Object Tracking in Video-Surveillance

D. Moroni and G. Pieri

Institute of Information Science and Technologies (ISTI), Italian National Research Council (CNR), Pisa, Italy
e-mail: {davide.moroni, gabriele.pieri}@isti.cnr.it

Abstract—This paper faces the automatic object tracking problem in a video-surveillance task. A previously selected and then identified target has to be retrieved in the scene under investigation because it is lost due to masking, occlusions, or quick and unexpected movements. A two-step procedure is used, firstly motion detection is used to determine a candidate target in the scene, secondly using a semantic categorization and Content Based Image Retrieval techniques, the candidate target is identified whether it is the one that was lost or not. The use of Content Based Image Retrieval serves as support to the search problem and is performed using a reference data base which was populated a priori.

1. INTRODUCTION

Recognizing and tracking people in real time video sequences is nowadays a very relevant and challenging task. There is a need for automatic tools to perform the identification and to follow the *target* under investigation. Actually most tools for performing automatic tracking need to be run under specific environment and/or constraints, such as complete visibility of the target, impossibility to recover lost targets in a scene, and eventually also privacy issues play an important role in the efficiency of a tool.

In this paper we firstly describe the global task of active video-surveillance, then the focus will be on the subtask of automatically recover a lost target while in an automatic object tracking task.

Global task of active video-surveillance. The general and more global problem of detecting and tracking a moving target in a video-surveillance application has been faced in [1], in particular processing infrared (IR) video. Acquired images from video are processed for the detection of the target to be tracked in the actual frame, subsequently *active tracking* is performed through a *Hierarchical Artificial Neural Network* (HANN) for recognizing the actual target. In case the target cannot be recognized and it is lost due to some specific reason (e.g. occlusion), the *automatic target retrieval* phase is performed to identify and localize the lost target, this search is performed also on an a priori populated database and it is based on the paradigm of Content-Based Image Retrieval (CBIR).

Active tracking in a real time video surveillance is performed through 2 subphases:

—target identification: to localize spatially the target within the current frame of the sequence,

—target recognition: to recognize whether or not the target is the correct one to follow.

As initialization to the tracking task a motion detection algorithm can be used to detect a moving target in the scene, this can be done efficiently by exploiting the thermal characteristics of a human target in comparison to general background (independently to illumination conditions). The target can be localized and extracted from the frame under analysis through a segmentation algorithm. Once the target has been identified a set of various features can be computed, and these features can be used to assign the target to a *semantic class* it belongs to (i.e., for humans: upstanding person, crawling, crouched, …).

In order to perform target recognition, the set of computed features, included the computed semantic class is sent as input to a HANN which has been trained a priori off-line. If the network recognize the target as belonging to correct semantic class active tracking is considered successful and is repeated in the following frame acquired from the video sequence. In case a wrong object has been identified, and it is not recognized, the automatic search for the lost target object is started, based on CBIR paradigm and supported by the reference database. The identification of a wrong object can be due to various reasons: i.e., a masking, a partial occlusion of the target, or a quick and unpredictable movement in an unexpected direction.

Using CBIR allows to access information at a perceptual level i.e., using visual features, automatically extracted with appropriate similarity model [6].

The outcome of the automatic target search can be positive (i.e., target object tracked again) or negative (i.e., target object not yet identified or recognized). In the first case active tracking from the next frame will be performed again, while in the second case there is a need of continuing to perform automatic search of the target in the following frame. A stop condition to this search after a number of frames is set and will be

