3-D Virtual Environments on Mobile Devices for Remote Surveillance

R. Vezzani, R. Cucchiara
D.I.I - University of Modena and Reggio Emilia
{vezzani.roberto,cucchiara.rita}@unimore.it

A. Malizia, L. Cinque
D.S.I. - University of Roma "La Sapienza"
{malizia,cinque}@di.uniroma1.it

Abstract

In this paper we present a distributed video-surveillance framework. Our end is the remote monitoring of the behavior of people moving in a scene exploiting a virtual reconstruction on low capabilities devices, like PDAs and cell phones. The main novelty of this system is the effective integration of the computer vision and computer graphics modules. The first, using a probabilistic frameworks, can detect the position, the trajectory and the posture of peoples moving in the scene. The second exploits the new possibility of both standard 3D graphics libraries on mobile (namely JSR184 and M3G graphic format) and new PDAs processing capability in order to reconstruct the remote surveillance data in real-time.

1. Introduction

In this paper we present a framework for remote surveillance. The proposed system aims to keep under control an environment using computer vision techniques to generate a compact representation of the scene and virtually reconstructions on mobile devices as final result.

Remote surveillance on mobile devices is becoming a wide market demand in many contexts. First, centralized control centers for visual surveillance have management costs much higher than a network of distributed and mobile control points. Centralized surveillance also has some limits in terms of efficacy, since people employed for watching tens of monitors and videos coming from hundreds of cameras cannot keep their attention on all the controlled scenes. Therefore, there is a high demand of connections between control centers and distributed mobile platforms to send in real-time surveillance data, images and videos.

In the last years the technology of robust video streaming on mobile devices has been improved but sometimes it is unfeasible or cannot be provided with an acceptable quality due to the unavailability of the connection or the lack of transmission stability. Moreover, the current standards GPRS or UMTS frequently insert a not negligible delay in streaming transmission, so that surveillance images cannot be received in real-time everywhere and every time. A possible solution is the intra-media transcoding of visual data to textual information with a manual or automatic extraction of surveillance knowledge from the videos. Examples are the detection of people, moving objects, or suspicious abandoned packs in the scene, the counting of how many people are moving, the estimation of their position and their behavior. The textual information can be easily transmitted in real-time to mobile devices. Computer vision techniques are now sufficiently robust to ensure reliable image understanding processes that automatically extract parts of the surveillance knowledge in real time. Therefore, many proposals of new generations of computer based surveillance systems have been proposed and developed [6, 7].

In this paper we propose to enrich the computer-vision based surveillance systems with computer graphics technology enabling a remote virtual reconstruction of the scene. The goal of the project is the definition of a virtual environment corresponding to the real controlled one where the useful surveillance data are kept. In such a manner the textual surveillance knowledge can be transmitted in an affordable and robust way to mobile devices where the visual information can be reconstructed. Geometric 3D data on the background scene are pre-loaded in the mobile device and only the dynamic information is transmitted in real time.

In the proposed architecture of remote surveillance platforms, the video streams are processed in real-time by local servers, which extract the surveillance knowledge on the environment, provide a semantic transcoding of the visual data and make the data available for mobile connection.

The semantic transcoding can be provided in intra-media or inter-media modality. In the first case videos are processed and modified in order to provide compression rates that are acceptable in remote mobile

