14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

ACTIVE VIDEO-SURVEILLANCE BASED ON STEREO AND INFRARED IMAGING

Gabriele Pieri, and Ovidio Salvetti

Institute of Information Science and Technologies
Via Moruzzi 1, 56124, Pisa, Italy
phone: + (39) 050-3153124, fax: + (39) 050-3152810, email: Ovidio.Salvetti@isti.cnr.it

ABSTRACT

Video-surveillance is a very actual and critical issue at the present time. Within this topic we address the problem of firstly identifying moving people in a scene through motion detection techniques, and subsequently categorising them in order to identify humans for tracking their movements. The use of stereo cameras, coupled with infrared vision, allows to apply this technique to images acquired through different and variable condition, and allows an a priori filtering based on the characteristics of such images to give evidence to objects emitting an higher radiance (i.e. higher temperature).

1. INTRODUCTION

Recognizing and tracking moving people in video sequences is generally a very challenging task and automatic tools to identify and follow a human – *target* – are often subject to constraints regarding the environment under investigation, the characteristics of the target itself and its full visibility with respect to the background.

Current approaches regarding real-time target tracking are based on (i) successive frame differences [1], using also adaptive threshold techniques [2] (ii) trajectory tracking, using weak perspective and optical flow [3], (iii) region approaches, using active contours of the target and neural networks for movement analysis [4], or motion detection and successive regions segmentation [5]. In recent years, thanks to the improvement of infrared (IR) technology and the drop of its cost, also thermal infrared imagery has been widely used in tracking applications [6, 7]. Besides, the fusion of visible and infrared imagery is starting to be explored as a way to improve the tracking performance [8].

Regarding specific approaches for human tracking, frame difference, local density maxima, and human shape models are used in [9] for tracking in crowded scenes, while face and head tracking by means of appearance-based methods and background subtraction are used in [10].

In this paper, the problem of detecting a moving target, and its tracking is faced by processing multi-source information acquired using a vision system capable of stereo and IR vision. Combining the two acquisition modalities assures different advantages consisting, first of all, of an improvement of target detection capability and robustness, guaranteed by the strength of both media as complementary vision modalities. Infrared vision is a fundamental aid when low-lighting conditions occur or the target has similar colour to the background. Moreover, as a detection of the thermal radiation of the target, the IR information can be manageably acquired on a 24-hour basis, under suitable conditions. On the other hand, the visible imagery has a higher resolution and can supply more detailed information about target geometry and localization with respect to the background.

The acquired multi-source information is firstly elaborated for detecting and extracting the target in the current frame of the video sequence. Then, the tracking task is carried on using two different computational approaches. A Hierarchical Artificial Neural Network (HANN) is used during active tracking for the recognition of the actual target; while, when the target is lost or occluded, a content-based retrieval (CBR) paradigm is applied on an a priori defined database to relocalize the correct target.

In the following sections, we describe our approach, demonstrating its effectiveness in a real case study: the surveillance of known scenes for unauthorized access control [11, 12].

2. PROBLEM FORMULATION

We face the problem of tracking a moving target distinguishable from a surrounding environment owing to a difference of temperature. In particular, we consider to overcome lighting and environmental condition variation using IR sensors.

Humans tracking in a video sequence consists of two correlated phases: *target spatial localization*, for individuating the target in the current frame, and *target recognition*, for determining whether the identified target is the one to be followed.

Spatial localization can be sub-divided into *detection* and *characterization*; while recognition is performed for an *active tracking* of the target, frame by frame, or for relocalizing it, by means of an *automatic target search* procedure.

The initialization step is performed using an automatic motion detection procedure. A moving target appearing in the scene under investigation is detected and localized using the IR camera characteristics. Then, the identified target is extracted from the scene by a rough segmentation.

Once segmented, the target is described through a set of meaningful multimodal features, belonging to *morphological*, *geometric* and *thermographic* classes, computed to obtain useful information on shape and thermal properties.

