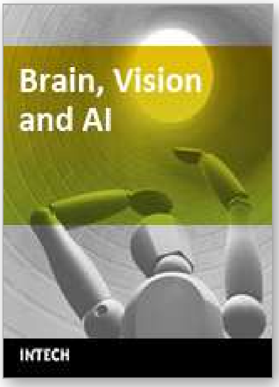
Li. X. R., Zhu, Y., Wang, J., Han, C., (2003). Optimal linear estimation fusion-Part I: Unified fusion rules, *IEEE Trans. Information Theory*. Vol. 49, No. 9, Sep.

R. R. Brooks and S. S. Iyengar, (1998). Multi-Sensor Fusion: Fundamentals and Applications with software, Prentice Hall.

H. K. Lee, and Hanseok Ko, (2005). Occlusion Activity Detection Algorithm Using Kalman Filter for Detecting Occluded Multiple Objects, *Computational Science*, pp.139-146, LNCS 3514, Springer, May. 2005.

H. K. Lee, and Hanseok Ko, (2005). Spatio-Temporal Attention Mechanism For More Complex Analysis To Track Multiple Objects, *Brain, Vision and Artificial Intelligence*, pp.447-456, LNCS 3704, Springer, October.

H. K. Lee, Jungho Kim, and June Kim, (2006). Decision Fusion of Shape and Motion Information Based on Bayesian Framework for Moving Object Classification in Image Sequences, *Foundations of Intelligent Systems, LNAI 4203*, pp.19-28, Springer, Sept.



Brain, Vision and AI
Edited by Cesare Rossi

ISBN 978-953-7619-04-6
Hard cover, 284 pages
Publisher InTech
Published online 01, August, 2008
Published in print edition August, 2008

The aim of this book is to provide new ideas, original results and practical experiences regarding service robotics. This book provides only a small example of this research activity, but it covers a great deal of what has been done in the field recently. Furthermore, it works as a valuable resource for researchers interested in this field.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821