HAL
open science

AI Techniques for VSIS Human Tracker

Nathanaël Rota, Robert Stahr, Monique Thonnat

To cite this version:

Nathanaël Rota, Robert Stahr, Monique Thonnat. AI Techniques for VSIS Human Tracker. RR-4138, INRIA. 2001. inria-00072488

HAL Id: inria-00072488

https://inria.hal.science/inria-00072488

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

AI techniques for VSIS human tracker

Nathanaël ROTA — Robert STAHR — Monique THONNAT

N° 4138

Mars 2001

THÈME 3

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4138--FR+ENG

AI techniques for VSIS human tracker

Nathanaël ROTA , Robert STAHR , Monique THONNAT

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Orion

Rapport de recherche n° 4138 — Mars 2001 — 31 pages

Abstract: The aim of this work is to track humans in a known environment. The input data are pre-segmented video streams captured with a fixed camera. In this paper we propose a new approach to the problem of simple hypothesis non-delayed human tracking. Furthermore the approach handles simultaneously and in a unified manner the actual similarity matching and the associated real-world problems of entries, exits, occlusions and false detections (noise). Results on visual surveillance applications for security in metro stations , security in banks and monitoring in offices are shown. Quantitative results of this method in terms of time performance are presented.

Key-words: Human Tracking, Heuristic Search, Diagnosis

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : 04 92 38 77 77 - International : +33 4 92 38 77 77 —Fax : 04 92 38 77 65 - International : +33 4 92 38 77 65

Techniques d'intelligence artificielle pour le suivi de personnes dans VSIS

Résumé : Le but de ce travail est l'élaboration d'un programme capable de suivre des personnes dans un environnement connu et stable. La donnée d'entrée de ce programme est un flux vidéo présegmenté. Dans ce rapport, nous proposons une nouvelle approche du problème du suivi de personne sans délais ni hypothèses multiples. De plus, cette approche gère simultanément, et d'une façon unifiée, les calculs de similarité et les problèmes réels tel que les entrées/sorties de personnes, les occultation et les fausses détections (bruit). Des résultats de cette approche en vidéo-surveillance dans les stations de métro, dans les agences bancaires et les bureaux sont exposés.

Mots-clés : Suivi de personne, methodes heuristiques, Diagnostique

1 Introduction

The aim of our research is the elaboration of programs able to recognise certain human behaviours in a known environment filmed by a fixed camera. Our approach is to compute a temporal description of the scene using the images of a video stream. Then a set of behaviour models can be recognised from the temporal description. The temporal description, in addition to the set of objects composing the scene, consists of a set of descriptions of humans in the scene. The process which compute this set of descriptions is called the human tracking. We present in this article a novel approach to the problem of single hypothesis non-delayed frame-to-frame person tracking. Furthermore the approach handles simultaneously and in a unified manner the actual similarity matching and the associated real-world problems of entries, exits, occlusions and false detections (noise). The general approach consists of two steps. Firstly, given a set of blobs obtained by a background subtraction method, we compute the number of humans and the location of each of them. Secondly, the recognised humans at two consecutive frames are matched.

To do this, two artificial intelligence techniques are used. A heuristic method is used to recognise humans in each frame and a diagnosis method is used to compute the match between two consecutive frame descriptions. In each case, we try to provide formal and generic methods which remain robust and realistic.

We will see, in section 2, how the human tracking problem has been solved the past ten years. Then, in section 3, we will detail the modelling of the problem we want to solve. In section 4 and 6, we will detail the proposed methods for human recognition and tracking. Results obtained on various environments will be shown in sections 5, 7 and 8.

2 State of the Art

During the past decade a lot of work has been done in human tracking for various applications and with various approaches. In every case, the human tracking consists in computing a temporal description of a scene with some humans filmed by one or more cameras. The problem of the transcription of such a video stream into a set of temporal descriptions of humans has been addressed for various tasks such as video compression, man-machine interfaces or scene analysis.

We will concentrate on work done for scene analysis. In this case, scene analysis means that the temporal description of the humans in the scene will be the input to other algorithms such as human counting [1, 2], human posture analysis [3, 4, 5], scene analysis [6], event recognition [7, 8] or scenario recognition [9].

The difference between those works also depend on the kind of scene in which the human tracking must be done. From this point of view, we can find various environments such as offices [10], labs [11, 3, 5], parking lots [12], metro stations [13], outdoor environments [14, 15] or even soccer stadiums [6, 16].

The second source of differences between the above mentioned works is the kind of model used to represent a human. Three types of models are used: 2D models, 3D models and

hybrid models. Several approaches using 2D models have been proposed such as the classical bounding box, cubic B-splines [14, 17], sets of ellipses [15] or sets of ribbons [18]. The second category of human models is the 3D model, i.e. volumetric models composed of a set of 3D volumes such as sets of cylinders [4], sets of ellipsoids [19] or non-parametric models [11]. The last category of models is the hybrid model which is generally a mixture of 2D and 3D models. The fact is that when the model gets more complex (i.e. when the number of degrees of freedom gets higher), the matching between two frames gets easier, but the recognition becomes more difficult.

The third source of differences between the approaches found in the literature is the technique used to recognise the models. I.e. the different approaches used to transform the video stream into a set of instantiated models. Two approaches can be found. The first approach consists in recognising in the current image the model instantiated in the previous frame [18, 20, 21, 22]. With this kind of approach, the match between two descriptions of a human seems to be more accurate, but the issue of the initialisation and the termination of the tracks is often ignored. The second approach consists in recognising in the current image a set of models without any other information and then compute the match between these recognised models and those from the previous frame [19, 15, 9]. With this approach, the problem is to find the optimal match between two sets of descriptions among all possible matches.

In this context, our approach deals with three kinds of scenes such as offices, metro stations and banks. We will use a simple hybrid model composed of a 2D box and a 3D cylinder, because we strongly believe that a model with a high number of degrees of freedom cannot be recognised with a good and constant accuracy under real world conditions. In terms of which technique is used to track humans, our work belongs to the second category.

3 Problem Modelling

3.1 Notation

Let $\bar{G}_t = (\bar{F}_t, \bar{A}_t)$ be a graph called the interpretation graph as shown in figure 1. $\bar{F}_t$ is the set of vertices representing all the timed concepts of the scene and $\bar{A}_t$ is the set of arcs representing all the binary relations between the concepts of the scene. We introduce on $\bar{G}_t$ three types of vertices noted $\bar{O}_t$, $\bar{P}_t$ and $\bar{V}_t$ and two types of arcs noted $\bar{T}_t$ and $\bar{R}_t$. I.e. $\bar{G}_t = (\bar{O}_t \cup \bar{P}_t \cup \bar{V}_t, \bar{T}_t \cup \bar{R}_t)$. Let $\tilde{G}_t = (\tilde{F}_t, \tilde{A}_t)$ be the graph called the partial interpretation graph defined by $\tilde{G}_t = (\bar{O}_t \cup \bar{P}_t \cup \bar{V}_{t-1}, \bar{T}_t \cup \bar{R}_{t-1})$. We define $\bar{O}_t$ (resp. $\bar{P}_t$, $\bar{V}_t$, $\bar{T}_t$, $\bar{R}_t$ and $\bar{G}_t$) as $\bar{O}_t = O_0 \cup ... \cup O_t$ where O_i (resp. P_i, V_i T_i, R_i and G_i) is the set of vertices (resp. vertices, vertices, arcs, arcs and graph) representing the concepts at time i.

3.2 Semantics

$\bar{O}_t$ is the set of vertices representing the objects of the scene at time t; either 2D areas or 3D equipment. $\bar{P}_t$ is the set of vertices representing the persons in the scene until time t. $\bar{V}_t$ is

the set of vertices representing the behaviours until time t. $\bar{T}_t$ is the set of arcs representing the temporal binary relations between two vertices f_1 and f_2: "f_1 at time $t-1$ is the same concept as f_2 at time t". $\bar{R}_t$ is the set of arcs representing the binary relations, between two vertices f_1 and f_2, "f_1 at time t refers to f_2".

We associate with each vertex $f \in \bar{F}_t$ a set of eight characteristics called attributes:

1. *name(f)* is a symbolic identifier,
2. *time(f)* is a temporal identifier,
3. *type(f)* is one of the categories: person, objects or one of the behaviour types,
4. *box(f)* is a 2D geometric description,
5. *hull(f)* is a 3D geometric description,
6. *velocity(f)* is a 3D velocity vector,
7. *properties(f)* is a set of symbolic characteristics,
8. *references(f)* is a set of pairs ($name$,$time$) for each binary arc in $\bar{R}_t$.

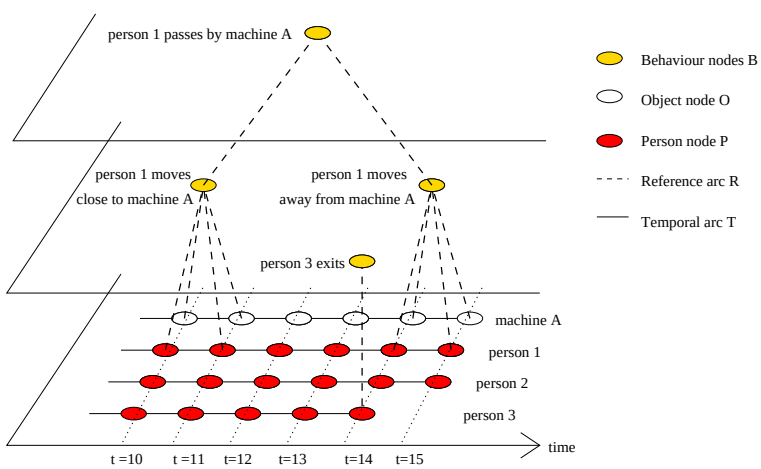

Figure 1: Example of interpretation graph

3.3 Properties

1. $\forall f_1, f_2 \in \bar{F}_t$: $f_1 = f_2 \Leftrightarrow name(f_1) = name(f_2)$ AND $time(f_1) = time(f_2)$

2. $\forall f_1, f_2 \in \bar{F}_t$: $\exists t(f_1, f_2) \in T_t \Leftrightarrow$ *name*(f_1) = *name*(f_2), that is to say that f_1 and f_2 represent the same real concept.

