3

Appearance-Based Retrieval for Tracked Objects in Surveillance Videos

Thi-Lan Le[1], Monique Thonnat[2] and Alain Boucher[3]
[1]*MICA Center, HUST - CNRS/UMI 2954 - Grenoble INP, Hanoi,*
[2]*PULSAR, INRIA Sophia Antipolis,*
[3]*IFI, MSI Team; IRD, UMI 209 UMMISCO; Vietnam National University,*
[1,3]*Vietnam*
[2]*France*

1. Introduction

Video surveillance is a rapidly growing industry. Driven by low-hardware costs, heightened security fears and increased capabilities, video surveillance equipment is being deployed more widely and with greater storage than ever. This provides a huge amount of video data. Associating to these video data, retrieval facilities become very useful for many purposes and many kinds of staff. Recently, several approaches have been dedicated to retrieval facilities for surveillance data (Le, Thonnat et al. 2009) (Zhang, Chen et al. 2009). Figure 1 shows how indexing and retrieval facility can be integrated in a surveillance system. Videos coming from cameras will be interpreted by the video analysis module. There are two modes for using the analysed results: (1) the corresponding alarms are sent to members of the security staff to inform them about the situation; (2) the analysed results are stored in order to be used in the future. In this chapter, we focus on analysing current achievements in surveillance video indexing and retrieval. Video analysis (Senior 2009) is beyond the scope of this chapter.

Video analysis module provides two main result types of result: objects and events. Thus, surveillance video indexing and retrieval approaches can divided into two categories: surveillance video indexing and retrieval at the object level (Calderara, Cucchiara et al. 2006; Ma and Cohen 2007; Le, Thonnat et al. 2009) and at the event level (Zhang, Chen et al. 2009; Velipasalar, Brown et al. 2010). As events of interest may vary significantly among different applications and users, this chapter focuses on presenting the work done for surveillance video indexing and retrieval at the object level.

The remaining of the chapter is organized as follows: In Section 2, we give a brief overview of surveillance object retrieval. Section 3 aims at analysing in detail appearance-based surveillance object retrieval. We first give some definitions and point out the existing challenges. Then, we describe the solutions proposed for two important tasks: object signature building and object matching in order to overcome these challenges. Section 4 presents current achievements and discusses about open problems in this domain.

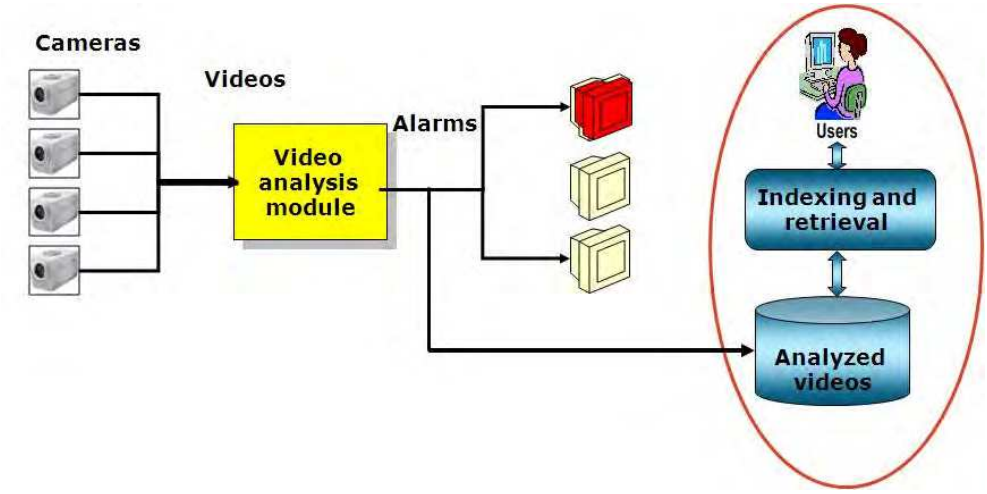


Fig. 1. Indexing and retrieval facility in a surveillance system. Videos coming from cameras will be interpreted by the video analysis module. There are two modes for using the analysed results: (1) the corresponding alarms are sent to security staffs to inform them about the situation; (2) the analysed results are stored in order to be used in the future.

2. Object retrieval for surveillance videos

This section aims to give an overview of existing approaches for object retrieval in surveillance videos.

2.1 Architecture

In the same way as video analysis systems which have two main architectures, i.e. centralized and decentralized architecture (Senior 2009), object video retrieval for surveillance systems has also two main modes: late fusion and early fusion modes. In the late fusion mode (cf. Fig. 2), the object detection and tracking are performed on the video stream of each camera. Then, the object matching compares the query and the detected objects for each camera. The matching results are fused to form the retrieval results. In the early fusion mode (cf. Fig. 3), the data fusion is done in the object detection and tracking module. We can see that the object retrieval method in this early fusion mode has more opportunities to obtain a good result because if an object is not totally observed by a camera, it may be well captured by other cameras. Most of the state of the art work belongs to the early fusion mode. However, the fusion strategy is not explicitly discussed except in the work of Calderara et al. (Calderara, Cucchiara et al. 2006).