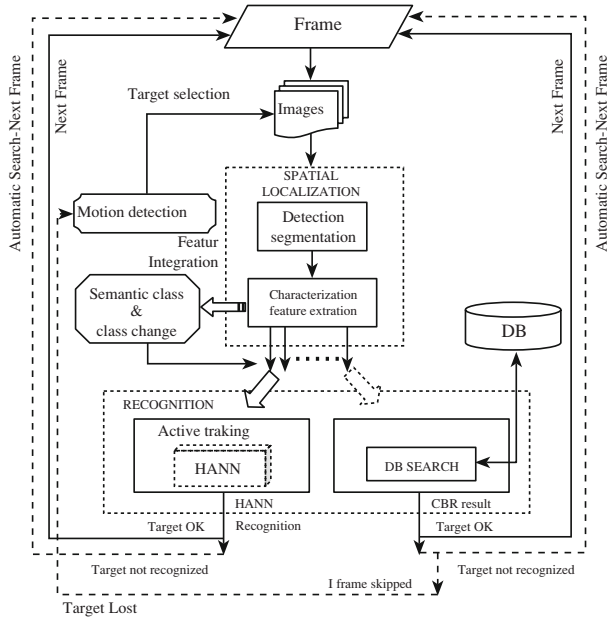

Fig. 1. The general algorithm for the trucking in the video-surveillance task [1].

defined in the following sections. When this stop condition is reached with no positive result the final decision and the control of the system are given back to the user warning about the lost target.

In Fig. 1 the general algorithm implementing the *on-line* approach for the global task of automatic tracking in video-surveillance is shown.

In order to be able to use the neural network, which needs to be trained over known samples, and the database, which needs to be populated for using the CBIR, both in real time, an *off-line* stage is performed tuning both components over the scene under surveillance. The database used for CBIR is organized into predefined semantic classes, and for each of the classes representative separate sets of shapes are stored to take into account many possible events to be recognized, such as target partial masking, and different orientation.

In the following sections the feature selection for target characterization and the automatic target search using CBIR are described in detail, finally some discussion and conclusion are drawn.

2. TARGET CHARACTERIZATION

The segmented target can then be characterized through a set of meaningful extracted features. The extracted features are grouped into 4 main categories due to their intrinsic characteristics that are related with:

—morphology;

—geometry;

—intensity;

—semantics.

The features are extracted from the region enclosed by the target contour that is defined by a sequence of N points (i.e., in our case $N = 16$) having coordinates $\langle x_s, y_s \rangle$, which can be described as being the target delimiting points.

Morphology. Extracting parameters from a shape, thus obtaining contour descriptors, represent a characterization of the morphology of the target. The shape can be obtained in an efficient way computing frames difference during the object segmentation. This frame difference is computed on a temporal window spanning 3 frames, this is made in order to prevent inconsistencies and problems due to intersections of the shapes. Let $\Delta(j, j-1)$ be the modulus of difference between actual frame F_j and previous frame F_{j-1}. Otsu's thresholding is applied to $\Delta(j, j-1)$ in order to obtain a binary image $Bin(j, j-1)$. Letting TS_j to be the target shape in the frame F_j, heuristically we have:

$$Bin(j, j-1) = TS_j \bigcup TS_{j-1}.$$

Thus, considering actual frame F_j, the target shape is approximated by the formula:

$$TS_j = Bin(j, j-1) \bigcap Bin(j+1, j).$$

Once the target shape is extracted two steps are performed: first an edge detection is performed in order to obtain a shape contour, second a computation of the normal in selected points of the contour is performed in order to get a better characterization of the target.

Two low level morphological features are computed following examples reported in [2]: the normal orientation, and the normal curvature degree. Considering the extracted contour, 64 equidistant points $\langle s_p, t_p \rangle$ are selected. Each point is characterized by the *orientation* Θ_p of its normal and its *curvature* K_p. To define these local features, a local chart is used to represent the curve as the graph of a degree 2 polynomial. More precisely, assuming without loss of generality that in a neighborhood of $\langle s_p, t_p \rangle$ the abscissas are monotone, the fitting problem

$$t = as^2 + bs + c$$

is solved in the least square sense. Then we define:

$$\Theta_p = \arctan\left(\frac{-1}{2as_p + b}\right), \quad (1)$$

$$K_p = \frac{2a}{\left(1 + (2as_p + b)^2\right)^{3/2}}. \quad (2)$$

More information can be obtained discretizing the normal orientation into 16 different bins corresponding to equiangular directions.

