Multimedia Target Tracking through Feature Detection and Database Retrieval

Maria Grazia Di Bono
Gabriele Pieri
Ovidio Salvetti
Institute of Information Science and Technologies, ISTI-CNR, Via Moruzzi 1, 56124 Pisa, ITALY

Abstract
The real-time detection and tracking of moving objects is a challenging task and automatic tools to identify and follow them are often subject to constraints regarding the environment under investigation or the full visibility of the targeted object. Exploiting the possibility of a multi-source acquisition in the targeted scene, firstly detection is performed by means of characteristic features extraction and storing in a database; secondly, the tracking task is approached using algorithms, where automatic search involves occluded or masked targets in the scene. This latter problem is solved through database retrieval, based on well-defined multi-modal features. The method has been tested on case studies regarding the identification and tracking of animals moving at night in an open environment (i.e. natural reserves or parks), and the surveillance of known scenes for unauthorized access control.

1. Introduction
According to the cognitive processes of the human perception (Milner & Goodale, 1995), a methodology has been developed which provides a way to realize object recognition and tracking in 3D real environments. In particular, this approach is based on the acquisition of multi-source information that is firstly elaborated for object detection and characterization, and then for its localization and active tracking. After target detection is achieved through an automatic segmentation, the characterization phase is performed through the description of multi-modal features (morphological, densitometric and semantic), which are extracted from the acquired multi-source information. Localization is realized using also features previously extracted and stored in a reference database. In order to improve the localization performance when only partial information is available (i.e. in case of lost or occluded targets), the implemented method is supported by a content-based retrieval (CBR) paradigm using an a priori defined multimedia (MM) database. This MM database is built using the multi-modal features extracted from a set of target examples organised on the basis of semantic classes defined on the specific environment under investigation.

Current approaches regarding real-time object tracking from videos are based on (i) successive frame differences (Fernandez-Caballero et al., 2003), using also adaptive threshold techniques (Fejes & Davis, 1999), (ii) trajectory tracking, using weak perspective and optical flow (Yau, Fu & Liu, 2001), (iii) region approaches, using active contours of the target and neural networks for movement analysis (Tabb et al., 2002), or motion detection and successive regions segmentation (Kim & Kim, 2003).

Regarding the CBR paradigm, techniques of shape retrieval in large databases are particularly interesting. Considering a shape of an object as a sequence of contour points, a method using both global and local features is discussed in (Wang, Yang & Acharya, 1998) while in (Wang, Chang & Acharya, 1999) retrieval is based on a hash table and a majority voting algorithm for an efficient estimation of shape similarity. Furthermore, another interesting approach considers a shape database structured as an M-tree of organised tokens, representing parts of the shape enclosed between contour points. Possible shapes are clustered into semantic classes, each belonging to an object typology defined in its environment (Berretti, Del Bimbo & Pala, 2000).

In this paper, the problem of moving target detection and tracking is faced by processing multi-source information acquired using cameras of different typology (Far-IR and visible). Object characterization is based on region segmentation and feature extraction processes. Object localization uses a CBR approach based on similarity functions defined for each multi-modal feature class.

The method has been applied to real case studies regarding the monitoring of animal movements during the night in an open environment (i.e. natural reserves or parks) and the surveillance of known scenes for unauthorised access control.

Unauthorized access control in both open and closed spaces (Pieri et al., 2004).

2. Problem definition

The precise identification of a defined target in a real video, frame by frame, is approached. The proposed methodology is based on recognition and spatial localization of the target: recognition is sub-divided into identification and characterization, while the spatial localization performs active tracking.

The multi-source information is acquired using a physical system composed of a thermo-camera and two stereo visible-cameras synchronized. Thus, we obtain a set of infrared (IR) images, which make the system more robust and invariant to light changes in the scene, corresponding to stereo grey level images.

A procedure has been defined based on two different stages:

- **Off-line stage**, in which the recognition phase is performed using selected examples belonging to a set of predefined semantic classes, in order to populate the reference MM database.
- **On-line stage**, in which the tracking is performed by applying recognition and spatial localization.

In deep details, during the recognition process, the identification phase consists of an automatic segmentation, based on edge detection using a gradient descent along 16 directions starting from a reference point internal to the target (centroid).

In the characterization phase, for each frame, the multi-source information is used in order to extract a target description from the scene. This is made through a feature extraction process performed on the three different images available for each frame in the sequence. In particular, the extraction of a depth index from the grey level stereo images, performed by computing disparity of the corresponding stereo points, is realized in order to have significant information about the target spatial localization in the 3D scene and the target movement along depth direction, which is useful for the determination of a possible static or dynamic occlusion of the target itself in the observed scene. Other features consisting in radiometric parameters measuring the temperature and visual features are extracted from the IR images. The visual features, grouped in morphological, densitometric and semantic classes, consist of shape contour descriptors, dominant colour discriminants, statistical parameters, computed on the regions enclosed by the contours (area, perimeter, average brightness, standard deviation, skewness, kurtosis, and entropy) and the semantic class to which the target belongs (i.e. human, small, medium and large animal, …). While the depth index and the visual features are automatically extracted from the images, the semantic classes of the observed targets are selected by the user among a predefined set of possible choices.

