HUMAN BEHAVIOUR VISUALISATION AND SIMULATION FOR AUTOMATIC VIDEO UNDERSTANDING

Van Thinh VU, **François BRÉMOND** and **Monique THONNAT**

Project ORION of I.N.R.I.A. Sophia Antipolis
2004 route des Lucioles, BP93-06902 Sophia Antipolis Cedex
France
e-mail: {Thinh.Vu, Francois.Bremond, Monique.Thonnat}@sophia.inria.fr
http://www-sop.inria.fr/orion/orion-eng.html

ABSTRACT

The objective of this work is the visualisation and simulation for automatic video interpretation. We have conceived a test framework that generates 3D animations corresponding to behaviours recognised by an automatic interpretation system or corresponding to behaviours described by an expert. Conceiving this test framework is essential in order to be able to develop and validate the interpretation process. The objective of our test framework is (1) to visualise the computation of the interpretation, (2) to be flexible (configurable) enough for testing the different configurations of the interpretation and (3) to be realist enough to understand what is interpreted. To solve this problem we have defined six types of model to represent all the information that is necessary for the interpretation. First, we propose a model of the scene context (containing the 3D geometry) and a model for the virtual camera. Second, we propose an articulated and hierarchical model for representing the human body given its sub parts. We propose two other hierarchical models for modelling human actions and scenarios, and also a model of scene-scenarios that gathers all previous models. We have defined a description language for representing these models. The obtained results are promising: we have developed a test system for a given interpretation system and started evaluating it by generating test animations.

Keywords: 3D visualisation, 3D animation, simulation, video understanding, modelling of the human body, human behaviours and scenes.

1. INTRODUCTION

This paper presents a modelisation framework for the visualisation and simulation of automatic video interpretation. The automatic video interpretation consists in recognising pre-defined scenarios describing human behaviours from video sequence. This framework (called *test framework*) has (1) to visualise the computation of the interpretation and scenarios described by an expert, (2) to be flexible enough (configurable) for testing the different configurations of the interpretation system and (3) to be realist enough to understand what is going on in a real scene. Another requirement of this framework is to verify the coherence between a given interpretation system and the test framework, so we can establish the limit and the robustness of the interpretation system. This test framework will be an efficient tool for the developers (e.g. experts in vision and in scenario recognition) and for the experts of the application domain (e.g. agents of security). To validate this framework, we have developed a test system for an automatic video interpretation system. Here, we are using VSIS system (Video Surveillance Intelligent System) as an example of interpretation system.

For 20 years, the problem of 3D scene visualisation has been approached. There are many laboratories ([3], [7], [13], [14]) who study the visualisation of a 3D scene from its description. For example, at the faculty of Computer Science of Toronto university [3], researchers generate 3D animations where many fishes and a swimmer evolve in the bottom of the sea. To visualise these animations, they have modelled the behaviours of individuals and fishes and theirs interactions in groups. In particular, they have modelled all the physical and biological rules

for fish to swim, eat, reproduce and perceive other fishes. At the Computer Graphics Lab of the Swiss Technology Institute of Lausanne ([7], [13]), researchers have modelled individuals evolving in a museum, in a street and in a supermarket. They have also modelled the crowd behaviours like the reaction of people in fire situations.

These laboratories have obtained many results in the domain of 3D animations from a scene description. However, there are few laboratories who study the *visualisation of scenarios recognised by an automatic video interpretation system*. For example, the Robotvis group at the research unit INRIA Sophia-Antipolis [10] visualises the tracking of the members (legs, arms,…) of an individual who is running. The Robotics Institute, at Carnegie Mellon University [12], computes 3D animations where a group of individuals enters/leaves the university site by taking as input the camera network surrounding the university. The goal of these animations is mainly to demonstrate the tracking of the group all around the university.

In our knowledge, we did not find any system that visualises the recognition of human behaviours from a video by an automatic interpretation system.

Our approach to describe human behaviours, consists in defining six generic models (i.e. meta-class): scene context, camera, human body, action, scenario and scene-scenario. Using these generic models, we can construct specific models (e.g. the scenario class "meeting at a coffee machine") described in libraries of models. Then these specific models are used to generate instances (e.g. scenario "individuals A and B meet at the coffee machine M") to visualise what is occurring in a given real scene (corresponding to videos or scene descriptions). We also propose a description language to build all these models.