3. $\forall f_1, f_2 \in \bar{F}_t$: $f_1 \in$ *references*$(f_2) \Leftrightarrow \exists r \in \bar{R}_t$ such that r is an arc between f_1 and f_2 . We will denote by *refname*(f_1, i) (resp. *reftime*(f_1, i)) the *name* (resp. the *time*) of the ith *reference* of f_1.

3.4 Interpretation Process

We define the global interpretation process as the computation of $\bar{G}_{t+1}$ using the knowledge of $\bar{G}_t$, an image taken from the video stream at time t and a set of behaviour models. In other terms the global interpretation process is the computation of G_{t+1} ($\bar{G}_{t+1} = \bar{G}_t \cup G_{t+1}$). The computation of G_{t+1} is the computation of O_{t+1}, P_{t+1}, V_{t+1}, T_{t+1} and R_{t+1}. The computation of O_{t+1} is trivial because the number of objects of the scene and their characteristics are known at time 0 and are constant. P_{t+1} and T_{t+1} are obtained by person tracking and V_{t+1} and R_{t+1} are obtained by behaviour recognition.

3.5 Computing P_t and T_t

In this global context, the present paper focuses only on the computation of the set of vertices P_t and the set of arcs T_t from $\bar{G}_{t-1}$ and a set of blobs B_t computed by image processing.

Let $P_0(\bar{G}_{t-1}, B_t) = (P_t, T_t)$ be this problem. To characterise the solution of $P_0(\bar{G}_{t-1}, B_t)$, we will denote by "human" the physical phenomenon we want to describe and by $\mathcal{H}_t$ the set of all humans present in the scene at time t.

We also define the operator "correspond", noted $\equiv$, between a human h and a vertex $p \in P_t$ by: $h \equiv p$ **if and only if** the distance between $hull(p)$ and the minimal bounding cylinder of the human h in the scene is minimal.

We say that (P_t, T_t) is the solution to $P_0(\bar{G}_{t-1}, B_t)$ **if and only if** :

1. $\forall p_{i,t} \in P_t$: $\exists\,!\ h \in \mathcal{H}_t$ such that $h \equiv p_{i,t}$. In other terms, for each element $p_{i,t} \in P_t$ there exists a human in the scene at time t corresponding to the representation given by $p_{i,t}$.

2. $\forall h \in \mathcal{H}_t$: $\exists\,!\ p_{i,t} \in P_t$ such that $h \equiv p_{i,t}$. In other words, for each human in the scene at time t, there exists a P-representation $p_{i,t} \in P_t$ corresponding to this human.

3. $\forall t(x, y) \in T_t$, $x \in P_{t-1}$, $y \in P_t$: $\exists h \in \mathcal{H}_{t-1} \cap \mathcal{H}_t$ such that $x \equiv h \wedge y \equiv h$. In other words, for each relation $t(x, y) \in T_t$ the representations $x \in P_{t-1}$ and $y \in P_t$ correspond to the same human h, x at time $t-1$ and y at time t.

4. $\forall h \in \mathcal{H}_{t-1} \cap \mathcal{H}_t$: $\exists t(x, y) \in T_t$ such that $x \in P_{t-1} \wedge y \in P_t \wedge x \equiv h \wedge y \equiv h$. In other words, for each human present in the scene both at time $t-1$ and time t, there exists a binary relation $t(x, y) \in T_t$ such that x corresponds to this human at time $t-1$ and y to the same human at time t.

Let us now transform $P_0(\bar{G}_{t-1}, B_t) = (P_t, T_t)$ into: $M(\bar{G}_{t-1}, C(B_t)) = (P_t, T_t)$, where $C(B_t)$ is called the clustering problem and $M(\bar{G}_{t-1}, C(B_t))$ the temporal matching problem. We will see in section 4 how to solve $C(B_t)$ and in section 6 how to solve $M(\bar{G}_{t-1}, C(B_t))$.

4 Clustering

4.1 Clustering Problem

Let $B_t = \{b_{1,t}, ..., b_{k,t}\}$ be the set of blobs at time t. Each blob is defined by a vector $(x_{2D}, y_{2D}, w_{2D}, h_{2D}, a_{2D}) \in \mathbb{R}^5$, where x is the average x_{2D}-coordinate and y_{2D} the minimum y_{2D}-coordinate of all the pixels composing the blob and w_{2D} is the width, h_{2D} the height and a_{2D} the number of pixel of the blob.

Let us define the clustering problem $C(B_t)$ as a set partitioning problem. We say that the partition $Q_t = \{q_{1,t}, ..., q_{k,t}\} = \{\{b_{\alpha,t}, ..., b_{\beta,t}\}, ..., \{b_{\delta,t}, ..., b_{\gamma,t}\}\}$ is a solution to $C(B_t)$ **if and only if** :

1. **IF** 2 blobs $b_{i,t}$ and $b_{j,t}$ come from the projection of the same human **AND** $b_{i,t} \in q_{\lambda,t}$ **THEN** $b_{j,t} \in q_{\lambda,t}$.
2. **IF** a blob $b_{i,t}$ comes from any kind of noise (image noise, scene noise, shadow, etc ...) **AND** $b_{i,t} \in q_{\lambda,t}$ **THEN** $|q_{\lambda,t}| = 1$

With the first point, we want to group all the blobs which are the projection of a particular human in the same partition. With the second point, we want to isolate the noise blobs in distinct partitions.

4.2 Solving the Clustering Problem

We propose to solve $C(B_t)$ by a heuristic search on the set of all the possible partitions of B_t. We define the heuristic search as follows:

$$\begin{cases} e_0, \text{ the initial state of the search, is a partition composed of } k \text{ singleton sets} \\ e_{i+1} \in A(e_i) \text{ such that } \kappa(e_{i+1}) = MIN_{e_k \in A(e_i)}(\kappa(e_k)) \wedge \kappa(e_{i+1}) < \kappa(e_i) \\ e_f, \text{ the final state of the search} \end{cases}$$

I.e. e_{i+1} is the partition which minimises the heuristic κ on the set of all admissible partitions $A(e_i)$. $A(e_i)$ is the set of partitions that can be obtained from e_i by moving only one blob from a subset to another. The value of the heuristic is defined as follows:

$$\begin{aligned} \kappa(e_i) &= \sum_{q_{k,t} \in e_i} f(q_{k,t}) \\ f(q_{k,t}) &= \begin{cases} \lambda & \text{if} \quad q_{k,t} = \emptyset \\ \frac{\Delta_m(q_{k,t})}{d(q_{k,t})} & \text{if} \quad q_{k,t} \neq \emptyset \end{cases} \end{aligned}$$

where $\Delta_m(q_{k,t})$ is the $\mathbb{R}^4$ distance of a given partition $q_{k,t}$ to an *a priori* model of a human defined by $(w^m_{2D}, h^m_{2D}, W^m_{3D}, H^m_{3D})$ and $d(q_{k,t})$ is the density of mobile pixels in $q_{k,t}$:

$$\begin{aligned}\Delta_m(q_{k,t}) &= \left[(w^m_{2D} - w_{2D}(q_{k,t}))^2 + (h^m_{2D} - h_{2D}(q_{k,t}))^2\right. \\ &\quad \left. +(W^m_{3D} - W_{3D}(q_{k,t}))^2 + (H^m_{3D} - H_{3D}(q_{k,t}))^2\right]^{1/2} \\ d(q_{k,t}) &= \frac{a_{2D}(q_{k,t})}{w_{2D}(q_{k,t}).h_{2D}(q_{k,t})}\end{aligned}$$

Then $C(B_t) = \{q_{k,t} \in e_f \quad | \quad \Delta_m(q_{k,t}) < \Delta_{max}\}$, where Δ_{max} is the threshold defining the maximal admissible distance to the *a priori* human model. We will see later the influence of the value of Δ_{max}.

Note that:

- The *a priori* human model is obtained by regression of a learning set to k vectors from $\mathbb{R}^4$ using the mobile centers algorithm ("Algorithme des nuées dynamiques"). ($k \simeq 10$).
- $d(q_{k,t})$ is enhanced by a gamma function in order to control the influence of the density on the heuristic.
$$d(q_{k,t}) = e^{\frac{1}{\gamma} ln(\frac{a_{2D}(q_{k,t})}{w_{2D}(q_{k,t})h_{2D}(q_{k,t})})}$$
- The value of λ is not constant during the heuristic search. Its value is computed for each e_{i+1} as follows:
$$\lambda = \frac{1}{k} \sum_{q_{k,t} \in e_i} \Delta_m(q_{k,t})$$

4.3 Analysis of the proposed solution

The influence of Δ_{max} on the correctness of the results is important. If Δ_{max} is big then a lot of subsets of the final state of the search will be kept. On the other hand, if Δ_{max} is small then only few subsets of the final state of the search will be kept. In other terms, if Δ_{max} is big then a lot of subsets even far from the human model will be considered as human, so the number of humans not recognised as such (failed recognition) will be small and the number of non-human entities considered as humans (false recognition) will be high. If Δ_{max} is small then the number of failed recognitions will be high and the number of false recognition will be small.

The influence of the value of γ in the gamma function on the density is also important. The smaller γ is, the more the influence of the density on the heuristic κ is high and the smaller the number of subsets with more than one blob will be in the final state of the search. At the limit ($\gamma \to 0$), the final state is equal to the initial state ($e_0 = e_f$). The larger γ is, the smaller the influence of the density is on the heuristic κ and the higher the number of empty subsets will be in the final state of the search. At the limit ($\gamma \to \infty$), the final state is a partition composed of $b - 1$ empty subsets and one subset with all the b blobs.

The complexity depends on two points. The first point is the cost of the passage from one state e_i to the next e_{i+1} and the second point is the number f of states e_i until the final state e_f. The cost of the passage from one state to the next is $|B_t|^2 = b^2$. So the complexity of our approach is fb^2.

We cannot strictly prove that our approach converges in all possible cases. We can only prove properties which enable us to say that the heuristic h will prefer certain good partitions to other less good partitions.

5 Clustering Results

We present in this section some results of our approach. In table 1, we present the results of the method compared with the aim of the method. I.e. we consider as an error every misclassified blob. If a blob corresponding to noise is in a subset of the partition corresponding to a human, this blob is misclassified. If a blob corresponding to some part of a human is not in the subset of the partition corresponding to the rest of this human, this blob is misclassified.

