第 29 卷 第 3 期
2003 年 5 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 29, No. 3
May, 2003

A Cooperative Multisensor System for Face Detection in Video Surveillance Applications

Luca Marchesotti Alessandro Messina Lucio Marcenaro Carlo Regazzoni
(*DIBE-University of Genoa, Genoa, Italy*)
(E-mail: carlo@dibe.unige.it)

Abstract This paper presents an innovative architecture for multisensor video surveillance characterized by two pan-tilt video cameras. The aim of the system is to track and to characterize moving objects in outdoor environments in real time, with a robust behaviour. In particular, the system here presented exhibits the capacity of extracting video shots by means of a mobile camera operating at multiple zoom levels. The mobile camera collects successive frames of face oval, segments them and track them over time automatically, starting from a cooperative initialisation of a static widefield camera performing automatic object tracking and classification.

Key words Multisensor video surveillance, multiple zoom levels, object tracking

1 Introduction

A surveillance system can be defined as a technological tool that provides humans with an extended perception and reasoning capability about events of interest[1]. Typically, human operators have always been given the tasks of monitoring of complex environments. The actual trend is to develop systems with enhanced capabilities, in terms of "extended" perception[2,3] on the monitored environment and on activities which take place in that environment.

This paper explores this last objective proposing a multicamera system[4,5] able to successfully locate, detect and track human faces in complex, outdoor environments as well as to collect facial video shots. A typical structure for a video surveillance system[6] based on static cameras involves a module for the image acquisition from the sensors, a change detection module that outputs a difference image[7] (computed pixel by pixel by subtracting the present image with the background image), an object localization estimator capable of identifying the position of moving objects and finally, an object tracker[8~11] that has the duty to keep track of the detected object until they get away from the monitored environment. To perform this, the system uses features provided by the lower levels which enable it to estimate 3-D position[12] and dimensions of objects tracked on the scene. The next step is to classify the objects[13~15] accordingly to their features such as shape attributes and to focus the attention of the system on people faces. To this end various techniques have been used for classification purposes in order to distinguish human walking in a car park from cars.

Systems which reach this level of analysis can be qualified as second-generation video surveillance systems whereas first-generation architectures are characterized by humans operators delegated to visually process data coming from sensors. The difference with the previous systems can be found exploiting a deeper analysis of the pre-processed data. Once moving objects in the scene are successfully detected and tracked, information such as position, trajectories and colour can be used to make assumptions on objects behaviour[16,17]. In addition, techniques specifically designed for face detection[18] in still images, based on non parametric[19,20] and parametric models[21,22], can be inserted in the architecture of vid-

Received December 23, 2002; in revised form April 16, 2003
收稿日期 2002-12-23; 收修改稿日期 2003-04-16

eo surveillance systems in order to collect video shots of each person. This will lead to an enhanced description of the monitored scene with both symbolic data and biometric data. In this way, in such architecture the operator is lightened from the burden of discerning between standard or abnormal situations.

2 System Overview

The purpose of the presented system is to track moving objects in real-time in a variety of different conditions with the aim of focusing the system's view to detect humans' face.

To perform this task, a multisensor approach has been chosen, composed of two pan-tilt video cameras. The proposed architecture can be divided into several distributed modules, specifically devoted to well defined tasks which belong to three different layers:

- Sensor Layer: calibration of sensors, positioning of cameras, collection of video data
- Image Processing Layer: video analisys and manipulation
- Data Fusion Layer : Interaction strategies, multicamera calibration and networking

To clarify tasks accomplished in each layer the system is described in terms of Physical and Logical architecture.

2.1 Physical Architecture

Physical architecture, outlined in Fig. 1, is mainly characterized by two sensors (video cameras) which are managed by two computational units (PC1/2) which are standard Pentium based PCs. The first camera (Cam1) is a CCD pan-tilt "Sony Evi-D31" which acts as static sensor. The camera zoom is set to lowest value in order to get a wide field view of the monitored scene. The second camera is a "Philips Envirodome G3" that is the active sensor in the system remotely controlled via RS-232. A client/server approach has then been used to enable the cooperation of the two sensors with standard TCP/IP communication channels. To do this PC1 and PC 2 are connected in a Local Area Network with wireless LAN 802.11 facilities in order to be able to freely place the sensors in the most appropriate position. In our case both cameras watch a car park with few entrances.