2. VISUALISATION AND SIMULATION FOR VIDEO UNDERSTANDING

An automatic video interpretation system contains three principal modules [8]: (1) individual detection, (2) individual tracking and (3) scenario (i.e. behaviour) recognition. It takes its inputs from a video camera and generates recognised scenarios as output. The system is represented in Figure 1.

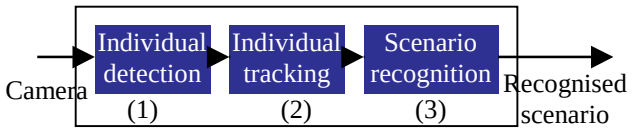

Figure 1: a video interpretation system contains three principal modules.

The test framework for a video interpretation system should contain the following functionalities:

a) *visualisation of scenarios recognised by an interpretation system and scenarios described by an expert*. It is important for the developer (e.g. expert in vision and scenario recognition) to visualise each step of the scenario recognition process. It is also important for the experts of the application domain (e.g. agent of security in a metro) to visualise the scenarios that they describe.

b) *evaluation of the couple interpretation-test system*: verify the coherence between the interpretation and test system.

c) *validation of interpretation system*: establish the limits and robustness of the interpretation system by simulating test videos.

For this purpose, we propose to define a test framework that allows the three following tasks (see Figure 2):

1) *generation of realistic 3D animations corresponding to the scenarios recognised by an interpretation system*. The generation of animations needs to be flexible enough to illustrate specific steps of the interpretation process. For example, to illustrate the tracking process, it is convenient to give a specific colour to each tracked individual.

2) *comparision of two animations, one coming from the interpretation of an initial video and the other one coming from the interpretation of a new video generated by the test system and corresponding to the interpretation of the initial video*. For a scenario recognised by an interpretation system from an initial video, the test system generates a first 3D animation and a second video that corresponds to the recognised scenario. We have to compare that the first animation is similar to the second one.

3) *automatic generation of a set of videos* corresponding to a scenario described by an expert. This set of videos should illustrate the variety of all possible instances of this scenario (e.g. variety due to different locations of individuals and due to different optical effects like illumination).

3. SCENE CONTEXT

3.1. Scene context and camera representation

The scene context contains a set of contextual information related to the environment of the scene (e.g. the 3D geometry of the scene) and used by the video interpretation and visualisation process. We use scene contexts containing the four following elements:

- a set of polygonal zones with semantic information: entrance zone, zone near the seat,…
- a set of areas of interest that gather the connected zones with the same semantic information: platform, metro tracks,…