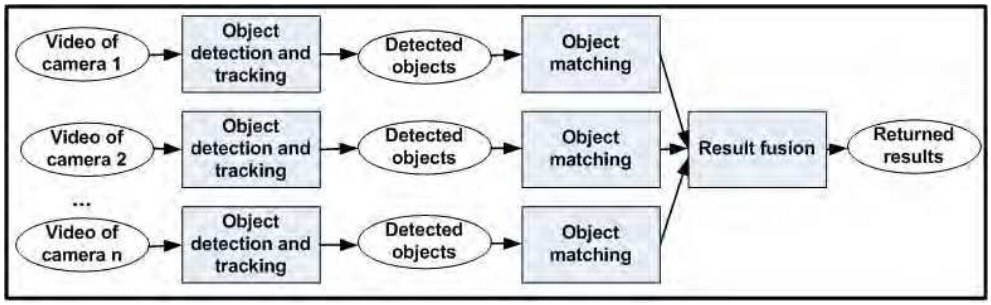


Fig. 2. Late fusion object retrieval approach: the object detection and tracking is performed on video stream of each camera. Then, the object matching compares the query and the detected objects of each camera. The matching result is then fused to form the retrieval results.

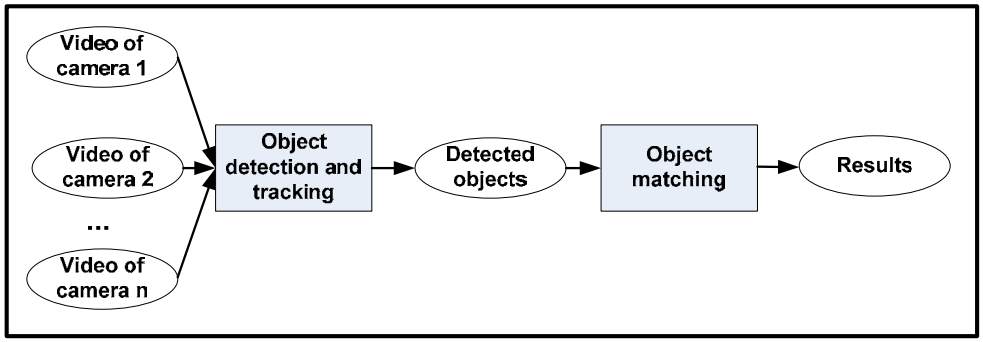


Fig. 3. Early fusion object retrieval approach.

2.2 Object feature extraction and representation

Since objects in video surveillance are physical objects (e.g. people, vehicles) that are present in the scene at a certain time, in general, they are detected and tracked in a large number of frames. Objects in videos possess two main characteristics named spatial and temporal characteristics. Spatial characteristics of an object may be its positions in frames (in 2D coordinates) and positions in scene (in 3D coordinates), its spatial relationships with other objects and its appearance. Temporal characteristics of an object contain its movement and its temporal relationships with other objects. Therefore, an object may be represented by one sole or several characteristics. However, among these characteristics, object movement and object appearance are the two most important characteristics and are widely used in the literature.

Concerning the object representation based on object movement, in the literature, a number of different approaches have been proposed for object movement representation and matching (Broilo, Piotto et al. 2010). Certain approaches directly use detected object positions across frames that are represented in trajectory form (Zheng, Feng et al. 2005). As object trajectory may be very complex, other authors try to segment an object trajectory into several sub-trajectories (Buchin, Driemel et al. 2010) with the purpose that each sub-

trajectory represents a relatively stable pattern of object movement. Other work attempts to move to higher levels of object trajectory representation, named symbolic level and semantic level. At symbolic level, (Chen, Ozsu et al. 2004; Hsieh, Yu et al. 2006; Le, Boucher et al. 2007) aim to convert object trajectory into a character sequence. The advantage is that they promote the applying of successful and famous methods in text retrieval such as the Edit Distance for object trajectory matching. The approaches dedicated to object trajectory representation at the semantic level try to learn the semantic meaning such as turn left, low speed from object movement (Hu, Xie et al. 2007). As results, the output is close to the human manner of thinking. However, they strongly depend on applications.

Object representation based on its appearance has attracted a lot of research interest. Appearance-based object retrieval methods for surveillance video are distinguished each other by two criteria. The first criterion is the appearance feature extracted on the image/frame where the object is detected and the second one is the way to create object signature from all features extracted over the object's life time and to match objects based on their signatures. In the next section, we describe in detail the object signature building and object matching methods. In this section, we only present the object appearance feature.

There is a great variety of object features used for surveillance object representation. In fact, all features that are proposed for image retrieval can be applied for surveillance object representation. Appearance object features can be divided into two categories: global and local. Global features are color histogram, dominant color, covariance matrix, just to name a few. Besides global features, local features such as interest points and SIFT descriptor can be extracted from the object's region.

In (Yuk, Wong et al. 2007), the authors have proposed to use MPEG-7 descriptors such as dominant colors, edge histograms for surveillance retrieval. In the context of one research project conducted by IBM research center[1], the researchers have evaluated a large number of color features for surveillance application that are standard color histograms, weighted color histograms, variable bin size color histograms and color correlograms. Results show color correlogram to have the best performance. Ma et Cohen (Ma and Cohen 2007) suggest to use the covariance matrix as object feature. According to the authors, the covariance matrix is appealing because it fuses different types of features and has small dimensionality. The small dimensionality of the model is well suited for its use in surveillance videos because it takes very little storage space. In our research (Le, Boucher et al. 2010), we have evaluated the performance of 4 descriptors which are dominant color, edge histogram, covariance matrix (CM) and SIFT descriptor for surveillance object representation and matching. The obtained results show that if the objects are detected while the background and context objects are not present in the object region, the used descriptors allow retrieving objects with relatively good results. For other cases, the covariance matrix is more effective than the other descriptors. According to our experiments, it is interesting to see that when the covariance matrix represents information of all pixels in a blob, the points of interest use only few pixels. The dominant color and the edge histogram use the approximate information of pixel color and edge. A pair of descriptors (covariance matrix and dominant color) or (covariance matrix and edge histogram) or (covariance matrix and SIFT descriptors) may be chosen as default descriptors for object representation.

