TRACKING MULTIPLE INDIVIDUALS FOR VIDEO COMMUNICATION

Alberto Avanzi, François Brémond and Monique Thonnat

INRIA Sophia Antipolis
2004, route des Lucioles, BP93 - 06902 Sophia Antipolis Cedex, France
{Alberto.Avanzi, Francois.Bremond, Monique.Thonnat}@sophia.inria.fr
http://www-sop.inria.fr/orion/personnel/Alberto.Avanzi/index.html

ABSTRACT

We propose a new interpretation platform dedicated to video communication. Basically, a video communication system takes an image flow from a camera and broadcasts it in a computer network. The goal of interpretation is to enable the video communication system to adapt automatically the broadcasted image flow, using image filtering, blurring and zooming. For that we need to detect and track individuals in office scenes, then to understand their behaviour. In this article, we focus on a new tracking method based on a 3D model of the scene, on explicit models of individuals and on the computation of several possible paths for each individual. The main issues of the tracking algorithm are presented. Finally we show the results of our algorithm for several sequences illustrating office activities in every-day life situations.

1. INTRODUCTION

In this article, we present a method for tracking multiple individuals in a cluttered scene (typically an office). The class of applications we are interested in, is multiple individuals tracking and scenarios recognition for video communication applications with a fixed wide-angle monocular colour camera.

The concept of video communication is dedicated to allow permanent and informal communication and group awareness between spatially separated individuals (including video-conference). Current video communication (VC) systems rise two problems. First, the static nature of the cameras interfere with the real word activity; i.e. the VC system should automatically follow and zoom at the main individual in the office. Second, the level of availability (protection of privacy) has to be set up automatically, if not studies show that VC systems will be less accepted [1]. For these reasons we propose to use an interpretation system to be a part of the global VC system to automate these tasks [2], [3].

Given image sequences of everyday office life, our proposed interpretation system is able to analyse the behaviour of individuals. This system is composed of three modules, as illustrated in Figure 1. First, from images the detection & classification module detects moving regions by subtraction to a background image and it classifies moving regions into mobile objects adding a label giving its type, such as *individual* or *noise*. Second, the tracking module associates the mobile objects with already tracked targets that correspond to real individuals. Third, the scenarios recognition module recognises the scenarios relative to the behaviour of tracked individuals. Finally, thanks to the scenario analysis, the VC system can decide what to send to the network a subimage of the camera image (zoom) or a filtered image (blur). For example, if the system recognises a work meeting situation, the VC system can automatically broadcast a blurred image to protect visitor's privacy. Moreover, if the scenario "user is writing on the blackboard" is recognised, the system can broadcast an image zoomed on the blackboard area.

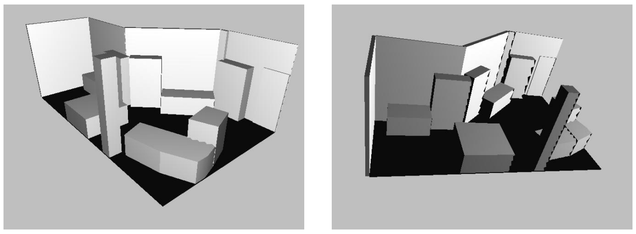


Fig. 2. Two VRML views of 3D scene model for the office.

In this article, we focus on the tracking module. The tracking problem is a central issue in scene interpretation, as the lost of a tracked object prevents to analyse its behaviour. The article presents a tracking method, based on a 3D model of the scene, on explicit models of individuals and of individual's trajectories and on the computation of several possible paths for each individual (using a time delay). Resulting algorithm is able to cope with severe occlusions, merging/splitting individuals, errors in detection and in classification.

First, we present the interpretation system. Then we show how to compute the different paths and the correspondences between paths and individuals. Then we explain how to solve ambiguous correspondences and we conclude by providing examples on how this method works on several real sequences. Related works can be found in [4], [5], [6], [7].