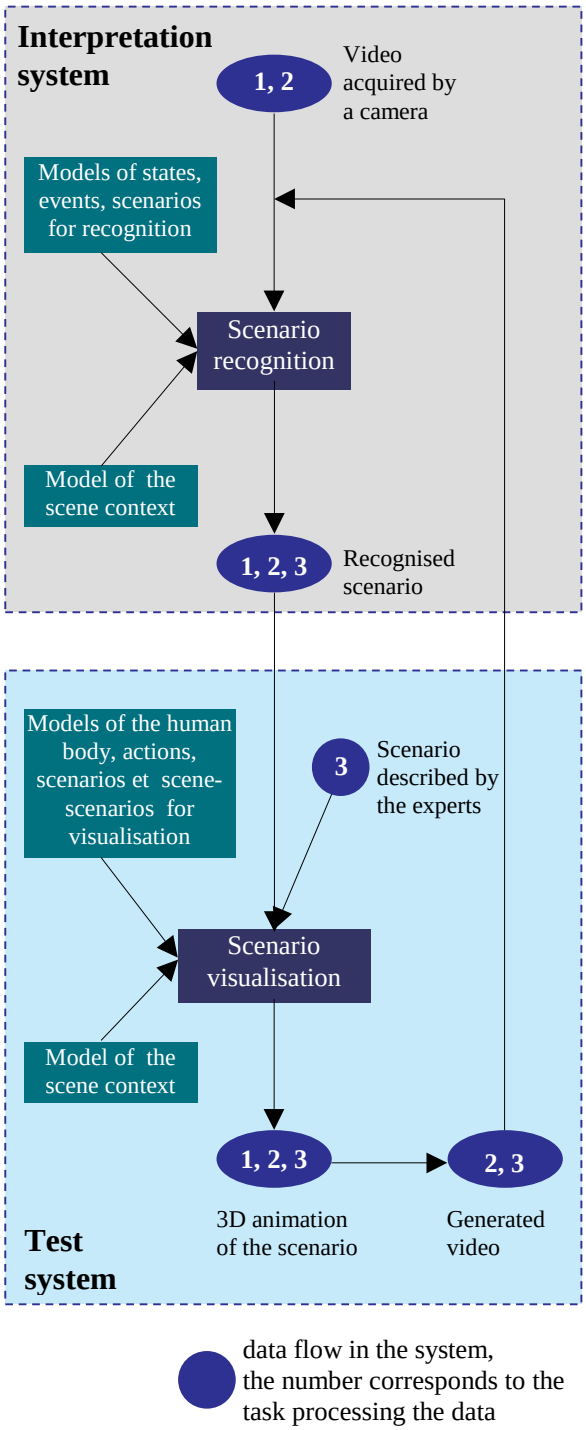

Figure 2: an interpretation system and its test system.

- a set of 3D objects of the environment which principally includes the equipment (e.g. a seat, a door).
- a calibration matrix of the scene containing the extrinsic parameters of the camera (position, direction and field of view - FOV).

We represent the 3D context objects of the environment using a meta-class (generic model of 3D context object) and we represent each object type as a class of this meta-class. Thus, we have built five classes of context objects that usually appear in metro scenes: class "seat", "trashcan", "validation machine", "ticket machine" and "door". To facilitate the building process, we have defined a description language where the meta-class *3D context object* enables the construction of a hierarchy of 3D context object classes. These classes contain five attributes:

- the *relative co-ordinates* that represent the position of the 3D object in the referential of the super part (super 3D object). For example, the leg of a chair is defined relatively to the chair.
- the *angular co-ordinates* of the 3D object in the object referential,
- the *size* of the 3D object along its referential axis,
- the *sub-parts* or/and *geometric primitives* that constitute the 3D object,
- the *colour* of the 3D object.

We use three types of geometric primitives: *sphere*, *truncated cone* and *parallelepiped*. For the truncated cone, its both sections can have different radius. The *geometric primitives* have the same attributes of the 3D object classes except the list of sub-parts.

3.2. Visualisation of the scene context

To visualise the scene context, we use GEOMVIEW (a free software for 3D visualisation) for visualising the polygonal zones and 3D objects. To use GEOMVIEW, it is first necessary to represent the objects of the scene context in OpenGL format (see Figure 3).

Figure 3: visualisation of a scene context description using GEOMVIEW.

For the polygonal zones, there are specific methods in GEOMVIEW to display them in 3D. For the 3D context objects we first compute the co-ordinates of the geometric primitives of the objects in the scene referential: we multiply the co-ordinates of the geometric primitives by the referential transformations of all containing super 3D objects. Then we use a GEOMVIEW method that constructs the vertices and the facets of the geometric primitives.

Using this representation we have built two scene contexts for metro station including a platform, a corridor and a hall: one for Yser station in Brussels and one for Segrada-Familia station in Barcelona. Figure 4 shows two images from Sagrada-Familia station.

