Automatic Target Retrieval in a Video-Surveillance Task

Davide Moroni and Gabriele Pieri

Institute of Information Science and Technologies (ISTI)
Italian National Research Council (CNR), Via Moruzzi 1, 56124, Pisa, Italy

Abstract. In this paper we face the automatic target search problem. While performing an object tracking task, we address the problem of identifying a previously selected target when it is lost due to masking, occlusions, or quick and unexpected movements. Firstly a candidate target is identified in the scene through motion detection techniques, subsequently using a semantic categorization and content based image retrieval techniques, the candidate target is identified whether it is the correct one (i.e. the previous lost target), or not. 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

The recognition and tracking of people in video sequences is a very challenging task and automatic tools to firstly identify and subsequently follow a human - *target* - are often subject to constraints regarding the environment under surveillance, the complete or uncomplete visibility of the target itself and eventually privacy issues.

In this paper we investigate and detail the single subtask of automatically recover a target within the more ample task of active video-surveillance as reported in [1].

Active video-surveillance task. In our previous work [1] we faced the global problem of detecting and tracking a moving target, in particular processing infrared (IR) video. The acquired information is elaborated for detecting and extracting the target in the current frame, then *active tracking* is performed using a *Hierarchical Artificial Neural Network* (HANN) for the recognition of the actual target. When the target is lost or occluded a Content-Based Image Retrieval (CBIR) paradigm is applied, with the support of an a priori populated reference database, in order to identify and localize the correct lost target in a phase called *automatic target retrieval*.

The active tracking in a real time video sequence is performed through 2 sub-phases:

- target spatial localization: to identify the target in the current frame of the sequence
- target recognition: to recognize whether or not the target is the correct one to follow

An initialization step is required to start the tracking task, in detail a moving target in the scene is detected by means of a motion detection algorithm, exploiting the thermal characteristics of the target in relation to the IR camera characteristics. Such identified target is firstly localized and extracted through a segmentation algorithm and then

Moroni D. and Pieri G. (2008).
Automatic Target Retrieval in a Video-Surveillance Task.
In *Image Mining Theory and Applications*, pages 113-121
DOI: 10.5220/0002340901130121

used to compute a set of different features containing useful information on shape and thermal properties of the target. First of all the combination of such features is used to assign the target a *semantic class* to which the it belongs to (i.e. upstanding person, crawling, crouched, . . .).

The HANN is used to recognize, through the extracted features, included the semantic class, if the target belongs to the same class it was in the previous frames or not. In the former case active tracking is considered successful and is repeated in the following frame acquired from the video sequence. In the latter case, a wrong object recognition happens, this can be due to either a masking, a partial occlusion of the target, or a quick and unpredictable movement. Following this condition the automatic search for the lost target, supported by the CBIR, is started.

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

The search of the target for the current frame can give a positive result (i.e. the target is grabbed back), or a negative one (i.e. target not yet recognized), in the former case starting from the successive frame the active tracking through HANN is performed again, in the latter the automatic search will be performed again in the successive frame, this condition stops after a computed number of frames is reached, in this case the control is given to the user of the system with a warning about the lost target.

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

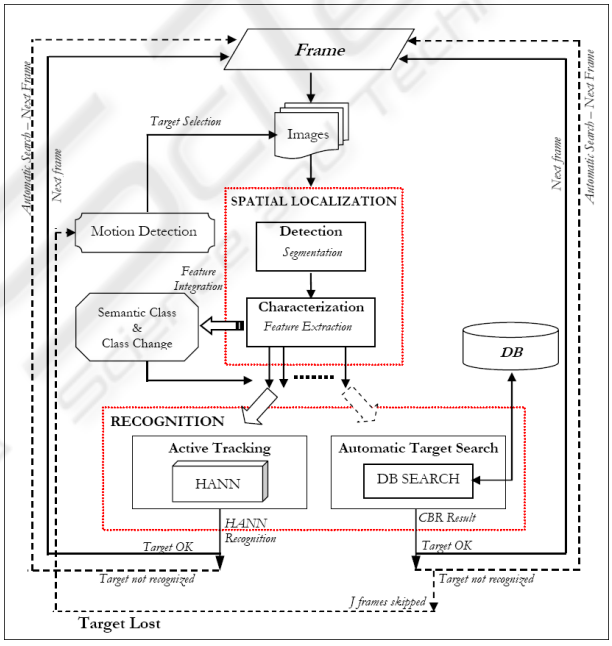

Fig. 1. The general algorithm for the tracking in the video-surveillance task [1].

The on-line active tracking process requires an *off-line* process in order to both: train the neural network over selected and known examples, and populate the database used during the CBIR for the automatic target search. The database is organized on the basis of the predefined semantic classes computed. For each of the defined classes sets of different shapes are stored, in order to take into account partial masking of the target and different orientations.
In the following sections the feature selection and the CBIR stage are described in detail.

