Tracking of Moving Targets in Video Sequences

MARCO BENVENUTI°, SARA COLANTONIO, MARIA GRAZIA DI BONO, GABRIELE PIERI
and OVIDIO SALVETTI

Institute of Information Science and Technologies (ISTI – CNR)
'Signals and Images' Laboratory
Via Moruzzi, 1 – 56124 Pisa
ITALY
Sara.Colantonio@isti.cnr.it, Maria.Grazia.Dibono@isti.cnr.it, Gabriele.Pieri@isti.cnr.it, Ovidio.Salvetti@isti.cnr.it

° TD Group SpA
Via Traversagna, 29 – Migliarino Pisano
ITALY
m.benvenuti@tdnet.it

Abstract: - A research has been carried out finalised to the definition of a methodology useful to detect and track moving targets in video sequences. Algorithms performing this task have been also developed for real time monitoring and surveillance purposes. Due to deformations occurring in the appearance of the target in the videos, a Hierarchical Artificial Neural Network (HANN) has been used to recognize target occlusion or masking, and to increase the normal tracking performance. Preliminary results are presented regarding both identification and tracking of animal moving at night in an open environment, and the surveillance of known scenes for unauthorized access control.

Key-Words: - Target Tracking, Hierarchical Artificial Neural Networks, Image Processing, Object Recognition.

1 Introduction

Real time target tracking from video sequences using an automatic robotized system in an open environment is still a challenging task. Current approaches are based on successive frame differences [1, 2], on trajectory tracking using weak perspective and optical flow [3], or region segmentation, defining active contours of the target objects, and neural networks to perform the movement analysis [4]. Also approaches using adaptive threshold techniques have been developed to detect the points that are moving in a coherent way through different frames [5], or to perform a motion detection, and a successive region segmentation [6]. Segmentation techniques have been also used to cluster pixels into regions corresponding to single objects on the basis of grey level and proximity, and then local motion estimation for region merging has been applied [7].

In this paper we present a methodology based on an algorithm for target tracking, combined with a HANN system for object recognition [8]. The HANN has been introduced to improve the tracking performance even when only partial information is available (i.e. lost or occluded targets). This approach is based on the acquisition of information that is firstly elaborated for target identification and characterization, and then for active tracking [9].

The proposed methodology 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 environmental surveillance in both open and closed spaces.

2 Problem Formulation

The performance and accuracy regarding the monitoring and tracking of moving targets in a free and open environment has been considered as a combination of two correlated steps: target detection, and target recognition. In particular, the following different phases have been pointed out:

- *Target selection*
- *Target characterization*
- *Target tracking (detection and recognition)*

The video sequences are acquired using infrared (IR) cameras, in order to have a more robust approach invariant to light changes in the scene. The acquired video sequences are then composed of grey level images so that the target has usually a higher temperature than the major part of the background.

The *target selection* phase, which starts the automatic tracking algorithm, consists of a first manual intervention of the user for initialisation, followed by the automatic extraction of the target from the scene through a rough segmentation.

This is to offer a major control to the user on the choice of the target, and also to exploit the clearly contrasted nature of the IR images.
At the beginning the user selects a *characterizing point*, representative of the target. During the automatic tracking the target will be then identified by this point in each frame.

The *target characterization* phase consists of describing the segmented target through a feature extraction process. In particular, the extracted visual features belongs to *morphological* (i.e. Shape Contour Descriptors), *geometric*, and *densitometric* classes.

The *target tracking* phase takes into account the movements of the robotized camera which are typically in the opposite direction with respect to the target motion (i.e. aiming at keeping the target centred). Algorithms for motion detection and tracking have been implemented, including also a *target recognition* process, based on a HANN, to improve the algorithm performance and take into account events in which the degree of uncertainty on the localization of the actual target is increased by a possible occlusion or masking in the scene.

3 Approach and Techniques

A scheme of the developed approach is shown in Fig.1.