2. 3D SCENE MODEL

The interpretation system uses a 3D scene model to compute accurately the relationship between individuals. This is an a priori contextual knowledge (cf [8]) containing semantic information and a tridimensional description of the geometry of the scene: walls, doors, and the set of furniture existing in the scene (desktop, shelves, cupboards...) as shown in Figure 2. It is used for example by the classification module to compute occlusions and by the tracking module in order to calculate the position of in/out areas (doors). A calibration matrix, used in order to obtain 3D co-

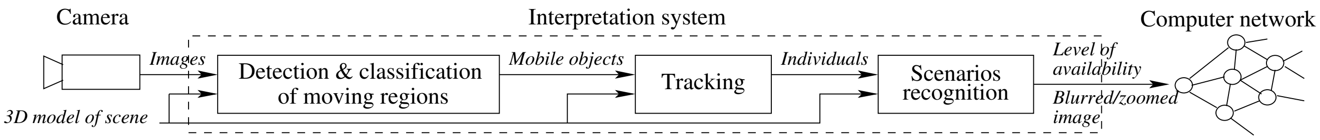


Fig. 1. The interpretation system is composed of three modules.

(2a) Green bounding box (2b) Red (left), yellow (right) boxes

Fig. 3. Three classes of mobile object type: *group* (green box, left image), *individual* (red box, bottom left on the right image) and *occluded_individual* (occluded by the desktop, yellow box, on the right in the right image).

ordinates from 2D coordinates on the image, completes the scene model.

3. THE INTERPRETATION SYSTEM

As described above, the interpretation system is composed of three modules. Because of noises (shadows, reflections, illumination changes) and occlusions, a first main issue consists of classifying correctly the moving regions. To improve the classification of moving regions corresponding to *noise, shadows* and *scene_objects* (they don't contain skin regions) compared to *individuals*, we propose to detect skin regions. They are sub-areas of moving regions representing skin. The detection module compares the RGB coordinates of each moving pixel with a given chromatic model of skin colour [9]. Our algorithm is based on three histograms computed during the initialisation phase: a skin histogram (H_s), containing the chromatic model of skin and computed using a library of skin images, a background histogram (H_b), containing the chromatic model of the scene (computed using background skinless images) and the histogram (H_r) obtained by the ratio of H_s and H_b (in order to reduce the noise present in H_s). Each histogram is computed over the three coordinates R,G,B normalised with respect to luminosity (R+G+B). The goal of normalisations is to reduce the effects of light intensity variations.

The detection module allows to discriminate skin areas as illustrated by Figure 4. Thanks to skin occurrence, human being model (e.g. 3D size) and 3D scene model, the classification module can in 95% of cases recognise individuals, individuals occluded by scene objects (i.e. desktop) and groups of people (merged persons) from noise (due to shadows and reflections) and from movements due to scene objects (i.e. the opening of a door), as shown in Figure 3.

Fig. 4. On the left the input image, on the right the detected skin areas (randomly coloured).

4. TRACKING

4.1. Path and individual models

The method used for tracking is based on the modelling of two entities: *paths* and *individuals*.

A *path* P_i represents a possible trajectory of one individual in the scene during the temporal interval $[t_f - T; t_f]$, where t_f is the time of the last (current) image processed and T is the size of the temporal window used to analyse paths. P_i is composed of a temporal sequence of mobile objects $M_i(t), t \in [t_f - T; t_f]$, fulfilling two conditions: $(a_1)\ \forall\ t\ \exists!\ M_i(t); M_i(t) \in P_i$ and (a_2) $dist_{2D}(M_i(t); M_i(t+1)) \leq D_{max}$. Condition (a_1) is introduced in order to limit the number of paths computed in case of over-detection (if several mobile objects represent different parts of the same person we choose the best one). (a_2) prevents the creation of paths with large spatial discontinuities. A quality coefficient Q_i qualifies each path P_i, taking into account three coefficients C_i:

- C_1: the coherency between each $M_i(t)$ of P_i and a model of human being (including 3D size and class of mobile objects, like *individuals* or *group*).
- C_2: the temporal coherency between the different mobile objects $M_i(t)$ composing P_i (e.g. 3D size, motion).
- C_3: the inverse of the 2D distance between $M_i(t)$ and $M_i(t+1)$ compared with a human movement model.