To cope with the uncertainty of the localization, increased by partial occlusions or masking, an HANN can be designed to process the set of features during an active tracking proce-

dure, in order to recognize the correctness of the detected target.

In case the HANN does not recognize the target, a wrong object recognition should happen, due to either a masking, partial occlusion of the person in the scene, or a quick movement in an unexpected direction. In this circumstance, the localization of the target is performed by an automatic search, supported by the CBR on a reference database. In this process, only a limited number of frames is considered and, if problems arise, the control is given back to the user.

The general algorithm implementing the above described approach is shown in Figure 1 and it regards its *on-line* processing. In this case, the system is used in real time to perform the tracking task. Extracted features from the selected target drive active tracking with HANN and support the CBR to resolve the queries to the database in case of lost target. Before this stage, an off-line phase is necessary, where known and selected examples are presented to the system so that the neural network can be trained, and all the extracted multimodal features can be stored in the database, which is organised on the basis of the defined semantic classes. For each defined target class, possible variations of the initial shape are also recorded, for taking into account that the target could be still partially masked or have a different orientation.

More details of the algorithm are described in the following.

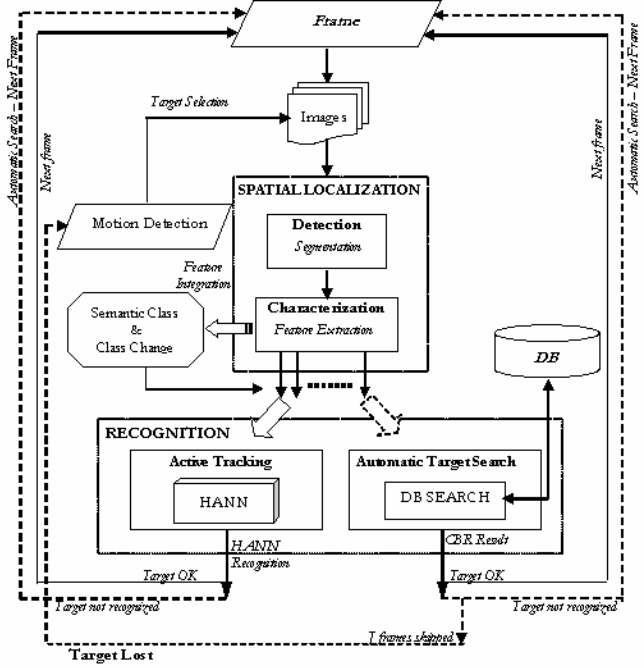

Figure 1 – Automatic tracking algorithm

3. TARGET SPATIAL LOCALIZATION

3.1 Target Detection

After the tracking procedure is started, a target is localized and segmented using the automatic motion detection procedure, and a reference point, called *centroid* C_0, internal to it is selected (e.g. the centre of mass can be used for the first step). This point is used in the successive steps, during the automatic detection, to represent the target. In particular, starting from C_0, a motion prediction algorithm has been defined to localize the target centroid in each frame of the video sequence. According to previous movements of the target, the current expected position is individuated, and then refined through a neighbourhood search, performed on the basis of temperature similarity criteria.

Let us consider the IR image sequence $\{F_i\}_{i=0,1,2,\ldots}$, corresponding to the set of frames of a video, where $F_i(p)$ is the thermal value associated to the pixel p in the i-th frame. The trajectory followed by the target, till the i-th frame, $i>0$, can be represented as the centroids succession $\{C_j\}_{j=0,\ldots,i-1}$. The prediction algorithm for determining the centroid C_i in the current frame can be described as shown in Figure 2.