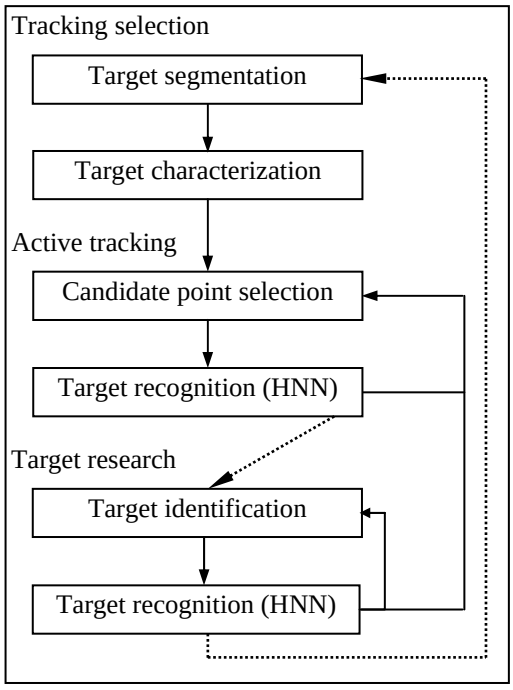

Fig.1. Scheme of the target tracking approach.

3.1 Target Selection

After the interactive selection on the first frame of an internal point to the target, an edge detection algorithm is applied to achieve an automatic coarse segmentation of its contour (Fig.2).
The edge detection algorithm performs a gradient descent along *N* predefined directions starting from the characterizing point (dark cross in Fig.2), obtaining *N* edge points. By interpolating these edge points the segmented contour of the target is approximated (*N* consecutive segments, named *tokens*).

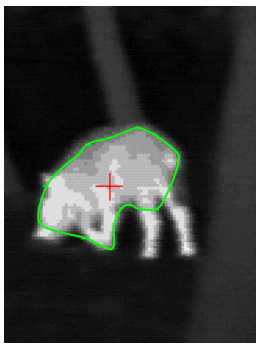

Fig.2. Interactive selection of the target and automatic coarse segmentation.

Contextually with the initial operation, also an interactive target class selection is performed choosing it among the ones used during an off-line training. This information is then used during both the active tracking, and the automatic target search to control if the target itself is correctly tracked.
The automatic tracking is then based on the characteristic features extracted from the target, described by its tokens, and the region enclosed by the tokens themselves.

3.2 Features Characterization

Target characterization consists of a feature extraction process able to define the salient characteristics of the actual segmented target.
The extracted features can be divided into the three following classes:

- *Morphological*, describing characteristics properly related to the shape of the target
- *Geometric*, related to general geometric characteristic derivable from the contour
- *Densitometric*, representative of statistical measures about pictorial indexes of the image region enclosed by the contour

The morphological features are derived extracting characterization parameters from the *N* tokens that compose the target contour. In particular, each token can be described through a couple of parameters (ω_k, p_{σ_k}), where ω_k represents the angle measuring the orientation of the *k-th* token, while p_{σ_k} is an index of the token curvature [10].
From the mathematical point of view, let

$$\varphi(t) = \{x(t), y(t)\} \qquad (1)$$

be the parameterisation of the *k-th* token according to its arc-length $t \in [0 \ldots 1]$. Then, the curvature of the token can be expressed as:

$$\psi(t) = \frac{x'(t)y''(t) - y'(t)x''(t)}{\left(x'^2(t) + y'^2(t)\right)^{3/2}} \qquad (2)$$

Assuming that $\psi(t)$ is a continuous function, a maximum value exists between the two minimum values corresponding to the edge points, enclosing the token itself. This maximum point is used as a curvature index. The parameter ω_k is defined as the orientation, in polar coordinates, of the vector connecting the median point of the token and the point p_{σ_k}, calculated with respect to an absolute reference system.

The considered geometric features reflect more general aspects related to the shape of the target and consist of the perimeter and area measures of the region enclosed by the target contour.

Finally, densitometric features deal with specific statistical measures of the considered image region, and in particular, average brightness, standard deviation, skewness, kurtosis, and entropy.