Such a histogram is invariant for scale transformation and, thus, independent from the distance of the target, hence it will be used for a more precise characterization of the semantic class of the target. This distribution represents an additional feature to the classification of the target e.g. a standing person will have a far different normal distribution than a crouched one. A vector $[\nu(\Theta_p)]$ of the normal for all the points in the contour is defined, associated to a particular distribution of the histogram data.

Geometry. Taking into account the shape and contour previously extracted other different features regarding the geometry of the segmented object can be computed. In particular two basis features are computed: the area and perimeter of the current shape.

$$\text{Area} = \left| \sum_{s=1}^{N} [(x_s y_{s+1}) - (y_s x_{s+1})] \right| /2,$$

$$\text{Perimeter} = \sum_{s=1}^{N} \sqrt{(x_s - x_{s+1})^2 + (y_s - y_{s+1})^2}.$$

Intensity. In our particular case study under investigation specific images were acquired, i.e., infrared images, thus we choose to exploit their specific characteristic of having a thermal radiation value of the target in each point of the image which also characterize the intensity of the image, to improve the efficiency of the system. Described below are more specific and IR-oriented features.

Average Intensity:

$$I = \frac{1}{\text{Area}} \sum_{t \in \text{Target}} F_j(t).$$

Standard deviation:

$$\sigma = \sqrt{\frac{1}{\text{Area} - 1} \sum_{t \in \text{Target}} (F_j(t) - I)^2}.$$

Skewness:

$$\gamma_1 = \mu_3 / \mu_2^{3/2}.$$

Kurtosis:

$$\beta_2 = \mu_4 / \mu_2^2.$$

Entropy:

$$E = - \sum_{t \in \text{Target}} F_j(t) \log_2(F_j(t)),$$

where $F_j(t)$ represents the thermal radiation value acquired for the frame j on the image point t, and μ_r are moments of order r of $F_j(t)$.

Semantics. In order to be able to reduce access time to the DB, the organization of the DB itself is made exploiting the predefined semantic classes, this allows to perform class-specific searches (i.e., on a reduced database set) which are more selective and efficient [5].

Each target, once it is identified through its extracted features, belongs to a specific semantic class (e.g., for a human: upstanding person, crouched, crawling, etc. …). This class can be considered as an additional feature computed considering the combination of the above defined features, among a predefined set of possible choices and assigned to the target.

Another characterization of the tracked object which regards its semantics, is what we call a *Class-Change* event, this occurs when, among different successive frames (i.e., during time), the semantic class assigned to the target changes. The Class-change is defined as a couple $\langle S_b ef, S_a ft \rangle$ that is associated with the target, and represents the modification from the semantic class $S_b ef$ computed before, and the semantic class $S_a ft$ computed in the actual frame. This event can be considered when more complex situations need to be evaluated in the context of video-surveillance. Due to the fact that the semantic classes are defined a priori, a matrix of class-changes can be defined, giving different specific weights to each class-change, e.g. a target recognized as a standing person before ($S_b ef$), and computed to be crouched in the current frame ($S_a ft$) could be given more attention (i.e., higher weight) than a different class-change.

The most important aspect for understanding class-changes is represented by morphology, in particular we can consider an index of the normal histogram distribution.

All the extracted information is passed to the recognition phase, in order to assess whether or not the localized target is correct.

3. TARGET SEARCH USING CBIR

In cases when the identified target has not been recognized (i.e., a wrong target recognition occurs), such as masking or occlusion, or quick movements, then the target search phase begins.

Systems based on CBIR generate low-level feature vectors which represent both the query and the image to retrieve, while in this system a semantic-based image retrieval is performed, hence a *semantic concept* is defined by means of sufficient number of training examples containing these concept [5]. Once the semantic concept is defined it is used to make a selective and more efficient access to the DB, and this *subset* is used to perform typical CBIR on low-level features.