Function *Prediction*(i, $\{F_i\}_{i=0,1,2,\ldots}$, s, fps);

//* *Compute the number of frames in* s *seconds*
$n = s*fps$;
//* *Control the movement of the target*
if $\|C_{i-n} - C_{i-1}\| > Threshold_1$
then
//* *Compute the expected target position* P_i^1 *in the current frame by interpolating the last n centroid positions*
P_i^1=INTERPOLATE($\{C_j\}_{j=i-n,\ldots,i-1}$);
//* *Compute the average length of the movements*

$$d = \left(\sum_{j=i-n}^{i-2} \|C_j - C_{j+1}\| \right) / i-1 ;$$

//* *Compute a new point on the basis of temperature similarity criteria in a circular neighbourhood* Θ_d *of* P_i^1 *of radius d*

$$P_i^2 = \arg\min_{P \in \Theta_d} [F_i(P) - F_{i-1}(C_{i-1})] ;$$

If $\|P_i^1 - P_i^2\| > Threshold_2$
then $P_i^3 = \alpha P_i^1 + \beta P_i^2$; //*where $\alpha + \beta = 1$

//* *Compute the final point in a circular neighbourhood* N_r *of* P_i^3 *of radius r*

$$C_i = \arg\min_{P \in N_i} [F_i(P) - F_{i-1}(P_{i-1})] ;$$

else $C_i = P_i^2$;
else //* *compute the new centroid according to temperature similarity in a circular neighbourhood* N_l *of the last centroid*

$$C_i = \arg\min_{P \in N_l} [F_i(P) - F_{i-1}(C_{i-1})]$$

Return C_i

Figure 2 – Prediction algorithm used to compute the candidate centroid in a frame

The coordinate of centroids referring to the last s seconds (s < 2sec, s dependent on the semantic class) are interpolated for detecting the expected position P_i^1. Then, in a circular neighbourhood of P_i^1 of radius equal to the average movement amplitude, an additional point P_i^2 is detected as the point having the maximum similarity with the centroid C_{i-1} of the previous frame. If $\|P_i^2 - P_i^1\| > Threshold_2$, then a new point P_i^3 is calculated as a linear combination of the previous determined ones. Finally, a local maximum search is again performed in the neighbourhood of P_i^3 to make sure that it is internal to a valid object. This search finds the point C_i that has the thermal level closest to the one of C_{i-1}:

Starting from the current centroid C_i, an automated edge segmentation of the target is performed, using a gradient descent along 16 directions, starting from C_i. Figure 3 shows

a sketch of the segmentation procedure and an example of its result.

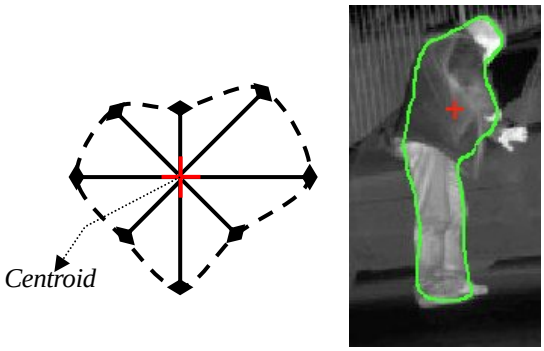

Figure 3 – Example of gradient descent procedure to segment a target (left) and its application to an example frame identifying a person (right)

3.2 Target Characterization

Once the target has been segmented, multi-source information is extracted in order to obtain a target description. This is made through a feature extraction process performed on the three different images available for each frame in the sequence. The sequence of images is composed of both grey levels images (i.e. frames or thermographs) of high temperature target (with respect to the rest of the scene) integrated with grey level images obtained through a reconstruction process [13].

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. There are four different groups of visual features which are extracted from the region enclosed by the target contour defined by the sequence of N_c (i.e in our case N_c =16) points having coordinates $<x_i, y_i>$:

- *Morphological*: shape contour descriptors;

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 [14].

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.