During target recognition all the features classes are inputs to the HANN system in order to improve the accuracy of the active tracking algorithm, also considering the case of lost or occluded targets.

3.3 HANN system architecture

The target recognition procedure has been realised using a hierarchical architecture of neural networks [8]. In particular, the architecture is composed of two independent network levels each using a specific network typology that can be trained separately to perform its basic task (Fig.3).

The recognition procedure consists of two levels: a) clustering of the different features extracted from the segmented target and b) final recognition based on the input results of the first phase.

The first level (*clustering level*) is composed of a set of classifiers, each corresponding to one of the aforementioned classes of features. These classifiers are based on an unsupervised SOM typology [11] and the training is performed with the aim of classifying the input features into crisp classes that would be representative for the set of possible semantic classes the target belongs to. At the end of the training each network is able to classify the values of the specific feature set. The output of the clustering level is an *m-dimensional* vector consisting in the concatenation of the output of the *m* SOMs (in our case *m*=3). This vector represents the input stimulus of the second level.

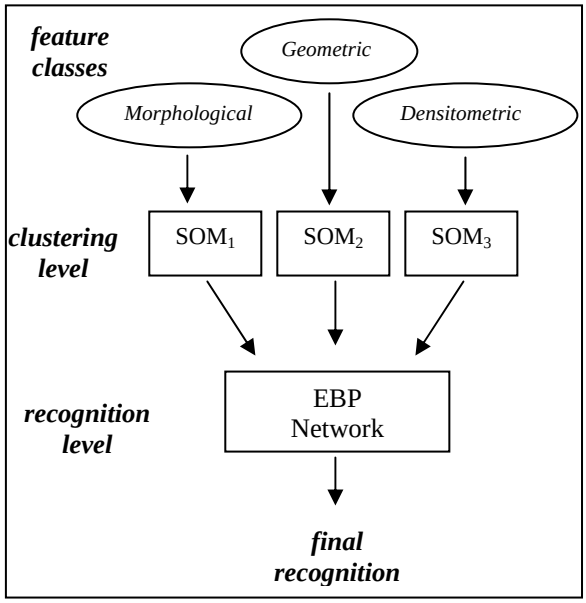

Fig.3. Architecture of the Hierarchical Artificial Neural Network.

The second level (*recognition level*) consists of a neural network classifier based on Error Back-Propagation (EBP) [12]. The input layer is composed of n_I neurons; the output layer is composed of n_O neurons corresponding to the number of semantic classes to be identified (Fig.4). The network has used no hidden layers due to the fact that experimental tests have shown no improvements of either the performance or the quality of results.

Once the EBP network has been trained, it is able to recognize the semantic class that can be associated with the target examined.

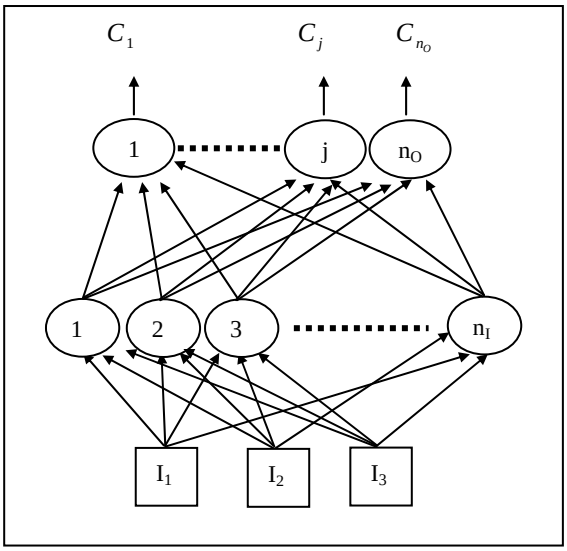

Fig.4. Architecture of the Error Back-Propagation based network.

HANN has a modular architecture that allows the insertion, if necessary, of new sets of features including new information useful for a more accurate recognition. The introduction of new features does not influence the training of the other SOM classifiers and only requires small changes in the recognition level. The modular architecture allows the reduction of local complexity and, at the same time, to implement a flexible system.