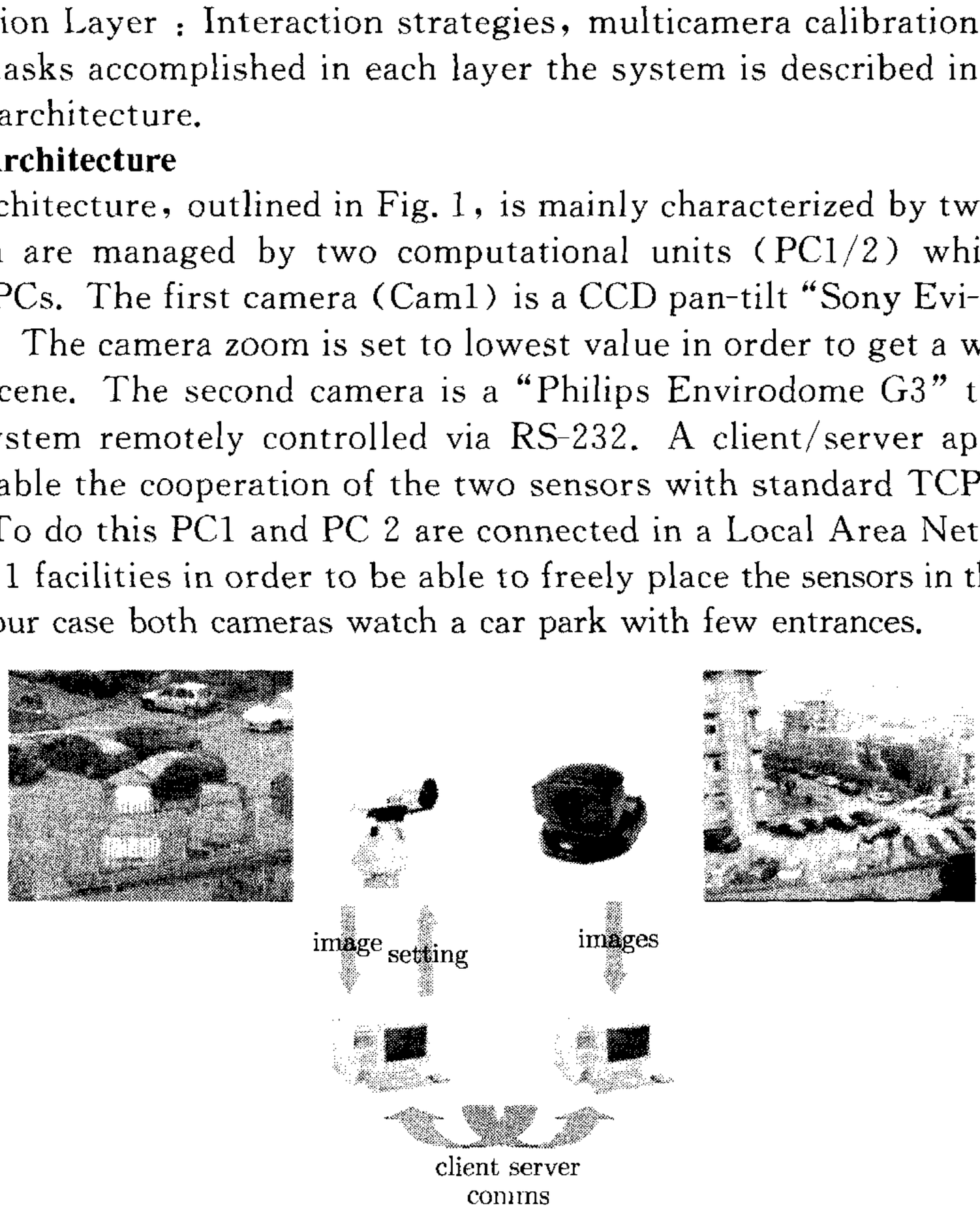


Fig. 1 Physical architecture of the system

2.2 Logical Architecture

The logical architecture is composed of different modules which can be mapped into the three layers previously outlined. The hierarchy of modules drawn in Fig. 2 represents a configuration of the system with respect to the two PCs and the combination of static and dynamic sensors.

As stated before, Cam1 is meant to statically monitor, with a wide field view, the

scene of interest; therefore modules such as "Change Detection", "Blob Extraction" and "Localization" can be seen as a typical configuration for a Video Surveillance system able to detect, follow and classify moving regions. The cluster of these modules represents the core part of the Image Processing Layer. To this end, the system can give under the form of "metadata", positioning information regarding objects which "populate" the guarded environment. Once this information gets the Sensor Layer, the "Localization" module outputs positions in terms of (x, y) image coordinates. In this case the reference image (X,Y plane) is represented by Fig. 3, that is a sample image grabbed by Cam1. After a moving region has been successfully detected and placed in a 3-D world coordinates, via camera calibration, the region has to be classified with respect to two main categories: "human" or "others". In the latter category, all non-rigid objects such as vehicles are included. Once the system knows the nature of the moving region, its location and, implicitly its speed vector (from trajectories), a cooperation strategy takes place into the Data Fusion Layer and the two computational units (PC1-2) begin to dialog and positioning the two sensors in turn. As it can be seen in Fig. 2 all these tasks directly depend on a logical unit called "Sensor Handler" which will be explained later. In particular, PC1 sends to PC2 transformed coordinates of the human, and, as a response, Cam2 points itself to that location with appropriately high zoom value. At this point, the system, through the "Face Detector" module extracts the face "oval" of the human and tries to evaluate, with a good degree of precision, its location and some biometric data such as face mass center and its area. Then with the "Face Tracker", system follows face over time. Due to the dynamic nature of the "target", the face of a human, a static tracker working only on high zoom images represents an unfeasible approach because the face could be followed just for few frames. For this reason, if the face is predicted to be lost soon, the camera moves following it, as an "active" mechanical tracker. Once the camera has been repositioned, it loops on normal "software" tracker. This approach is continued until a sufficiently large set of video frames representative of multiple poses is collected. Then the mobile camera is ready to start again this operation with a new human.