During the off-line stage, all the multi-modal feature information is stored in the MM database, organised on the base of semantic classes. This information is used during the on-line spatial localization process, in particular in the automatic target retrieval which acts as a support during the active tracking in case of partial occlusion or quickly direction changes of the target.

For each defined target class, possible variations of the initial shape, taking into account that the target could be still partially masked or have a different orientation, are recorded together with the other multi-modal features as it is shown in Figure 1.

![Figure 1. Example of actual targets in the MM database, grouped according to the classes “small animal” (CSA), “medium animal” (CMA) and “large animal” (CLA).](image-url)
The tracking algorithm is performed on the IR image sequence, in order to build a system which can be used both at night and daylight.

The segmented target from the first IR image of the sequence is then tracked automatically in the following frame. The features used for the automatic tracking are local maxima, movement prediction (on the basis of the movements of the previous steps), temperature and a priori knowledge about the specific class the object belongs to. For each frame, the algorithm performs the steps to correctly identify the target and to follow it.

Firstly, a candidate characterizing point \( P_1 \) of the target is selected in its centroid, in the actual frame. The selection follows criteria of brightness local maximum, inside the contour segmented in the previous frame; \( P_1 \) is the point having the maximum similarity with the centroid \( P_p \) of the previous frame.

In a second step, the algorithm takes into account the previous movements of the centroid. The trajectory is stored and then used in the computation of the actual step, locating a new candidate point \( P_2 \). If \( P_2 \) is not coincident with \( P_1 \) then a new point \( P_3 \) is calculated as:

\[
P_3 = \alpha P_1 + \beta P_2
\]

(1)

where \( \alpha \) and \( \beta \) represent the weight assigned, and \( \alpha + \beta = 1 \). These parameters are empirically defined and can be adjusted by the user.

Again, a local maximum search is performed in the neighbourhood of \( P_3 \) to make sure that it is internal to a valid object. This search finds the point \( P_N \) that has the grey level closest to the one of \( P_p \), so that \( P_N \) is the centroid chosen for the actual frame. Starting from this point, the edge detection is performed and the object new contour is segmented.

In each frame, a first control is made trying to avoid a wrong object recognition, due to either a masking, partial occlusion of the object in the scene or to a quick movement in an unexpected direction. This control takes into account the above mentioned statistical parameters computed on the region enclosed by the contour, without using CBR paradigm in order to optimise the number of accesses to the database. If there are parameters exceeding \( p \) times (\( p \) is defined a priori) the standard deviation of the same parameters computed over the last \( n \) frames, the database search for the correct target is started. This search is based on the CBR paradigm; the multi-modal features of the candidate target are compared to the ones recorded in the MM database. A similarity function is considered for each feature class. In particular, we used similarity functions, as in (Tzouveli et al., 2004), for colour matching, using percentages and colour values, and shape matching, using the cross-correlation criterion. In order to obtain a global similarity measure, each similarity percentage is associated to a pre-selected weight, using the reference semantic class as a filter to access the MM information. 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 distance between \( P_p \) and the edge point of the image along the search direction, divided by the average velocity of the target previously measured in the last \( n \) frames (Eq. 2).

\[
j = \frac{\text{Dist}(P_p; E_r)}{\text{Vel}}
\]

(2)

where \( \text{Dist}(x, y) \) is the Euclidean distance between points \( x \) and \( y \); \( E_r \) is the point crossing the edge of the frame along the search direction \( r \) determined by the last \( n \) centroids; and \( \text{Vel} \) is

\[
\text{Vel} = \frac{\sum_{i=0}^{n-1} \text{Dist}(P_p^i; P_p^{i+1})}{n}
\]

(3)

where \( P_p^i \) is the centroid \( i \) steps before the actual.

The sketch of the methodology described is shown in Figure 2.

**Figure 2.** Recognition and description of a target object (on-line process).
3. Results and Conclusions

The method implemented has been applied to real case studies: (i) to track animal movements in an open environment during the night, for the fauna monitoring in natural parks, and (ii) for video surveillance of known scenes both at night and daylight to control unauthorized access (see Figure 3).

![Figure 3. Examples of thermo images regarding human (left and centre, two different views and shapes) and animal (right) targets (crosses are the centroids).](image)

Regarding the first case, due to the environmental conditions, only the thermo-camera has been used.

The videos were acquired using a thermo-camera in the 8-12µm wavelength range, mounted on a moving structure covering 360° pan and 90° tilt, and equipped with 12° and 24° optics to have 320x240 pixel spatial resolution.

Both the thermo-camera and the two stereo visible-cameras have been positioned in order to explore a scene 100 meters far, sufficient in our experimental cases.

In the fauna monitoring experimental case, during the off-line stage, the MM database has been built taking into account different image sequences relative to different classes of the monitored animals. In particular, three main semantic classes have been determined. The *large-animal* class counting all the monitored animals of a large size like deer, the *medium-animal* class including animals of medium size like boars and the *small-animal* class considering other kind of animals like rabbits or badgers. For each outlined semantic class, different positions have been considered. In more details, four different positions for boars, rabbits and other small animals and six for deer have been registered.

In the video-surveillance case, the *human* class has been composed taking into account six different pose conditions for three different people typology.

The acquired images are pre-processed to reduce the noise, the algorithm has shown an effective performance and seems promising in the lights of further improvements regarding for example the integration with audio information, coming from different aligned microphones installed in the scene, and aiming at the same direction of the cameras.

References