A preliminary filtering on the DB is obtained using the reference semantic class to access the information

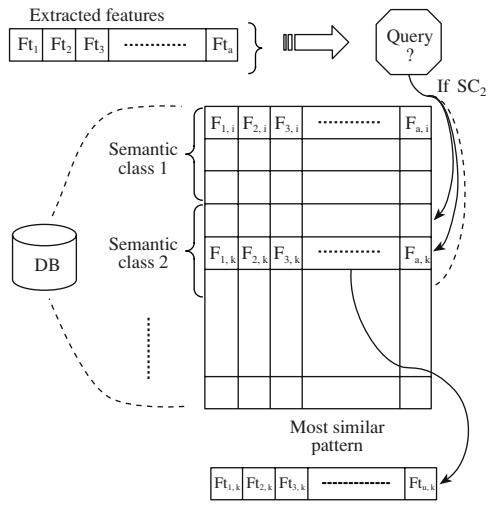

Fig. 2. Automatic target search with the support of Che Content-Based Image Retrieval and driven by the semantic class feature.

Fig. 3. The original frame (top), shape extraction by frames difference (bottom). Left and right represent two different postures of a tracked person.

in a more efficient manner [7]. The features extracted from the identified target are compared to the ones recorded in the reference DB using a quick similarity function for each feature class [3]. In particular, we considered intensity values matching, using percentages, shape matching, using the cross-correlation criterion, and the vector $[v(\Theta_p)]$ representing the distribution histogram of the normal. A weight is defined associated to each similarity value, in such a way a final single (and global) similarity measure can be obtained. Possible variations of the initial shape are recorded for each semantic class. In particular, the shapes to perform the similarity comparison are retrieved in the database using information in a set obtained considering the shape information stored at the time of the initial target selection joined with the one of the last valid shape of the target object. If the identified target has a final similarity value which has a distance below a fixed tolerance threshold, to at least one of the shapes in the obtained set, then it can be considered valid. Otherwise the target search starts again in the next frame [4].

In Fig. 2 a sketch of the CBIR in case of automatic target search is shown considering with the assumption that the database was previously defined (i.e., off-line), and considering a comprehensive vector of features $\langle Ft_k \rangle$ for all the above described groups.

Thus, when a pattern of the DB is found through CBIR, having a higher similarity value than the prefixed threshold, then the automatic target search has reached its goal and the target has been grabbed back and continuing in next frame automatic active tracking can be performed. If this does not happens then automatic target search is performed again in the next frame, not considering the last frame, but again the last valid shape of the searched target (i.e., the last correct target) as a starting point.

A stop condition has to be set for this automatic search, that is when after m frames the correct target has not yet been retrieved. In such a condition the control is given back to the user with a warning regarding the lost target situation. The value of m is computed when the target search starts considering the Euclidean distance between the *centroid* C of the last valid target object and the edge point of the frame E_r along the search direction r (the search direction vector is computed on the basis of the last $f = 5$ frames movements of the target), divided by the average speed of the target previously measured in the last f frames:

$$j = \|C - E_r\| / \left(\frac{\sum_{i=1}^{f} \text{Leap}_i}{f} \right), \tag{3}$$

where Leap_i represents the distance covered by the target between frames i and $i - 1$.

4. CONCLUSIONS

The object recognition and tracking problem has been described in the frame of a general case study of video surveillance in the control of accesses to restricted areas. In particular more focus has been given to the sub-task of recovering a target which is lost while in real time tracking something happens, such as occlu-

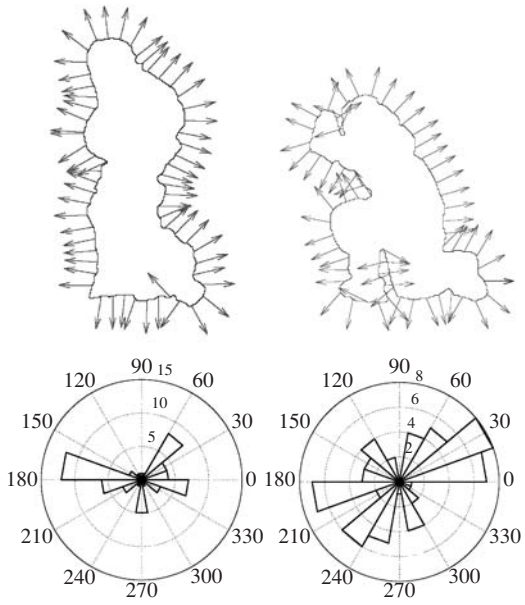

Fig. 4. Shape contour with normal vector on 64 points (top), distribution histogram of the normal (bottom). Left and right epresent two different postures of a tracked person (same frames as in Fig. 3).