[1] https://researcher.ibm.com/researcher/view_project.php?id=1393

3. Appearance-based object retrieval in surveillance videos

In this section, we firstly give some definitions and point out the existing challenges for appearance-based object retrieval in surveillance videos. Then, we describe the solutions proposed for two important tasks: object signature building and object matching in order to overcome these challenges.

3.1 Definitions

Definition 1: An **object blob** is a region determined by a minimal bounding box in a frame where the object is detected.

The minimal bounding box is calculated by the object detection module in video analysis and an object has one sole minimal bounding box. Fig. 4 gives some examples of detected objects and their corresponding blobs.

Fig. 4. Detected objects and their blobs (Bak, Corvee et al. 2010).

Definition 2: Object representation

In surveillance applications, one object is in general detected and tracked in a number of frames. In other words, a set of object blobs is defined for an object. Therefore, an object can be represented as:

$$O = \{B_i\}, i \in 1, N \tag{1}$$

where O is object, B_i is the i^{th} object blob, N is the total number of blobs of object O.

It is worth noting that object blobs can be non-consecutive since an object may not be detected in certain frames and the value of N varies depending on the object life time in the scene. Fig. 5 gives an example of an object that is represented by its blobs. As we can notice, with poor object detection, several object blobs do not cover well the object appearance.

Fig. 5. An object is represented by its blobs.

3.2 Challenges in appearance-based object retrieval for surveillance videos

This section aims at pointing out existing challenges in appearance-based object retrieval for surveillance videos. As object indexing and retrieval take the output of video analysis as its input (cf. Fig. 1), the quality of the video analysis has a huge influence on object indexing and retrieval. Current achievements on surveillance video analysis show that video analysis is far from perfect since it is hampered by issues in low resolution, pose and lighting variations and object occlusion. In this section, we point out the challenges in appearance-based object retrieval by analyzing the effect of two modules of video analysis on the object indexing and retrieval quality: the object detection and the object tracking modules.

The object detection module is the module that allows to determine the object blobs. An object detection module is good if all blobs of a detected object (1) cover totally this object and (2) do not contain other objects. However, these constraints are not always met. Object retrieval has to address three difficult cases as shown in Fig. 6. In the first case, the object is not present at all in the blob (Fig. 6a). With the second case, the object is partially present in the blob (Fig. 6b) while with the third case, the blob of the detected object covers totally this object, however, it contains also other objects (Fig. 6c and Fig. 6d).

Concerning the object tracking quality, two metrics that are widely used for evaluating the performance of object tracking in the video surveillance community are *object ID persistence* and *object ID confusion* (Nghiem, Bremond et al. 2007). The *object ID persistence* metric helps to evaluate the ID persistence. It computes over the time how many tracked objects (output of the object tracking module) are associated to one ground-truth object. On the contrary, the *object ID confusion* metric computes the number of objects per detected object (having the same ID). A good object tracking algorithm obtains a small value for these two metrics (minimum is 1).

Fig. 6. Examples of object detection quality (a) The object is not present in the blob; (b) The object is partially present in the blob; (c) and (d) The object is totally present in the blob.

However, the obtained results in several video surveillance benchmarks show that current achievement on object tracking is still limited (object ID persistence and object ID confusion metrics are generally much greater than 1). Fig. 7 shows an example of the object ID persistence problem: two tracked objects created for one sole ground-truth object, therefore object ID persistence is equal to 2. Fig. 8 illustrates an example of object ID confusion: three ground-truth objects IDs associated to one sole detected object (object ID confusion = 3).

Fig. 7. An example of the object ID persistence problem: two tracked objects created for one sole ground-truth object (object ID persistence = 2).

Based on the above-mentioned analysis, the main challenge in surveillance object indexing and retrieval is the poor quality of object detection and tracking. An object indexing and retrieval algorithm is robust if it can work with different quality of the object detection and tracking.

With the object representation as defined in Eq. 1, we believe that object indexing and retrieval methods can address the poor quality of object detection and tracking problem if they have an effective object signature building and a robust object matching.

Fig. 8. An example of object ID confusion: three ground-truth object IDs associated to one sole detected object (object ID confusion = 3).

3.3 Object signature building

Object signature building is a process that aims at calculating one or a set of descriptors, named object signature, from a set of object blobs.

The calculated signature should (1) be able to represent all object appearance aspects, (2) be distinctive and (3) be as compact as possible. Among these characteristics, the two first characteristics ensure the robustness of the retrieval part. The third characteristic relates to the effectiveness of the indexing part. If the signature is compact, it does not require much storage.