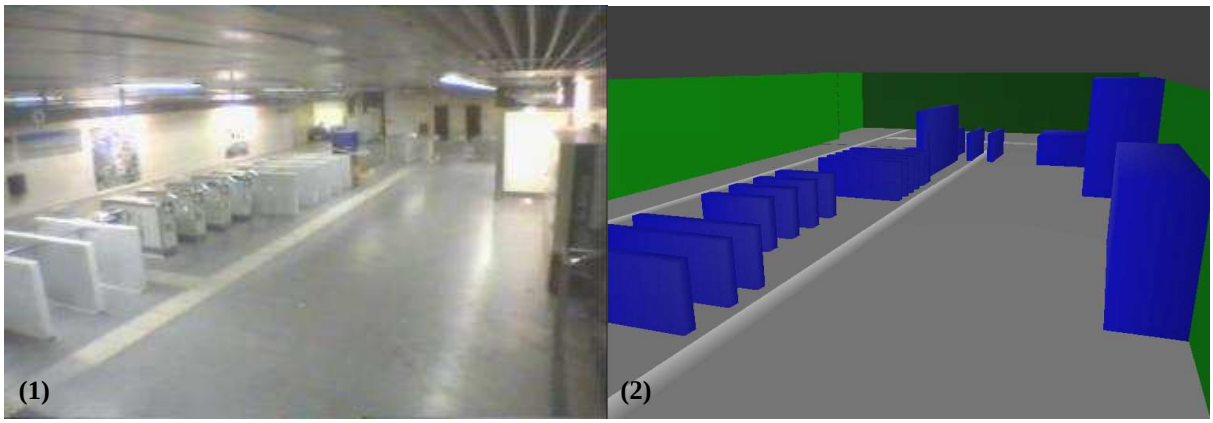

Figure 4: Sagrada-Familia station in Barcelona: (1) raw image taken by a camera and (2) scene context model corresponding to this image.

4. HUMAN BODY

We use a hierarchical and articulated model for the *generic model* of human body parts [10]. A human body part is composed by sub-parts or geometric primitives. These primitives are the same one used for 3D context objects: *spheres*, *truncated cones* and *parallelepipeds*. Figure 5 shows the 26 geometric primitives composing the human body. We represent classes of human body parts (and the whole human body) using a **generic model** similar to the generic model of 3D context object.

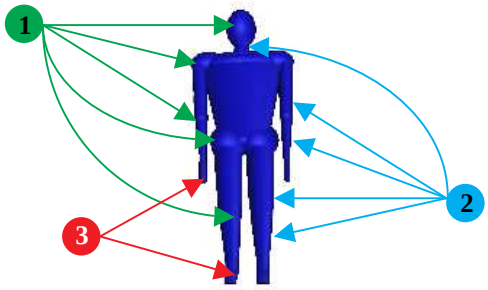

Figure 5: hierarchical and articulated model of the human body using three types of primitives (1) spheres, (2) truncated cones and (3) parallelepipeds.

In the description language we have defined 14 classes for modelling human body parts: the whole human body, the head, the arms, the legs, the neck, the shoulders, the hips, the trunk, the foot and the hand. Figure 6 shows the human body from different view points.

Using these models, there are two ways of visualising the human body. First, we can visualise an individual from its description by an expert. Second, we can visualise an individual detected by an interpretation system from a video sequence. In both cases, we visualise the body part by displaying the geometric primitives composing the body part through GEOMVIEW.

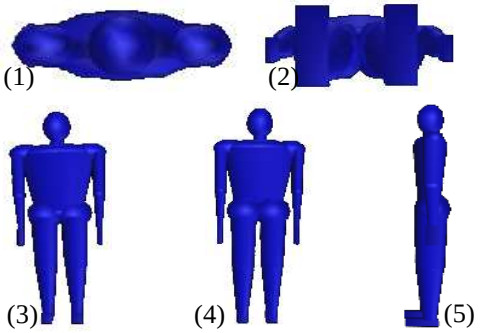

Figure 6: visualisation of the 3D model of the human body: (1) top view, (2) bottom view, (3) front view, (4) back view and (5) view from the left.

5. HUMAN BEHAVIOUR

5.1. Human behaviours for interpretation systems

In interpretation systems, the notions of state, event and scenario [8] are used to recognise human behaviours. A state characterises at a given instant, the situation of an individual detected by a camera. An event defines a change of state at two successive instants. A scenario defines a combination of events.
In the interpretation system that we are testing, eight states are defined: posture (e.g. lying, crouching, standing), direction (e.g. towards the right, towards the left, leaving, arriving), velocity (e.g. stopped, walking, running), location w.r.t. a zone (e.g. inside, outside), proximity w.r.t. a context object (e.g. close, far), relative location w.r.t. another individual (close, far), relative posture w.r.t. a context object (e.g. seated, any) and relative walk w.r.t. another individual (e.g. coupled, any). By using these eight states, eighteen events are defined: for example, the event “falling” is defined from the change of posture

from “standing” to “lying”. Combining these events, several scenarios are defined such as “two persons meet at a coffee machine” for office applications and “graffiti on wall” for metro station applications. A scenario is a set of spatio-temporal constraints on the individuals of the scene, on the context objects and/or on the previously recognised sub scenarios (or events). The temporal constraints are expressed by equations that combine the instants when the events are detected.