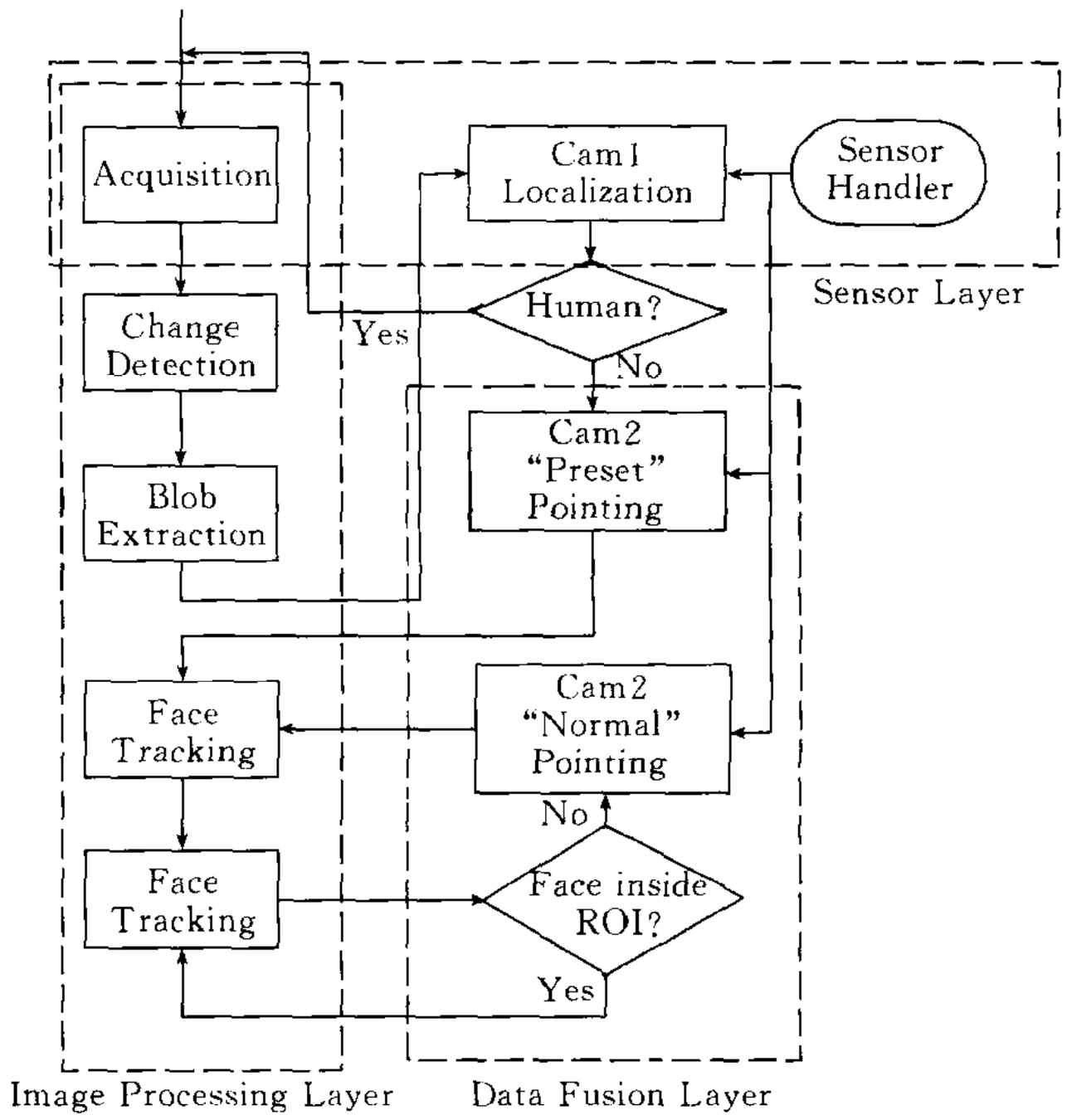


Fig. 2 Logical architecture of the system

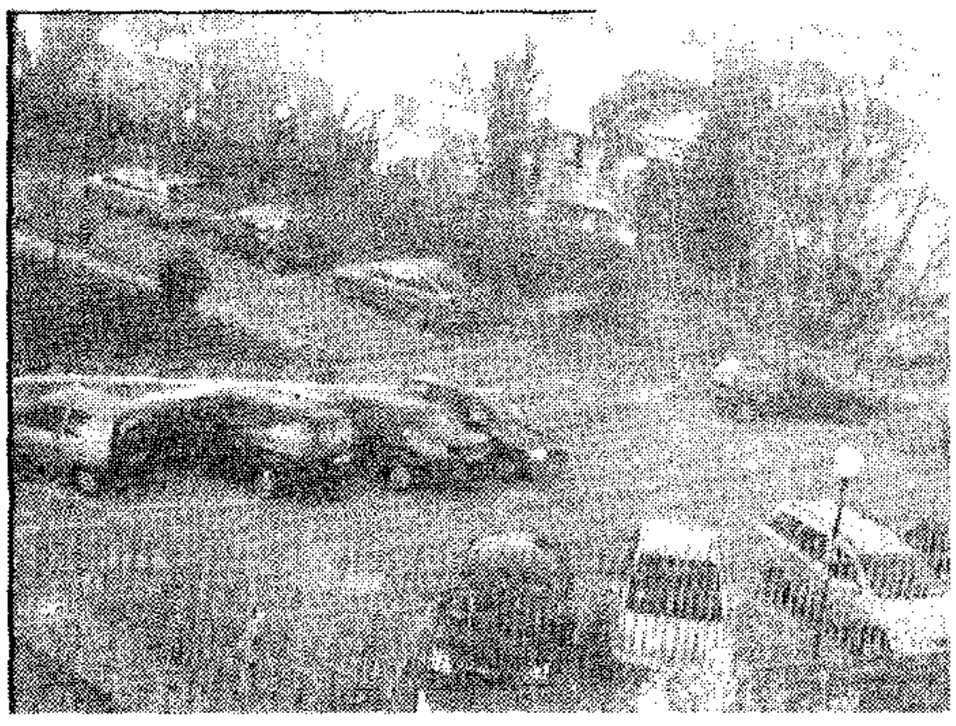


Fig. 3 Output of Cam1, wide field image

2.3 Sensors Control Handler

The Sensor Control handler is a dedicated module that has the duty to directly control

the two cameras and information flow between them. For this reason the "handler" can be seen as an interface between the Sensor Layer of the system and the Image processing Layer. In fact, starting from the positioning of the sensors, it has to provide the higher levels (IP levels) with images acquired from sensors and eventually negotiate future positions among them. In particular with respect to Fig. 2 the handler has to:

• position the sensor that is used statically (Camera 1) in the preset pan-tilt coordinates which enable it to have a wide field view on the scene. This position will give as output for the whole functioning of the system, the Fig. 3 image.

• position the active sensor (Camera 2) to given pan-tilt coordinates and zoom level.

• Synchronize requests of positioning between Cameras 1～2.

• Collect video data from Cameras 1～2.

The first action is typically performed when the system boots up and the sensors have to be positioned in the predefined locations. This step has to be accurately performed because camera calibration (discussed in the following paragraph) is unique for each camera view. Action 2 is taken in real-time during normal functioning of the system when an object is detected and there's the need to focalise the "attention" of the system on it. In our case, given a moving object, main target is represented by the face of the detected person. Action 3 is performed in order to preserve the sensors from timing errors; video cameras are controlled via serial port using RS-232 standard and camera-dependant commands are sent to them. Problems arise when command rate is above certain threshold and the sensor enters a fault state. Considering that requests for positioning are generally asynchronous and randomly distributed over time, an entity to manage this situation is therefore needed. For this reason Control Handler can receive request for commands, it schedules them and after having performed some buffering activities it sends them to the appropriate sensor with the correct rate.

2.4 Camera Calibration

Camera calibration is the process by which optical and geometrical features of cameras can be determined. Generally, these features are addressed as intrinsic and extrinsic parameters and they allow to estimate a correspondence between coordinates in the Image Plane and in Real Word. Calibration of sensors is generally used to deduce 3-D information from 2-D data, but in some cases can be applied to get 2-D image coordinates from 3-D data. Camera calibration we use is based on classic Tsai method[12]. In the presented system, both sensors have been calibrated because they need to intensively exchange position information of moving objects among them. Therefore, a common calibration strategy has been evaluated. In Fig. 5 the chosen approach is outlined. First of all Cam1 and Cam2 are calibrated reffeving to images in Fig. 3 and Fig. 4. Then a common reference point has to be found in order to make the system able to swich between the two reference systems.