Object signature building methods for surveillance video are divided into two approaches. The first object signature building approach is based on the following observation: Surveillance objects are generally detected and tracked in a large number of frames. Consequently, an object is represented by a set of blobs. Due to errors in object detection, using all these blobs for object indexing and retrieval is irrelevant. Moreover, it is redundant because of the similar content between blobs (two consecutive blobs of an object are closely similar). Based on this observation, methods belonging to the first approach try to select the most relevant and representative blobs from a set of blobs and then to compute object features on these blobs. This process is defined by Eq. 2. This approach is composed of two steps. The first step, called representative blob detection, chooses from the object blobs the most relevant and representative ones that represent significantly the object appearance while the second step computes the object features mentioned in Section 2.2 from the calculated representative blobs.

$$\left\langle \{B_i\}, i \in 1,N \right\rangle \xrightarrow{(1)} \left\langle \{Br_j\}, j \in 1,M \right\rangle \xrightarrow{(2)} \left\langle \{F_j\}, j \in 1,M \right\rangle \quad \text{with } N \gg M \tag{2}$$

where:

- $\left\langle \{B_i\}, i \in 1,N \right\rangle$: set of original blobs for the object O determined by using object detection output.

- $\left\langle \{Br_j\}, j \in 1,M \right\rangle$: set of representative blobs detected for the object O.
- $\left\langle \{F_j\}, j \in 1,M \right\rangle$: set of features extracted on the representative blobs. The extracted feature can be color histogram, dominant color, etc.

Instead of calculating only the representative blobs, several authors compute a set of pairs: the representative blob and its associating weight while the weight associated with a representative blob shows the importance of this blob. With this, the first approach is defined as follows:

$$\left\langle \{B_i\}, i \in 1,N \right\rangle \xrightarrow{(1)} \left\langle \{Br_j, w_j\}, j \in 1,M \right\rangle \xrightarrow{(2)} \left\langle \{F_j, w_j\}, j \in 1,M \right\rangle$$
$$\text{with } N \gg M \text{ and } \sum_{j=1}^{M} w_j = 1 \tag{3}$$

Fig. 9 shows an example of the first object signature building approach. From a large number of blobs (905 blobs), the object signature building method selects only 4 representative blobs. Their associated weights are 0.142, 0.005, 0.016 and 0.835.

Fig. 9. An example of representative blob detection: 4 representative blobs are extracted from 905 blobs.

The methods presented in (Ma and Cohen 2007) and in (Le, Thonnat et al. 2009) are the most significant ones of the first object signature building approach. These methods are distinguished each from the other by the way to define the representative blobs.

The representative blob detection method proposed by Ma et Cohen (Ma and Cohen 2007) is based on the agglomerative hierarchical clustering and the covariance matrix extracted from the object blobs. This method is composed of the three following steps:

Step 1. Do agglomerative clustering on the original set of object blobs based on the covariance matrix.
Step 2. Remove clusters having a small number of elements.
Step 3. Select representative blobs.

The first step aims at forming clusters of similar blobs. The similarity of two blobs is defined by using the covariance matrix. The covariance matrix is built over a feature vector f, for

each pixel, that is: f(x,y)=[x, y, R(x, y), G(x, y), B(x, y), ∇R^T (x, y), ∇G^T (x, y),∇B^T (x, y)] where R, G, B are the colorspace axes and x, y are the coordinates of the pixel contributing to the color and the gradient information. The covariance matrix is computed for each detected blob as follows:

$$C = \sum_{x,y} (f - \bar{f})(f - \bar{f})^T \tag{4}$$

The covariance matrices for blobs of different sizes have the same size. In fact, the covariance matrix is a N×N matrix while N is the dimension of the feature vector f.

The distance between two blobs is calculated as:

$$d(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C_i, C_j)} \tag{5}$$

For the agglomerative clustering, the distance $d(A,B)$ between two clusters A and B is computed by average linkage as:

$$d(A,B) = \frac{1}{|A|.|B|} \sum_{A_i \in A} \sum_{B_j \in B} d(A_i, B_j) \tag{6}$$

where $d(A_i, B_j)$ is defined in Eq. 5.

The objective of the second step is to detect and remove outliers that are clusters containing a small number of elements. The final step determines one representative blob for each cluster. For a cluster B, the representative blob B_l is defined as:

$$l = \arg\min_{j=1,...,|B|, j \neq i} \sum_{i=1,...,|B|} d(B_i, B_j) \tag{7}$$

where $d(B_i, B_j)$ is the blob distance defined in Eq. 5.

Fig.10 gives an example result of Ma and Cohen method (Ma and Cohen 2007): (a) original sequence of blobs; (b) clustering results having valid cluster and invalid cluster; (c) representative frame for the second cluster in (b); (d) representative frame for the third cluster in (b). We can see that this method can dominate errors of the object detection if they occur in a small number of frames. However, if the detection error occurs in a large number of frames, the cluster containing the blobs of these frames will be defined as valid cluster by this method (the validity of clusters is decided by their sizes).

