SMART CAMERAS WITH REAL-TIME VIDEO OBJECT GENERATION

Alessio Del Bue[1] *Dorin Comaniciu*[2] *Visvanathan Ramesh*[2] *Carlo Regazzoni*[1]

[1]Department of Biophysical and Electronic Engn.
University of Genova
Via All'Opera Pia 11a, 16145 Genova, Italy

[2]Vision and Modeling Department
Siemens Corporate Research, Inc.
755 College Road East, Princeton NJ 08540

ABSTRACT

This paper presents a system for video object generation and selective encoding with applications in surveillance, mobile videophones, and automotive industry. Object tracking and MPEG-4 compression are performed in real-time. The system belongs to a new generation of intelligent vision sensors called **smart cameras**, which execute autonomous vision tasks and report events and data to a remote base-station. A detection module signals the object of interest presence within the camera field of view, while the tracking part follows the target to generate temporal trajectories. The compression is MPEG-4 compliant and implements the Simple Profile of the standard, capable of encoding up to four video objects. At the same time, the compression is selective, maintaining a higher quality for the foreground objects and a lower quality for the background representation. This property contributes to bandwidth reduction while preserving the essential information of foreground objects. The system performance is demonstrated in experiments that involve objects representing faces and vehicles seen from both static and moving cameras.

1. INTRODUCTION

The 3G digital cellular technology [11] will soon provide increased bandwidth: up to 384K bit/sec when a device is stationary or moving at pedestrian speed, 128K bit/sec in a car, and 2M bit/sec in fixed applications. By combining this new communication framework with powerful vision algorithms, better sensors, and DSP chips with increased computational power and memory capacity, the concept of *smart cameras* becomes a reality. A smart camera is an autonomous vision-based device capable to perform intelligent tasks such as surveillance or obstacle detection while reporting to its base station events and data.

This paper presents a prototype real-time system that generates video objects of interest and encodes them selectively. Our system represents a step forward towards the implementation of smart cameras for surveillance [3,8], mobile videophones [6,7], and intelligent vehicles.

For these scenarios, the system task is to transmit to its base station high quality foreground data, while trading-off the quality of the background data. For surveillance, the base station will forward the data stream to a central processing unit for person recognition/re-identification. In the case of videophone the base station will transmit the data to another device. In automotive applications, communication is also needed between smart cameras.

The paper is organized as follows. Section 2 presents on overview of the system and Section 3 discusses the object detection and tracking module. The compression module is presented in Section 4. Section 5 shows experimental results.

2. SMART CAMERA OVERVIEW

The block diagram of a smart camera is presented in Figure 1. The detection and tracking module signals the object of interest presence within the camera field of view and provides the 2-D coordinates of the detected object and the estimated scales to the compression module. Based on the foreground and background data, this module generates MPEG-4 [12] compliant compressed video objects. Our software implementation is modular, involving multiple threads that are synchronized for the tasks of grabbing, detection, tracking, camera control, compression, and visualization.

When active cameras are used, the control module initiates commands that ensure the centering of the target in the camera view. Appropriate control of the pan, tilt, and zoom is an important phase of the tracking process. The camera should execute fast saccades in response to sudden and large movements of the target while providing a smooth pursuit when the target is quasi-stationary. We implemented a control mechanism that resembles the human visual system. The fovea sub-image occupies laterally about 6 degrees of the camera's 50 degrees field of view, at zero zoom. The communication with the Sony EVI-D30 camera is achieved through a standard RS-232C interface.

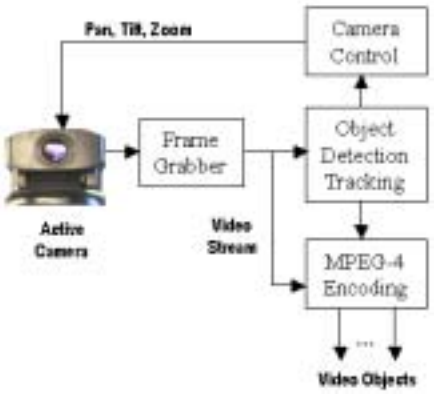


Figure 1. Block diagram of a smart camera with real-time video object generation and encoding.

However, contrary to other tracking systems that suspend the processing of visual information during the saccade movements [1], our visual tracker is sufficiently robust to deal with the large amount of blurring resulting from camera motion. Thus, the tracking is a continuous process, not interrupted by the servo commands.

