Tracking Groups of Pedestrians in Video Sequences*

Jorge S. Marques, IST / ISR, Lisbon, Portugal
Pedro M. Jorge, ISEL / IST, Lisbon, Portugal
Arnaldo J. Abrantes, ISEL, Lisbon, Portugal
J. M. Lemos, INESC-ID / IST, Lisbon, Portugal

Abstract

This paper describes an algorithm for tracking groups of objects in video sequences. The main difficulties addressed in this work concern total occlusions of the objects to be tracked as well as group merging and splitting. A two layer solution is proposed to overcome these difficulties. The first layer produces a set of spatio temporal strokes based on low level operations which manage to track the active regions most of the time. The second layer performs a consistent labeling of the detected segments using a statistical model based on Bayesian networks. The Bayesian network is recursively computed during the tracking operation and allows the update of the tracker results everytime new information is available. Experimental tests are included to show the performance of the algorithm in ambiguous situations.

1 Introduction

Video surveillance aims to track and classify human activities, providing the human operator with augmented perception capability [9]. This task is usually divided into several processing layers, e.g., region detection, object tracking and activity classification.

Efficient algorithms have been developed to detect active regions in video sequences e.g., using statistical models of the background image [11, 10]. Nonlinear operations are then used to discard small regions and cluster neighboring pixels into connected active regions. Higher level operations use the output of the low level layer to track moving objects present in the scene and to classify their behavior. This is done by associating regions with similar properties detected in consecutive frames (region tracking). Region sequences are then classified using statistical pattern recognition techniques e.g., Hidden Markov Models [8].

Most region trackers perform well in simple cases. However, in practice, the objects to be tracked interact forming groups: a single track may correspond to several objects. Therefore the tracking algorithm must be able to cope with group formation and splitting. Splitting is a very difficult problem since the tracker must distinguish which object belongs to each subgroup. A second difficulty concerns temporarily occlusions produced either by static or by interacting objects. In this case, it is not possible to detect the objects trajectories but only a set of non connected segments. The estimation of the objects trajectories from the observed segments is a difficult problem since several hypothesis have to be considered.

Figure 1: Group formation

Tracking with occlusions has been thoroughly studied by several authors (e.g. see [11, 4]). On the other hand, tracking groups of objects is much less studied, being still an open issue (e.g., see [2, 6]). Most tracking systems solve both difficulties (occlusions and groups) using hard decisions based on instantaneous rules. This approach works well in simple cases but the performance of these systems is poor in large problems with complex interactions. For example, when an object is occluded for a long time or when a group is split a large ambiguity is created which can not be instantaneously removed.

A different approach was recently proposed in [1] by formulating trajectory estimation as a labelling problem, modelled by Bayesian networks. Bayesian networks are able to incorporate the appearance information extracted from the video stream and to propagate the uncertainty associated with the labeling decisions. This allows to deal with am-

*this work was partially supported by FCT and POCTI in the scope of project LTT 37844.

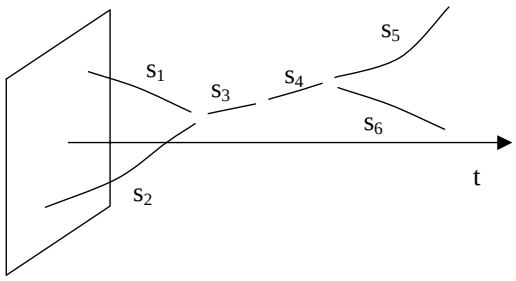

Figure 2: Detected strokes

biguous situations in which multiple interpretations have to be considered.

This paper extends the ideas proposed in [1] by addressing the problem of tracking groups of objects with Bayesian networks, i.e., each active region detected in the image may contain several objects. The main question which has to be considered is: *which objects are associated with each segment detected in the video stream?* The system must be able to provide the most probable labeling sequences as well as their confidence degrees.

A second contribution concerns the way splitting trajectories are modeled. Instead of using competitive links, this paper proposes the use of restriction nodes. This allows a simplification of the probabilistic models which become symmetric and reduces the amount of memory and computation effort associated with the inference task.