Our work presented in (Le, Thonnat et al. 2009) is an improvement of Ma and Cohen work (Ma and Cohen 2007), based on two remarks. The first remark is that the drawback of Ma and Cohen's method is that it cannot work well with imperfect object detection since it processes all object blobs including relevant and irrelevant ones. We can resolve this drawback by removing all irrelevant blobs before doing the agglomerative clustering. The second remark is that one blob of an object is relevant if it contains this object or objects belonging to the same class of this object. For example, one blob of a detected person is relevant if it represents somehow the person class. With these analyses, we add two

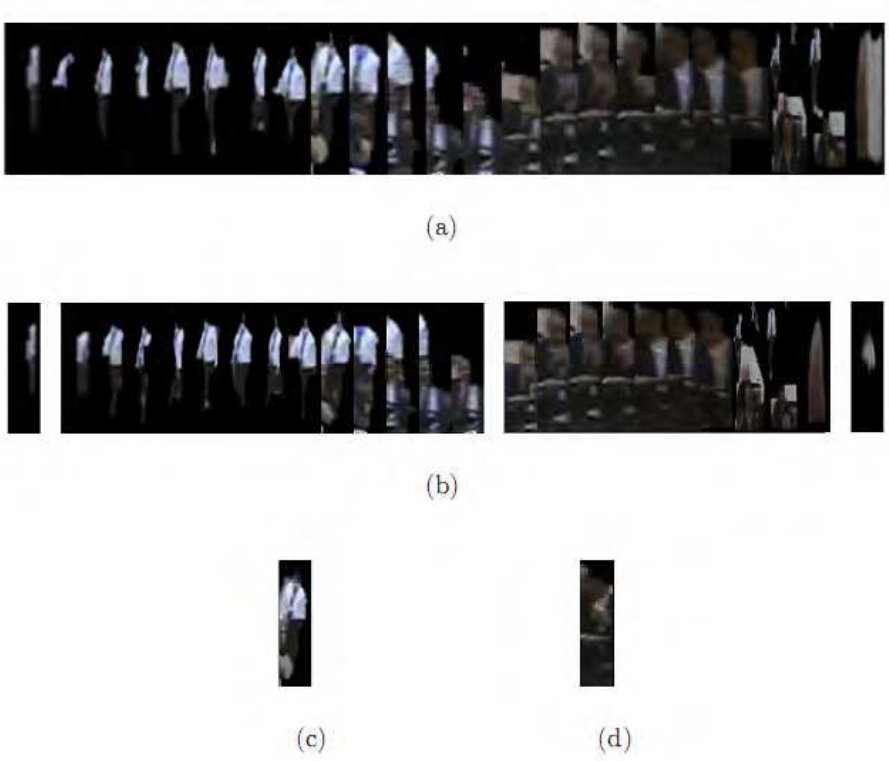


Fig. 10. Example result of Ma and Cohen method (Ma and Cohen 2007): (a) original sequence of blobs; (b) clustering results having valid clusters and invalid clusters; (c) representative frame for the second cluster in (b); (d) representative frame for the third cluster in (b).

preliminary steps in Ma and Cohen's work. These steps will be performed before the first step of Ma and Cohen's work.

Step 0. Classify blobs of all objects into relevant (with the object of interest) and irrelevant blobs (without object of interest) by a two-class SVM classifier with radial basis function (RBF) kernel using edge histograms (Won, Park et al. 2002).

Step 1. Remove irrelevant blobs from the set of blobs for each object.

It is worth noting that the appearance of tracked objects may vary but their blobs usually have some common visual characteristics (e.g. human shape characteristics for the blobs of different tracked persons). As we can see, the two added steps allow to remove irrelevant blobs before agglomerative clustering. Therefore, this object signature building method is robust while working with poor quality object detection.

The second object signature building approach does not perform explicitly the representative blob detection. It attempts to sum up all object appearances into one sole signature. This approach is defined as follows:

$$\left\langle \{B_i\}, i \in 1, N \right\rangle \rightarrow \left\langle \tilde{F} \right\rangle \tag{8}$$

The work presented in (Calderara, Cucchiara et al. 2006) belongs to the second object signature building approach. In this work, the authors have proposed three notations that are person's appearance (PA), single camera appearance trace (or SCAT in short) and multicamera appearance trace (or MCAT in short). SCAT of the person P on camera C_i is composed of all the past person's appearance (PA) of P at instant time t:

$$SCAT_i^P = \left\{ PA_i^P(t) \mid t = 1, \ldots N_i^P \right\} \tag{9}$$

where t represents the samples in time in which the person P was visible from the camera C_i and N_i^P is the total number of frames in which he was visible and detected.

MCAT for a person P is composed of all the $SCAT_i^P$ for any camera C_i in which, at the current moment, the person P has been detected at least for one frame. We can see that SCAT is equivalent to MCAT if the surveillance system has only a camera and SCAT is equivalent to $\left\langle \{B_i\}, i \in 1, N \right\rangle$ in our definition.

The object signature building based on mixture of Gaussians is performed as follows:

Step 1. Using the first PA in the MCAT, the ten principal modes of the color histogram are extracted;

Step 2. The Gaussians are initialized with a mean μ equal to the color corresponding to the mode and a fixed variance σ^2; weights are equally distributed for each Gaussian;

Step 3. successive PA belonging to the MCAT are processed to extract again the ten main modes that are used to update the mixture; then, for each mode:

- (a) its value is checked against the mean of each Gaussian and if for none of them the difference is within 2.5σ of the distribution, the mode generates a new Gaussian (using the same process reported above) replacing the existing Gaussian with the lowest weight;
- (b) the Mahalanobis distance is computed for every Gaussian satisfying the above-reported check, and the mode is assigned to the nearest Gaussian; the mean and the variance of the selected Gaussian are updated with the following adaptive equations:

$$\begin{aligned} \mu_t &= (1-\alpha)\mu_{t-1} + \alpha X_t \\ \sigma_t^2 &= (1-\alpha)\sigma_{t-1}^2 + \alpha (X_t - \mu_t)^T (X_t - \mu_t) \end{aligned} \tag{10}$$

where X_t is the vector with the values corresponding to the mode and α is the fixed learning factor; the weights are also updated by increasing that of the selected Gaussian and decreasing those of the other Gaussians consequently.