Video id.	frames	misclassifieds	blobs	% misclassified
st1-23	190	17	951	1.78
c02-2	535	136	4268	3.18
mc2-17	197	290	6073	4.77
va2-7	55	25	451	5.54
va2-4	340	364	5888	6.18
B008	570	365	5840	6.25
c07-2	137	49	590	8.30
mc1-22	153	106	1140	9.29
va2-6	322	275	2363	11.63
TOTAL	2499	1627	27564	5.90

Table 1: Results of the method in terms of misclassifications

We can give two explanations to these errors. A human entering or exiting the scene is only partially visible and does not correspond to our *a priori* model of a human. The second source of errors are the little pieces of shadow which are often found close to a human. The corresponding blobs are often misclassified even if those misclassifications do not really change the description of the corresponding human.

The tables 2 and 3 present the results of the method in terms of what we expect. That is to say a description of the humans present in the scene at each frame. This description can be judged at two levels. The first level is the number of resulting partitions (i.e. the number of humans). The second level is the correctness of the description of each human. The first level is very important because a wrong number of resulting partition can seriously corrupt the results of the matching step. Table 2 presents the results in terms of failed recognitions

(false negative), when there is no subset in the partition corresponding to a given human. Table 3 presents the results in terms of false recognitions (false positive), when there is no human corresponding to a given subset of the partition.

Video id.	frames	# of failed	humans	% failed
c02-2	535	3	906	0.33
st1-23	190	1	180	0.55
c07-2	137	1	120	0.83
mc2-17	197	4	354	1.12
mc1-22	153	5	217	2.30
B008	570	21	782	2.68
va2-7	55	4	138	2.89
va2-6	322	16	345	4.63
va2-4	340	65	304	21.38
TOTAL	2499	120	3346	3.58

Table 2: Results of the method in terms of failed recognitions

The only source of error causing failed recognition is when a partition corresponding to a given human is too far from our *a priori* model. This is often the case when a partition corresponds to more than one person (groups, occlusion, etc.).

Video Description	frames	# of false	humans	% false
mc2-17	197	0	354	0.00
B008	570	0	782	0.00
va2-7	55	0	138	0.00
c02-2	535	1	906	0.11
va2-6	322	1	345	0.28
st1-23	190	1	180	0.55
c07-2	137	1	120	0.83
va2-4	340	10	304	3.28
mc1-22	153	10	217	4.60
TOTAL	2499	24	3346	0.71

Table 3: Results of the method in terms of false recognitions

The principal cause of false recognition is an excess of blobs corresponding to noise. We can see that the results in terms of false recognitions are better than those in terms of failed recognitions. This point is important because the problems caused by the false and failed recognitions are very different for the temporal matching problem and the method described in the following section is more robust to failed recognitions than to false ones.

6 Temporal Matching

6.1 Definitions of the Matching Functions

We define the notion of a matching function Φ_j as a function from either P_{t-1}, Q_t or $P_{t-1} \times Q_t$ onto $P_t \times T_t$ representing a precise configuration involving a human $h \in \mathcal{H}_{t-1} \cap \mathcal{H}_t$. The sets Q_t, P_{t-1}, P_t and T_t are defined as follows:

$Q_t = \{q_{1,t}, ..., q_{m,t}\} = C(B_t)$ is the solution to the clustering problem (i.e. a partitioning of the set of blobs). We will denote by m the cardinality of Q_t. We associate to any $q \in Q_t$ two representations *box*(q) and *hull*(q) where *box*(q) is a rectangle in the image directly defined by $x_{2D}(q), y_{2D}(q), w_{2D}(q)$ and $h_{2D}(q)$ and *hull*(q) is a 3D cylinder obtained by the projection of *box*(q) into the scene O_t. We also extend the definition of the operator $\equiv$ given in section 3 for a vertex $p \in \bar{P}_t$ to a set $q \in Q_t$ with the exception that we will permit two different humans $h_1, h_2 \in \mathcal{H}_t$ to correspond to the same set $q \in Q_t$ in the case of occlusions.

P_t (resp. P_{t-1}) is the set of vertices of $\bar{G}_t$ of type *person* at time t (resp. $t-1$). We will denote by n the cardinality of P_{t-1}.

$T_t = \{t(p_{\alpha,t-1}, p_{\beta,t}), ..., t(p_{\gamma,t-1}, p_{\delta,t})\}$ is the set of arcs of $\bar{G}_t$ between an element of P_{t-1} and an element of P_t.

In the following we will define a set of seven functions Φ_j corresponding to possible configurations of the temporal matches. In each case, we will specify the result of the function Φ_j as a pair $(p_{k,t}, t(p_{i,t-1}, p_{k,t}))$ where $p_{k,t} \in P_t$ and $t(p_{i,t-1}, p_{k,t}) \in T_t$. We will also specify the particular conditions of application of Φ_j in terms of correspondence between the physical phenomena and the different kinds of representations we have. Finally, we will specify a way of evaluating whether those conditions are verified. We will denote by $e(\Phi_j) \in [0, 1]$ the evaluation of Φ_j.

In the general case, $\Phi_j(p_{i,t-1}, q_{j,t}) = (p_{k,t}, t(p_{i,t-1}, p_{k,t}))$ such that:

$$
\begin{array}{lcl}
\mathit{name}(p_{k,t}) & = & \begin{cases} \mathit{name}(p_{i,t-1}) & \text{if} \quad t(p_{i,t-1}, p_{k,t}) \in T_t \\ \text{a new name} & \text{if} \quad t(p_{i,t-1}, p_{k,t}) \notin T_t \end{cases} \\
\mathit{time}(p_{k,t}) & = & t \\
\mathit{type}(p_{k,t}) & = & \mathit{person} \\
\mathit{box}(p_{k,t}) & = & \mathit{box}(q_{j,t}) \\
\mathit{hull}(p_{k,t}) & = & \mathit{hull}(q_{j,t}) \\
\mathit{velocity}(p_{k,t}) & = & \text{estimated 3D velocity vector} \\
\mathit{properties}(p_{k,t}) & = & \emptyset \\
\mathit{references}(p_{k,t}) & = & \emptyset
\end{array}
$$

In other words: in the general case $p_{k,t}$ is the temporal continuation in frame t of the track $p_{i,t-1}$ using the information (the match) $q_{j,t}$ from frame t.

The seven matching functions are MATCH, LOST, NOISE, ENTRY, EXIT, HIDDEN and APPEARS and they are defined as follows:

- Φ_1: MATCH$(p_{i,t-1}, q_{j,t}) = (p_{k,t}, t(p_{i,t-1}, p_{k,t}))$ **if and only if** $p_{i,t-1}$ and $q_{j,t}$ correspond to the same human at times $t-1$ and t. We determine whether the conditions of

this case are verified by the similarity between $p_{i,t-1}$ and $q_{j,t}$: $e(\text{MATCH}(p_{i,t-1}, q_{j,t})) = S(p_{i,t-1}, q_{j,t})$ where S is the similarity function detailed in section 6.2.

$$\text{MATCH}(p_{i,t-1}, q_{j,t}) = (p_{k,t}, t(p_{i,t-1}, p_{k,t})) \textbf{ if and only if } \exists h \in \mathcal{H}_{t-1} \cap \mathcal{H}_t \text{ such that } p_{i,t-1} \equiv h \wedge q_{j,t} \equiv h$$

- Φ_2: $\text{LOST}(p_{i,t-1}) = (p_{k,t}, t(p_{i,t-1}, p_{k,t}))$ **if and only if** there is no $q_{j,t} \in Q_t$ corresponding to the human described by $p_{i,t-1}$. The significance of $p_{k,t}$ is explained at the end of this section. This case is always possible so $e(\text{LOST}(p_{i,t-1})) = 1$.

$$\text{LOST}(p_{i,t-1}) = (p_{k,t}, t(p_{i,t-1}, p_{k,t})) \textbf{ if and only if } \exists h \in \mathcal{H}_{t-1} \cap \mathcal{H}_t \text{ such that } p_{i,t-1} \equiv h \wedge \forall q \in Q_t : q \not\equiv h$$

- Φ_3: $\text{NOISE}(q_{j,t}) = (\emptyset, \emptyset)$ **if and only if** $q_{j,t}$ does not correspond to any human. This case is always possible so $e(\text{NOISE}(q_{j,t})) = 1$

$$\text{NOISE}(q_{j,t}) = (\emptyset, \emptyset) \textbf{ if and only if } \forall h \in \mathcal{H}_t : q_{j,t} \not\equiv h$$

- Φ_4: $\text{EXIT}(p_{i,t-1}) = (\emptyset, \emptyset)$ **if and only if** the human corresponding to the vertex $p_{i,t-1}$ leaves the scene at time t, so there is no corresponding $q_{j,t}$. We evaluate the conditions of this case by comparing the location of $p_{i,t-1}$ in the scene with a set of predefined areas where humans can leave the scene. In other terms, $e(\text{EXIT}(p_{i,t-1})) = 1$ if the predicate $\text{IsInIO}(p_{i,t-1}) = \text{TRUE}$.

$$\text{EXIT}(p_{i,t-1}) = (\emptyset, \emptyset) \textbf{ if and only if } \exists h \in \mathcal{H}_{t-1} \setminus \mathcal{H}_t \text{ such that } p_{i,t-1} \equiv h$$

- Φ_5: $\text{ENTRY}(q_{j,t}) = (p_{k,t}, \emptyset)$ **if and only if** the human corresponding to the partition $q_{j,t}$ was not present in the scene at time $t-1$. We evaluate the conditions of this case by comparing the location of $q_{j,t}$ in the scene with a set of predefined areas where humans can enter the scene. In other terms, $e(\text{ENTRY}(q_{j,t})) = 1$ if the predicate $\text{IsInIO}(q_{j,t}) = \text{TRUE}$.

$$\text{ENTRY}(q_{j,t}) = (p_{k,t}, \emptyset) \textbf{ if and only if } \exists h \in \mathcal{H}_t \setminus \mathcal{H}_{t-1} \text{ such that } q_{j,t} \equiv h$$