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

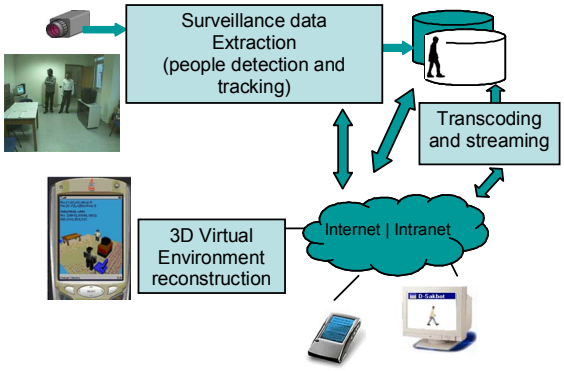

Figure 1: system architecture

connections without loosing important data. As in [1,8] background and moving objects can be coded at different compression rates, sent separately and reconstructed at the client-side. Instead, in the intra-media modality the scene is virtually reconstructed with computer graphic techniques, allowing the transmission of few textual data only. This second modality is less realistic but has several peculiar advantages. Firstly, the 3D scene is reconstructed so that all 3D information is available, new views can be provided, different from the field of view of the camera and a virtual interactive navigation is allowed. Secondly, there is the possibility to filter out some critical data that should not be transmitted for privacy issues. For instance, current laws of many countries do not allow the use of surveillance data in commercial sites such as offices or supermarkets. With a virtual reconstruction the privacy protected data (e.g., the face) can be eliminated. Therefore we reconstruct the presence of people in an environment, their motion, their trajectory, and their possible interactions without any individual or biometric information. In this way, security employers equipped with mobile devices can manage the multimedia information in real-time interacting with the environment and in a total respect of privacy issues.

The main novelty of this system is the effective integration of the computer vision and computer graphics modules. The first, using a probabilistic frameworks, can detect the position, the trajectory and the posture of peoples moving in the scene. The second exploits the new possibility of both standard 3D graphics libraries on mobile (namely JSR184 and M3G graphic format) and new PDAs processing capability in order to reconstruct in real-time the remote surveillance data. In this manner, the computational load is efficiently decoupled between the computer-vision based server and the mobile client supporting 3D graphics libraries; the bandwidth requirements are strongly reduced since few textual information are sent instead of severe visual streams.

2. System Architecture

Traditional Closed Circuit TeleVision (CCTV) systems acquire streaming videos from static or Pan-Tilt-Zoom cameras, collect and process them in video server and transmit snapshots and videos to remote control centers. New multimedia and mobile platforms allow a ubiquitous handling of surveillance data, by receiving images, video and audio streams over IP on small devices. Up to now, the computational, display and bandwidth limitations of the devices had required a severe computational load on the server-side, that should process video in real-time in order to extract surveillance data, annotate and store useful information and transmit compressed video. A typical architecture is reported in Fig.1 where the video server is coupled with the computer-vision module and it is responsible to all the computation capability. Standard personal computers as well as mobile devices can be used as interfaces for human operators. In this work, we want to exploit the new computation capability of the last generation of PDAs and handle phones and of the new graphic library to move some computations from the server to the client side and reconstruct the scene starting from few textual transmitted data. The computer vision module provides people segmentation and tracking.

Some metrical information about the environment and the camera calibration allows the construction of homographic view and the extraction of some parameters about the people that are in the scene.

Moreover, each object O_j detected in the time t has a set of temporal attributes: position $(x(t),y(t))$ in the ground plane; the bounding box BB in the image plane, appearance image AI, that is the color aspect and a $P(t)$ probabilistic mask that take into account the reliability of the detection given the field of view and the possible occlusions. The objects recognized as a person inherit some additional attributes, that are $Status(t) \in \{moving, still\}$ and $Posture(t) \in \{standing, sitting, crawling, laying\}$.

This extracted information is annotated in a text file, that is the basis of a standard annotation for stored video server that uses MPEG7 standard to keep information for historic and content-based search. The textual data can be downloaded or transmitted in a streaming manner over standard http-based connections to mobile clients. The mobile module stores pre-existing information about the background, the 3D environment and the mesh shapes, processes the dynamic information and provides 3D real-time rendering according with its computational and display capabilities.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

3. Surveillance data extraction

We have developed a two stage algorithm composed of a motion segmentation (based on background subtraction) followed by a probabilistic and appearance based tracking [3]. At each frame *t* the blobs can be segmented with standard background suppression algorithms. Since shadows and background changes can frequently occur, we included in the segmentation process a shadow suppression module and we adopt a dynamically adaptive model of the background [2]. Thus, after the segmentation and the tracking, in addition to the current blob B (Fig. 3), the appearance image AI (or temporal template) and the probability mask PM of the object (Fig. 3) are available. Probabilistic and appearance based tracking is frequently adopted for objects with non rigid motion and variable shape, like people. Details are reported in [3]. The approach is able to cope with large occlusions due to static objects as in Fig. 3 or to other moving people as in Fig. 2. Finally, the system roughly classifies people as adult or child and detects the main posture as in [3].

