Intelligent Video Storage of visual Evidences on Site in fast Deployment

Xavier Desurmont, Arnaud Bastide, Jean-François Delaigle
Multitel A.S.B.L., Av Copernic, 1, B-7000 Mons, Belgium.
www.multitel.be

ABSTRACT

In this article we present a generic, flexible, scalable and robust approach for an intelligent real-time forensic visual system. The proposed implementation could be rapidly deployable and integrates minimum logistic support as it embeds low complexity devices (PCs and cameras) that communicate through wireless network. The goal of these advanced tools is to provide intelligent video storage of potential video evidences for fast intervention during deployment around a hazardous sector after a terrorism attack, a disaster, an air crash or before attempt of it. Advanced video analysis tools, such as segmentation and tracking are provided to support intelligent storage and annotation.

Keywords: real-time, forensic video, intelligent visual surveillance.

1. INTRODUCTION

Video security is becoming more and more important today, as the number of installed cameras can attest. CCTV (closed circuit television) domain, after broadcast television is currently migrating from analog to digital. Due to theses innovations, miniaturizations and standardizations, it is now possible to deploy easily and rapidly CCTV in site for a permanent or temporary uses. Recent bomb attacks in Europe (Madrid 11th March 2004) have shown again the risk of terrorist attacks in public areas like trains, metros and stations. Examples of challenging surveillance applications [1] are monitoring metro stations [2] detection of loitering or abandoned objects [3]. The requirements for these systems are to be network-connected, multi-cameras, modular, generic, flexible, real-time and robust (low false alarm rate and low missed alarm rate). Our goals are to obtain an efficient system that can meet these strong industrial requirements. A computer cluster based approach with network connections is the innovative solution proposed. The main advantage of this approach is its flexibility. Since mobile objects are important in video-surveillance, the system will include image analysis tools such as segmentation and object tracking (both vehicles or people). Change detection [4] is based on frame differencing and/or reference images and segmentation based on shape, color or texture analysis.

The paper is organized as follows: section 2 describes the global system and its main characteristics for understanding each underlying module, section 3 is devoted to the image analysis module, section 4 illustrates the interest of the system in a real case study, which is the “outside building surveillance”. Section 5 concludes and indicates future works.

2. SYSTEM OVERVIEW

The major components of the architecture are presented in fig. 1. Basically, the system is composed of computers connected together through a typical network (wire of wireless). The various cameras are plugged either in an acquisition card in a PC or directly in the local network hub for IP cameras. A human computer interface and a storage space are also plugged on this system. 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. Complexity is a major issue for IVS systems, as shown in [5], especially when real-time IVS systems require dedicated hardware. It is very difficult to embed IVS algorithms directly in the camera, while keeping the performance constant. When it is not possible to do so, an additional computer vision module could be plug in the cluster as shown in Fig. 1. Fig. 2 shows a screen shot of the graphical user interface (GUI). The system is currently running nine video streams, some of them are indoor scenes and other ones are outdoor scenes. The right part of the picture shows the virtual VCR commands.
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,...). 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 configurations parameters to all the PC units. Changes in the distribution of the different tasks between the units could be performed dynamically during the run-time of the system.

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, a frame loss in the video acquisition process, etc. A powerful data management handles the distribution of information along the whole system.

The various modules of the software part of the system are explained hereunder. We explain the acquisition, the coding-decoding (codec), the network (with a section dedicated to security) and the storage modules. Section 3 will describe the image analysis module.

2.1 Acquisition
There are two main reasons to handle a large variety of video inputs. First, many industries cannot afford the cost of changing all their installed video surveillance equipment from CCTV to full digital cameras distributed on a network. Second, we 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 (analog, wifi) and composite (PAL, NTSC, SECAM). A time-stamp is attached to each frame at grabbing time. This information is very useful in the subsequent processing stages.