5.2. Human behaviours for the test framework

In the test framework, the notions of posture, action, scenario and scene-scenario are defined to visualise behaviours recognised by an interpretation system or described by an expert. A posture corresponds to all body parameters of an individual to be visualised at one instant. An action characterises an individual motion when one (or several) of its body parameters change(s). Behaviours are represented by scenarios. A scenario combines the individuals of the scene and the context objects with sub scenarios which are relevant to the same activity. An elementary scenario is an action. A scene-scenario combines and instantiates all previously defined scenarios.

In our formalism, an action (or scenario) can be visualised at different speeds which indicates how many frames per second are displayed. An action (or scenario) can have a departure/arrival position which locates the individual at the beginning and the end of action (or scenario). The temporal constraints are expressed by intervals (named *periods*) that correspond to the duration of an action (or scenario). The interval of a sub action (or sub scenario) is defined relatively to the period of the containing action (or scenario).

Because our purpose is to conceive a test framework for automatic video interpretation systems, we do not consider more precise actions such as “balance the arm” and “move the finger” which are difficult to detect by interpretation systems.

5.3. Action

5.3.1. Generic model of actions

An action is relative to the motion of one body part (or the whole human body) which is characterised by the changes of the body part parameters. These changes are mainly rotations around the body part axis. An action is described by a *hierarchical model*: an action can be decomposed into *sub action(s)* describing the motion of sub part(s) (see Figure 7). In our formalism, to ease the description of actions by experts, it is possible to indicate the departure/arrival position in the case where the body part is the whole individual. There are two types of actions: periodic (e.g. “walking”) and non-periodic (e.g. “move close to”). For non-periodic actions, the period corresponds to the duration of the action. For periodic actions, the number of periods is defined in the containing action and the duration is obtained by multiplying the number of periods times the action period.

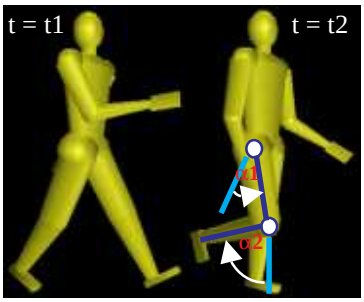

Figure 7: in the action “walking” during interval [t1,t2], the right leg rotates with angle $\alpha 1$ around the hip; and in its sub action “the right leg up”, the lower part of the leg rotates with angle $\alpha 2$ around the knee.

To represent actions, we propose a generic model with the following attributes:

- the *concerned part* of human body.
- the *fixed part of human body on the ground* (see section 5.3.2: visualisation of an action).
- the *global period* of the action.
- the *variation of angles* of rotation around the part referential.
- the *speed* of the action.
- the *departure/arrival position* (optional, used only when the part is the whole individual).
- the *list of sub actions* with:
  + their *relative period*,
  + the *concerned sub part* of human body,
  + the *variation of angles of rotation* around the sub part referential.

5.3.2. Visualisation of actions

An action is visualised by displaying the individual performing the action at regular instants. In the case where the test framework visualises the actions recognised by an interpretation system, the individual posture to be visualised is obtained by the posture detected by the interpretation system. Therefore the test framework just needs to display the individual where it has been detected. If the visualisation frequency is greater than the frequency of the input video, then it is necessary to interpolate linearly the intermediary positions of the individual. Knowing the global position of the individual, we calculate the vertices of the geometric primitives of the individual body in the scene referential and display the primitives through GEOMVIEW in the same way with the visualisation of 3D context objects.

