

IP-DISTRIBUTED COMPUTER-AIDED VIDEO-SURVEILLANCE SYSTEM

B Georis¹ X Desurmont² D Demaret² S Redureau² JF Delaigle¹ B Macq¹

¹Université Catholique de Louvain, Belgium

²Multitel ASBL, Belgium

ABSTRACT

In this article we present a generic, flexible and robust approach for an intelligent real-time video-surveillance system. The proposed system is a multi-camera platform that is able to handle different standards of video inputs (composite, IP, IEEE1394). The system implementation is distributed over a scalable computer cluster based on Linux and IP network. Data flows are transmitted between the different modules using multicast technology, video flows are compressed with the MPEG4 standard and the flow control is realized through a TCP-based command network (e.g. for bandwidth occupation control). The design of the architecture is optimized to display, compress, store and playback data and video flows in an efficient way.

This platform also integrates advanced video analysis tools, such as motion detection, segmentation, tracking and neural networks modules. The goal of these advanced tools is to provide help to operators by detecting events of interest in visual scenes and store them with appropriate descriptions. This indexation process allows one to rapidly browse through huge amounts of stored surveillance data and play back only interesting sequences.

We report here some preliminary results and we show the potential use of such a flexible system in third generation video surveillance system. We illustrate the interest of the system in a real case study, which is the surveillance of a reception desk.

Keywords: multi-camera, distributed architecture, real-time, multi-threading, multicast, computer vision, intelligent visual surveillance, intelligent storage

1. INTRODUCTION

In this article we present a generic, flexible and robust approach for an intelligent real-time video-surveillance system. Video security is becoming more and more important today, as the number of

installed cameras can attest. Nevertheless, there is still a need for complete and generic systems that can be inserted in an existing camera network (analogic or numeric) and still handle the various ways of video transmission (Firewire, IP, BT,...). Examples of challenging applications are monitoring metro stations (in Cupillard et al (1)) or detecting highways traffic jam, intelligent content access, detection of loitering,...Marcenara et al (2) call this type of platform third-generation video surveillance systems. The requirements for these systems are to be network-connected, multi-cameras, modular, the display must be user-friendly, the vision modules should be plug-and-play and the overall system must be highly reliable and robust.

The work reported here has both research and industrial motivations. Our goals are first to obtain an efficient system that can meet the strong industrial requirements and second to have a system that allow researchers to develop new vision algorithms. Such a system must for example include evaluation facilities, such in Jaynes et al (3).

The proposed system is a multi-camera platform that is able to handle different standards of video inputs (composite, IP, IEEE1394) so that it can be combined with existing systems (e.g., CCTV). A computer cluster based approach with fast ethernet network connections is the innovative solution proposed to address the increasing needs of computing power at an affordable cost. The main advantage of this approach is the flexibility and the robustness.

Many systems usually require big pipe and MJPEG over ATM techniques to design the architecture of huge monitoring systems. We show here how the MPEG4 standard and a classical IP architecture allows the system to monitor a high number of cameras while ensuring small transmission delays. A careful interconnection of small ethernet networks allow us to fit to the required size of the system.

We also introduce a first intelligent vision module that is able to track people and cars in varying conditions and to raise an alarm in predefined quite simple scenarios.

We report here some preliminary results and we show the potential use of such a flexible system in third generation video surveillance system. We illustrate the interest of the system in a real case study, which is the surveillance of a reception desk.

The paper is organised as follows: section 2 describes the global system and its main characteristics, section 3 goes deeper in the understanding of each underlying module, section 4 is devoted to the image analysis module, section 5 shows the performance obtained with this powerful approach and section 6 concludes and indicates future work.

2. OVERALL SYSTEM OVERVIEW

The major components of the physical architecture are presented in figure 1. Basically, the system is composed of computers connected together through a typical fast Ethernet network (100 Mb/s). The various cameras are plugged either on an acquisition card on a PC or directly on the local network hub for IP cameras. A human computer interface and a storage space are also plugged on this system. A web server will soon be added to remotely browse stored events. The main advantage of such an architecture is the flexibility. Future needs in computing power will be simply addressed by adding a PC in the cluster. A new camera can be plugged and configured easily. Finally, we can design a system of any number of cameras by connecting two or more basic architectural elements through IP and routers.

The logical architecture has been designed in a modular way to allow a fair resource allocation over the cluster. In its implementation, each software module is dedicated to a specific task (e.g., coding, network management,...) and will be

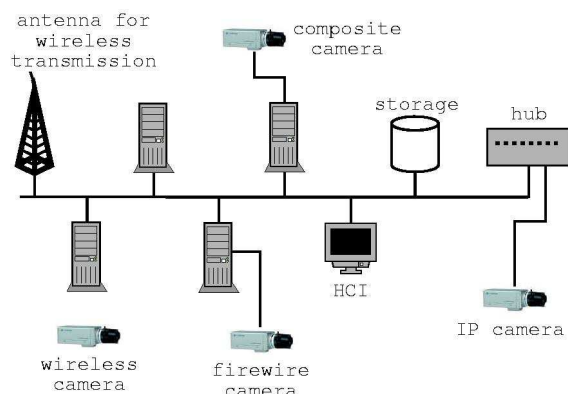


Figure 1 : The physical architecture.

discussed in section 3. The various treatment modules are distributed on the PC units according to a configuration file. In this file, the operator will define the architecture of the distributed system and moreover, he will be able to customize the action of the different modules (e.g., the vision processing units require a lot of parameters). A master process will be assigned the task of communicating the configuration to all the PC units. A change in the distribution of the different tasks between the units will simply require to change the configuration file (no need to compile again).