2 System Overview

Given a video sequence, low level processing provides a set of segments each of them associated with the evolution of an active region in the image [10]. We wish to link segments in order to estimate the trajectories of each object present in the scene. If a given segment corresponds to a group it will belong to several trajectories.

Every time new segments are detected (e.g., due to occlusions, new objects or group splitting), we need to know which objects belong to the new segments (Fig. 1). This ambiguity is solved by building a model for each object previously observed by the system. The likelihood of the new segment is computed considering all the admissible models. The Bayesian network is used to model object interaction as well their visual appearance.

To obtain the object trajectories a label x_k must be assign to each segment s_k. Each label identifies all the objects in the segment i.e., if the segment corresponds to a single object, the label is the object identifier. If the segment corresponds to a group, the label is a set of identifiers of all the objects inside the group.

Let $s_1, ..., s_N$ be the sequence of detected segments.

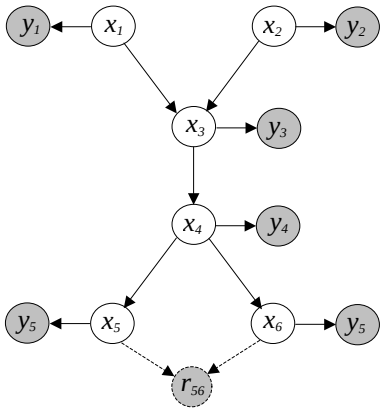

Figure 3: Bayesian network

A set of probabilistic labels $x = (x_1, ..., x_N)$ is estimated from the object trajectories and visual features $y = (y_1, ..., y_N)$, detected in the video stream e.g. using the dominant colors to characterize each detected segment. The most probable labels are obtained using the MAP method

$$\hat{x} = \arg\max_x p(y, x) \tag{1}$$

Bayesian networks are used in this paper to represent the joint probability distribution $p(x, y)$. It is assumed that each variable x_i is a hidden node of a Bayesian network and each observation y_i is a visible node connected to x_i. Object interaction (trajectory geometry) is encoded in the network topology. Two hidden nodes x_i, x_j are connected if the j-th segment starts after the end of the i-th segment. Additional restrictions are used to reduce the number of connections as discussed in Section 4.

A third type of nodes (restriction nodes r_{ij}) are included every time a hidden node has more than one hidden child. These nodes are used to model the competitive interaction among the hidden children (the same label can not be assigned to more than one child). Fig. 3 shows the Bayesian network associated with the example of Fig. 2 showing the three types of nodes.

Three issues have to be considered in order to specify a Bayesian network for a tracking problem:

- computation of the network architecture: nodes and links;
- choice of the admissible labels L_i associated to each hidden node;
- the conditional distribution of each variable given its parents.

The last two items depend on the type of application. Different solutions must be adopted if one wants to track isolated objects or groups of objects. Group tracking leads

to more complex networks since each segment represents multiple objects.

These topics are addressed in the next sections. Section 3 describes low level processing. Section 4 describes the network architecture. Section 5 describes the Bayesian network for tracking multiple isolated objects. Section 6 describes the Bayesian networks for group tracking. Section 7 presents experimental results and section 8 presents the conclusions.

3 Low level processing

The algorithm described in this paper was used for long term tracking of groups of pedestrians in the presence of occlusions. The video sequence is first pre-processed to detect the active regions in every new frame. A background subtraction method is used to perform this task followed by morphological operations to remove small regions [10].

Then region linking is performed to associate corresponding regions in consecutive frames. A simple method is used in this step: two regions are associated if each of them selects the other as the best candidate for matching [12]. The output of this step is a set of strokes in the spatial-temporal domain describing the evolution of the regions centroids during the observation interval.

Every time there is a conflict between two neighboring regions in the image domain the low level matcher is not able to perform a reliable association of the regions and the corresponding strokes end. A similar effect is observed when a region is occluded by the background. Both cases lead to discontinuities and the creation of new strokes.