3.4 Active Tracking

During the automatic tracking a chain of successive steps is performed in each frame of the video sequence in order to identify and track the target.
Firstly, a candidate characterizing point C_1 is selected in the actual frame. C_1 is internal to the region segmented in the previous frame, and it is selected on the basis of local maximum criterion, considering the brightness value similarity with respect to the characterizing point in the previous frame.

Then, the trajectory of the target in the previous frames is considered, to make the selection more precise. The trajectory of the previous characterizing points is stored and used in the computation of the actual step as a weighted average of the directions and magnitudes, in such a way that another candidate point C_2 is obtained.
Assigning weights α and β (with $\alpha + \beta = 1$) to one point compared to the other, a new candidate point C_3 is computed (3):

$$C_3 = \alpha C_1 + \beta C_2 \qquad (3)$$

Furthermore, in order to guarantee that C_3 belongs to a valid target, a final candidate point C_N is then selected in a circular fixed neighbourhood of C_3, as the one with the closest brightness value to the previous characterizing point.
Starting from C_N, the new edge points, tokens, and contour for the target are finally computed.

Following the selection of the characterizing point in the actual frame, the target characterization is performed extracting the features as mentioned above.

At last, a control is needed to recognize whether the segmented target is the correct one, or not. This control is performed through the HANN, whose inputs are the computed features. The output of the HANN is the class the actual target belongs to; if the class is the same as the one selected during the initial target selection, then a correct tracking was performed and the active tracking starts again with the next acquired frame. Otherwise, if the HANN gives an output which is a different class, then an event of wrong target recognition happens, and the algorithm for the automatic search of the target, trying to forecast its motion, is performed.

The events of wrong target recognition usually correspond either to an occlusion (or partial occlusion or masking) of the target in the scene, or to a quick movement in an unexpected direction. The automatic search tries to resolve the first event.

3.5 Automatic Target Search

This phase is divided into two sub-phases:

1. Search of the characterizing point of the actual target
2. HANN recognition and confirmation.

The first step follows the hypothesis that the occluded target is still moving towards a similar direction as it was previously. Considering the last valid characterizing point C_L, and performing an estimation of the search direction along the interpolated trajectory of the last *n* characterizing points (not counting the actual one supposed to be wrong), a *movement vector M* with direction and average magnitude is obtained:

$$M = \langle Dir, Avg \rangle \qquad (4)$$

Searching along the direction *Dir*, from point C_L, the algorithm selects the first point C_A having brightness value most similar to C_L.

Starting from C_A, the automatic segmentation is performed following the same approach as in the active tracking, and the characterizing features are computed.

To recognize if this target is the one initially selected, the previously described recognition procedure is performed. If the HANN recognizes this target as correct, then the method returns back to the active tracking phase, whereas the search starts again from point 1. acquiring a new frame from the video sequence.
If the correct target is not identified within *j* frames, the control of the system is given back to the user for a new target selection. The value *j* can be estimated hypothesizing that the target, moving along the direction *Dir*, reached the boundary of the frame:

$$j = Dist(C_L; E_{Dir}) / Avg \qquad (5)$$

Where
$Dist(x, y)$ is the Euclidean distance between points *x* and *y*;
E_{Dir} is the point crossing the boundary of the frame along direction *Dir*.

4 Results

The developed approach has been applied to track animal movements in an open environment during the night (fauna monitoring in natural parks), and to environmental video surveillance both in night and daylight (Fig.5).
The dimension of the features vectors which are sent as input to the clustering level, are defined by the dimension of each features class:

- $2 \cdot N$ inputs for the morphological class (i.e. to SOM_1)
- *2* inputs for the geometric class (i.e. to SOM_2)
- *5* inputs for the densitometric class (i.e. to SOM_3)

The value of n_I for EBP was fixed to 15, which resulted to be the better choice among various values tested, and the value of n_O corresponds to the number of different semantic classes existing in each test (e.g. small size animal, medium size animal, human, car, etc.).
The video sequences 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.