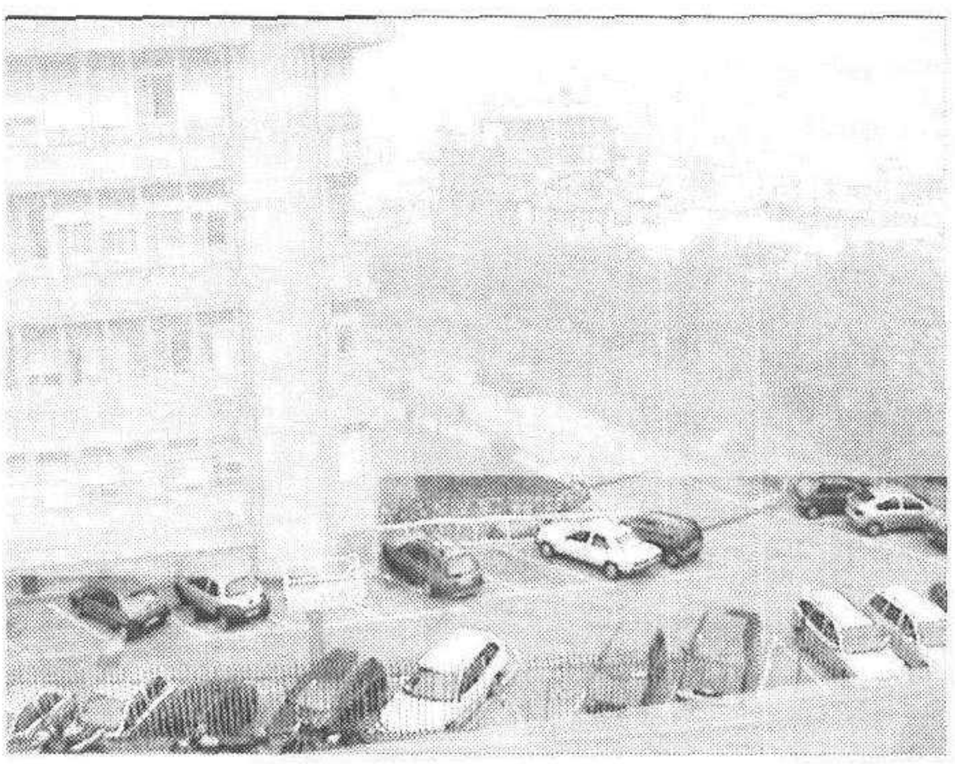


Fig. 4 Output of Cam2 in preset position

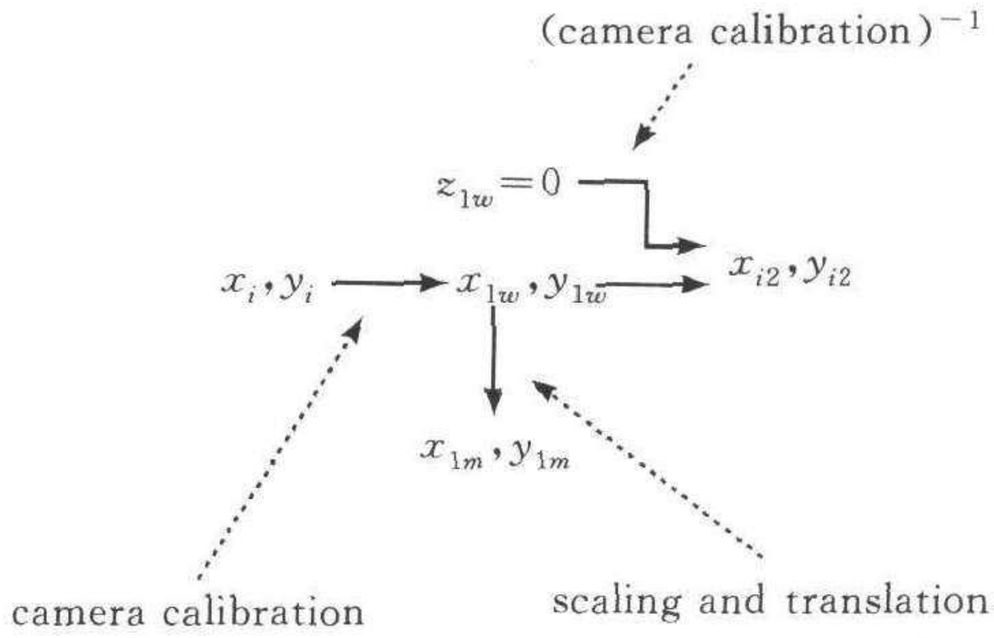


Fig. 5 Steps involved in coordinates conversion with camera calibration

Origin of world coordinates represents a good choice because it is common to the two cameras. Therefore steps involved in the conversion of coordinates from Cam1 to Cam2 are the ones sketched in Fig. 5. Image Coordinates for Cam1 (x_{i1}, y_{i1}) are converted in World Coordinates (x_w, y_w). Then (x_w, y_w) have to be placed in a 2-D map (x_m, y_m) with the following transformation: $x_m = \frac{x_w}{s} + \Delta x$, $y_m = \frac{y_w}{s} + \Delta y$, with $\Delta x, \Delta y$ shift values used to traslate coordinates whereas s is a scale factor to convert "World Coordinates" from *mm* to *pel*. The next step is to convert (x_w, y_w) to Image Coordinates for Cam2. To do this z_w is set to zero because the assumption of projection on the ground-plane has been made, then with Cam2 intrinsic and extrinsic parameters coordinates (x_{i2}, y_{i2}) can be calculated.

3 Sensor positioning

To focus the attention of the system and to give higher detection rates, optimal views of people present in the scene have to be collected. For this reason position of the active sensor has to be modified and optimised. The aim is to follow and track faces of detected people over time.

To do this, Cam2 (active camera) can be controlled in two modalities:

1) Target Pointing(P1)

2) Tracking Pointing(P2)

The two modalities are needed because the general problem of tracking faces can be decomposed on two sub-problems, the first is concerned with the detection and the localization of humans in the wide field camera (cam1) and the second consists on succesfully tracking humans' faces over time in the high resolution (narrow-field image). During normal functioning, the system correclty detects and classifies humans in cam1 and it passes their positions to PC2 which has to point cam2 on the given human's oval (P1). Then cam2 position has to be "smoothly" updated (P2) in order to follow face over time. To achieve this, the two modes require different mechanical "behaviours" of camera; in P1, camera has to be as fast as possible not to loose the target (human), whereas in P2 precision is what really matters. To solve this, as active camera, a "Philips Envirodome G3" CDD-video camera has been used; this sensor is characterized by a two operational modes, a "normal mode" and a "preset mode". In preset mode the camera positions itself on the coordinates previously stored in its hardware memory whereas in normal mode it moves step by step according to command messages passed via RS-232. The camera, as it can be seen in table 1, appears to be consistently faster in preset mode.

Table 1 Performances of Preset/Normal mode

Mode	Pan Speed(°/s)	Tilt Speed(°/s)
Normal	120	120
Preset	360	360

Considering that the total time for pointing has to be shorten as much as possible because the risk of loosing the target is high, preset mode comes in help to handle pointing mode situations. To avoid this problem, prediction methods have been used in the past in order to estimate future positions of moving targets using Kalman-filter based techniques. This approach is generally functional to the problem, however it is an overhead in computation and it becomes quite unstable dealing with rapid movements.