At the end of this process, ten Gaussians and the corresponding weights for each MCAT are available and are used as object signature.

3.4 Object matching

Object matching is the process that computes the similarity/dissimilarity between two objects based on their signatures calculated by above-mentioned approaches. In information

retrieval in general and in surveillance object retrieval in particular, with a given query, the system will (1) compute the similarity between this query and all elements in the database and (2) return the retrieved results which are a list of elements sorted by their similarity with the query. The number of returned results will be decided for each application.

Corresponding to the two approaches for object signature building, there are two approaches for the object matching. Object matching for the first object signature building approach is expressed in Eq. 11. In this equation, object O_q and O_p are represented by $\left\{(F_i^q, w_i^q) \mid i \in 1, M^q\right\}$ and $\left\{(F_j^p, w_j^p) \mid j \in 1, M^p\right\}$ respectively. The object matching methods allow to define a similarity/dissimilarity between two sets of blobs. These sets may have different sizes. It is worth noting that we can always compute the similarity/dissimilarity of a pair of blobs based on visual features such as color histogram, covariance matrix.

$$\begin{aligned} &\left\langle \left\{(Br_i^q, w_i^q) \mid i \in 1, M^q\right\}, \left\{(Br_j^p, w_j^p) \mid j \in 1, M^p\right\} \right\rangle \to Dis, Dis \in \Re \text{ or} \\ &\left\langle \left\{(F_i^q, w_i^q) \mid i \in 1, M^q\right\}, \left\{(F_j^p, w_j^p) \mid j \in 1, M^p\right\} \right\rangle \to Dis, Dis \in \Re \end{aligned} \tag{11}$$

In (Ma and Cohen 2007), the authors define a similarity measure between two objects O_q and O_p using the Hausdorff distance (Eq. 12). The Hausdorff distance is the maximum distance of a set to the nearest point in the other set.

$$\begin{aligned} Dis &= Hausdorff\left(\left\{(F_i^q, w_i^q) \mid i \in 1, M^q\right\}, \left\{(F_j^p, w_j^p) \mid j \in 1, M^p\right\}\right) \\ &= \max_{i \in M^q} \min_{j \in M^p} d(F_i^q, F_j^p) \end{aligned} \tag{12}$$

where $d(F_i^q, F_j^p)$ is the distance between two blobs by using the covariance matrix.

The above object matching allows to take into consideration multiple appearance aspects of the object being tracked. However, the Hausdorff distance is not relevant when working with object tracking algorithms having a high value of object ID confusion because this distance is extremely sensitive to outliers. If two sets of points A and B are similar, all the points are perfectly superimposed except only one single point in A which is far from any point in B, then the Hausdorff distance determined by this point.

In (Le, Thonnat et al. 2009), we propose a new object matching based on the EMD (Earth Mover's Distance) (Rubner, Tomasi et al. 1998). This method is widely applied with success in image and scripted video retrieval.

$$Dis = EMD\left(\left\{(F_i^q, w_i^q) \mid i \in 1, M^q\right\}, \left\{(F_j^p, w_j^p) \mid j \in 1, M^p\right\}\right) \tag{13}$$

Computing the EMD is based on a solution to the old transportation problem. This is a bipartite network flow problem which can be formalized as the following linear programming problem: Let I be a set of suppliers, J a set of consumers, and c_{ij} the cost to ship a unit of supply from $i \in I$ to $j \in J$. We want to find a set of flows f_{ij} that minimizes the overall cost:

$$\sum_{i\in I}\sum_{j\in J} f_{ijc_{ij}} \tag{14}$$

subject to the following constraints:

$$\begin{aligned} &f_{ij} \geq 0, i \in I, j \in J \\ &\sum_{i\in I} f_{ij} = y_j, j \in J \\ &\sum_{j\in J} f_{ij} \leq x_i, i \in J \\ &\sum_{j\in J} y_j \leq \sum_{i\in I} x_i \end{aligned} \tag{15}$$

where x_i is the total supply of supplier i and y_j is the total capacity of consumer j. Once the transportation problem is solved, and we have found the optimal flow $F^* = \{f^*_{ij}\}$, the EMD is defined as:

$$EMD = \frac{\sum_{i\in I}\sum_{j\in J} f^*_{ij} c_{ij}}{\sum_{j\in J} y_j} \tag{16}$$

When applied to surveillance object matching, the cost c_{ij} becomes the distance of two blobs and the total supply x_j and y_j are the blob weights. c_{ij} can be various descriptor distance between two blobs such as color histogram distance, covariance matrix.

In comparison with the matching method based on the Hausdorff distance (Ma and Cohen 2007), our matching method based on the EMD distance possesses two precious characteristics. Firstly, it considers the participation of each blob in computing the distance based on its similarity with other blobs and its weight. Thanks to the representative blob detection method, blob weight expresses the important degree of this blob in object representation. The proposed matching method ensures a minor participation of irrelevant blobs produced by errors in object tracking because these blobs are relatively different from other blobs and have a small weight. Therefore, the matching method is robust when working with object tracking algorithms having a high value of *Object Id Confusion*. Secondly, the proposed object matching allows partial matching.