4. 3D virtual environment

In this paragraph we present an application of our system, which consists of video surveillance of an indoor environment. First of all, the 3D virtual indoor environment has been built by using the JSR 184 software environments as described above. The background model is a 3D model made of mesh objects that refer to static elements presents in the scene: table, chairs, cabinet with TV and pavement. For representing the scene we used two ambient lights sources (instances of the ambient node class) in order to obtain a photorealistic rendering of the indoor environment. The human figures are instances of the correspondent scene graph node, and include simple human textures. In fact a photorealistic representation of the human figures is less crucial (in this application case study) than the actions taking places in the virtual environment. The

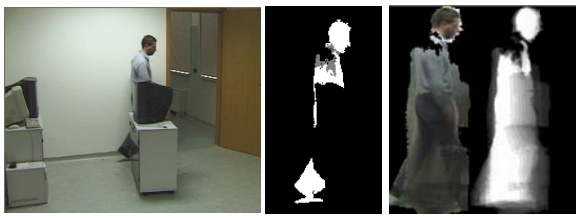

Figure 3: A person occluded by furniture, his detected blob, his appearance image AI, and his probability map PM

Figure 2: Blobs belonging to multiple objects are split by the probabilistic tracker

application takes as input a textual data stream coming from the video surveillance extraction module. These data stream contains the extracted positions and postures of the tracked human figures. The 3D virtual environment reconstruction module takes these data and renders the human figures placing them in the environment according to their position and posture data.

The rendered postures are taken as classes from the extraction module, in the presented system four major classes are considered: standing, crawling, laying and sitting. In order to enhance the system performances the animations are executed at human figures positioning level and not on the models themselves. The system is already capable of implementing animations interpolations as already written in the past paragraph, but it is seems more reasonable to use them for other kind of domains like children surveillance (where the running speed could be a measure of the actual danger and it's crucial to visually represent that data).

The 3D virtual environment also has many advantages, especially the possibility of changing the point of view. We support three different points of view:

- *Standard view*: this view is basically the same as the calibrated cameras positioned in the real environment. It is very useful in order to have a general view of the scene.

- *Bird eye view*: this view is very suitable for identifying the distances among the human figures and the objects in the scene. Spatial positioning is an important synthetic information to be used as a visual clue for detecting which actions are taking place in the environment.

- *Interactive view*: this view, not only presents the scene with a certain angle (usually the same as the standard view) but it is able to navigate the scene by using three different camera motion modalities. These modalities are: move (allow translations with respect to *z* and *x* planes), rotate (rotation among all the planes), and float (allow translations with respect to *y* and *x* planes).

The 3D virtual environment thus could represent and effective approach for easily visualize and detect people and action in a video-surveilled environment. The use of multiple synthetic points of view clearly helps users in identifying different situations and by different angles of view. Since our system is extensible, many different points of view could be imple-

mented, for instance for certain domains, a first person view could be useful for understanding the danger or damages occurred in the environment (elderly people surveillance, and care).

5. Software components

The software components used in our approach are based on M3G (Mobile 3D Graphics) library. This library consists of represents a high-level API (Application Programming Interface) for the implementation of graphics rendering on mobile device. This choice leads us to develop a system in which the 3D scene rendering is carried out by on client side (mobile devices).

The mobile 3D graphics libraries used in our system are included in the JSR 184 standard specification, thus enabling the support by a wide set of devices; moreover this is a general approach because of the standard adoption of the M3G file format specification.

The M3G file format represents all the objects present inside a three-dimensional scene through the use of a tree structure called scene graph. Every node of this structure describes and defines any physical or abstract object of three-dimensional worlds (cameras, lights, meshes, animations).

The root of this tree is always a World object. The following figure 4 shows a reduced version of the tree structure used for building the presented 3D virtual environment.

In order to simplify the diagram, in the above scene graph we have not inserted the Mesh objects that refer to static elements presents in the scene: table, chairs, cabinet with TV and pavement. In our scene model two different classes of lights are defined: spot and ambient. The Background object is represented by a background node with a colour attribute in the format RGBA (Red, Green, Blue, and Alpha for transparencies). The interactive camera is a child of a Group object. This object has the task of managing the camera translation and rotation on the y axis. The camera object manages rotation on the *x* axis. With this separation all the changes on *x* axis don't have effect on camera movements and rotation on *y* axis.