The role of the Bayesian network is to perform a consistent labeling of the strokes detected in the image i.e., to associate strokes using high level information when the simple heuristic methods fail. Every time a stroke begins a new node is created and the inference procedure is applied to determine the most probable label configuration as well as the associated uncertainty.

4 Network Architecture

The network architecture is specified by a graph, i.e., a set of nodes and corresponding links. Three types of nodes are used in this paper: the hidden nodes x_i representing the label of the i-th segment, the observation nodes y_i which represent the features extracted from the i-th segment and binary restriction nodes r_{ij} which are used to avoid labeling conflicts. The restriction node r_{ij} is created only if x_i and x_j share a common parent.

A link is created from a hidden node x_i to x_j if x_j can inherit the label of x_i. Physical constrains are used to determine if two nodes are linked (e.g., the second segment must start after the end of the first and the average speed during during the occlusion gap is smaller than the maximum velocity specified by the user).

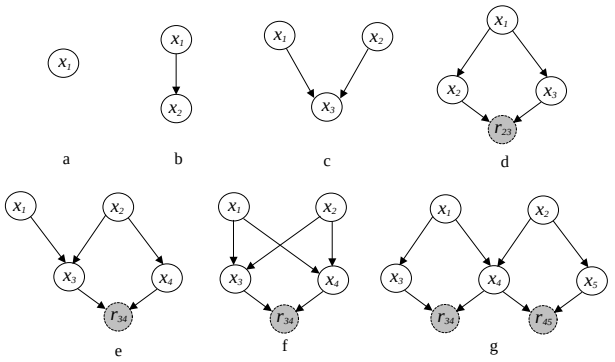

Figure 4: Basic structures (grey circles represent restriction nodes).

Furthermore, we assume that the number of parents as well as the number of hidden children of each node is limited to 2. Therefore, seven basic structures must be considered (see Fig. 4). These structures show the restriction nodes r_{ij} but the visible nodes y_i are omitted for the sake of simplicity.

When the number of parents or children is higher than two, the network is pruned using link elimination techniques. Simple criteria are used to perform this task. We prefer the connections which correspond to small spatial gaps.

5 Tracking Isolated Objects

A stroke s_i is either the continuation of a previous stroke or it is a new object. The set of admissible labels L_i is then the union of the admissible labels L_j of all previous strokes which can be assigned to s_i plus a new label corresponding to the appearance of a new object in the field of view. Therefore,

$$L_i = \left[\bigcup_{j \in I_i} L_j \right] \cup \{l_{new}\} \qquad (2)$$

where I_i denotes the set of indices of parents of x_i. See Table 1 which shows the labels associated to the hidden nodes of the Bayesian network of Fig. 3.

The Bayesian network becomes defined once we know the graph (see Section 4) and the conditional distributions $p(x_i|p_i)$ for all the nodes, where p_i are the parents of x_i.

Seven cases have to be considered (see Fig.4). The distribution $p(x_i|p_i)$ for each of these cases are defined following a few rules. It is assumed that the probability of assigning a new label to x_i is a constant P_{new} defined by the user.

k	L_k
1	1
2	2
3	1 2 3
4	1 2 3 4
5	1 2 3 4 5
6	1 2 3 4 6

Table 1: Admissible labels (isolated objects)

k	L_k
1	1
2	2
3	1 2 (1,2) 3
4	1 2 (1,2) 3 4
5	1 2 (1,2) 3 4 5
6	1 2 (1,2) 3 4 6

Table 2: Admissible labels (groups of objects)

Therefore,

$$p(x_i = l_{new} | x_j = k) = P_{new} \tag{3}$$

All the other cases are treated on the basis of a uniform probability assignment. For example in the case of Fig. 4c, x_i inherits the label of each parent with equal probability

$$p(x_i | x_p, x_q) = (1 - P_{new})/2, \tag{4}$$

for $x_i = x_p$ or $x_i = x_q$.