For this reason, in the proposed architecture preset and normal functioning modalities have been allocated as follows:

Target Pointing→	Preset Mode +(Normal Mode)
Tracking Pointing→	Normal Mode

Accordingly to logical chain outlined in Fig. 2, once a human is detected in position (x_w, y_w, z_w) "Sensor Handler" translates to $v=(x_i, y_j)$ and Cam 2 is redirected in "Target Pointing" to v. Given preset p_k defined as (x_k^p, y_k^p) with $1<k<K$ (K=total number of presets) and a target specified by (x_i, y_j) with x_k^p, y_k^p and x_i, y_j belonging to wide field image, the system has to perform the following steps:

1)Pointing camera to the most appropriate preset p_k(Preset Mode)

2)If $(x_k^p, y_k^p)\neq(x_i, y_j)$ perform a "fine tuning" of camera position (Tracking Mode)

Of course it is preferable to avoid step 2 because in "Tracking Mode"camera is slower; this implies accurate allocation of presets on Image Plane (X, Y). The allocation can be performed with a Homogeneous or Non-Homogeneous distribution in the Image Plane. A homogeneous distribution imposes to allocate the total number of presets (K) on the image plane. Therefore, given an $H\times W$ image and a total number of presets equal to N, a homogeneous allocation creates a grid of presets with :

$$d_r=\frac{H}{R},\quad d_c=\frac{W}{C}\quad \text{with:}$$

$R\cdot C\leqslant K$, d_r, d_c = displacement between rows and columns

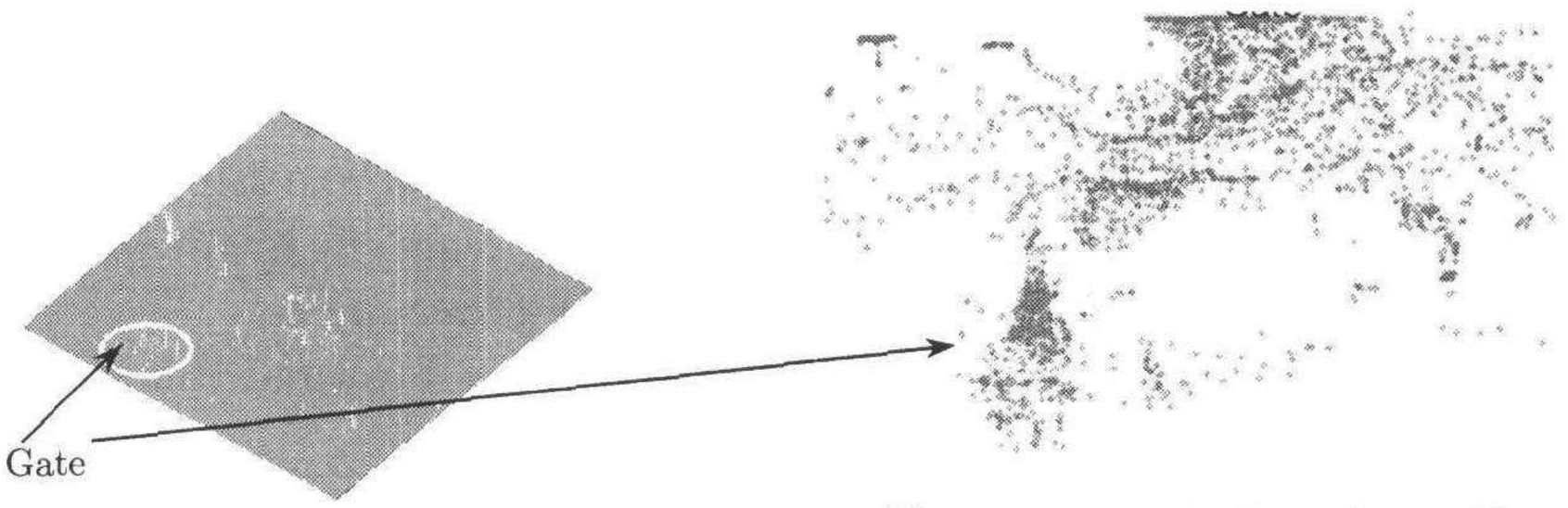
$\frac{H}{W}=\frac{R}{C}$, R, C = total number of Rows and Columns

To achieve a given horizontal and vertical displacement d_c, d_r between presets, i. e. : $d_r=d_c=10$ in an Image Plane of 352×288, the total number of required presets K is approximately equal to 980, too much for a typical commercial camera. For this reason, non homogeneous distribution can be used; in this case, pointing positions (x_k^p, y_k^p) are not equally spread into the whole image. Pointing positions generally represent location of human. This assumption dramatically restricts the region which has to be covered with presets. In fact, as it can be seen in Fig. 4, shaded regions are not taken into account because they are not allowed locations for people. For this reason, presets have to be allocated elsewhere in the remaining part of the image. The criteria followed in the assignment of presets is based on test sessions carried out to evaluate common trajectories of human. Tests have been performed on recording positioning data of people walking in the car park with cam1 for an approximate period of 2 hours. In Figs. 6(a),(b) the histogram distribution of people typical positions are reported with respect to the Image Plane of Camera 2:

$$f(v_r)=\sum_{p=1}^{P}\delta[v_p-v_q]$$

With: v_p=position of detected human (target), v_q=position in the image, P=total number of positions of detected humans.

As it can be seen from the projection of the 3-D distribution on the image plane (Fig. 6(b)) distribution concentrates on two main areas, one at the top of graph and a second



(a) 2-D Trajectory distribution (b) Trajectory projection in Image Plane

Fig. 6 Histogram distribution of people typical positions

that has been highlighted with a circle. This location in the graph corresponds to the gate that can be seen at the bottom-center of Fig. 6(b). This is reasonable because people usually transit on the car park and pass through the gate. Given the distribution, the allocation of the presets can be more dense in the gate region and also in the center of the car park where a marginal but not neglecting number of trajectories is found.

4 Face Detection

Face detection in the context of the proposed architecture, represents the final step of the logical chain represented in Fig. 2. The general aim of FD is formalized in [18] as the task of determining whether or not there are any faces in a given image, and if present return the image location and extent of each face. What we intend here for face detection is strongly influenced by contextual information that enables us to simplify the task of pure detection; in fact in our case face detection reduces to face "localisation" because we can assume the presence of a face in the given image. In fact FD is performed on images coming from Cam2 after it has been pointed to a human target location as it is automatically classified by the static camera. To perform FD authors of [18] state that more than 150 different approaches are available and he cathegorizes them in 4 different groups: Knoledge-Based methods[23], Feature Invariant approaches[24], Template Matching methods[25] and Appearance Based methods[26].

For the presented architecture, the chosen approach exploits colour because it guarantees a real-time functioning being computationally light.

4.1 The model

Several techniques have developed exploiting different colour spaces; some of them exploit a neural based approach, very reliable but also non-real time. The model we use is based on the work outlined[1]. This model uses skin colours by calculating normalized RGB components. RGB chromatic domain represents not only colour but also intensity in a given image that has shown to be more effective in the discrimination between skin and non-skin pixels. Therefore, the first step is a component normalization for r and b channels in the given image: $r=\frac{R}{R+G+B}, b=\frac{B}{R+G+B}$. With: $R=$value of RED channel, $G=$value of Green channel, $B=$value of Blue channel. Green component is not evaluated because, with the assumption of normalization, it can be found with r and b values: $r+b+g=1, g=1-b-r$.