This quality coefficient is given by

$$Q_i(t) = C_1 \cdot C_2 \cdot C_3 \cdot Q_i(t-1)$$

A new path is created when a mobile object is detected in the in/out area. This 2D area, which corresponds to the office door, is computed using a 3D model of the scene. A path is updated using the new mobile object detected at time t_f and is deleted when its quality coefficient Q_i is too low.

An *individual* I_j is a structure representing a real person moving in the scene at time $t_f - T$, where T is the delay between the update of paths and of individuals (see Figure 5). It represents the output of the tracking module. Each individual is associated with a

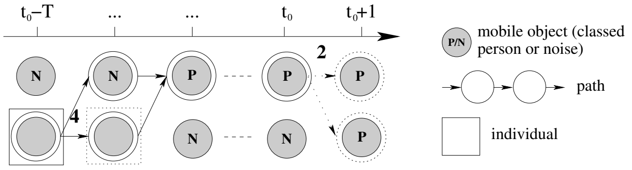


Fig. 5. Path (2) and individual (4) updates

set of paths, representing all its possible future trajectories. A path corresponds to a track of a potential person whereas the structure individual corresponds to a real person.

A new individual is created when we are sure it corresponds to a real person (when a path P_i, not yet associated with any individual, fulfils the following four conditions): (1) P_i starts in the in/out area, (2) ends outside the in/out area, (3) is composed by T mobile objects and (4) does not overlap a path P_j already associated with another individual. The goal of these four conditions is to avoid to create a new individual using a wrong path due to detection errors (e.g. shadows). Time delay T is to enssure that the individual choose the correct path. An individual is updated with the most compatible path (see below). It is deleted when its last mobile objects are in the in/out area and when there is no more mobile object newly detected (occurs when the real person has left the room).

4.2. Main tracking issues

The first issue consists in limitating the combinatorial explosion of the number of computed paths, due to the tracking module trying to extend all existing paths using all new mobile objects. Our solution consists in splitting the set of paths existing at time t_f into $N+1$ subsets S_j, where N is equal to the number of individuals existing at time $t_f - T$. Each subset S_j gathers all the paths associated with the individual I_j. The additional subset S_0 corresponds to individuals just entering the scene (not yet created, but associated with partially completed paths). The tracking module allows an equal number $N_{ext}(t)$ of extentions for each subset S_j; each path P_{ji} (belonging to subset S_j) receives a number of extensions $N_{ji}(t)$ proportional to its quality $Q_{j,i}(t)$ over the total quality coefficient $Q_j(t)$ of the set S_j:

$$N_{ji}(t) = N_{ext}(t)\frac{Q_{j,i}(t)}{Q_j(t)}, \quad with \quad Q_j(t) = \sum_{k \in S_j} Q_{j,k}(t).$$

In any case the maximum number of extensions for each path is equal to the number of new detected mobile objects, $O(t)$. If $N_{ji}(t) < O(t)$ then only the $N_{ji}(t)$ mobile objects giving the best extensions are used for extending the path P_{ji}.

The second issue consists in the optimal and global association between paths and individuals. The problem is to avoid the association of two individuals (crossing each other) with two paths strongly overlapped and corresponding to the same isolated individuals. To solve this problem, we compute four relationships between paths and individuals:

- D_1: a path P_k and an individual I_q are **associated** if they share the same mobile object at $t_f - T$. This path represents a possible future trajectory for individual I_q. We quantify this relationship by computing the coefficient of compatibility $D_1(P_k; I_q)$ which takes into account the compatibility of the sizes and directions of I_q and P_k.
- D_2: two paths **overlap** each other if they share several mobile objects. This relation is quantified by a coefficient $D_2(P_q; P_r)$ which exponentially decreases with the number of shared mobile objects:

  $$D_2(P_q; P_r) = e^{-\left(\frac{N_{ov}}{K}\right)^2}$$

  where N_{ov} is the number of mobile objects (not classified *group*) shared by P_q and P_r and K is a normalisation coefficient.
- D_3: it is a boolean measure telling whether two individuals are **crossing** each other (i.e. they are associated with paths which are overlapping each other).
- D_4: a couple individual-path is **classified** *person* or *group* depending on the class of the path's mobile objects and on the fact that the individual is isolated (not sharing a mobile object at time $t_f - T$ with another individual).