Every time two nodes x_i, x_j have a common parent, a binary node r_{ij} is included to avoid conflicts i.e., to avoid assigning common labels to both nodes. The conditional probability table of the restriction node is defined by

$$\begin{array}{l} p(r_{ij} = 1 / x_i \cap x_j = \emptyset) = 1 \\ p(r_{ij} = 1 / x_i \cap x_j \neq \emptyset) = 0. \end{array} \tag{5}$$

It is assumed that $r_{ij} = 0$ if there is a labeling conflict i.e., if the children nodes x_i, x_j have a common label; $r_{ij} = 1$ otherwise. To avoid conflicts we assume that r_{ij} is observed and equal to 1.

Inference methods are used to compute the most probable configuration (label assignment) as well as the probability of the admissible labels associated with each node. This task is performed using the Bayes Net Matlab toolbox [7].

Each stroke detected in the image is characterized by a vector of measurements y_j. In this paper y_j is a set of dominant colors. The dominant colors are computed applying the LBG algorithm to the pixels of the active region being tracked in each segment. A probabilistic model of the active colors is used to provide soft evidence about each node [3].

Each label is also characterized by a set of dominant colors. This information is computed as follows. The first time a new label is created and associated to a segment, a set of dominant colors is assigned to the label.

The probability of label $x_j \in L_j$ given the observation y_j is defined by

$$P(x_j / y_j) = \left(\begin{array}{c} N \\ n \end{array} \right) P^n (1 - P)^{N-n} \tag{6}$$

where n is the number of matched colors, N is the total number of colors (N=5 in this paper) and P is the matching probability for one color.

6 Group Model

This section addresses group modeling. Three cases have to be considered: group occlusions, merging and splitting.

Fig. 2 shows a simple example in which two persons meet, walk together for a while and separate. This example shows three basic mechanisms: group merging, occlusion and group splitting. These mechanisms allow us to model more complex situations in which a large number of objects interact forming groups. After detecting the segments using image processing operations each segment is characterized by a group label x_i. A group label is a sequence of labels of the objects present in the group. A Bayesian network is then built using the seven basic structures of Fig. 4.

Let us now consider the computation of the admissible labels. The set of admissible labels L_k of the k-th node is recursively computed from the sets of admissible labels of its parents L_i, L_j, starting from the root nodes. This operation depends on the type of connections as follows:

occlusion:

$$L_k = L_i \cup l_{new} \tag{7}$$

merging:

$$L_k = L_i \cup L_j \cup L_{merge} \cup l_{new} \tag{8}$$

$$L_{merge} = \{a \cup b : a \subset L_i, b \subset L_j, a \cap b = \emptyset\} \tag{9}$$

splitting:

$$L_k = L_j = \mathcal{P}(L_i) \cup l_{new} \tag{10}$$

where $\mathcal{P}(L_i)$ is the partition of the set L_i, excluding the empty set. In all these examples, l_{new} stands for a new label, corresponding to a new track.

Table 2 shows the set of admissible labels for the example of Fig. 3. Labels 1,2 correspond to the objects detected in the first frame and labels 3-6 correspond to new objects which may have appeared.

Conditional probability distributions must be defined for all the network nodes, assuming that the parents labels are known. Simple expressions for these distributions are used based on four parameters chosen by the user:

- P_{occl} - occlusion probability

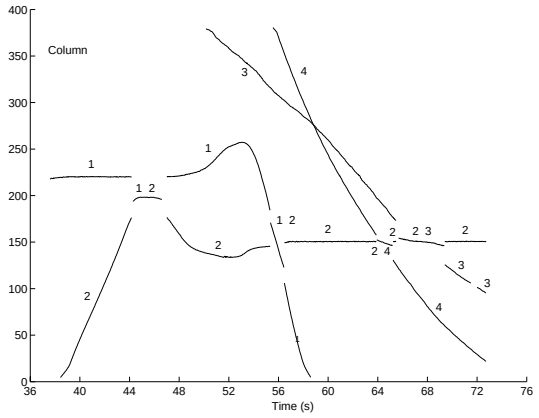

Figure 5: Stroke detection and the most probable labels

- P_{merge} - merging probability
- P_{split} - splitting probability
- P_{new} - probability of a new track

These parameters are free except in the case of the occlusion (Fig. 4b). In this case, the conditional probability of x_k given x_i in given by

$$P(x_k/x_i) = \begin{cases} 1 - P_{new} & x_k = x_i \\ P_{new} & x_k = l_{new} \end{cases} \quad (11)$$

The spliting and merging conditional distributions are defined in appendix.