Once normalized RGB components are calculated, some considerations on skin pixels have to be made. Colour distribution on RGB space of skin pixels for different people is quite compact and localized into the $r-b$ plane. This distribution can be well approximated by a 2-D Gaussian distribution: $p(\bar{x})=e^{-\frac{1}{2}(x-\bar{m})^{T}C^{-1}(x-\bar{m})}$. With: $C=$Covariance Matrix, $\bar{x}=(r,b), \bar{m}=(\bar{r},\bar{b})$. The aim is to build a robust model of skin by evaluating characteristic parameters of the bidimensional Gaussian in terms of C covariance matrix and vector $\bar{m}$.

These values have been "estimated" using a 768×576 image made with "patches" of skins derived from faces of people walking in the car park. Pixels belonging to this picture have been used as "training set" to build the model in terms of values for the covariance matrix C and $\bar{m}$: $C=\begin{bmatrix} 1.422 & -0.101 \\ -0.101 & 0.090 \end{bmatrix}, \bar{m}=(0.395, 0.290)$. The resulting distribution image is reported in Fig. 7. As can be seen from Fig. 7(a) the Gaussian distribution is well separated from the r,b plane. However considering its 2-D projection in Fig. 7(b) it can be seen that equal height contour lines are elliptic with quite elongated shapes.

1) http://www-cs-students.stanford.edu/~robles/ee368/skincolor.html

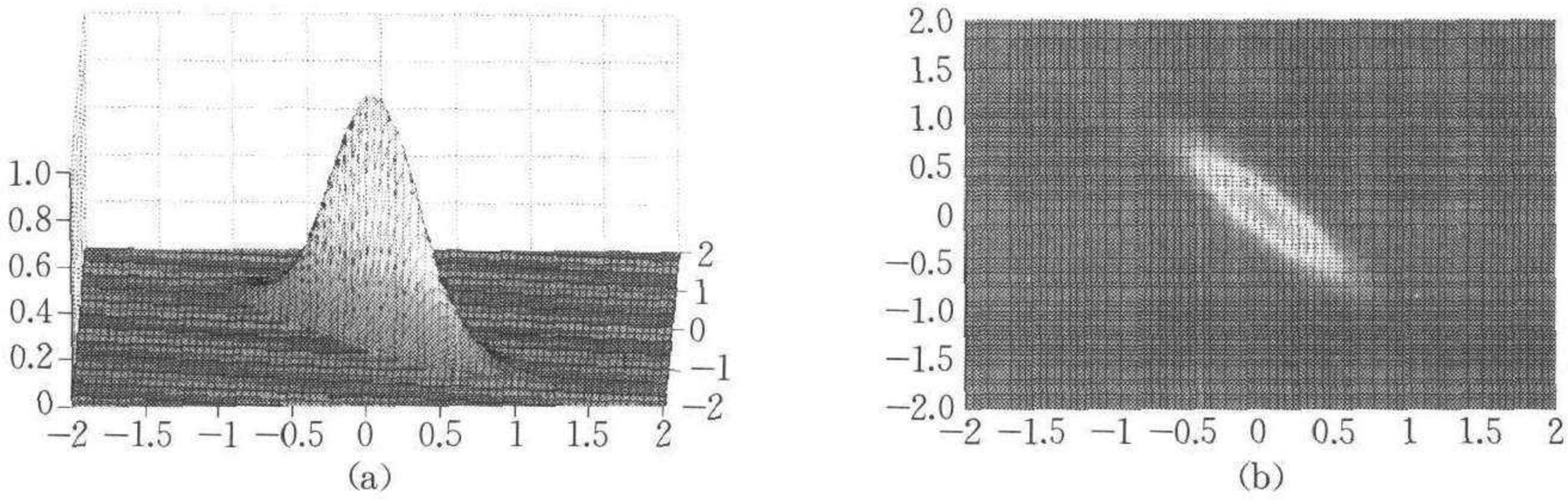


Fig. 7 Gaussian distribution for SkinModel

4.2 The classification

Given the model with the Gaussian distribution in Fig. 7, the problem of classification consists of evaluating a threshold that can be used to recognize, for every pixel in the image, between skin or non-skin pixels. To do this, some considerations have to be made on probability error we aim at and on a priori knowledge. In particular if we indicate with $H1$ the event of having a skin pixel and $H0$ the event of a non skin pixel we can assert that the distribution sketched in Fig. SkinModel is equivalent to: $p(r,b/H1=skin)$, as the probability of having the pixel (x_i,y_j) with values of red and blue channels respectively equal to r and b. Error on classification of pixels can be evaluated by calculating the probability error p_{err}: $p_{\text{err}}=p(H0)\cdot p_{fa}+p(H1)\cdot p_{md}$. Unfortunately, in the considered case $p(r,b/H1=skin)$ can be used to compute p_{md} once we have fixed a threshold, but we have no information on $p(H0)$, $p(H1)$, p_{fa} and $p(r,b/H1=skin)$.

p_{fa} gives the probability that a pixel is classified as skin pixel even if it belongs to non-skin cathegory, however the chosen approach that exploits contextual information implicitly reduces this kind of error. In particular, faces are generally searched in the high zoom image where at least a face is meant to be; in fact this image is generated on the basis of the positioning information of blobs extracted from camera 1 (static, wide field camera). This reduces probability of finding skin pixels (p_{fa}). For this reason the value that becomes quite descriptive in terms of performances is the p_{ma} that can be written by definition as:

$$p_{ma}=\iint_{th} f(r,b)\,\mathrm{d}r\mathrm{d}b$$

To get a reasonable error in the estimation of skin and non skin pixels a threshold on p_{ma} equal to 0.9 has been set.

5 Tracking faces

Once a human has been successfully detected, placed in the common 2-D map and sensor has been pointed to the believed center of mass of the oval of the face, a temporal correspondence between successive frames has to be preserved. To do this, tracking techniques have been applied to successfully estimate position of the face in time. In particular, "Camshift"[1] algorithm is used, a non-parametric technique for climbing density gradients to find the mode (peak) of probability distributions. This algorithm derives from the "Mean Shift" algorithm developed by Comaniciu *et al.*[27]. The algorithm takes as input the colour probability distributions outlined in paragraph 5.1, and an initial search window to begin the investigation from.

As output, it provides face centroid and size. To do that, zeroth moment is calculated as follows: $M_{00}=\sum_x\sum_y I(x,y)$. The first moment for x and y: $M_{10}=\sum_x\sum_y xI(x,y)$,

1) http://www.intel.com/technology/itj/q21998/articles/art-2g.htm

$M_{01}=\sum_x\sum_y yI(x,y)$. Then the mean search window location (the centroid) is: $x_c=\frac{M_{10}}{M_{00}}$, $y_c=\frac{M_{01}}{M_{00}}$, where $I(x,y)$ is the pixel (probability) value at position (x,y) in the image, and x and y range over the search window. Camshift has been shown to be robust to face distractors such as other people present in the scene, occlusions with other rigid and non-rigid objects (hands) or rapid movements.