- Φ_6: $\text{HIDDEN}(p_{i,t-1}, q_{j,t}) = (p_{k,t}, t(p_{i,t-1}, p_{k,t}))$ **if and only if** the human corresponding to the vertex $p_{i,t-1}$ is occluded by another human described by $q_{j,t}$. Note that in general an instance of $\text{HIDDEN}(p_{i,t-1}, q_{j,t})$ is accompanied by an instance of

MATCH($p_{i',t-1}, q_{j,t}$) corresponding to the *occluding* person (described by $q_{j,t}$). The two tracks $p_{i,t-1}$ and $p_{i',t-1}$ are thus matched to the same description $q_{j,t}$ at time t and will continue together (as a HIDDEN/MATCH pair) until none of the two humans occludes the other. We evaluate the conditions of this case by comparing the locations of $p_{i,t-1}$ and $q_{j,t}$. In other words $e(\text{HIDDEN}(p_{i,t-1}, q_{j,t})) = 1$ if the predicate OCCLUDES($p_{i,t-1}, q_{j,t}$)) = TRUE.

$$\boxed{\begin{array}{c}\text{HIDDEN}(p_{i,t-1}, q_{j,t}) = (p_{k,t}, t(p_{i,t-1}, p_{k,t})) \textbf{ if and only if} \\ \exists h_1, h_2 \in \mathcal{H}_{t-1} \cap \mathcal{H}_t \text{ such that} \\ p_{i,t-1} \equiv h_1 \wedge p_{i',t-1} \equiv h_2 \quad \wedge \quad q_{j,t} \equiv h_1 \wedge q_{j,t} \equiv h_2\end{array}}$$

- Φ_7: APPEARS($p_{i,t-1}, q_{j,t}$) = ($p_{k,t}, \emptyset$) **if and only if** the human corresponding to the partition $q_{j,t}$ was hidden at time $t-1$ by another human described by $p_{i,t-1}$. Note that this case only applies when two persons have entered the scene occluding each other and then separates. At the end of an occlusion that has been correctly handled by previous instances of HIDDEN, the two superimposed tracks will separate without the need for an instance of APPEARS. We evaluate the conditions of APPEARS by comparing the locations of $p_{i,t-1}$ and $q_{j,t}$. In other words $e(\text{APPEARS}(p_{i,t-1}, q_{j,t})) = 1$ if the predicate OCCLUDES($p_{i,t-1}, q_{j,t}$)) = TRUE.

$$\boxed{\begin{array}{c}\text{APPEARS}(p_{i,t-1}, q_{j,t}) = (p_{k,t}, \emptyset) \textbf{ if and only if} \\ \exists h \in \mathcal{H}_{t-1} \cap \mathcal{H}_t \text{ such that } q_{j,t} \equiv h \wedge \forall p \in P_{t-1} : p \not\equiv h\end{array}}$$

Note: We said that the *hull* and *box* characteristics of $p_{k,t}$ resulting from a matching function are computed with the help of a subset $q \in Q_t$ of blobs, but in the case of LOST (Φ_2), there is no q associated to this function. In this case the *hull* and *box* characteristics of $p_{k,t}$ must be extrapolated using the *hull* and *box* of $p_{i,t-1}$. To accomplish this there are several strategies. One strategy, called the stationary strategy, is to use $hull(p_{k,t}) = hull(p_{i,t-1})$ and $box(p_{k,t}) = box(p_{i,t-1})$. Another strategy, called the predictive strategy, consists in using as $hull(p_{k,t})$ (resp. $box(p_{k,t})$) the result of a Kalman filter on the previous $hull(p_{k,t-j})$ (resp. $box(p_{k,t-j})$).

The predicate IsInIO($p_{i,t-1}$) is defined w.r.t. the elements of O_t and the predicate OCCLUDES($p_{i,t-1}, q_{j,t}$) is defined w.r.t. the relative location of $p_{i,t-1}, q_{j,t}$ and the camera.

6.2 The computation of the similarity function

We define the similarity function $\mathcal{S}$ between a set $q \in Q_t$ of blobs and a vertex p by:

$$
\begin{aligned}
S(p,q) &= \sum_i \omega_{i,2D} e^{\frac{-d_{2D}(box(q),D_i(box(p)))}{\sigma_{i,2D}}} + \sum_i \omega_{i,3D} e^{\frac{-d_{3D}(hull(q),D_i(hull(p)))}{\sigma_{i,3D}}} \\
\sum_i \omega_i &= 1
\end{aligned}
$$

where d_{2D} (resp. d_{3D}) is the 2D distance (resp. 3D distance) between $box(q)$ (resp. $hull(q)$) and $D_i(box(p))$ (resp. $D_i(hull(p))$) where D_i is the ith way of estimation (average, median, Kalman, ...).

6.3 Matching Diagnosis Problem

Let $I(Q_t, P_{t-1})$ be the problem defined by $I(Q_t, P_{t-1}) = \{\Phi_\alpha(p_{\beta,t-1}, q_{\gamma,t}), ..., \Phi_\delta(p_{\mu,t-1})\}$ such that $\Phi_\alpha(p_{\beta,t-1}, q_{\gamma,t}) \cup ... \cup \Phi_\delta(p_{\mu,t-1}) = M(Q_t, P_{t-1})$. In other words, the problem I is to find a set of applications of functions from $\{\Phi_1, ..., \Phi_7\}$ to the elements of P_{t-1} and Q_t such that the union of the results of those applications is the solution to $M(Q_t, P_{t-1})$. We will denote by diagnosis any set of applications of functions from $\{\Phi_1, ..., \Phi_7\}$ to the elements of P_{t-1} and Q_t. We also define the evaluation of a diagnosis $\mathcal{I}_t = \{\Phi_\alpha(p_{\beta,t-1}, q_{\gamma,t}), ..., \Phi_\delta(p_{\mu,t-1})\}$ as the weighted sum of the evaluations of the matching functions of $\mathcal{I}_t$. That is to say that: $e(\mathcal{I}_t) = \lambda_\alpha e(\Phi_\alpha(p_{\beta,t-1}, q_{\gamma,t})) + ... + \lambda_\delta e(\Phi_\delta(p_{\mu,t-1}))$ where $\{\lambda_1, ..., \lambda_7\}$ represent the importance associated to each case. E.g. we always prefer having as many MATCH(x, y) as possible (i.e. $\forall k \neq 1 : \lambda_1 > \lambda_k$) and we always prefer having as few LOST(x) and NOISE(x) as possible (i.e. $\forall k \in \{1, 2, 3, 4, 5\} : \lambda_6 < \lambda_k \wedge \lambda_7 < \lambda_k$).

Figure 2 is an example of a problem $I(Q_t, P_{t-1})$ at time $t = 11$ with $Q_t = \{q_{1,11}, q_{2,11}, q_{3,11}\}$ and $P_{t-1} = \{p_{1,10}, p_{2,10}, p_{3,10}\}$. In this case:

$$
\begin{aligned}
I(Q_t, P_{t-1}) &= \{\text{MATCH}(p_{3,10}, q_{3,11}), \text{MATCH}(p_{1,10}, q_{2,11}), \\
&\quad \text{HIDDEN}(p_{2,10}, q_{2,11}), \text{ENTRY}(q_{1,11})\} \\
\text{MATCH}(p_{3,10}, q_{3,11}) &= (p_{1,11}, t(p_{3,10}, p_{1,11})) \\
\text{MATCH}(p_{1,10}, q_{2,11}) &= (p_{2,11}, t(p_{1,10}, p_{2,11})) \\
\text{HIDDEN}(p_{2,10}, q_{2,11}) &= (p_{3,11}, t(p_{2,10}, p_{3,11})) \\
\text{ENTRY}(q_{1,11}) &= (p_{4,11}, \emptyset)
\end{aligned}
$$

I.e. the solution to the temporal matching problem is $M(Q_t, P_{t-1}) = (P_t, T_t)$ with $P_t = \{p_{1,11}, p_{2,11}, p_{3,11}, p_{4,11}\}$ and $T_t = \{t(p_{3,10}, p_{1,11}), t(p_{1,10}, p_{2,11}), t(p_{2,10}, p_{3,11})\}$.

We will now define the notion of a valid or invalid diagnosis. We say that a diagnosis is invalid when two or more matching functions have incompatible conditions. For example $\{$EXIT(a), HIDDEN$(a)\}$ is an invalid diagnosis because the first function applies when the vertex $a \equiv h$ with h non-present in the scene at time t and the second function applies when the vertex $a \equiv h$ with h present (but occluded) in the scene at time t. We define the notion of a valid diagnosis as any non-invalid diagnosis.

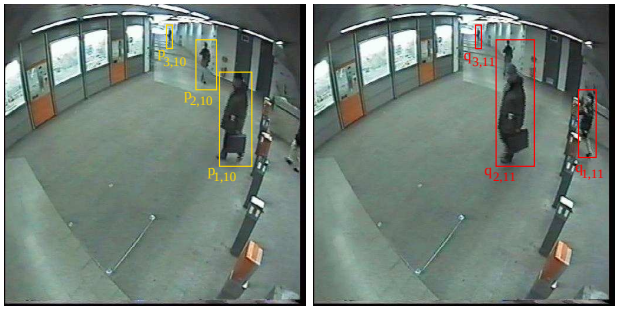

Figure 2: Example of the problem $I(Q_t, P_{t-1})$ at time $t = 11$

6.3.1 Theorem 1

Let $\mathcal{I}_t$ be the solution to $I(Q_t, P_{t-1})$. $\mathcal{I}_t$ is a valid diagnosis **if and only if** $\forall a, c \in P_{t-1}, a \neq c$ and $\forall b, d \in Q_t, b \neq d$, none of the following cases exist:

1. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{MATCH}(a, d), ...\}$
2. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{HIDDEN}(a, d), ...\}$
3. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{LOST}(a), ...\}$
4. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{EXIT}(a), ...\}$
5. $\mathcal{I}_t = \{...,\text{HIDDEN}(a, b),\text{LOST}(a), ...\}$
6. $\mathcal{I}_t = \{...,\text{HIDDEN}(a, b),\text{EXIT}(a), ...\}$
7. $\mathcal{I}_t = \{...,\text{LOST}(a),\text{EXIT}(a), ...\}$
8. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{MATCH}(c, b), ...\}$
9. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{APPEARS}(c, b), ...\}$
10. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{NOISE}(a), ...\}$
11. $\mathcal{I}_t = \{...,\text{MATCH}(a, b),\text{ENTRY}(a), ...\}$
12. $\mathcal{I}_t = \{...,\text{APPEARS}(a, b),\text{NOISE}(b), ...\}$
13. $\mathcal{I}_t = \{...,\text{APPEARS}(a, b),\text{ENTRY}(b), ...\}$
14. $\mathcal{I}_t = \{...,\text{NOISE}(b),\text{ENTRY}(b), ...\}$

6.3.2 Proof

We can prove this theorem in three steps:

1. Firstly, we can prove that each item provides an invalid diagnosis.
2. Then, we can prove that the cardinality of a valid diagnosis in bounded.
3. Finally, we can prove that only the forteen previous cases lead to invalid diagnosis.