2 Feature Selection

Once segmented, the target is described through a set of meaningful extracted features. The features are divided into 4 main categories:

- morphological
- geometric
- thermographic
- semantic

These four different groups of visual features which are extracted from the region enclosed by the target contour that is defined by a sequence of N_c points (i.e in our case $N_c = 16$) having coordinates $< x_s, y_s >$, which can be described as being the outer *limit points* of the target.

Morphological: Morphological features represent shape contour descriptors obtained by means of characterization parameters extracted from a shape which is computed through frames difference during the 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 $\delta(i-1,i)$ be the modulus of difference between actual frame F_i and previous frame F_{i-1}. Otsu's thresholding is applied to $\delta(i-1,i)$ in order to obtain a binary image $Bin(i-1,i)$. Letting TS_i to be the target shape in the frame F_i, heuristically we have:

$$Bin(i-1,i) = TS_{i-1} \bigcup TS_i$$

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

$$TS_i = Bin(i-1,i) \bigcap Bin(i,i+1)$$

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 $< s_m, t_m >$ are selected. Each point is characterized

by the *orientation* of its normal and its *curvature* K_m. 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 $< s_m, t_m >$ the abscissas are monotone, the fitting problem

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

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

$$\Theta_m = atan\left(\frac{-1}{2as_m + b}\right) \tag{1}$$

$$K_m = \frac{2a}{(1 + (2as_m + b)^2)^{3/2}} \tag{2}$$

Moreover the histogram of the normal orientation, discretized into 16 different bins, corresponding to 16 equi-angular directions, is extracted.

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 $[v(\Theta_m)]$ of the normal for all the points in the contour is defined, associated to a particular distribution of the histogram data.

Geometric: Geometric features are computed on the basis of the contour, and thus the shape, previously extracted. Two main features are computed: Area and Perimeter of the actual shape to be characterized.

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

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

Thermographic: Considering the nature of the acquired images in our video-surveillance task, i.e. infrared images, we choose to exploit the specific characteristics, such as the thermal radiation value of the target in each point of the image, to improve the efficiency of the system. Thus, more specific and IR oriented features are extracted and described below.

– Average Temp:
$$\mu = \frac{1}{Area} \sum_{p \in Target} F_i(p)$$

– Standard deviation:

$$\sigma = \sqrt{\frac{1}{Area - 1} \sum_{p \in Target} (F_i(p) - \mu)^2}$$

– Skewness:

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

– Kurtosis:

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

– Entropy:

$$E = - \sum_{p \in Target} F_i(p) log_2(F_i(p))$$

Where $F_i(p)$ represents the thermal radiation value acquired for the frame i on the image point p, and μ_r are moments of order r.

Semantic Class: Organizing the images in the DB through predefined semantic classes enables the possibility of categorizing them and hence allowing to perform class-specific searching that is more selective and efficient [5].

The semantic class the target belongs to (e.g. upstanding person, crouched, crawling, etc...) can be considered as an additional feature and is automatically selected, considering combinations of the above defined features, among a predefined set of possible choices and assigned to the target.

Moreover a *Class-Change* event is defined which is associated with the target when its semantic class changes in time (different frames). This event is defined as a couple $< SC_b, SC_a >$ that is associated with the target, and represents the modification from the semantic class SC_b selected before and the semantic class SC_a selected after the actual frame. This transition is considered in order to evaluate more complex situations in the context of a surveillance task, e.g. a person who was standing before and is crouched in the next frames, when identified by its class change, can be given a specific weight different from other class changes.

In order to retrieve when the semantic class of the target changes, important features to consider are the morphological features, and in particular 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 Automatic Target Search

When wrong target recognition occurs, due to masking or occlusion, or quick movements in unexpected directions, the automatic target search starts.

In CBIR systems low-level feature vectors are generated to represent the query and the images 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, onto which typical CBIR for images is performed on the low level features.

The features of the candidate target are compared to the ones recorded in a reference database using a similarity function for each feature class [3]. In particular, we considered color matching, using percentages and color values, shape matching, using the cross-correlation criterion, and the vector $[v(\Theta_m)]$ representing the distribution histogram of the normal. First of all the reference semantic class is used as a filter in order to access the database information in a more efficient manner [7]. Each of these similarity values is associated to a predefined weight, hence a single global similarity measure is computed. For each semantic class, possible variations of the initial shape are recorded. In particular, the shapes to compare with 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. If the candidate target shape has a similarity measure which has a distance below a fixed tolerance threshold, to at least one shape in the obtained set, then it can be considered valid. Otherwise the search starts again in the next frame acquired [4].

In Figure 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 $< Ft_k >$ for all the above mentioned categories.

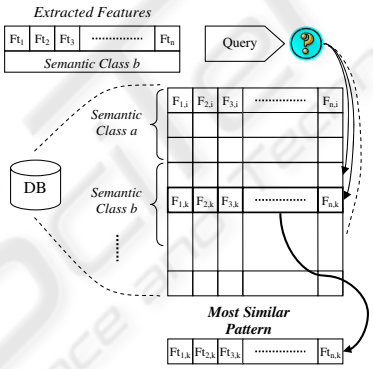

Fig. 2. Automatic target search with the support of the Content-Based Image Retrieval and driven by the semantic class feature.