7 Results

The proposed algorithm was used for long term tracking of multiple objects in video sequences. The tests were performed using PETS 2001 database as well as video sequences of an university campus. In both cases the sequences were digitalized at 25 fps and contain less than 20 active regions (pedestrians and vehicles) to be tracked.

These tests allowed to evaluate the performance of the proposed algorithms in the presence of occlusions and small groups of 2 to 4 objects. Similar results were obtained in both cases. The Bayesian network managed to correctly solve most ambiguous situations.

Fig. 5 shows the segments detected in one of the PETS sequences. Each object is easily tracked if it appears isolated in the image but the trajectories are broken everytime there is an occlusion or a group change (merging or splitting). Fig. 6 illustrates some of these difficulties. Figs. 6a,b correspond to a merge. Figs. 6b,c is a group split. Fig. 6d-e is a merge and split. These events can also be observed in Fig. 5.

The Bayesian network is recursively updated from the output of the low level module. Everytime a new segment is

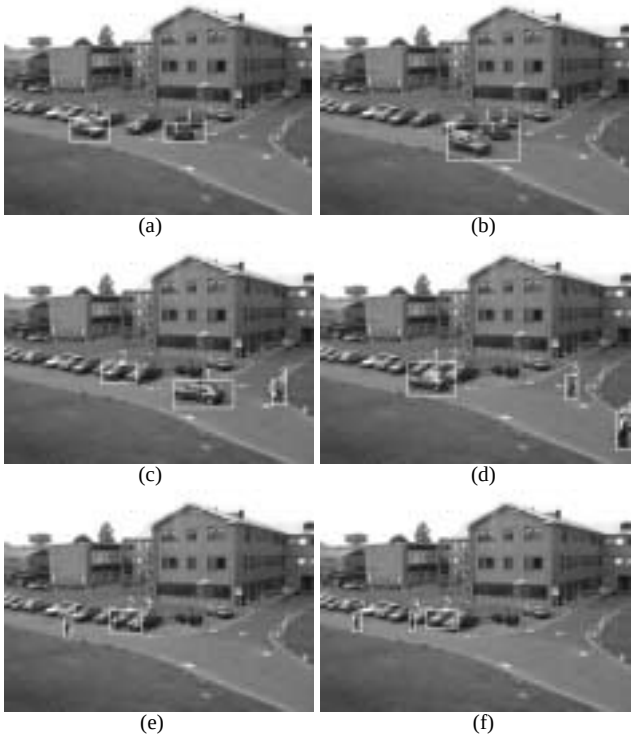

Figure 6: Tracking results at time 42, 45, 54, 56, 68, 70s

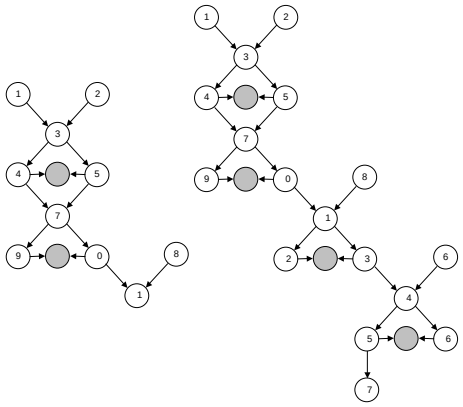

Figure 7: Bayesian network at two time instants

created the network is updated with a new hidden node. Fig. 7 shows the evolution of the Bayesian network at two time instants. The most probable labels obtained by the MAP method are displayed in Fig. 5. A consistent interpretation of the data was achieved by the tracker.