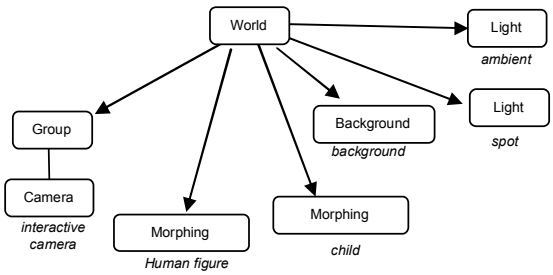

Figure 4. Tree structure of the 3D virtual environment

The more complex objects are the two MorphinMesh that contain all necessary data for the visualization and animation of classes of human figures (e.g., men, children). The MorphingMesh class allows to define models animated through morphing techniques: it requires to define the key models of an animation and an automatic interpolation procedure will compute shapes vertexes transformations needed for a smooth animation. By using the M3G file format and supporting the JSA 184 standards we noticed that the application jar file including models images and textures, testing data and source code does not exceed the 90kb.

More in detail M3G is a new standard file format, with extension m3g, used for storing all the scene graph data and the information for loading them in a custom application. In this manner the data of a scene, comprised the animations, can be created using common existing three-dimensional modelling programs. However, there is not a standard approach for creating and exporting this kind of files. Thus we decided to use a particular technique to import models in obj (wide spread computer graphics format) data type.

Our approach for creating M3G files is inspired by Andrew Davison work [5], which is based on the Java3D API. The approach consists of converting three-dimensional models in a Shape3D object from which is possible extract the vertexes, normal vectors and coordinates of texture lists. These lists are then optimized through the use of triangle strip. Finally a class is generated containing all the necessary methods to create and manage (starting from the acquired lists) a three-dimensional object in M3G.

6. Experiments

We provide several experiments in both indoor and outdoor environments. The real-time performance of the served-side depends on the number of cameras, the number of people and their size in the image space. In outdoor environments with cameras mounted in a high position tens of people can be captured from 4 cameras at about 10fps. The same performance is achieved indoor with 2 or 3 people as in the previous examples.

Figure 5. Example of different views of a3D virtual environment. a) standard view, b) interactive view, c) bird-eye view.

An example of the output of the indoor system is reported in Fig. 5, where the three different views (standard view, interactive view, and bird-eye view) are shown. The correspondent input frame together with the output of the vision module is depicted in Fig. 6.

7. Conclusions

In this paper we present a distributed video-surveillance framework. Our end is the remote monitoring of the behavior of people moving in a scene exploiting a virtual reconstruction on low capabilities devices, like PDAs and cell phones. The main novelty of this system is the effective integration of the computer vision and computer graphics modules.

The automatic video surveillance module extracts sufficient information to allow a virtual reconstruction of the environment on low capabilities devices, and, differently than a video stream, the bandwidth required to transmit this data is affordable and the system is working in real time.

8. References

[1] M. Bertini, R. Cucchiara, A. Del Bimbo, A. Prati, "An Integrated Framework for Semantic Annotation and Transcoding" in Multimedia Tools and Applications - Kluwer Academic Publishers, vol. 26, n. 3, pp. 345-363, 2005

Figure 6. An input frame from a monitored room (left) and the correspondent output of the video surveillance module (right). Blobs, positions, and postures of the people detected are superimposed.

[2] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts and shadows in video streams". *IEEE Trans. on PAMI*, 25(10):1337–1342, 2003

[3] R. Cucchiara, C. Grana, G. Tardini, R. Vezzani, 2004, "Probabilistic People Tracking for Occlusion Handling", in *Proc. of ICPR,* vol. 1, Cambridge, UK, pp. 132-135, Aug, 23-26 2004

[4] R. Cucchiara, C. Grana, A. Prati, R. Vezzani, "Probabilistic Posture Classification for Human Behaviour Analysis" in IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, vol. 35, n. 1, pp. 42-54, 2005.

[5] A. Davison, "Killer Game Programming in Java", O'Reilly Media, May 2005.

[6] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: real-time surveillance of people and their activities," IEEE Transactions on PAMI, 22, n. 8, pp. 809–830, (2000).

[7] A. Mihailidis, B. Carmichael, J. Boger, "The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home", IEEE Transactions on Information Technology in Biomedicine, 3(8), pp.238-247, 2004.

[8] A. Vetro, T. Haga, K. Sumi, and H. Sun;. Object-based coding for long-term archive of surveillance video. In Proc. of IEEE Int'l Conference on Multimedia & Expo, volume 2, pp. 417-420, 2003.