Thus, if a pattern is found through CBIR in the database, which has a similarity measure higher than the prefixed threshold, then the automatic search has success and the target is grabbed back for the active tracking in the next frame of the video sequence. Otherwise automatic search will be performed again in the next frame, considering the last valid shape (i.e. the target shape of the last correct target) as a starting point.

If after j frames the correct target has not yet been grabbed, the control is given back to the user. The value of j is computed considering the Euclidean distance between the *centroid* C_v of the last valid shape and the edge point of the frame E_r along the search

direction r, divided by the average speed of the target previously measured in the last f frames:

$$j = \|C_v - E_r\| / \left(\frac{\sum_{i=1}^{f} step_i}{f} \right) \quad (3)$$

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

4 Results and Conclusions

The automatic target search problem has been faced in a general real case study for video surveillance application to control unauthorized access in restricted access areas. The videos were acquired using a thermocamera in the $8 - 12\mu_m$ wavelength range, mounted on a moving structure covering 360° pan and 90° tilt, and equipped with 12° and 24° optics to have 320×240 pixel spatial resolution.

The database for the CBIR was built taking into account different image sequences relative to different classes of the monitored scenes. In particular, the human class has been composed taking into account three different postures (i.e. upstanding, crouched, crawling) considering three different people typologies (short, medium, tall).

The estimated number of operations performed for each frame when tracking persons consists of about $5 \cdot 10^5$ operations for the identification and characterization phases, while the active tracking requires about $4 \cdot 10^3$ operations. This assures the real time functioning of the procedure on a personal computer of medium power. 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. Of course, the number of operations depends on the relative dimension of the target to be followed, i.e. bigger targets require a higher effort to be segmented and characterized. The acquired images are pre-processed to reduce the noise. In Figure 3 example frames of video sequences are shown (top) together with their relative shape extraction.

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

A methodology has been proposed for the Content Based Image Retrieval problem in a video-surveillance task, when the tracking target is lost or occluded, and needs to be grabbed back again. Target recognition during active tracking has been performed, using a Hierarchical Artificial Neural Network (HANN). In case of automatic searching of a masked or occluded target, a Content-Based Image Retrieval paradigm has been used for the retrieval and comparison of the currently extracted features with the previously stored in a reference database. The reference database has been implemented and populated following an approach based on semantic categorization of the information.

Acknowledgements

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. Moroni, D., Pieri, G.: Active video-surveillance based on stereo and infrared imaging. J. Applied Signal Processing, (in press). (2007)
2. Berretti, S., Del Bimbo, A., Pala, P.: Retrieval by shape similarity with perceptual distance and effective indexing. IIEEE Transactions on Multimedia, Vol. 4(2). (2000) 225–239
3. Tzouveli, P., Andreou, G., Tsechpenakis, G., Avrithis, Y.: Intelligent visual descriptor extraction from video sequences. Lecture Notes in Computer Science - Adaptive Multimedia Retrieval, Vol. 3094. (2004) 132–146
4. Di Bono, M.G., Pieri, G., Salvetti, O.: Multimedia target tracking through feature detection and database retrieval. In: 22^{nd} International Conference on Machine Learning - ICML 2005. Bonn, Germany. August (2005) 19–22
5. Rahman, M.M., Battarcharya, P., Desai, B.C.: A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. IEEE Trans. on Inform. Tech. in Biomedicine, Vol. 11(1). (2007) 58–69
6. Smeulder, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. on Pattern Analysis and Mach. Intell., Vol. 22(12). (2003) 1349–1380
7. Chen, Y., Wang, J.Z., Krovetz, R.: CLUE: cluster-based retrieval of images by unsupervised learning. IEEE Trans. on Image Proc., Vol. 14(8). (2005) 1187–1201

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

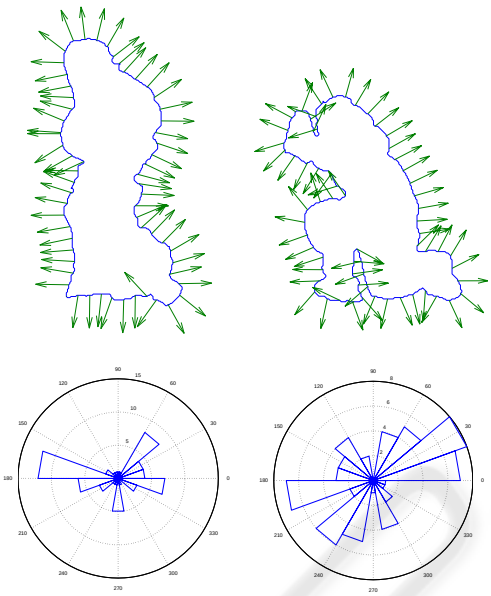

Fig. 4. Shape contour with normal vector on 64 points (top), distribution histogram of the normal (bottom). Left and right epresent two different postures of a tracked person (same frames as in Figure 3.