For the first step, we will only give the proof for the last item which is $\mathcal{I}_t = \{...,\text{NOISE}(b),\text{ENTRY}(b),...\}$ and certify that the other proofs are similar.

According to the definition of the matching functions NOISE and ENTRY, we know that NOISE(b) can be applied **if and only if** $\forall h \in \mathcal{H}_t : b \not\equiv h$ and ENTRY(b) can be applied **if and only if** $\exists h \in \mathcal{H}_t \setminus \mathcal{H}_{t-1}$ such that $b \equiv h$. Those two conditions are incompatible, so NOISE(b) and ENTRY(b) cannot be applied at the same time. In other words, $\mathcal{I}_t = \{...,\text{NOISE}(b),\text{ENTRY}(b),...\}$ is an invalid diagnosis.

To prove the second step, we note that for a given $p \in P_{t-1}$ (resp. $q \in Q_t$), p cannot be involved in two matching functions among MATCH, HIDDEN, LOST or EXIT (resp. MATCH, APPEARS, NOISE or ENTRY) in the same diagnosis. Now let a be the number of instances of MATCH, n the cardinality of P_{t-1} and m the cardinality of Q_t. The cardinality of a valid diagnosis is then $a + (m - a) + (n - a)$. If $a = 0$, the cardinality of a valid diagnosis is $n + m$. If $a = \text{MIN}(m, n)$ the cardinality of a valid diagnosis is $n + m - \text{MIN}(n, m) = \text{MAX}(n, m)$. We can therefore conclude that the cardinality of a valid diagnosis $\mathcal{I} = I(Q_t, P_{t-1})$ is bounded by the interval $[\text{MAX}(m, n), m + n]$.

The last step of the proof is carried out using the second point. The fact that the cardinality of a valid diagnosis is bounded implies that the number of diagnosis is finite and known. We do not want to enumerate all the valid diagnosis, but we certify that only the forteen previous cases are invalid.

6.4 Solving the Matching Diagnosis Problem

In this subsection, we propose a numerical schema to solve the Matching Diagnosis Problem. Solving the Matching Diagnosis Problem is to find the best evaluated diagnosis $\mathcal{I}$ among all the valid diagnoses. In other words, solving the Matching Diagnosis Problem is to find, among all the valid diagnoses, the diagnosis

$$\begin{aligned}\mathcal{I} \;=\; & \{\Phi_{\alpha_1}(p_{\beta_1,t-1}, q_{\gamma_1,t}), \ldots, \Phi_{\alpha_\rho}(p_{\beta_\rho,t-1}, q_{\gamma_\rho,t}),\\ & \Phi_{\zeta_1}(q_{\xi_1,t}), \ldots, \Phi_{\zeta_\sigma}(q_{\xi_\sigma,t}),\\ & \Phi_{\delta_1}(p_{\mu_1,t-1}), \ldots, \Phi_{\delta_\omega}(p_{\mu_\omega,t-1})\}\end{aligned}$$

such that the evaluation

$$\begin{aligned}&\lambda_{\alpha_1} e(\Phi_{\alpha_1}(p_{\beta_1,t-1}, q_{\gamma_1,t})) + \ldots + \lambda_{\alpha_\rho} e(\Phi_{\alpha_\rho}(p_{\beta_\rho,t-1}, q_{\gamma_\rho,t})) +\\ &\lambda_{\zeta_1} e(\Phi_{\zeta_1}(q_{\xi_1,t})) + \ldots + \lambda_{\zeta_\sigma} e(\Phi_{\zeta_\sigma}(q_{\xi_\sigma,t})) +\\ &\lambda_{\delta_1} e(\Phi_{\delta_1}(p_{\mu_1,t-1})) + \ldots + \lambda_{\delta_\omega} e(\Phi_{\delta_\omega}(p_{\mu_\omega,t-1}))\end{aligned}$$

is maximal.
Let χ be the application from $(P^*_{t-1} \times Q^*_t)$ to the set of all the diagnoses such that: if $n > m$

$$\begin{aligned}P^*_{t-1} &= \{p_1, \ldots, p_n, q^*_1, \ldots, q^*_m, q^{**}_1, \ldots, q^{**}_m, q^{***}_1, \ldots, q^{***}_m, q^{****}_1, \ldots, q^{****}_{2(n-m)}\}\\ Q^*_t &= \{q_1, \ldots, q_m, p^*_1, \ldots, p^*_n, p^{**}_1, \ldots, p^{**}_n, p^{***}_1, \ldots, p^{***}_n\}\end{aligned}$$

if $n < m$

$$\begin{aligned}P^*_{t-1} &= \{p_1, \ldots, p_n, q^*_1, \ldots, q^*_m, q^{**}_1, \ldots, q^{**}_m, q^{***}_1, \ldots, q^{***}_m\}\\ Q^*_t &= \{q_1, \ldots, q_m, p^*_1, \ldots, p^*_n, p^{**}_1, \ldots, p^{**}_n, p^{***}_1, \ldots, p^{***}_n, p^{****}_1, \ldots, p^{****}_{2(m-n)}\}\end{aligned}$$

if $m = n$

$$\begin{aligned}P^*_{t-1} &= \{p_1, \ldots, p_n, q^*_1, \ldots, q^*_m, q^{**}_1, \ldots, q^{**}_m, q^{***}_1, \ldots, q^{***}_m\}\\ Q^*_t &= \{q_1, \ldots, q_m, p^*_1, \ldots, p^*_n, p^{**}_1, \ldots, p^{**}_n, p^{***}_1, \ldots, p^{***}_n\}\end{aligned}$$

and

$$\chi(x,y) = \begin{cases} \text{MATCH}(p_i, q_j) & \text{if} \quad x = p_i \quad \text{and} \quad y = q_j\\ \text{HIDDEN}(p_i, q_j) & \text{if} \quad x = p_i \quad \text{and} \quad y = p^*_i\\ \text{APPEARS}(p_i, q_j) & \text{if} \quad x = q^*_j \quad \text{and} \quad y = q_j\\ \text{LOST}(p_i) & \text{if} \quad x = p_i \quad \text{and} \quad y = p^{**}_i\\ \text{NOISE}(q_j) & \text{if} \quad x = q^{**}_j \quad \text{and} \quad y = q_j\\ \text{EXIT}(p_i) & \text{if} \quad x = p_i \quad \text{and} \quad y = p^{***}_i\\ \text{ENTRY}(q_j) & \text{if} \quad x = q^{***}_j \quad \text{and} \quad y = q_j\\ \emptyset & \text{otherwise} \end{cases}$$

where the p_i are the elements of P_{t-1}, the q_j are the elements of Q_t, the $q_i^*, q_j^{**}, q_j^{***}, q_j^{****}$ are some notation artefacts and the $p_i^*, p_i^{**}, p_i^{***}, p_i^{****}$ are some notation artefacts too.
We represent the application χ by the matrix $\mathcal{M}_\chi$ shown in figure 3.
In the same manner, let $e\chi$ be the application from $(P_{t-1}^* \times Q_t^*)$ to $\mathbb{R}$ such that $e\chi(x,y) = e(\chi(x,y))$. We also represent the application $e\chi$ by the matrix $e\mathcal{M}_\chi$.

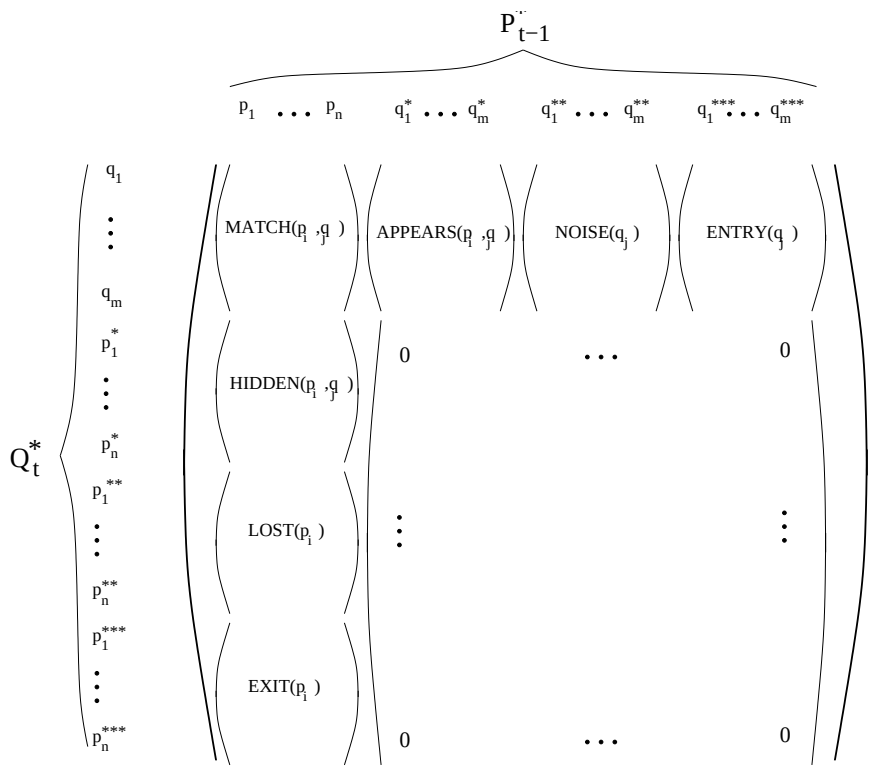

Figure 3: $\mathcal{M}_\chi$: matricial representation of the application χ (case $m = n$

Let $d\chi$ be the application which associates to any bijection f from P_{t-1}^* to Q_t^* a diagnosis $d\chi(f)$ defined by:

$$d\chi(f) = \{\chi(x, f(x)) \quad \forall x \in P_{t-1}^*\}$$

6.4.1 Theorem 2

$d\chi(f)$ is a valid diagnosis for every bijection f from P_{t-1}^* to Q_t^* **AND** for every valid diagnosis $\mathcal{I}$, there exists a bijection f from P_{t-1}^* to Q_t^* such that $d\chi(f) = \mathcal{I}$.