The robustness of the overall system is provided by the logical architecture. It manages the various problems that can arise: a network packet loss, a network transmission interruption, a hard drive stopping... Moreover, we can improve the overall system performances if we dedicate a vision task to a co-processor.

3. SYSTEM COMPONENTS DESCRIPTION

The various modules of the software part of the system are explained hereunder. We successively explain the acquisition module, the codec module, the network module, the storage module and the human computer interface module. The section 4 will depict the image analysis module.

3.1. Acquisition

There are two main reasons to handle a large variety of video inputs. First, many industries can not afford the cost of changing all the installed video surveillance equipment from CCTV to full digital cameras distributed on a network. Second, researchers have to test their video analysis tools on various inputs in order to prove their algorithms or to make them more robust. We are currently able to handle several protocols: IP (JPEG and MJPEG), IEEE1394 (raw and DV), wireless (analogic) and composite (PAL, NTSC, SECAM).

In order to test our algorithm on recorded particular sequences, an AVI player has also been designed just as a stream-builder from JPG images (some well-known test sequences are given with this format e.g., PETS 2001).

A time stamp is attached to each frame at grabbing time. This information is very useful in the subsequent processing stages (e.g., the tracker).

3.2. Codec

The goal of the coding process is to have a good compromise between the compression ratio and the bandwidth occupation. We propose here a MPEG4 compression scheme since it outperforms classical MJPEG encoding.

The main disadvantage of MJPEG is that it does not use any temporal redundancy to increase the compression factor. In our experience MPEG4 and MJPEG compression factor are in a ratio of 1:10 for the same quality. Indeed videosurveillance scenes are quite static when cameras are fixed. Compression methods suppressing temporal redundancy, such as MPEG4, are therefore more efficient.

In counterpart, we can not access images independently anymore. We have to resynchronize the flow on a I-picture (intra picture) each time a network transmission occurs. The number of I pictures per second is determined with respect to the VCR that replay the stored sequences. Moreover, in order to limit the delay between the encoding time and the display time, we do not encode pictures as B-pictures (bidirectional encoded pictures) to keep a causal decompression scheme.

This technique allows us to transmit up to 20 CIF (352x288) video flows at 25 fps on a typical 100baseT network. Thanks to this compression scheme, we are able to transmit video flows on new mobile phone networks (e.g., when an alarm is generated).

3.3. Network

Having a distributed system implies an efficient use of the bandwidth. We have seen that the various modules related to a video input can be dispatched on several computers. For example, we can have the acquisition on computer 1, the storage on computer 2 and the display on computer 3. We have chosen a multicasting technique to solve the bandwidth occupation problem. Each video source has an associated multicast channel. This multicast channel is accessed through a UDP connection by every module that needs the video input. Since UDP does not offer a quality of service (QoS), we have developed a protocol that can detect when a transmission failure occurs. We guarantee small delays because the network load is controlled in order to avoid buffer queuing. This delay is small enough to be imperceptible for the user.

3.4. Storage

The storage module has to deal with the enormous quantity of data to store. It must allow a 24 hours a day storage. This module has two levels: level 0 is a classical storage process with the MPEG4 technology. This level stores a CIF video flow at 25 fps for three days on a 40 Gb hard drive. We can further improve this number if we allow a two passes encoder to have a constant quality stream. Up to now, we have a constant bandwidth stream.

Level 1 is an intelligent storage process. It stores only interesting events that the user has defined. This level saves a tremendous storage space. Moreover, it allows a fast search to retrieve a stored sequence.

3.5. Graphical User Interface

From an end-user point of view, the graphical user interface (GUI) is one of the most important modules. It must be very easy to use but efficient. We have designed such a powerful GUI for a demonstrator.

First, it provides various modes to display the video inputs currently available on the network. For example, mode 1 displays simultaneously all video inputs, mode 2 displays a full screen for a specific camera and mode 3 only displays cameras selected by the user.

Second, it owns a VCR so that a user can make a request to play back a stored sequence. In such a case, the request is handled instantly through the TCP-based command network. An acquisition module is launched and the stored video is transmitted to the GUI. Classical options are available: play, pause, stop, fast forward, rewind.

Third, it allows a user to define contextual information either globally (information corresponding to many cameras viewing the same scene) or specifically. These contextual information is represented by means of 2D polygons on the image, each of them having a list of attributes: IN_OUT zone, NOISY zone, OCCLUSION zone, AREA_OF_INTEREST zone,... This type of information is fed to the image analysis module to help the scenario recognition process and the alarms management.