Fig. 8 shows tracking results obtained at the university campus. The figure displays the evolution of the active regions (column of the mass center) and the labels obtained using the Bayesian network. The Bayesian network associated with this example was automatically built from the video stream as before and it is shown in Fig. 9. Fig. 10

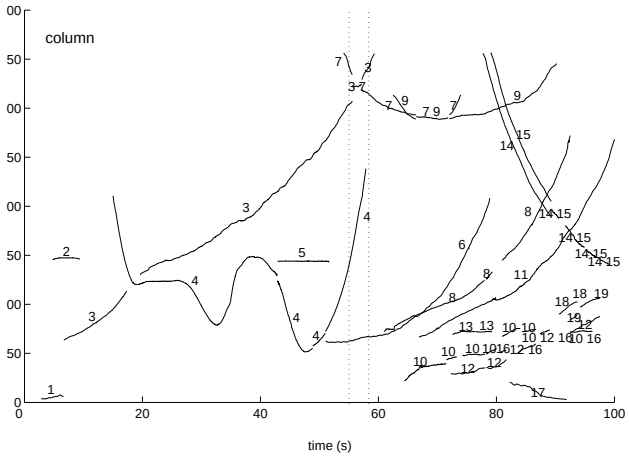

Figure 8: Stroke detection and the most probable labels

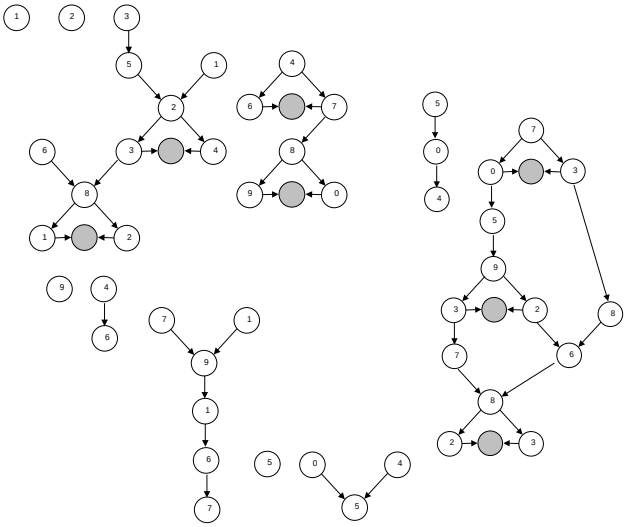

Figure 9: Bayesian network

shows four frames which illustrate the performance of the system in the tracking of multiple objects with occlusions and groups.

8 Conclusion

This paper presents a system for long term tracking of multiple objects in the presence of occlusions and group merging and splitting. The system tries to follow all moving objects present in the scene by performing a low level detection of spatio-temporal segments followed by a labeling procedure which attempts to assign consistent labels to all the segments associated to the same object. The interaction among the objects is modeled using a Bayesian network which is automatically built during the surveillance

Figure 10: Tracking results in the time interval shown in Fig. 8)

task. This allows to formulate the labeling problem as an inference task which integrates all the available information extracted from the video stream allowing to update the interpretation of the detected tracks every time new information is available. This is a useful feature to solve ambiguous situations such as group splitting and occlusions in which long term memory is needed.

Appendix

Fig. 4 depicts the set of net configurations which has to be considered to define all the probability conditional distributions in the Bayesian network. The first structure (Fig. 4a) corresponds to the case of an isolated track. This corresponds to a new object in the scene. A new label is assigned to this object.

The second case (occlusion), shown in Fig. 4b, was already considered in equation (11).

The remaining cases, corresponds to situations involving merging and splitting, being the conditional distributions defined as follows.