2.2 Codec
The goal of the video coding process is to reach a good compromise between the compression ratio (bandwidth occupation) and processing resources. Many CCTV systems are using proprietary video encoding schemes, which limits the interoperability and the emergence of low cost solutions. We propose here a MPEG4 SP 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 SP and MJPEG compression factor are in a ratio of 1:10 for the same quality. Indeed, video-surveillance scenes are quite static when cameras are fixed. Compression methods suppressing temporal redundancy, such as MPEG4 or H263, are therefore more efficient. New codecs are now emerging as such MotionJPEG 2000 and MPEG-4 AVC/H264, the two latest video/image coding standards coming from ISO/IEC/ITU.
2.3 Networking
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, the system can perform the acquisition on computer A, the storage on computer B and the display on computer C. Network implies delays. The decrease of the delay could be improved by passing through the ISO networks layers and use IP packets with the knowledge of its implementation on the physical layer. This delay is small enough to be imperceptible for the user (i.e. less than 150 ms). We are using both wire and wireless architecture. IEEE 802.11 Wireless Local Area Networks (WLAN) as an extension to the existing wired infrastructure, is rapidly growing, offering the mobility and easy deployment of equipment. In order to provide video services on such network, an application must be able to find these and to understand their capabilities. In addition, wireless connections face frequent changes and losses of physical connections. There is a need to study and to develop mechanisms to cope with that. Currently most WLANs are used for data transfer, but the high bandwidth provided will increase their use for multimedia transmission such as video technology.

2.4 Security
Security is a crucial problem. There are two security aspects that are investigated within the deployed system. The first one is the storage of a copy of the video content in a geographically different place. Indeed, in disasters, the storage unit in the site could be destroyed, but also it is known that dishonest people in the local security team could maliciously erase the evidence directly in site after some events.

The second security aspect is the protection of infrastructure and content from external ill-intentioned people. Security issues are currently the drawback of wireless networks [6] and current work has been undertaken by IEEE (802.11i) in order to solve the current weaknesses. The first problem to be addressed is the security of the communications. The WEP protocol (Wire Equivalent Privacy) has been cracked, while using the RC4 algorithm and it is now possible to eavesdrop on private communications [6]. Sometime the WEP protocol is not even not activated by the user! Some examples have already proven this lack. The security of 802.11 is deeply under consideration and more secure solutions will emerge in the future. The Wi-Fi Alliance Security roadmap is well defined. 2004 is the year of emergence 802.11i. Besides the security of the network, the security of the content itself is crucial. Indeed, content in digital form can easily be accessed, manipulated, copied and distributed at negligible cost. In the case of surveillance applications, security of the content will ensure the integrity of the system and the privacy of the users. Therefore, it is of the utmost importance to protect content from illegal access, copying and tampering. Pertinent problems include the control the access to the content and the determination of the content integrity, namely that it is authentic and has not been tampered with. MPEG-4 and JPEG 2000 have also developed specifications, namely IPMP Extensions and JPSEC respectively, which provide hooks to enable the use of a content security system. We are now investigating such solutions and others like fragile watermarking.

2.5 Storage
The storage module has to deal with the enormous quantity of data to store. It must allow a 24-hours per 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 7 days on an 80 GB hard drive. We can further improve this number if we allow a second-pass 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 tremendous storage space. Moreover, it allows a fast search to retrieve a stored sequence. The intelligent storage needs image analysis; we describe it in next section.

3. IMAGE ANALYSIS MODULE
The architecture of the vision part of the system is divided in three main levels of computation that achieve the interpretation (Fig. 3): Image level (image filtering, background evaluation and segmentation), blob level (description, blobs filtering, matching, tracking description and filtering), event level (tracking analysis, finite state machine, performance evaluations). We will only describe here the image and object levels as the event level is quite application dependant.
3.1 Pre-processing and segmentation
After acquisition, the current image is filtered in the pixel domain to decrease the spatial high-frequency noise. The image is then downsized to reduce the need in computational resources of the segmentation process. Many reference image models could be used for representing the backgrounds as they are implemented in the system. Thus the segmentation process is divided in two stages: update and estimation. The segmentation is fully configurable as one can choose between three types of background models (e.g. mixture of Gaussian [7], low pass temporal recursive filter [8], temporal median filter) and foreground extraction processes, even for non-background segmentation like frame differencing. For the application shown in section 4, the system will only use the mixture of Gaussians background, as it is quite robust to common noises such as monitor flattery or branches moving in trees. The update stage will revise the mixture of Gaussian for each pixel background, and the estimation stage will extract the a priori foreground, filter it via morphology processing and then label objects.

![Diagram of the vision system components]

Fig.3. Design of the vision system components.

3.2 Blobs description and filtering
The aim of blobs description and filtering is to make the interface between segmentation and tracking and simplify the information. The description process translates video data into a symbolic representation (i.e. descriptors). The goal is to reduce the amount of information to what is necessary for the tracking module. The description process calculates, from the image and the segmentation results at time t, the $k$ different observed features of a blob $i$: 2D position in image, 3D position in the scene, bounding box, mean RGB color, 2D visual surface, inertial axis of blob shape, extreme points of the shape, etc. At this point, there is another filtering process to remove small blobs, blobs in an area of the image not considered, etc. Other model-based vision descriptors could be also integrated for specific application such as vehicle or human 3D model parameters.