In the case where the test framework takes as input the actions modelled by an expert, we visualise an action in three steps:

1) *calculation of the current posture from the previous instant.* By using the posture of the previous instant and the angular variations of the action, we calculate the new angular co-ordinates of each sub part of the body at the current instant. From the new angular co-ordinates, we can calculate the new vertices of the primitives of each body part by multiplying their co-ordinates by the referential transformation matrix. This transformation matrix is defined for each body part and enable to compute co-ordinates in body party referential to co-ordinates in its containing body part referential. By this way we obtain the vertices of the body part defined relatively to the global position of the individual. These new co-ordinates define the new posture of the individual in the individual referential.
2) *calculation of the global position of the individual*. To calculate all positions of the individual, we make the following hypothesis: at each moment, there is a fixed point of a body part on the ground (see Figure 8). Currently, the actions we are interested in, are actions where the individual has a fixed part on the ground (e.g. "walking", "running"). In the near future, we are planning to extend our formalism to handle actions such as "jumping above a barrier". To calculate the global position, we first compute the distance between two successive fixed points on the ground (if the fixed point of the action has changed since last instant). Second, we compute the motion of the referential point of the individual relatively to the current fixed point. These two points (referential/fixed points) are defined by the expert. By applying the transformation corresponding to this motion to the vertices of primitives defining the individual, we obtain the new co-ordinates of these vertices that correspond to the current posture of the individual. There are other approaches to calculate the position of an individual from its motion description. In [10] the authors have proposed a method to calculate the trajectory of individuals based on the combination of human body contour points. In ([7], [13]) the authors describe the motion by mathematical equations (based on experimental data) and calculate the position of individuals by solving the equation system.
3) *visualisation*: after computing the geometric primitives of the human body relatively to the new global position of the individual, we display all the primitives with GEOMVIEW.

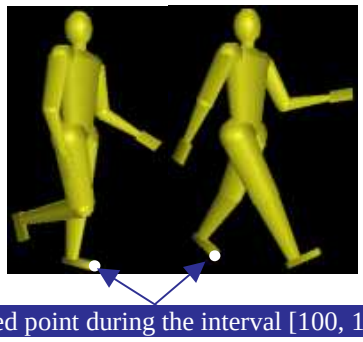

Figure 8: one of the fixed points while the individual is walking.

5.4. Scenario

5.4.1. Generic model of scenarios

A scenario combines the individuals of the scene and the context objects which are relevant to the same activity with more elementary sub scenarios. An elementary scenario is an action that corresponds to the motion of the whole human body of the involved individuals. The model of scenarios is defined as the model of actions. It is a hierarchy of sub scenarios. Each sub scenario is ordered in time thanks to intervals (called *periods*) that correspond to the duration of the sub scenarios defined relatively to the global period of the main scenario. Unlike actions, a scenario has an attribute corresponding to the list of actors and context objects involved in the scenario. At the level of scenarios, an actor (or a context object) is represented by a variable that corresponds to the role of the actor in the scenario.

5.4.2. Visualisation of scenarios

We visualise a scenario in three steps. First, we link all actors and context objects of the scene involved in the scenario to the variables defined in the actions composing the scenario. Second, we order these actions in time: for each action, we calculate its duration (start and end point) relatively to the scenario period, defining when the action is active (is displayed). Third, at each instant, we display all actors involved in active actions using GEOMVIEW. Figure 9 presents the visualisation of the scenario "two persons meet at a coffee machine" between the instants 80 and 240.

5.5. Scene-scenario

5.5.1. Generic model of scene-scenarios

A scene-scenario combines and instantiates all previously defined scenarios. To represent a scene-scenario we use a generic model that has five attributes:

1) the *scene context* includes the list of context objects involved in the scene. The expert describing the scene can change the default attributes of the context objects (e.g. their colour).

2) the *virtual camera information* that corresponds to the viewpoint from where the 3D animation is visualised. This information includes the 3D position, the direction and the field of view (FOV) of the camera.
3) the list of *actors involved* in the scene with their initial position, size, posture and colour. If this information is not provided, default values are used.
4) a set of *scenarios occurring* in the scene. For each scenario, we first specify which actor corresponds to which role defined in the scenario and we also specify the scenario period relatively to the global period of the scene-scenario.
5) the *visualisation speed* of the scene.