6.4.2 Proof

f is a bijection is equivalent to say that there do not exist an element a of P^*_{t-1} such that $f(a) = b$ and $f(a) = d$ AND there do not exist an element b of Q^*_t such that $f(c) = b$ and $f(c) = b$.

In the first case, we won't have a corresponding diagnosis like : {..., MATCH(a, b), MATCH(a, d), ...} or {..., MATCH(a, b), HIDDEN(a, d), ...} or {..., MATCH(a, b), LOST(a), ...} or {..., MATCH(a, b), EXIT(a), ...} or {..., HIDDEN(a, b), LOST(a), ...} or {..., HIDDEN(a, b), EXIT(a), ...} or {..., LOST(a), EXIT(a), ...}. All those cases are invalid (cf. theorem 1). In the second case, we won't have a corresponding diagnosis like : {..., MATCH(a, b), MATCH(c, b), ...} , or {..., MATCH(a, b), APPEARS(c, b), ...} or {..., MATCH(a, b), NOISE(a), ...} or {..., MATCH(a, b), ENTRY(a), ...} or {..., APPEARS(a, b), NOISE(b), ...} or {..., APPEARS(a, b), ENTRY(b), ...} or {...,NOISE(b),ENTRY(b), ...}. All those cases are invalid too (cf. theorem 1). That is to say that f is a bijection is equivalent to the fact that $d\chi(f)$ is a valid diagnosis.

Finally, the best evaluated valid diagnosis is found by searching the bijection f in the set $\mathcal{B}$ of all possible bijections from P^*_{t-1} to Q^*_t such that:

$$e(d\chi(f) = \mathrm{MAX}_{b\in\mathcal{B}}(e(d\chi(b)))$$

The rows and columns of $e\mathcal{M}_\chi$ that are all zeros are removed by pair. If the dimension of this square matrix is denoted d, there are $d!$ different bijections. The optimal bijection is found by a branch and bound method that in general evaluates much less than $d!$ different bijections.

For example, on the case shown in figure 2, the matrix $e\mathcal{M}_\chi$ is:

$$\begin{array}{c} \\ q_1 \\ q_2 \\ q_3 \\ p_1^* \\ p_2^* \\ p_3^* \\ p_1^{**} \\ p_2^{**} \\ p_3^{**} \\ p_1^{***} \\ p_2^{***} \\ p_3^{***} \end{array}
\begin{array}{c}
\begin{array}{cccccccccccc} p_1 & p_2 & p_3 & q_1^* & q_2^* & q_3^* & q_1^{**} & q_2^{**} & q_3^{**} & q_1^{***} & q_2^{***} & q_3^{***} \end{array} \\
\left(\begin{array}{cccccccccccc}
10 & 1 & 1 & 0 & 0 & 0 & 10 & 0 & 0 & 50 & 0 & 0 \\
70 & 1 & 1 & 0 & 25 & 0 & 0 & 10 & 0 & 0 & 0 & 0 \\
1 & 5 & 80 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 50 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right)
\end{array}$$

Where $\lambda_1 = 100$, $\lambda_2 = \lambda_3 = 25$, $\lambda_4 = \lambda_5 = 10$ and $\lambda_6 = \lambda_7 = 50$. After the removal of 0 rows/columns, the matrix $e\mathcal{M}_\chi$ become:

$$\begin{array}{c} \\ q_1 \\ q_2 \\ q_3 \\ p_2^* \\ p_1^{**} \\ p_2^{**} \\ p_3^{**} \\ p_3^{***} \end{array} \begin{array}{c} \begin{array}{cccccccc} p_1 & p_2 & p_3 & q_2^* & q_1^{**} & q_2^{**} & q_3^{**} & q_1^{***} \end{array} \\ \begin{pmatrix} 10 & 1 & 1 & 0 & 10 & 0 & 0 & 50 \\ 70 & 1 & 1 & 25 & 0 & 10 & 0 & 0 \\ 1 & 5 & 80 & 0 & 0 & 0 & 10 & 0 \\ 0 & 25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 50 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

The maximal bijection is then $e\chi(p_1, q_2)+$ $e\chi(p_2, p_2^*)+$ $e\chi(p_3, q_3)+$ $e\chi(q_1^{***}, q_1)+$ some terms equal to 0 where

$$\begin{aligned} e\chi(p_1, q_2) &= e(\mathrm{MATCH}(p_1, q_2)) \\ e\chi(p_2, p_2^*) &= e(\mathrm{HIDDEN}(p_2, q_2)) \\ e\chi(p_3, q_3) &= e(\mathrm{MATCH}(p_3, q_3)) \\ e\chi(q_1^{***}, q_1) &= e(\mathrm{ENTRY}(q_1)) \\ \mathcal{I} &= \{\mathrm{MATCH}(p_3, q_3), \mathrm{MATCH}(p_1, q_2), \\ &\quad \mathrm{HIDDEN}(p_2, q_2), \mathrm{ENTRY}(q_1)\} \end{aligned}$$

which is by the way, the Matching Diagnosis we want to find.

7 Matching Results

We present in this section some results of our approach. In table 4, we present for each test video sequence the error made by our approach. We consider as an error every false diagnosis, i.e. each frame where the diagnosis does not correspond to "the ground truth" (a hand-made diagnosis). We distinguish between two types of errors (two types of false diagnosis). The first type is the "one frame persons" (OFP) corresponding to a vertex p of the interpretation graph $\bar{G}_t$ without any arcs. This kind of error is not really a problem, because they can easily be identified and the associated vertices discarded from $\bar{G}_t$. The second type of errors is all other kinds of errors. These can be considered as real errors, because they change the structure of the interpretation graph $\bar{G}_t$.

Video id.	frames	# of OFP	# of failed	Failed diagnosis
st1-23	190	0	1	frm. 162: NOISE as APPEAR
c02-2	535	17	2	frm. 290: HIDDEN as LOST
				frm. 74: NOISE as ENTER
mc2-17	197	1	1	frm. 18: HIDDEN as EXIT
va2-7	55	0	2	frm. 29: LOST as EXIT
				frm. 33: LOST as EXIT
va2-4	340	6	0	
B008	570	0	0	
c07-2	137	1	1	frm. 99: NOISE as ENTER
mc1-22	153	0	1	frm. 79: NOISE as ENTER
va2-6	322	0	0	

Table 4: Results of the method in terms of failed diagnosis

Note that, even if there are no failed diagnosis in the sequence va2-4, the trajectories of the persons were so corrupted that we cannot consider this particular result as good.

In the following, we will detail three video sequences from different environments to illustrate both the errors we make and the difficulties we deal with. Each figure is composed of two parts: the left is the input image from the video stream and the right is our reconstruction. The different parts of the environment ($o_{i,t} \in O_t$) are represented in grey, green and orange. The different persons at time t ($p_{j,t} \in P_t$) are represented by brown cylinders and the tracks ($t \in \bar{T}_t$) are represented by red lines.

Figures 4, 5, 6, 7 and 8 illustrate the results of our approach in a metro station (video id: VA2-7), figures 9, 10, 11, 12 and 13 illustrate results from a bank (video id: MC2-17) and figures 14, 15, 16, 17 and 18 illustrate results obtained in an office environment (video id: C02-2).

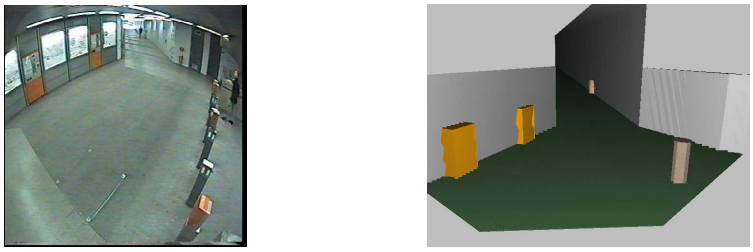

Figure 4: On the platform of the Nuremberg Metro Station there are two humans h_1 and h_2. h_1 is in the far end of the corridor and h_2 has just entered on the right.

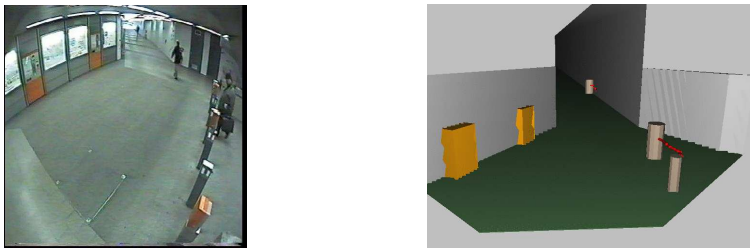

Figure 5: A new human h_3 enters the scene from the right.

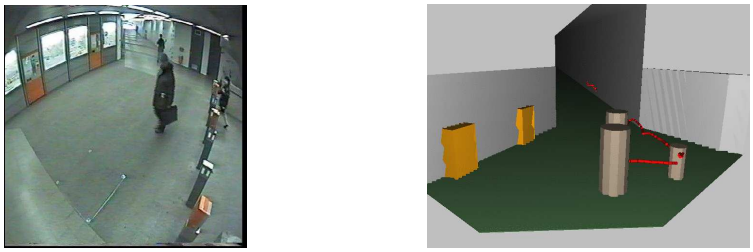

Figure 6: A new human h_4 enters the scene from the right. At this moment, h_3 occludes h_2, but we can see on the reconstruction that h_2 is not lost. We can also see that h_1 is lost both because he is not detected and because he is in an IO area.