As it can be seen from Fig. 8, tracker follows faces for a fixed number of frames to edit a small video clip of each oval. Once the clip has been collected, camera is ready to return to the reference position and satisfy other requests.

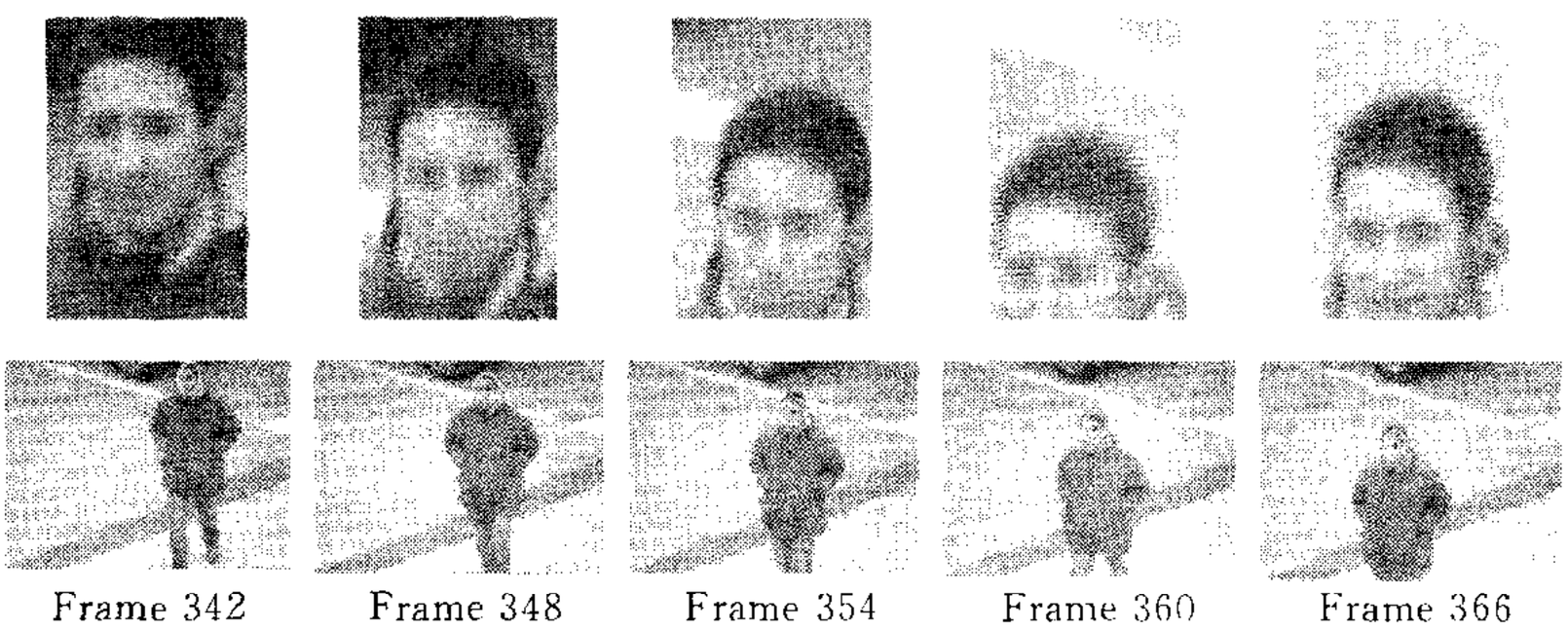


Fig. 8 Sample sequence for face tracking and recorded face clip

Statistical tests have been performed on the correct localization of face centroid; in particular tests have analyzed the following two events: F_0 = Face not present in the given high-resolution frame; F_1 = Face present in the given high-resolution frame. The following quantities have been evaluated:

• $\hat{P}_d=P\langle u=1|F_1\rangle$ Estimation of Correct Detection probability: face is present in the given high resolution frame and it is correctly localized ($u=1$);

• $\hat{P}_{FA}=P\langle u=0|F_1\rangle$ estimation of Miss Detection probability: face is present in the given high resolution frame and but it is not correctly localized ($u=0$);

• $\hat{P}_{MA}=P\langle u=1|F_0\rangle$ estimation of False Alarm probability: face is not present in the given high resolution frame and but it is not correctly localized ($u=0$).

Tests have been performed of three sequences containing a total number of frames equal to 1324; as it can be seen from Table 2, $\hat{P}_d$ is around 92% whereas the probability of false alarm is around 10%. This value should be equal to zero considering that the face is not present in the frame and camshift should not return any value. Anyhow, due to errors on skin-color probability distribution induced by the Gaussian model, Camshift outputs erroneous positions.

Table 2 Performances of face tracking

$\hat{P}_d$	$\hat{P}_{FA}$	$\hat{P}_{MA}$
92.2%	10.1%	7.8%

6 Conclusions

An innovative architecture for multisensor video surveillance has been presented along with some preliminary results taken upon the basis of different video sequences. These video sequences are available on the web at http://spt.dibe.unige.it/ISIP/sequencesLuca.html. The system shows cooperative behaviour between sensors which are able to track at different levels of zoom moving objects. In addition, the system is also able to "focus" its attention in order to extract biometric information about humans in the monitored environment, in particular faces of people can be detected, segmented and tracked. The cur-

rent version of the system gives global encouraging results. However the model for skin detection can be improved by using a more robust approach (eventually exploiting a feature alternative to colour such as shape) and enhancing the current model with a more consistent training set. For what concerns techniques for pointing the sensor, some metrics of goodness of pointing have to be developed whereas at tracking level, even if camshift performs well, a combined approach exploiting also shape information will be implemented. Nevertheless, the developed system has quite low set-up time thanks to some self-calibration mechanism here not described and this makes it attractive for the use in many different applications.