3. OBJECT DETECTION AND TRACKING

The modules performing object detection and tracking are based on our recent work described in [4,5].

In the case of faces, a color model is obtained by computing the mean histogram of face samples recorded in the morning, afternoon, and at night. The dissimilarity between the face model and the face candidates is measured by a metric based on the Bhattacharyya coefficient. The gradient ascent mean shift procedure is employed to guide a fast search for the best face candidate in the neighborhood of a given image location. For more details, please see [5].

4. MPEG-4 MODULE

The MPEG-4 module [2,12] is based on the software recently made public by the International Organization for Standardization. We use a Simple Profile encoder [9] capable of processing up to four video objects of rectangular shape. The reference software implements motion estimation with full search (16 x 16 pixels) block-matching algorithm with forward prediction. It is not optimized, however, achieving only 15 fps on a QCIF (176 x 144 pixels) stream with two video objects, processed on a 900 MHz PC.

Nevertheless, using optimization at all levels, including new algorithms for intra and inter video object plane encoding, fast motion estimation, and MMX technology, much better performance is possible. A frame rate of 70 fps is reported for CIF resolution video (352 x 288 pixels) on 800 MHz PC with similar quality as the reference software [13].

There are already dedicated chips that perform real time MPEG-4 compression (see Matsushita or TI-DSC24). Since the detection and tracking modules can be easily implemented in DSP, the natural step forward is the DSP implementation of our entire system. We are currently investigating the VLSI solution [10].

5. EXPERIMENTS

The performance of the system is assessed in this section by analyzing experiments that involve both static and moving cameras.

5.1 Static Camera with Automatic Pan and Tilt

The first experiment was performed in an office environment with daylight (coming from a large window in the background) and artificial light. A human subject walks through the office and executes large and sudden movements. Only two QCIF video objects are created in this experiment, the subject's face and background. The entire sequence, called *Alessio_1* has about 300 frames. Four frames containing the composition of the two reconstructed video objects are presented in Figure 2. Observe that face data is decoded with much higher accuracy in comparison to the background data.

The detection, tracking, video object formation, and selective encoding are performed at a frame rate of 15 fps on a 900 MHz PC. Since the decoder merges together the video objects according to the segmentation mask, the reconstructed stream is a composition of a high quality video object (the face) and a low quality video object (the background). We use a texture quantization step of 4 for the face and 30 for the background.

As an objective dissimilarity measure we employ the Peak Signal to Noise Ratio (PSNR) between the original and reconstructed frames:

$$PSNR = 20\log_{10}\left(255/RMSE\right) \quad (1)$$

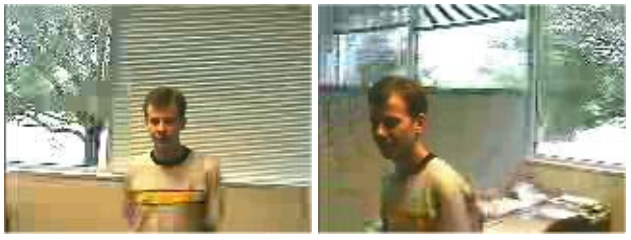
where RMSE denotes the Root Mean Squared Error, expressed by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - X_i^R\right)^2} \quad (2)$$

In equation (2) the original image value at position i is denoted by X_i, while X_i^R is the value of the decoded image and n is the number of pixels. For color images the formula (2) is applied for each color plane.

The PSNR values for each reconstructed frame are shown in Figure 3a for the face video object and in Figure 3b for the background. The PSNR of the background object varies significantly in time, about 6dB. The reason of the variation is due to both changes in the scene composition (regions with and without texture) and to camera motion. On the other side, the quality of the reconstructed face is remarkably constant over time, which

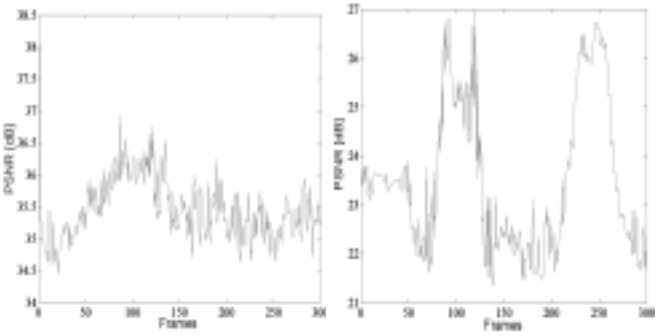
strengthens our conjecture that a recognition module can be successfully employed after the decoding.