sion, unexpected quick movements. The specific camera used to acquire the videos was a thermocamera working in the 8–12μ_m wavelength range. The camera was mounted on a robotic moving structure covering 360° pan and 90° tilt, and equipped with 12° and 24° optics. Finally spatial resolution of the video images to process was 320×240 pixels.

In our case study the off-line stage to build the database for the CBIR was performed taking into account various image sequences relative to different classes of the scenario under surveillance. Regarding in particular the choice of semantic classes for a *human* class, it has been composed defining three different postures (i.e., upstanding, crouched, crawling) while also taking into account considering three possible people categories (short, medium, tall).

An estimation of the number of operations computed for each frame in real time tracking yields the following values: about 5×10^5 operations for the identification and characterization phases, about 4×10^3 operations for the active tracking. This assures the real time functioning of the procedure on a personal computer of medium power due to the fact also that the heaviest operations are performed while in off-line stage. The automatic search process can require a higher number of operations, but it is performed when the target is partially occluded or lost due to some obstacles, so it can be reasonable to spend more time in finding it, thus loosing some frames. Among the many variables on which the number of operations depends, the relative dimension of the target to be followed is an important amount, i.e., bigger targets require a higher effort to be segmented and characterized, even if they prove not to be the correct ones.

In Fig. 3 example frames of video sequences are shown (top) together with their relative shape extraction.

While in Fig. 4 the same frames are processed, and shape contour with distribution histogram of the normal are shown.

A methodology has been proposed for object tracking and target recovery based on the Content Based Image Retrieval paradigm in a video-surveillance task. Target identification during active tracking has been performed, using a Hierarchical Artificial Neural Network (HANN). In case of automatic searching, when the tracking object is lost or occluded, and needs to be grabbed back again, a CBIR database search and compare has been used for the retrieval and comparison of the currently extracted features with the previously reliable stored. In order to optimize the storing and search in the reference database an approach based on semantic categorization of the information has been defined.

ACKNOWLEDGMENTS

This work was partially supported by EU NoE MUSCLE—FP6-507752. We would like to thank Eng. M. Benvenuti, head of the R&D Dept. at TDGroup S.p.A., Pisa, for allowing the use of proprietary instrumentation for test purposes.

REFERENCES

1. D. Moroni and G. Pieri, "Active Video-Surveillance Based on Stereo and Infrared Imaging," Eurasip J. Appl. Signal Proc. **2008** (Article ID 380210), 8 (2007).
2. S. Berretti, A. Del Bimbo, and P. Pala, "Retrieval by Shape Similarity with Perceptual Distance and Effective Indexing," IEEE Trans. Multimedia **4** (2), 225 (2000).
3. P. Tzouveli, G. Andreou, G. Tsechpenakis, and Y. Avrithis, "Intelligent Visual Descriptor Extraction from Video Sequences. Lecture Notes in Computer Science," Adaptive Multimedia Retrieval **3094,** 132 (2004).
4. M. G. Di Bono, G. Pieri, and O. Salvetti, "Multimedia Target Tracking through Feature Detection and Database Retrieval," in *22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, 2005* (2005), pp. 19–22.
5. M. M. Rahman, P. Battarcharya, and B. C. Desai, "A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques with Relevance Feedback," IEEE Trans. Inform. Tech. Biomed. **11** (1), 58 (2007).
6. A. Smeulder, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis Mach. Intell. **22** (12), 1349–1380 (2003).
7. Y. Chen, J. Z. Wang, and R. Krovetz, "CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning," IEEE Trans. Image Proc. **14** (8), 1187 (2005).