References

1 Collins R T, Lipton A J, Kanade T. Introduction to the special section on video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22** (8): 1～3

2 Stauffer C, Grimson W L. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Special Section on Video Surveillance*, 2000, **22**(8): 45～53

3 Haritaoglu I, Harwood D, Davis L S. W4S: A real-time system for detecting and tracking people in 2-1/2 D. In: Proceedings of Europeen Conference on Computer Vision, Germany: Sprinder, 1998

4 Dockstader S, Tekalp A M. Multiple camera tracking of interacting and occluded human motion. *Proceedings of the IEEE*, 2001, **89**(10): 1441～1455

5 Varshney P. Multisensor data fusion. *Electronic and Communication Engineering Journal*, 1997, **9**(2): 245～253

6 Regazzoni C S, Ramesh V, Foresti G L eds. Proceedings of IEEE(Special Issue on 3rd Generation Surveillance Systems), 2001

7 Foresti G L, Regazzoni C S. A change-detection method for multiple object localization in real scenes. In: Proceedings of the IECON 1994, Bologna Italy, 1994. 984～987

8 Regazzoni C S, Marcenaro L. Object detection and tracking in distributed surveillance systems using multiple cameras. In: Advanced Studies Institute on Multisensor Data Fusion, Hyder A ed. Kluwer Academic Publishers, 2000

9 Wren C, Azarbayejani A, Darrel T, Pentland A. Pfinder: Real time tracking of the human body. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 1997, **19**(7): 780～785

10 Tsap L V, Goldof D B, Sarkar S P. Non rigid motion analisys based on dynamic refinement of finite models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(5): 526～532

11 Marcenaro L, Oberti F, Regazzoni C S. Multiple objects color-based tracking using multiple-cameras in complex time-varying outdoor scenes. In: Proceedings of the 2nd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Kauai, Hawaii: IEEE Press, 2001

12 Tsai R Y. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, 1987, **3**(4): 323～344

13 Marcenaro L, Oberti F, Regazzoni C S. On line self-organizing non-rigid shape description in multiple objects scenes. In: Workshop on Real-Time Image Sequence Analysis, RISA2000, Oulu, 2000. 63～78

14 Teschioni A, Oberti F, Regazzoni C S. A neural network approach for moving objects recognition in color image sequences for surveillance applications. In: Nonlinear Signal and Image Processing NSIP'99, Antalya: Turkey, 1999. 28～32

15 Haritaoglu, Harwood D, Davis L S. Hydra: Multiple people detection and tracking using silhouettes. In: Proceedings of the Workshop on Visual Surveillance, Fort Collins, CO, 1999. 6～13

16 Theil A, Kemp R, Romeo K, Kester L, Bosse E. Classification of moving objects in surveillance algorithms. In: Proceedings of the 1st IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS'2000, France: Grenoble, 2000. 80～84

17 Chai D, Ngan K N. Locating Facial Region of a Head-and-Shoulders Color Image. In: Proceedings of the 3rd International Conference of Video Automatic Face and Gesture Recognition, 1998

18 Ming-Hsuan Yang, Member, David J Kriegman, Narendra Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Aanalysis and Machine Intelligence*, 2002, **24**(1): 34～58

19 Saxe D, Foulds R. Toward robust skin identification in video images. In: Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition, 1996. 379～384

20 Swain M J, Ballard D H. Color Indexing. *International Journal of Computer Vision*, 1991, **7**(1): 11～32

21 Yang M H, Ahuja N. Detecting human faces in color images. In: Proceedings of IEEE International Conference on Image Processing, IEEE Press, 1998. 127～130

22 Cai J, Goshtasby A, Yu C. Detecting human faces in color images. In: Proceedings of 1998 International Workshop Multi-Media Database Management Systems, 1998. 124～131

23 Yang G, Huang T S. Human face detection in complex background. *Pattern Recognition*, 1994, **27**(1): 53～63

24 Dai Y, Nakano Y. Face-texture model based on SGLD and its application in face detection in a color scene. *Pattern Recognition*, 1996, **29**(6): 1007～1017

25 Craw I, Tock D, Bennett A. Finding face features. In: Proceedings of the 2nd European Conference on Computer Vision, 1992. 92～96

26 Osuna E, Freund R, Girosi F. Training support vector machines: An application to face detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1997. 130～136

27 Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis Machine Intell.*, 2002, **24**(5): 603～619

Luca Marchesotti Doctor candidate in Electronic Engineering and Computer Science at University of Genoa and a member of the Signal Processing & Telecommunications Group in the same University. Received his bachelor degree in telecommunication engineering from the University of Genoa in 2001 with a master thesis dealing with the study of a distributed agent society for scene understanding in video-surveillance applications. His main research interests include intelligent agents architectures and vision (tracking, face detection).

Alessandro Messina He is an master's student at University of Genoa. He joined the Signal Processing & Telecommunications Group in the same University in 2002 working in multicamera video surveillance. His main interests include computer vision, tracking and face detection.

Lucio Marcenaro Doctor candidate in electronic engineering and computer science at University of Genoa and is a member of the Signal Processing & Telecommunications Group in the same University. Received the his bachelor in electronic engineering with telecommunication and telematic specialization in 1999 with a thesis about flexible models for human motion to analyze images from multiple videosurveillance cameras. His main research interests include image and sequence processing for video- surveillance systems and statistical pattern recognition.

Carlo Regazzoni(Senior Member, IEEE) Since 1998 he is professor of telecommunication systems in the Engineering Faculty of University of Genova. He received the bachelor degree in electronic engineering and the Ph. D. degree in telecommunications and signal processing from the University of Genoa, in 1987 and 1992, respectively. Since 1998 he is responsible of the Signal Processing and Telecommunications (SP&T) Research Group at the Department of Biophysical and Electronic Engineering (DIBE), University of Genova, that he joined in 1987. Dr. Regazzoni is involved in research on Multimedia Surveillance systems since 1988. He has been co-organizer and chairman of the first two International Workshops on Advanced Video Based Surveillance, held in Genova, Italy, 1998 and Kingston, UK, 2001. He has also organized several Special Sessions in the same field at International Conferences (Image Analysis and Processing, Venice 1999 (ICIAP99). European Signal Processing Conf. (Eusipco2000), Tampere Finland, 2000). Dr. Regazzoni has been project responsible in several EU research and development projects dealing with video surveillance methodologies and applications in the transport field (ESPRIT Dimus, Athena, Passwords, AVS-PV, AVS-RIO); he has been also responsible of several research contracts with italian industries; he served as a referee for international journals, and as reviewer for EU in different research programs. He is a consultant of the EU Commission for the definition of the 6th research framework program in the Ambient Intelligence domain. His main current research interests include multimedia and non-linear signal and video processing, signal processing for telecommunications, multimedia broadband wireless and wired telecommunications systems.

视觉监控应用中多传感器协作的人脸检测系统

Luca Marchesotti Alessandro Messina Lucio Marcenaro Carlo Regazzoni

(*DIBE-University of Genoa, Genoa*, 意大利)

(E-mail: carlo@dibe.unige.it)

摘 要 提出了一种新颖的由两个可控摄像机组成的多传感器视觉监控系统,旨在实现户外环境下的实时跟踪与特征化运动目标.特别地,该系统利用一个在多个缩放级别上可操作的移动摄像机在连续视频帧中自动获取与跟踪人脸.配合它的是一架能执行自动目标跟踪与分类的固定广域摄像机.

关键词 多传感器视觉监控,多个缩放级别,目标跟踪

中图分类号 TP391.41