1. Two merging parents (Fig. 4c):

$$P(x_k/x_i, x_j) = \begin{cases} P_{merge} & x_k = x_i \cup x_j \\ P_{occl_i} & x_k = x_i \\ P_{occl_j} & x_k = x_j \\ P_{new} & x_k = l_{new} \end{cases}$$

2. A single parent with a split (Fig. 4d):

$$P(x_k/x_i) = \begin{cases} P_{split}/(2^{N_i} - 2) & x_k \subset \mathcal{P}(x_i) \backslash x_i \\ P_{occl_i} & x_k = x_i \\ P_{new} & x_k = l_{new} \end{cases}$$

where N_i is the cardinality of x_i (number of objects in segment s_i). The expression for $N = 1$ has some minor modifications.

3. Two parents, but only one of them may have a split (Fig. 4e):

$$P(x_k/x_i, x_j) = \begin{cases} P_{split}/(2^{N_j} - 2) & x_k \subset \mathcal{P}(x_j) \backslash x_j \\ P_{merge}/(2^{N_j} - 1) & x_k \subset \mathcal{M}_{ij} \\ P_{occl_i} & x_k = x_i \\ P_{occl_j} & x_k = x_j \\ P_{new} & x_k = l_{new} \end{cases}$$

where $\mathcal{M}_{ij} = \{a \cup b : a = x_i, b \subset \mathcal{P}(x_j), a \cap b = \emptyset\}$.

4. Two parents, both with splits (Figs. 4f-g):

$$P(x_k/x_i, x_j) = \begin{cases} P_{split}/(2^{N_i} - 2) & x_k \subset \mathcal{P}(x_i) \backslash x_i \\ P_{split}/(2^{N_j} - 2) & x_k \subset \mathcal{P}(x_j) \backslash x_j \\ \frac{P_{merge}}{(2^{N_i}-1)(2^{N_j}-1)} & x_k \subset \mathcal{M}_{ij} \\ P_{occl_i} & x_k = x_i \\ P_{occl_j} & x_k = x_j \\ P_{new} & x_k = l_{new} \end{cases}$$

$\mathcal{M}_{ij} = \{a \cup b : a \subset \mathcal{P}(x_i), b \subset \mathcal{P}(x_j), a \cap b = \emptyset\}$.

References

[1] A. Abrantes, J. Marques, J. Lemos, Long Term Tracking Using Bayesian Networks, *IEEE Int. Conf. Image Processing*, 609-612, 2002.

[2] F. Bremond, M. Thonnat, Tracking Multiple Nonrigid Objects in Video Sequences, *IEEE Trans. Circuits, Systems and Video Technology*, 8, 585-591, 1998.

[3] F. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.

[4] I. Haritaoglu, D. Harwood, L. Davis, W4: Real-Time Surveillance of People and Their Activities, *IEEE Trans. PAMI*, 22, 809-830, 2000.

[5] M. Jordan, *Learning in Graphical Models*, MIT Press, 1998.

[6] S. McKenna S. Jabri, Z. Duric, A. Rosenfeld, H. Wechsler, Tracking Groups of People, *Journal of Comp. Vision Image Understanding*, 80, 42-56, 2000.

[7] K. Murphy, The Bayes Net Toolbox for Matlab, *Computing Science and Statistics*, 33, 2001.

[8] N. Oliver, B. Rosario, A. Pentland, A Bayesian Computer Vision System for Modeling Human Interactions, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22, 8, 831-843,2000.

[9] C. Regazzoni, P. Varshney, Multi-Sensor Surveillance Systems, *IEEE Int. Conf. Image Processing*, 497-500, 2002.

[10] C. Stauffer, W. Grimson, Learning Patterns of Activity Using Real-Time Tracking, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22, 8, 747-757, 2000.

[11] C. Wren, A. Azabayejani, T. Darrel, A. Pentland, Pfinder: Real Time Tracking of the Human Body, *IEEE Transactions on Patter Analysis and Machine Intelligence*, 19, 780-785, 1997.

[12] S. Ullman, R. Basri, Recognotion by Linear Combination of Models, *IEEE Transactions on Patter Analysis and Machine Intelligence*, 13, 992-1006, 1991.