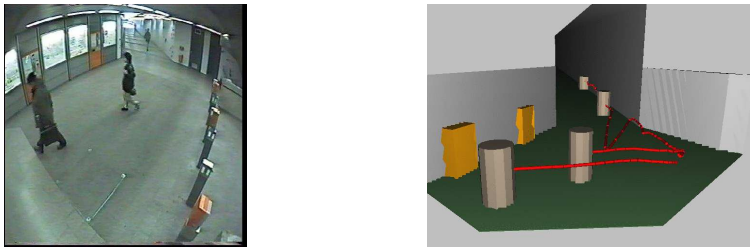

Figure 7: Even if h_2 has not been lost, his trajectory is partially corrupted. h_1 is now detected again and a new track has been created.

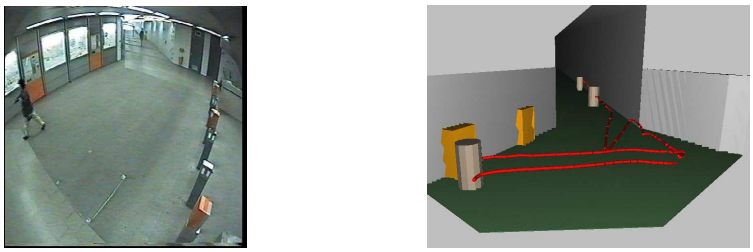

Figure 8: h_3 and h_4 exit the scene without problems.

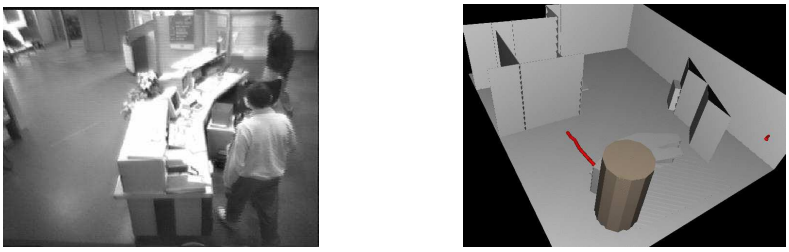

Figure 9: In the FNCA Bank, two clerks enter to sit in their chairs. h_1 (in the front) occludes h_2 in the back. There is at this time a lack of information concerning h_2 and he is considered as having exited.

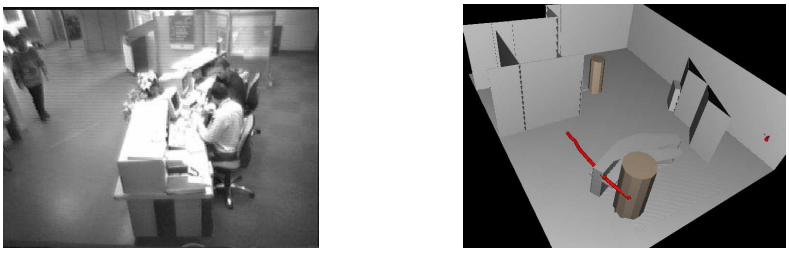

Figure 10: h_3 enters the scene and is correctly detected.

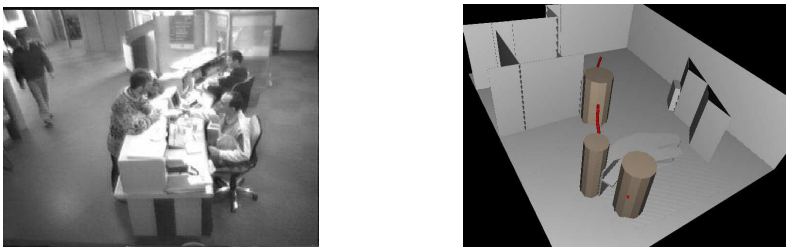

Figure 11: h_4 enters the scene. We see a typical error of location of h_3 due to shadows.

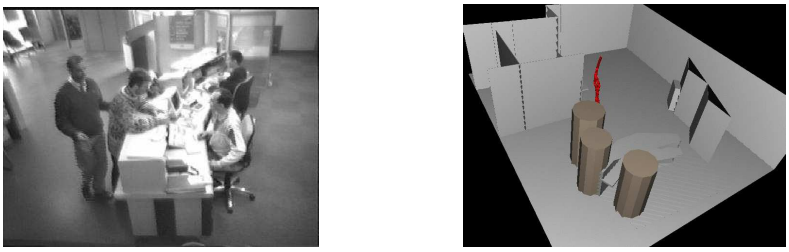

Figure 12: h_3 and h_4 touch each other, i.e. there is only one set $q \in Q_t$ of blobs to represent two humans, but we see in the reconstruction that both h_3 and h_4 are still considered as separate persons.

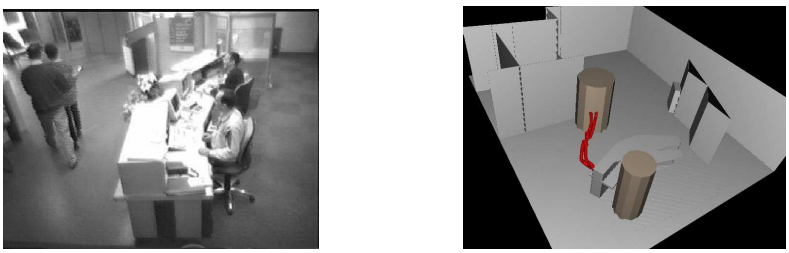

Figure 13: The lack of information (one set of blobs for two humans) continues in time, but both h_3 and h_4 are still there (superimposed in the reconstruction).

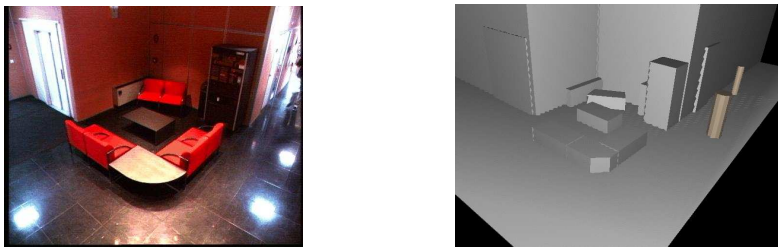

Figure 14: h_1 enters the scene from the right. A mix of reflexions and shadows on the floor is recognised as human in an IO area.

Figure 15: h_1 and h_2 are correctly located. In the reconstruction we see that the door of the elevator on the left is considered as a human.

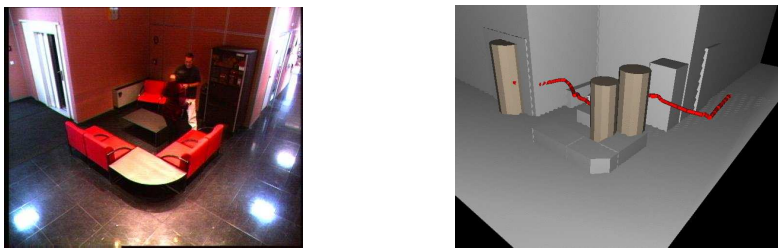

Figure 16: h_2 occludes h_1. Both are still correctly recognised.

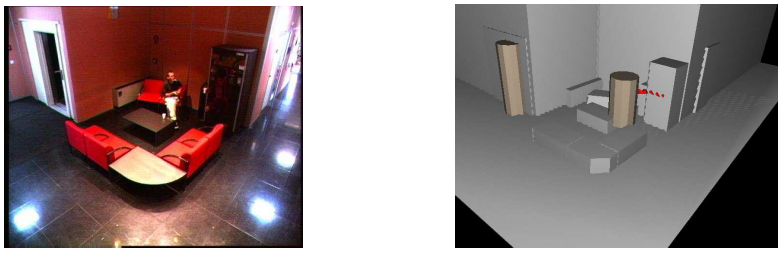

Figure 17: h_2 looks similar to the background and is not recognised (there are not even any blobs associated with her). She is lost at this time.

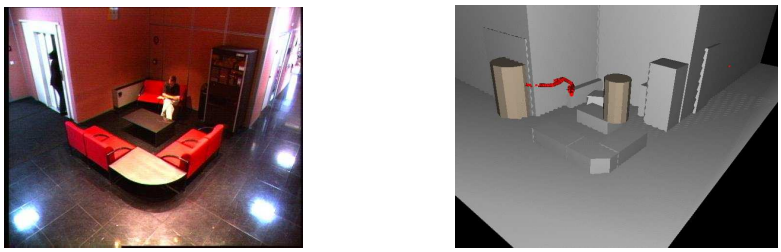

Figure 18: h_2 leaves the scene by the elevator. h_1 finishes his coffee and his cigarette.

8 Time Performance

This section presents the temporal performance of our approach. We have computed for each frame of the video sequence the duration of the complete interpretation process: i.e. background segmentation, resolution of the clustering problem, resoltion of the temporal matching problem (computation of P_t and T_t) and finally behaviour recognition (computation of V_t and R_t). In the same graph we represent 4 curves showing the cumulative duration of the processing per frame after each step.

The duration of the background extraction (on red) and the duration of the behaviour recognition (on pink) are not directly related to the purpose of this paper, but enable us to compare with the duration of the resolution of the clustering problem (on green) and the duration of the resolution of the temporal matching problem (on blue).

Figure 19 shows the time performance on the metro station sequences with 512×512 images. The average duration of a complete cycle of our algorithm is about 500 milliseconds.