3.3 Tracking algorithm
The tracking part of the system is flexible and fully parametrical. The set-up should be done for a trade-off between computational resources, needs of robustness and segmentation behavior. It is divided in four steps that follow a straightforward approach: estimation, cost matrix computation, matching decisions, tracks updates. Note that there are multiple predictions and cost matrices when the last matching decision is not unique, and there are only multiple matching decisions for some matching algorithms in MHT (multiple hypothesis tracking [9]). Fig. 4 explains briefly the architecture.

The tracking filtering is processed at the tracking description output. It is just as necessary as the other filters of the vision system. As usual the filter is used to decrease the noise. At this level of processing, it can use the temporal
consistency. We described above some types of filters that can be used in chain. Because the tracking description is a construction built piece by piece during the progression of the video sequence, it can process on-line or off-line. One filter detects and removes tracks that last for less that a fixed duration. This kind of noise comes when the segmentation detects noise in the image as an object. Another filter simplifies tracks by removing samples of blobs that give poor information (e.g. If the blob moves slowly). It could be seen as a dynamic re-sampling algorithm.

Fig. 4. Basic architecture of Tracking.

4. CASE STUDY

The developed system has a lot of high-level processing modules (detection of activity, abandoned object, missing object, people running too fast, car park management, etc.) that can be run as options in accordance with the end-user application. Here we decide to focus on one: “outside building surveillance”. The objective is the surveillance of the external perimeter of the building and especially the backside. The video camera could visualize the area described in Fig. 5. A typical application is the monitoring of a site or a building after a catastrophe (fire, explosion), when one wants to detect any human presence after the deployment of the camera network. Such a surveillance system is also useful when it is too expensive or dangerous to rent operators to monitor a site locally or remotely.

Fig. 5. The site to secure.

The human detection is done from the video stream given by the camera. This analysis generates events when a person is detected in the secure zone and more precisely in the three detection zones near the picture windows and when an activity is detected at the terrace roof near the emergency door. Each event stores the time and duration of the detection, the estimated size of the target and the concerned zone, the type of event. According to the set-up, the event should be repeated at a given rhythm as soon as the presence of object is still detected. The outdoor analysis has been created to cope only with a sufficient lighting. Therefore a suitable lighting should be added in the site for night processing. The zones are defined graphically by users.
4.1 Processing of human detection module
First mobile objects are described as shown in previous section (background model, extraction of blobs, description of
blobs) and in fig. 6. The human detection module takes as inputs the description done by the segmentation and gives as
outputs the event of person detection. To detect if an object is in the secure zone, the system checks if the position of the
object is in the zone, and then does an hysteresis for this zone for the time and size aspects of the object. Fig. 7 shows
the heuristic of detection.

Fig. 6. Left: the current image. Bottom-right: the background model. Top-right: the segmentation map.

Fig. 7. Decision heuristic for the human presence detection in a 2D zone.
\( \sigma_{\text{high}} \) and \( \sigma_{\text{low}} \) are spatial thresholds (surface size of mobile objects) for the change of state (active/inactive).

- \( \sigma_{\text{high}} \) is the surface threshold over which an object in the region could become active if it stays for a sufficient duration.
- \( \sigma_{\text{low}} \) is the surface threshold below which an active object becomes inactive if it stays for a sufficient duration.

\( T_{\text{high}} \) and \( T_{\text{low}} \) are the temporal thresholds for change of state.

- \( T_{\text{high}} \) is the delay threshold for detection when the surface is over \( \sigma_{\text{high}} \).
- \( T_{\text{low}} \) is the delay threshold for stopping detection when the surface is below \( \sigma_{\text{low}} \).

When a zone is active, an alarm is triggered and the video sequence is sent to the operator. The robustness of this module is good and has a few false detections. This is very important to have few false detections, because if there are too many false detections, the operator could decide to directly ignore all alarms or deactivate the whole system.

5. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed an original approach for a third-generation video-surveillance platform [10] that can provide the strong efficiency requirements of industrial applications.

We are currently investigating new vision modules, (e.g. better segmentation and tracking methods). Moreover, other extensions and improvements will be made on the global final product system. The encoding of content meta-data description, will be integrated according to the MPEG-7 philosophy. Because IPv6 offers richer services than IPv4, such as hierarchical addressing for routing optimizations, it could be interesting to integrate it.

Acknowledgements: this work has been granted by the Walloon Region under the FEDER project 171.

REFERENCES