We analyze here an example of these object matching methods: We want to compute the similarity/dissimilarity between object O_q with 4 representative blobs and object O_p with 5 representative blobs (Fig. 12). The *Object Id Confusion* values of the object tracking module for the first object and the second object are 2 and 1 respectively.

In order to carry out object matching, firstly, we need to compute the distance of each pair of blobs. Tab. 1 shows the distance of each pair of blobs computed on covariance matrix distance (cf. Eq. 5) while Fig. 12 presents the result of object matching methods. Hausdorff-based object matching is determined by the distance between blob 1 of object O_q and blob 5 of object O_p (dot line) while EMD-based object matching search for an optimal solution with the participation of each blob. This example shows how the EMD-based object matching method overcomes the poor object tracking challenge.

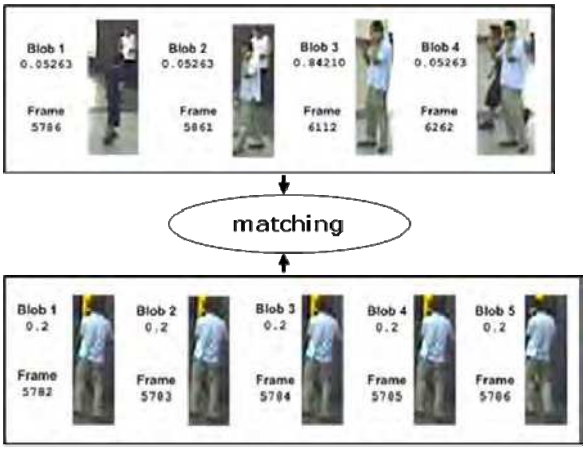


Fig. 11. Matching between object O_q with 4 representative blobs and object O_p with 5 representative blobs.

	Object O_p					
	Blob	Br_1^p	Br_2^p	Br_3^p	Br_4^p	Br_5^p
Object O_q	Br_1^q	3.873	3.873	3.873	3.873	3.361
	Br_2^q	2.733	2.733	2.733	2.733	2.161
	Br_3^q	2.142	2.142	2.142	2.142	1.879
	Br_4^q	2.193	2.193	2.193	2.193	2.048

Table 1. Distance of each pair of blobs of O_q and O_p based on covariance matrix distance

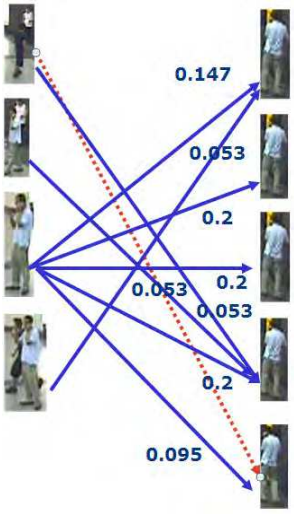


Fig. 12. Hausdorff-based and EMD-based object matching methods. Hausdorff-based object matching is determined by the distance between blob 1 of object Oq and blob 5 of object O_p (dot line) while EMD-based object matching search for an optimal solution.

With the output of the second object signature building approach, the object matching is relatively simple.

4. Surveillance object retrieval results

4.1 Databases

Despite the fact that a number of surveillance video systems have been deployed, very few surveillance databases are available. One reason is that surveillance videos concern to human and organization privacy. Recently, several surveillance video databases such as CAVIAR, i-LIDS, CARETAKER have been released for research purpose. CAVIAR (Context Aware Vision using Image-based Active Recognition) is a project funded by the EC's Information Society Technology's programme project IST 2001 37540. This project addresses two surveillance applications: city centre surveillance and marketers. Corresponding to these applications, two databases are available. Video clips in the first database were filmed with a wide angle camera lens in the entrance lobby of the INRIA Labs at Grenoble (France) while those of the second database are filmed with a wide angle lens along and across the hallway in a shopping centre in Lisbon (Portugal). Moreover, videos of these databases are annotated. 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) is a data set with multiple camera views from a busy airport arrival hall (Zheng, Gong et al. 2009). In the context of CARETAKER (Content Analysis and REtrieval Technologies to Apply Extraction to massive Recording), a video surveillance database is available. This project aims at studying, developing and assessing multimedia knowledge-based content analysis, knowledge extraction components and meta data management sub-systems in the context of automated situation awareness, diagnosis and decision support. During this project, a real testbed sites inside the metro of Roma and Torin, involving more than 30 sensors (20 cameras and 10 microphones) have been provided.

4.2 Surveillance object retrieval results

In recent years, a number of surveillance video retrieval results have been published. However, with the lack of common benchmarks and databases, the comparison of these results is difficult (even impossible). Two preliminary comparisons of three object signature building and object matching methods with CAVIAR and CARETAKER dataset have been presented in (Le, Thonnat et al. 2009a) (Le, Thonnat et al. 2009). However, these comparisons are done with a relatively small dataset.

5. Conclusions

In this chapter, firstly a brief overview of surveillance object retrieval is given. Then, current work dedicated to appearance-based surveillance object retrieval are analysed in detail. The analysis shows that preliminary and promising results have been obtained for surveillance object retrieval. However, it is still a challenging issue. This issue needs more work and contributions on surveillance video analysis, feature extraction and common benchmark for surveillance object retrieval evaluation.