4. IMAGE ANALYSIS MODULE

A first tracking module has been added to our platform. This work is derived from the method proposed by Piater and Crowley (4).

First of all, a number of zones in the image where moving objects can appear are defined by the user through the GUI.

Second, a detection module takes the list of zones as input and compute a list of blobs. For each image zone, blobs are created by applying a clustering step on the corresponding zone of the detection image D . This detection image D is the thresholded difference between the current image I and the background image B , computed according to the following equation:

$$D = thresh(\min(\sum_i |I_i - B_i|, I_{max}), t) \quad i=r,g,b \quad (1)$$

where t is a parameter of the algorithm. The background is updated to take small illumination changes into account:

$$B_i = \alpha I_i + (1 - \alpha) B_{i-1} \quad (2)$$

where α is a parameter of the algorithm.

Third, each blob parameters (first and second spatial moments) are fed to a kalman filter. This filter obtain a prediction for these parameters and a gaussian region of interest (ROI) is defined, which is centered at the predicted location of the blob. These ROI are added to the list of zones the detector has to segment.

Fourth, a careful analysis is conducted on the list of blobs obtained in the current frame to resolve ambiguities that can appear. This analysis try to handle split and merge cases, start and end of tracks, static and dynamic occlusion.

The main advantage of this approach is its robustness to noise and its computational efficiency. Results are presented in the next section.

5. RESULTS AND PERFORMANCE

We report hereunder the results obtained with a real case study which is the surveillance of a reception desk. The actual system implementation is composed of four P4 processors cadenced at 1.5 GHz, running C++ code under Linux. The software modules implementation relies on POSIX threads for multitasking scheduling. Three different video inputs were used in this experiment: IEEE1394, composite and IP. All computers are network connected with a 100 Mb/s bandwidth pipe. Up to nine cameras are acquired, stored, tracked and displayed at 25 fps. These experiments were conducted on several sequences.

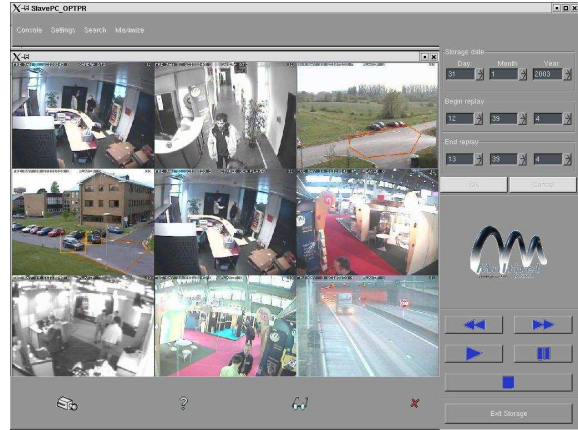


Figure 2 : A view of the GUI.

Figure 2 shows a screen shot of the GUI in mode 1. The system is currently running nine cameras, some of them are indoor scenes and other ones are outdoor scenes. The right part of the picture shows the VCR started by the user. Once he has encoded the date of the sequence he want to replay, the system will start the replay in a specific window.



Figure 3 : Alarm generation for a human intrusion in a predefined area of interest.

Figure 3 is a zoom of what is happening on a particular camera since an alarm has been generated. The region of interest defined by the user is the front of a reception desk and it is represented on the picture by the green polygon. The bounding box of a detected person is represented in red. The system has generated an alarm since the human with identifier number 0 has entered the region of interest without stopping in it (to contact the receptionist). So, the system has concluded that the person is entering a private zone. In such a case, the trajectory of the human is also displayed (in red).

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an original approach for a third generation video surveillance platform that can provide the flexibility needed by researchers and that can meet the strong efficiency requirements of industrial applications. This first version of the system is shown to be very efficient and it has been validated on various sequences. We are currently investigating new vision modules e.g., better tracking methods. Other extensions and improvements will be made on the global system. For example, we are adding a message manager for a better communication and a better collaborative or concurrent processing between the various modules.

Acknowledgements: this work has been granted by the Walloon Region under the FEDER project 171, the FIRST SPIN OFF and the FIRST EUROPE programs.

REFERENCES

1. Cupillard F, Brémond F and Thonnat M, "Tracking groups of people for video surveillance", 2nd European Workshop on AVBS Systems, 1,
2. Marcenara L, Oberti F, Foresti L and Regazzoni C, "Distributed architectures and logical-task decomposition in multimedia surveillance systems", Video Communications Processing and Understanding for 3GSS, Proceedings of the IEEE, 89, 1419-1440
3. Jaynes C, Webb S, Steele R and Xiong Q, "An open development environment for evaluation of video surveillance systems", 3rd Int. Workshop on PETS, 1, 32-39
4. Piater J and Crowley J, "Multi-modal tracking of interacting targets using gaussian approximations", 2nd Int. Workshop on PETS, 1,