- *Geometric*:

$$\text{Area} = \left| \sum_{i=1}^{N_c} \left[(x_i y_{i+1}) - (y_i x_{i+1}) \right] \right| / 2$$

$$\text{Perimeter} = \sum_{i=1}^{N_c} \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2}$$

- *Thermographic*:

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

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

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

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

Entropy: $E = - \sum_{p \in \text{Target}} F_i(p) \log_2 (F_i(x, y))$

where μ_r are moments of order *r*

The s*emantic class* the target belongs to (i.e. upstanding, crouched or crawling person) 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.

All the extracted information is passed to the recognition phase, in order to assess if the localized target is correct.

3.3 Target Recognition

The target recognition procedure is realised using a hierarchical architecture of neural networks. In particular, the architecture is composed of two independent network levels each using a specific network typology that can be trained separately.

The first level focuses on clustering the different features extracted from the segmented target; the second level performs the final recognition, on the basis of the results of the previous one

The *clustering level* is composed of a set of classifiers, each corresponding to one of the aforementioned classes of features. These classifiers are based on unsupervised *Self Organizing Maps* (SOM) and the training is performed to cluster the input features into classes representative of the possible target semantic classes. 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 of the concatenation of the *m* SOMs outputs (in our case *m*=3). This vector represents the input of the second level.

The *recognition level* consists of a neural network classifier based on Error Back-Propagation (EBP). Once trained, such network is able to recognize the semantic class that can be associated to the examined target. If the semantic class is correct, as specified by the user, the detected target is recognized and the procedure goes on with the active tracking. Otherwise, a wrong target recognition occurs and the automatic target search is applied to the successive frame, in order to find the correct target.

3.4 Automatic Target Search

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

The multi-modal features of the candidate target are compared to the ones recorded in a reference database. A similarity function is applied for each feature class [15]. In particular, we considered *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 database information.

For each semantic class, possible variations of the initial shape are recorded. In particular, the shapes to compare with are retrieved in the MM 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 distance, from at least one in the obtained set, below a fixed tolerance threshold, then it can be considered valid. Otherwise the search starts again in the next frame acquired [11].

Furthermore, the information related to a semantic class change is used as a weight for possible candidate targets; this is done considering that a transition from a semantic class SC_b to another class SC_a has a specific meaning (e.g. a person who was standing before and is crouched in the next frames) in the context of a surveillance task, which is different from other class changes.

The features of the candidate target are extracted from a new candidate centroid, which is computed starting from the last valid one (C_v). From C_v, considering the trajectory of the target, the same algorithm as in the target detection step is applied so that a candidate centroid C_i in the current frame is found and a candidate target is segmented.

With respect to the actual feature vector, if the most similar pattern found in the database has a similarity degree higher than a prefixed threshold, then the automatic search has success and the target tracking for the next frame is performed through the active tracking. Otherwise in the next frame the automatic search is performed again, still considering the last valid centroid C_v 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 C_v 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 $\{C_j\}_{j=0,\ldots,v}$ (Eq. 3)

$$j = \left\| C_v - E_r \right\| / \left(\sum_{j=v-n}^{v-1} \left\| C_j - C_{j+1} \right\| / f \right) \qquad (3)$$

4. RESULTS

The method implemented has been applied to a real case study for video surveillance to control unauthorized access in restricted access areas.

Due to the nature of the targets to which the tracking has been applied, using IR technology is fundamental. The temperature that characterizes humans has been exploited to enhance the contrast of significant targets with respect to a surrounding background.

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 were positioned in order to explore a scene 100 meters far, sufficient in our experimental cases. The frame acquisition rate ranged from 5 to 15 fps.

In the video-surveillance experimental case, during the off-line stage, the database 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, middle, tall).

A set of surveillance videos were taken during night time and positioned in specific areas, such as a closed parking lot, and an access gate to a restricted area, for testing the efficiency of the algorithms.

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.

Examples of persons tracking and class identification are shown in Figures 4 and 5.

The acquired images are pre-processed to reduce the noise.