After computing these relationships, the tracking module first updates all individual crossing nobody else ($D_3 = 0$) with the path giving the best compatibility coefficient. Then we calculate all sets of individuals crossing each other ($D_3 = 1$), called E_h. The tracking module then selects the couple $I_j - P_k$ of E_h with the best compatibility coefficient (D_1). This couple is updated and if it is classified as a person ($D_4 = person$), the compatibility coefficient of other couples $I_l - P_m$ is decreased for each path P_m overlapping P_k (D_2):

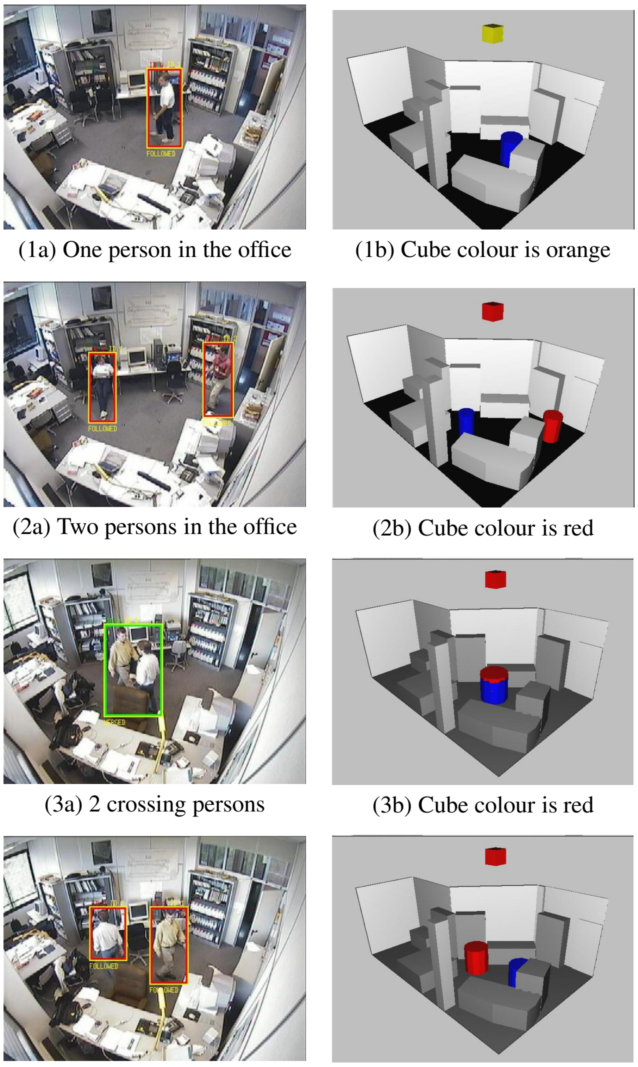
$$D_{1new}(I_l; P_m) = D_{1old}(I_l; P_m) \cdot D_2(P_k; P_m).$$

Then the tracking module reselects the second best couple $I_{j'} - P_{k'}$ with the second best compatibility coefficient (D_1). This algorithm is iterated until all couples from the set E_h of individuals crossing each other are processed.

The third issue is to handle non detections when the mobile object corresponding to a real person who has not been detected. To solve this problem, we propose to introduce virtual mobile objects M_{lost}. Each time the tracking algorithm extends a path P_i, a virtual mobile object is taken into account as one possibility of extention. The quality coefficient $Q_{i,lost}$ in this case is based on a numeric constant representing the lack of information for the mobile object M_{lost} (as a virtual mobile object, it doesn't provide any information). The extended path $P_{i,lost}$ has a lower quality coefficient than any other path P_i extended with a detected mobile object, but in following images this $P_{i,lost}$ path can obtain the best quality.