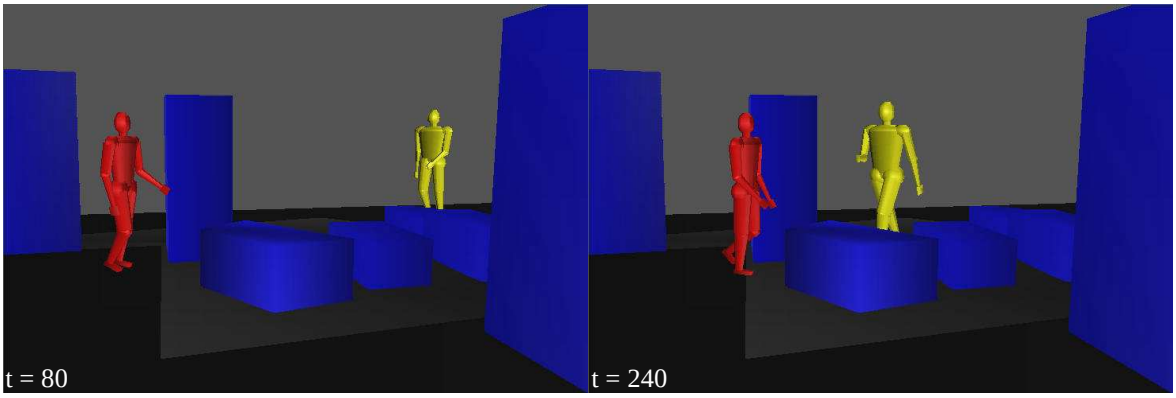

Figure 9: visualisation of the scenario "two persons meet at a coffee machine" between the instants 80 and 240.

5.5.2. Visualisation of scene-scenarios

We display a scene-scenario in three steps. First, we initialise and connect the actors and the context objects to the scenarios defined in the scene. Second, we calculate the parameters of the virtual camera of GEOMVIEW. Third, we display all active scenarios composing the scene at each instant. The visualisation frame rate can be specified either at the level of the scene-scenario or at the level of the scenarios or actions.

6. RESULTS

To validate this framework, we have developed a test system for VSIS [8] (Video Surveillance Intelligent System), an automatic video interpretation system taken as an example of interpretation system. Thanks to this test system, we have realised three 3D animations that visualise the results of VSIS from real videos of metro taken for the ADVISOR European project. Figure 10 shows (1) an image that illustrates the individuals detected by VSIS and corresponds to the output of VSIS and (2) an image that illustrates the 3D animation generated by the test system (named "animation 1"). Currently, VSIS is not able to detect the posture and the orientation of the individuals (front view, lateral view). By default, the 3D animation shows the front view of the individuals.

Thanks to the test system, we have also realised seven 3D animations from scene-scenarios described by an expert and then generated the corresponding videos taken from the view point of the real camera. Moreover we were able to verify the coherence between the interpretation and the test system. For that, we have first generated a 3D animation (named "animation 1") corresponding to a recognised scenario. Then we have generated a video from "animation 1", processed this second video by the interpretation system and generated a second 3D animation. As shown on Figure 10, these two animations are almost identical which indicates that the interpretation system does not make any difference between real videos and videos generated by the test system.

All these animations and videos can be found on the WEB site
http://www-sop.inria.fr/orion/personnel/Thinh.Vu/TestVSIS/.

7. CONCLUSION

We have proposed a framework for the visualisation and the simulation of automatic video interpretation systems. Thanks to this framework, we were able to build a test system that generates the 3D animations corresponding to scenarios recognised by an automatic video interpretation system, or scenarios described by an expert. To realise this framework, we have defined *six original models* for modelling the virtual camera, the visualisation of the scene geometry, the human body, the actions, the scenarios and the scene-scenarios of individuals evolving in the scene.

These encouraging results open many perspectives. We are planning three main extensions of the framework. First, we plan to add functionalities to help the developer (e.g. expert of vision or scenario recognition) to understand the influence of algorithm parameters setting, and to test the robustness of the

interpretation. It will be interesting to generate test videos (animations) with noise phenomena (e.g. shadow) for simulating more realistically the input video of interpretation systems. Second, we plan to extend the description language for the expert of the application domain (e.g. security agent) to be able to describe more complex scenarios and to visualise scenario variations, for example w.r.t. the variation of actor location.

Finally, we would like to define an unified framework using the same models for the interpretation and the test system (e.g. models of individual, action and scenario).