5. CONCLUSION

A methodology has been proposed for detection and tracking of moving people in real time video sequences acquired with

two stereo visible cameras and an IR camera mounted on a robotized system.

Target recognition during active tracking has been performed, using an *Hierarchical Artificial Neural Network* (HANN). The HANN system has a modular architecture which allows the introduction 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.

In case of automatic searching of a masked or occluded target, a *Content-Based Retrieval* paradigm has been used for the retrieval and comparison of the currently extracted features with the previously stored in a reference database.

The achieved results are promising for further improvements as the introduction of additional new characterizing features and enhancement of hardware requirements for a quick response to rapid movements of the targets.

Figure 4 – Example of an identified and segmented person during video-surveillance on a gate

Figure 5 – Example of an identified and segmented person during video-surveillance in a parking lot

REFERENCES

1. A. Fernandez-Caballero, J. Mira, M.A. Fernandez, A.E. Delgado, "On motion detection through a multilayer neural network architecture" *Neural Networks*, 16, pp. 205–222, 2003.
2. S. Fejes, L.S. Davis, "Detection of Independent Motion Using Directional Motion Estimation" *Computer Vision and Image Understanding*, Vol. 74 (2), pp. 101–120, 1999.
3. W.G. Yau, L.-C. Fu, D. Liu, "Robust Real-time 3D Trajectory Tracking Algorithms for Visual Tracking Using Weak Perspective Projection" *in Proc. of the American Control Conference*, 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, pp. 145–172, 2002.
5. J.B. Kim, H.J. Kim, "Efficient region-based motion segmentation for a video monitoring system" *Pattern Recognition Letters*, Vol. 24, pp. 113–128, 2003.
6. M. Yasuno, N. Yasuda, M. Aoki. "Pedestrian Detection and Tracking in Far Infrared Images" *in Proc. of the Conference on Computer Vision and Pattern Recognition Workshop*, June 2004, pp.125–131.
7. J. Zhou, J. Hoang, "Real Time Robust Human Detection and Tracking System" *in Proc. of the 2nd Joint IEEE International Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum*, San Diego, CA, USA, June 20 2005.
8. B. Bhanu, X. Zou, "Moving humans detection based on multi-modal sensory fusion" *in Proc. IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS'04)*, pp. 101–108, July 2004.
9. C. Beleznai, B. Fruhstuck, H. Bischof, "Human tracking by mode seeking" *in Proc. of the 4th Int. Symp. On Image and Signal Processing and Analysis (ISPA 2005)*, 15-17 September 2005, pp. 1–6.
10. A. Utsumi, N. Tetsutani, "Human tracking using multiple-camera-based head appearance modeling" *in Proc. Sixth IEEE Int. Conference on Automatic Face and Gesture Recognition (FGR'04)*, 17-19 May 2004, pp. 657–662.
11. M.G. Di Bono, G. Pieri, O. Salvetti, "Multimedia Target Tracking through Feature Detection and Database Retrieval" *in Proc. of the 22nd ICML*, Bonn, Germany, 11th August 2005, pp. 19–22.
12. S. Colantonio, M.G. Di Bono, G. Pieri, O. Salvetti, M. Benvenuti, "Object tracking in a stereo and infrared vision system" *in Proc. of the 8th AITA Conference*, Rome, Italy, 7-9 September, 2005, pp. 113.
13. M. Sohail, A. Gilgiti, T. Rahman, "Ultrasonic and stereo vision data fusion" *in Proc. of 8th International Multitopic Conference INMIC 2004*, Lahore, Pakistan, 24-26 December 2004, pp. 357–361.
14. 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, pp. 225–239, 2000.
15. P. Tzouveli, G. Andreou, G. Tsechpenakis, Y. Avrithis, S. Kollias, "Intelligent Visual Descriptor Extraction from Video Sequences", *Lecture Notes in Computer Science – Adaptive Multimedia Retrieval*, Vol. 3094, pp. 132–146, 2004.