5. RESULTS

This tracking module has been tested on several office sequences. The longest sequence is 15 minutes time long (more than 3900 frames). Some of these sequences are available on the web, at the author's address shown on the first page. These sequences show people entering the office together, crossing and overlapping each other. On these sequences the tracking module is able to cope with the static occlusions of individuals (e.g. partially hidden by a desktop) and non-detection of individuals during at least 4 frames (e.g. due to lack of contrast). The tracking module is also able to handle situation where different individuals are merging and crossing



(1a) One person in the office (1b) Cube colour is orange

(2a) Two persons in the office (2b) Cube colour is red

(3a) 2 crossing persons (3b) Cube colour is red

(4a) After crossing (4b) Cube colour is red

Fig. 6. On the left the images illustrate the tracking results, whereas on the right they illustrate the scenarios recognition results in VRML form (the colour of the flag (cube) represents the level of availability). Coloured cylinders represent the individual identifiers detected and tracked in the office. Each person has its own colour. Images 3a-4a show a situation where two individuals are crossing each other.

each other. Currently the two main limitations of the system are the mix of individual identifiers in difficult cases (e.g. long crossings) and to handle situations where individuals are mixed with mobile objects corresponding to noise (e.g. a chair moved by a person).

An example of scenarios recognition "two people are meeting in the office" is given in the first four images in Figure 6. This scenario illustrates the case when guests are visiting the office and when the availability level should decrease (we don't want to broadcast guest pictures). Guests can be recognised according to the place where they sit. A flag (cube), showing the availability level, changes its colour following the number of people in the office: green (nobody in the office, maximum of availability), orange (low probability of guests in the office, medium level of availability) and red (high probability of guests in the office, meeting situation, no availability).

6. CONCLUSIONS

We have presented an algorithm for tracking multiple individuals in an office scene which to be used in video communication systems. The quality and the robustness for long sequences of the tracking algorithm enable high level interpretation. A scenario recognition module uses the output of the tracker to recognise interesting situations to protect automatically the privacy of users of video communication. Currently we are planning to extend the model of individual to allow person recognition. To recognise office owners, one possibility is to register automatically in the morning the dominant colour of each owner's clothes. Moreover, this technic would help the system to keep track of the correct identifier after long crossings. Second, we are planning to improve and complete the scene model that we are using. Specially we are thinking to introduce a model for scene objects, which handles 3D scene objects able to move (like chairs and doors). This new model will prevent paths to be mixed with mobile objects corresponding to the motion of these scene objects.

7. REFERENCES

[1] J.Coutaz et al., "Early experience with the mediaspace comedi," *Engineering HCI*, 1998.

[2] F. Brémond and M. Thonnat, "Object tracking and scenario recognition for video-surveillance," in *the poster sessions of the 15th Int'l Joint Conference on Artificial Intelligence (IJCAI)*, Nagoya (Japan), Aug. 1997.

[3] M. Thonnat and N. Rota, "Image understanding for visual surveillance application," in *Third international workshop on cooperative distributed vision CDV-WS'99*, UK British Machine Vision Conference, Guildford, Ed., Kyoto, Japan, Nov. 1999, pp. 51–82.

[4] A. Baumberg and D. Hogg, "An adaptive eigenshape model," in *proc. of the British Machine Vision Conference (BMVC)*, Birmingham, Sept. 1995.

[5] I. Cox and S. Hingorani, "An efficient implementation of reid's Multiple Hypothesis Tracking algorithm and its evaluation for the purpose of visual tracking," in *IEEE Transactions on pattern analysis and machine intelligence*, Sept. 1996, vol. 18.

[6] F. Meyer and P. Bouthemy, "Region-based tracking in an image sequence," in *Proc. of European Conference on Computer Vision (ECCV)*, May 1992, pp. 476–484.

[7] Z. Zhang, "Token tracking in a cluttered scene," Research report 2072, I.N.R.I.A., Sophia Antipolis, Oct. 1993.

[8] F. Brémond and M. Thonnat, "Issues in representing context illustrated by scene interpretation applications," in *proc. of the Int'l and Interdisciplinary Conf. on Modeling and Using Context (CONTEXT-97)*, Rio de Janeiro, Feb. 1997.

[9] J.L. Crowley K. Schwerdt and J.B. Durand, "Robustification of detection and tracking of faces," *Computer Vision and Mobile Robotics Workshop*, September, 1998.