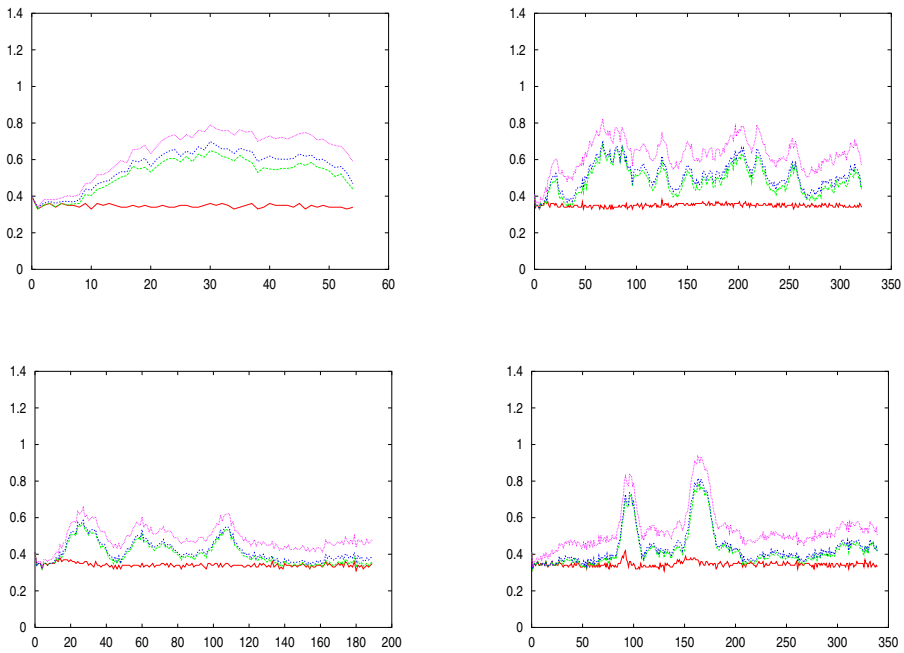

Figure 19: Time performance on metro station video sequences va2-7, va2-6, st1-23 and va2-4.

Figure 20 shows the time performance on the bank sequences with 440×334 images. The average duration of a complete cycle of our algorithm is about 1100 milliseconds.

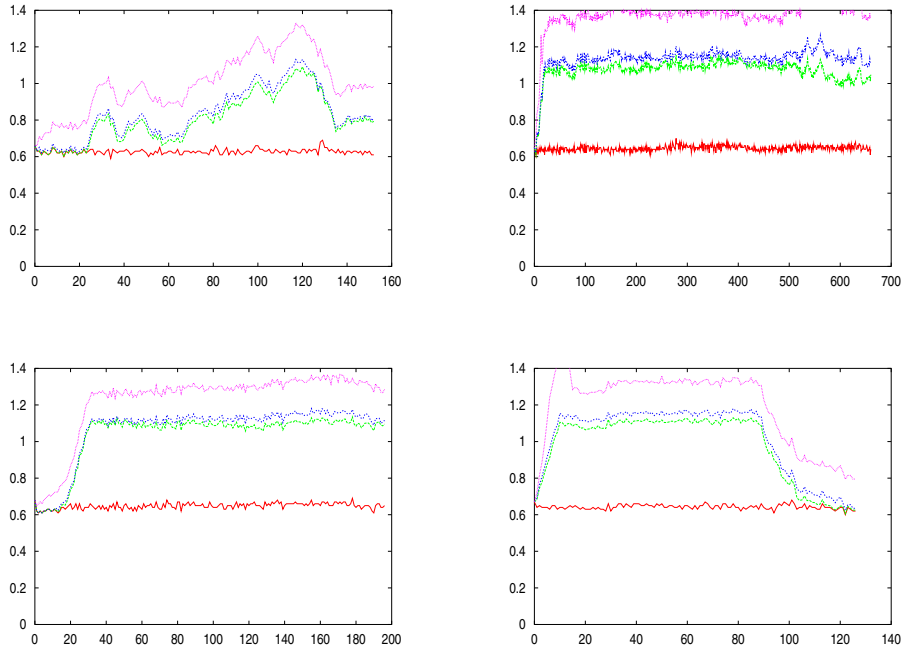

Figure 20: Time performance on bank video sequences mc1-22, mc1-30, mc2-17 and mc2-18.

Figure 21 shows the time performance on the office sequence with 512×384 and 440×334 images. The average duration a complete cycle of our algorithm is about 700 milliseconds.

9 Conclusion

In this paper we have presented new methods for human tracking. Firstly, given a set of blobs obtained by a background extraction method, we compute the number of humans and the location of each of them in the image and in the scene. Then, the recognised humans at two consecutive frames are matched. To accomplish this, two artificial intelligence techniques are used. A heuristic method is used to recognise humans in each frame and a diagnosis method is used to compute the match between descriptions in consecutive frames.

The advantage of these methods is to stay formal and robust at the same time. In fact, we think that the presented methods can easily be extended to other environments or to other tasks, adding new matching functions or changing the heuristics function given in

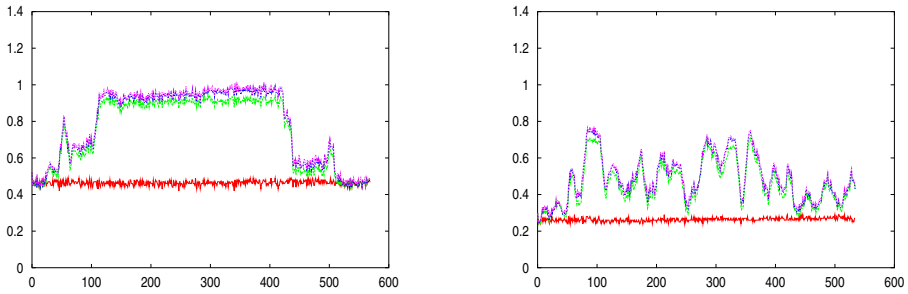

Figure 21: Time performance on cooperative work environment video sequences B008 and c02-2.

sections 4 and 6. In terms of robustness, 5, 7 and 8 provide some examples of what can be done with the presented methods.

References

[1] A. Sato, K. Mase, A. Tomono, and K. Ishii, "Pedestrian counting system robust against illuminastion changes," in *Visual Communication and Image Processing*, Massachusetts, 1993.

[2] J. Heikkila and O. Silven, "A real-time system for monitoring of cyclists and pedestrians," in *2nd International Workshop on Visual Surveillance*, Fort Collins, Colorado, June 1999.

[3] K. Akita, "Image sequence analysis of real world human motion," *Pattern recognition*, vol. 17, no. 1, pp. 73 – 83, 1984.

[4] D.M. Gavrila and L.S. Davis, "Tracking of humans in action: a 3-d model-based approach," in *ARPA Image Understanding Workshop*, Feb 1996.

[5] C.R. Wren and A.P. Pentland, "Dynamic models of human motion," in *3th International Conference on Face and Gesture Recognition*, Nara, Japan, April 1998, pp. 22 – 27.

[6] S. Intille and A. Bobick, "Closed world tracking," in *5th International Conference on Computer Vision*, Cambridge, 1995.

[7] D. Ayers and M. Shah, "Monitoring human behavior in an office environment," in *Computer Society Workshop on Interpretation of Visual Motion*, 1998.

[8] S. J. McKenna, S. Jabri, Z. Duric, and H. Wechsler, "Tracking interacting people," in *4th International Conference on Face and Gesture Recognition*, Grenoble, France, March 2000, pp. 348 – 353.

[9] N. Chleq and M. Thonnat, "Realtime image sequence interpretation for videosurveillance," in *International Conference on Image Processing*, IEEE, Ed., Lausanne, Switzerland, 1996, pp. 801 – 804.

[10] S. Dettmer, A. Seetharamaiah, L. Wang, and M. Shah, "Model-based approch for recognizing human activities from video sequences," in *Workshop on Motion of Non-Rigid and Articulated Objects*, June 1998.

[11] L. Davis, E. Borovikov, R. Cutler, D. Harwood, and T. Horprasert, "Multi-perspective analysis of human action," in *Third International Workshop on Cooperative Distributed Vision*, Kyoto, Japan, November 1999.

[12] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," in *DARPA Image Understanding Workshop*, Monterey, November 1998.

[13] N. Rota, R. Stahr, and M. Thonnat, "Tracking for visual surveillance in vsis," in *First Workshop on Performance Evaluation of Tracking and Surveillance*, Grenoble, March 2000.

[14] A. Baumberg and D. Hogg, "Learning flexible models from image sequences," Tech. Rep. 93.36, University of Leeds, October 1993.

[15] I. Haritaoglu, D. Harwood, and L.S. Davis, "Hydra: Multiple people detection and tracking using silhouettes," in *2nd International Workshop on Visual Surveillance*, Fort Collins, Colorado, June 1999, pp. 6 – 13.

[16] E. André, G. Herzog, and T. Rist, "On the simultaneous interpretation of real world image sequences and their natural language description: The system soccer," in *8th European Conference of Artificial Intelligence*, Munich, 1988, pp. 449 – 454.

[17] N.J. Byrne, A. Baumberg, and D. Hogg, "Using shape and intensity to track non-rigid objects," Tech. Rep. 94.14, University of Leeds, May 1994.

[18] K. Rohr, "Toward model-based recognition of movement in image sequences," *Computer Vision, Graphique and image processing: image understanding*, vol. 59, no. 1, pp. 94–115, jan 1994.

[19] A. Pentland, "Machine understanding human action," in *7th International Forum on of Frontier of Telecommunication Technology*, Tokyo, 1995.

[20] S. Intille and A. Bobick, "Visual tracking using closed-world," Tech. Rep., M.I.T Media Laboratory Perceptual Computing Section, Cambridge, MA 02139, November 1994.

[21] P. Huttenlocker, J.J. Noh, and W. Rucklidge, "Tracking non-rigid objects in complex scenes," Tech. Rep. 1320, Computer Science Department Cornell University, Ithaca, NY 14853, 1992.

[22] G. Rigoll, S. Eickeler, and S. Muller, "Person tracking in real-wolrd scenarios using statisical methods," in *4th International Conference on Face and Gesture Recognition*, Grenoble, France, March 2000, pp. 342 – 347.

Contents

1 Introduction **3**

2 State of the Art **3**

3 Problem Modelling **4**
3.1 Notation 4
3.2 Semantics 4
3.3 Properties 5
3.4 Interpretation Process 6
3.5 Computing P_t and T_t 6

4 Clustering **7**
4.1 Clustering Problem 7
4.2 Solving the Clustering Problem 7
4.3 Analysis of the proposed solution 8

5 Clustering Results **9**

6 Temporal Matching **12**
6.1 Definitions of the Matching Functions 12
6.2 The computation of the similarity function 14
6.3 Matching Diagnosis Problem 15
6.3.1 Theorem 1 16
6.3.2 Proof 17
6.4 Solving the Matching Diagnosis Problem 18
6.4.1 Theorem 2 19
6.4.2 Proof 20

7 Matching Results **22**

8 Time Performance **27**

9 Conclusion **28**

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot St Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, B.P. 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399