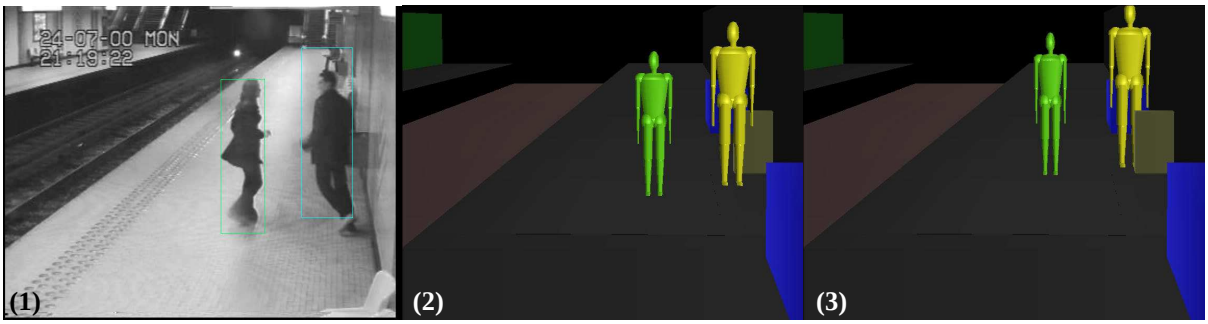

Figure 10: illustration of the test system results: (1) detection of individuals corresponding to the output of VSIS, (2) generation of a first animation and a second video corresponding to the output of VSIS and (3) generation of a second animation corresponding to the second video processed a second time by VSIS.

REFERENCES

[1] Badler NI, Smoliar SW: Digital Representation of Human Movement, *ACM Computer Survey March issue*, pp19-38, 1979.

[2] Bruderlin A, Calvert TW: Goal Directed, Dynamic Animation of Human Walking, *Proc. SIGGRAPH'89, Computer Graphic*, Vol. 23,3.

[3] D. Terzopoulos: Artificial life for computer graphics, *Communications of the ACM*, 42(8), August, 1999, 32-42.

[4] F. Brémond: Environnement de résolution de problèmes pour l'interprétation de séquences d'images, *PhD thesis*, INRIA-Université de Nice Sophia-Antipolis, 1997.

[5] F. Brémond, M. Thonnat: Issues of representing context illustrated by video-surveillance application, *International Journal of Human-Computer Studies Special Issue on Context*, pp375-391, 48, 1998.

[6] François Bresson et Maurice de Montmollin: La simulation du comportement humain, *Collection sciences du comportement*, Dunod, Paris, 1969.

[7] Laurent Bezault, Ronan Boulic, Nadia Magnenat-Thalmann, Daniel Thalmann: An interative Tool for the Design of Human Free-Walking trajectories, *Computer Animation '92*, pp.87-104.

[8] Monique Thonnat et Nathanael Rota: Image Understanding for Visual Surveillance, *Proceeding of the Third International Workshop on Cooperative Distributed Vision*, Kyoto, Japan, 19-20/11/1999.

[9] N. Chleq, M. Thonnat: Realtime Image Sequence Interpretation for Video-Surveillance Applications, *International Conference on Image Processing (ICIP96)*, Laussance, Suisse, Septembre 1996.

[10] Quentin Delamarre and Olivier Faugeras, RobotVis Projet - INRIA-Sophia Antipolis: 3D Articulated Models and Multi-View Tracking with Sihouettes, *ICCV 99*.

[11] Richard W. Pew and Anne S. Mavor, National Research Council: Modeling Human and Organizational Behavior, *National academy press*, Washington, D.C., 1998.

[12] Robert T. Collins and Takeo Kanade, Robotics Institute Carnegie Mellon University: Multi-camera Tracking and Visualisation for Surveillance and Sports, *Proc. Of the Fourth International Workshop on Cooperative Distributed Vision*, 22-24/03/2001.

[13] Ronan Boulic, Nadia Magnenat-Thalmann, Daniel Thalmann: A global human walking model with real-time kinematic personification, *Visual Computer*, 6(6), pp 344-358 (1990).

[14] S. Donikian, F. Devillers, G. Moreau: The kernel of a scenario language for animation and simulation, *Eurographics workshop on animation and simulation*, Springer Verlag, Milano, Italia, September 1999.