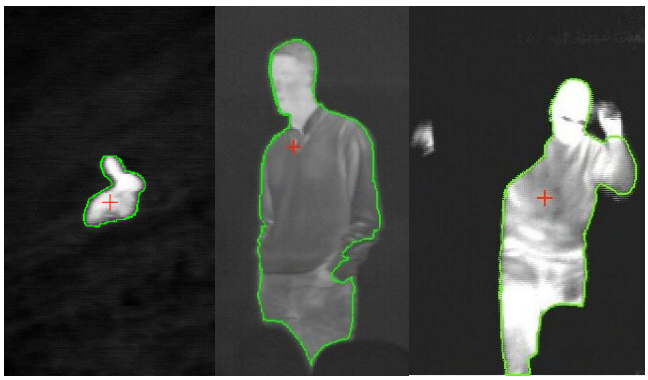

Fig.5. Examples of thermo images acquired on animal (left) and human (centre and right, different cases) targets.

5 Conclusion

A methodology has been proposed for detection and tracking of moving targets in real time video sequences acquired with an IR camera mounted on a robotized system. The problem of real time object tracking has been faced, and robust algorithms performing this task have been implemented. Target recognition during both tracking and searching of a masked or occluded target has been also considered using a Hierarchical Artificial Neural Network, able to recognize whether the target is the correct one or not.

The achieved results have shown to be effective and promising for further improvements, mainly regarding the introduction of new characterizing features, and the enhancement of hardware requirements for quick response to rapid movements of the targets and robustness to very noisy environments.

References:

[1] A. Fernandez-Caballero, J. Mira, M.A. Fernandez, A.E. Delgado, On motion detection through a multi-layer neural network architecture, *Neural Networks*, Vol.16, 2003, pp. 205-222.

[2] B.C. Arrue, A. Ollero, J.R. Martinez de Dios, An Intelligent System for False Alarm Reduction in Infrared Forest-Fire Detection, *IEEE Intelligent Systems*, Vol.15, No.3, 2000, pp. 64-73.

[3] W.G. Yau, L.-C. Fu, D. Liu. Robust Real-time 3D Trajectory Tracking Algorithms for Visual Tracking Using Weak Perspective Projection, *American Control Conference*, 25-27 June, Arlington VA, 2001.

[4] K. Tabb, N. Davey, R. Adams, S. George, The recognition and analysis of animate objects using neural networks and active contour models, *Neurocomputing*, Vol.43, 2002, pp. 145-172.

[5] S. Fejes, L.S. Davis, Detection of Independent Motion Using Directional Motion Estimation, *Computer Vision and Image Understanding*, Vol.74, No.2, 1999, pp. 101-120.

[6] J.B. Kim, H.J Kim, Efficient region-based motion segmentation for a video monitoring system, *Pattern Recognition Letters*, Vol.24, 2003, pp. 113-128.

[7] J. Badenas, M. Bober, F. Pla, Segmenting traffic scenes from grey level and motion information, *Pattern Analysis and Applications*, Vol.4, 2001, pp. 28-38.

[8] S. Di Bona, H. Niemann, G. Pieri., O. Salvetti, Brain volumes characterization using hierarchical neural networks, *Artificial Intelligence in Medicine, Vol.*28, No.3, 2003, pp. 307-322.

[9] M.G. Di Bono, G. Pieri, O. Salvetti, Multimedia Target Tracking through Feature Detection and Database Retrieval. *22nd International Conference on Machine Learning (ICML 2005)*, Bonn, Germany, 7-11 August, 2005.

[10] S. Berretti, A. Del Bimbo, P. Pala, Retrieval by Shape Similarity with Perceptual Distance and Effective Indexing. *IEEE Transactions on Multimedia*, Vol.2, No.4, 2000, pp. 225-239.

[11] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, second ed., vol. 30, Berlin: Springer, 1997.

[12] D.E. Rumelhart, J.L. McClelland, *Parallel distributed processing: explorations in the microstructure of cognition*, vol.1 and 2, MIT Press, Cambridge (MA): Bradford, 1986.