6. References

Bak, S., E. Corvee, et al. (2010). Person Re-identification Using Spatial Covariance Regions of Human Body Parts. *AVSS*.

Broilo, M., N. Piotto, et al. (2010). Object Trajectory Analysis in Video Indexing and Retrieval Applications. *Video Search and Mining Studies in Computational Intelligence*, Springer Berlin Heidelberg. 287: 3–32.

Buchin, M., A. Driemel, et al. (2010). An Algorithmic Framework for Segmenting Trajectories based on Spatio-Temporal Criteria. *18th ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems (ACM GIS)*.

Calderara, S., R. Cucchiara, et al. (2006). Multimedia Surveillance: Content-based Retrieval with Multicamera People Tracking. *ACM International Workshop on Video Surveillance & Sensor Networks (VSSN'06)*. Santa Barbara, California, USA: 95-100.

Chen, L., M. T. Ozsu, et al. (2004). Symbolic Representation and Retrieval of Moving Object Trajectories. *MIR'04*.

Hsieh, J. W., S. L. Yu, et al. (2006). "Motion-Based Video Retrieval by Trajectory Matching." *Proc IEEE Trans. on Circuits and Systems for Video Technology* 16(3).

Hu, W., D. Xie, et al. (2007). "Semantic-Based Surveillance Video Retrieval." *IEEE Transactions on Image Processing* 16(4): 1168–1181.

Le, T.-L., A. Boucher, et al. (2007). *Subtrajectory-Based Video Indexing and Retrieval*. The International MultiMedia Modeling Conference (MMM'07), Singapore.

Le, T.-L., A. Boucher, et al. (2010). *Surveillance video retrieval: what we have already done?* ICCE, Nha Trang, VietNam.

Le, T.-L., M. Thonnat, et al. (2009)a. *Appearance based retrieval for tracked objects in surveillance videos*. ACM International Conference on Image and Video Retrieval 2009 (CIVR 2009), Santorini, Greece.

Le, T.-L., M. Thonnat, et al. (2009). "Surveillance video indexing and retrieval using object features and semantic events." *International Journal of Pattern Recognition and Artificial Intelligence, Special issue on Visual Analysis and Understanding for Surveillance Applications* 23(7): 1439-1476

Ma, Y. and B. M. a. I. Cohen (2007). Video Sequence Querying Using Clustering of Objects' Appearance Models. *International Symposium on Visual Computing (ISVC'07)*: 328–339.

Nghiem, A.-T., F. Bremond, et al. (2007). ETISEO, performance evaluation for video surveillance systems. *In Proceedings of International Conference on Advanced Video and Signal Based Surveillance (AVSS'07)*. London, United Kingdom.

Rubner, Y., C. Tomasi, et al. (1998). A metric for distributions with applications to image databases. *ICCV'98*: 59–66.

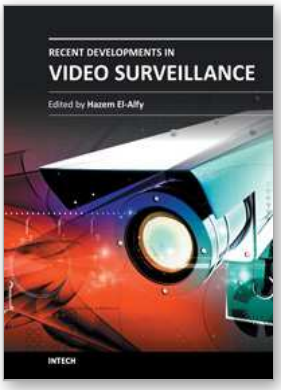
Senior, A. (2009). An Introduction to Automatic Video Surveillance *Protecting Privacy in Video Surveillance*: 1-9.

Velipasalar, S., L. M. Brown, et al. (2010). "Detection of user-defined, semantically high-level, composite events, and retrieval of event queries " *Multimedia Tools and Applications* 50(1): 249-278.

Won, C. S., D. K. Park, et al. (2002). "Efficient use of mpeg-7 edge histogram descriptor." *ETRI Journal* 24: 23–30.

Yuk, J. S. C., K. Y. K. Wong, et al. (2007). Object-Based Surveillance Video Retrieval System with Real-Time Indexing Methodology. *International Conference on Image Analysis and Recognition (ICIAR'07)*: 626-637.

Zhang, C., X. Chen, et al. (2009). "Semantic retrieval of events from indoor surveillance video databases." *Pattern Recognition Letters* 30(12): 1067-1076.



Recent Developments in Video Surveillance

Edited by Dr. Hazem El-Alfy

ISBN 978-953-51-0468-1
Hard cover, 122 pages
Publisher InTech
Published online 04, April, 2012
Published in print edition April, 2012

With surveillance cameras installed everywhere and continuously streaming thousands of hours of video, how can that huge amount of data be analyzed or even be useful? Is it possible to search those countless hours of videos for subjects or events of interest? Shouldn't the presence of a car stopped at a railroad crossing trigger an alarm system to prevent a potential accident? In the chapters selected for this book, experts in video surveillance provide answers to these questions and other interesting problems, skillfully blending research experience with practical real life applications. Academic researchers will find a reliable compilation of relevant literature in addition to pointers to current advances in the field. Industry practitioners will find useful hints about state-of-the-art applications. The book also provides directions for open problems where further advances can be pursued.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Thi-Lan Le, Monique Thonnat and Alain Boucher (2012). Appearance-Based Retrieval for Tracked Objects in Surveillance Videos, Recent Developments in Video Surveillance, Dr. Hazem El-Alfy (Ed.), ISBN: 978-953-51-0468-1, InTech, Available from: http://www.intechopen.com/books/recent-developments-in-video-surveillance/appearance-based-retrieval-for-tracked-objects-in-surveillance-videos

INTECH
open science | open minds

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