#17 #76

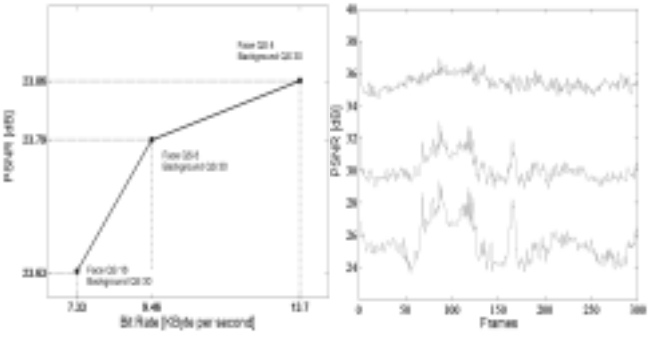
#119 #172

Figure 2. Reconstructed *Alessio_1* sequence. Compression ratio is 63.06.



(a) (b)

Figure 3. PSNR of the reconstructed data for *Alessio_1* sequence: (a) Face. (b) Background.



(a) (b)

Figure 4. Performance obtained by applying different quantization steps for the face object. (a) PSNR of the entire sequence function of the bit rate. (b) PSNR of the face only, function of the frame number.

The bit rate at the output of the encoder for various quantization steps (4, 8, and 16, respectively) applied to the face object is represented in Figure 4a, with the quantization step for the background maintained unchanged, equal to 30. The corresponding compression ratio is 63.06, 91.32, and 117.9, respectively. Figure 4b shows the resulting PSNR values computed only for the face object.

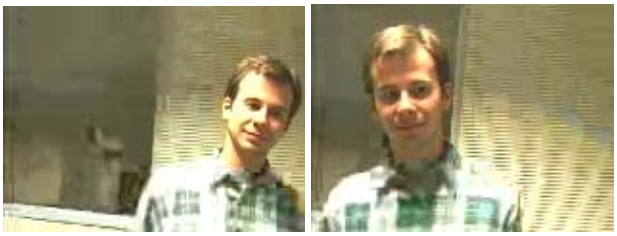
5.2. Moving Camera

We present next two other experiments, however, this time the camera is moving. In the first experiment the camera is hand held, while in the second experiment the camera is installed in a car and the tracker follows the vehicle from the front of the camera.

5.2.1 Walking Person Sequence

The original sequence *Alessio_2* contains 300 frames grabbed in an office with artificial light. The camera and the subject are moving simultaneously, uncorrelated with each other.

#1 #159



#268 #279

Figure 5. Reconstructed *Alessio_2* sequence. Compression ratio is 53.74. The PSNR is 35.5 dB for the face and about 26dB for the background.

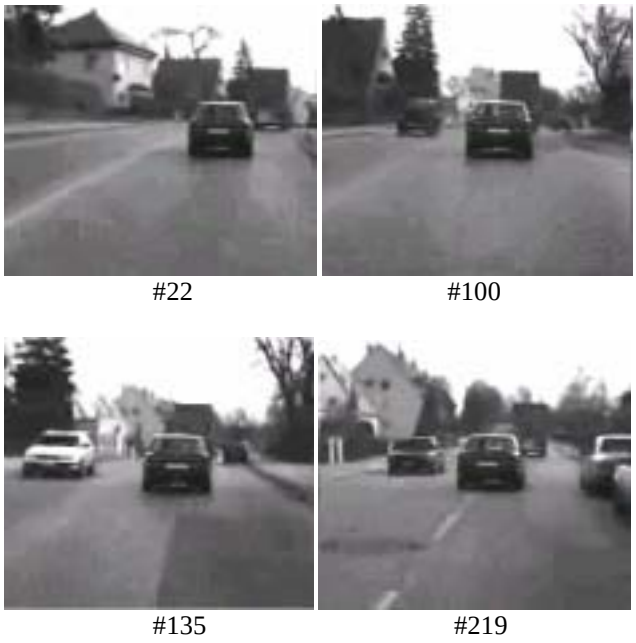
Observe the preservation of the face details in the reconstructed frames shown in Figure 5. This property of the system is remarkable, taking into account that the camera underwent large and sudden motions.

5.2.2 Vehicle Pursuit Sequence

Finally, we use as input the *Pursuit* sequence, containing about 300 frames grabbed in a moving vehicle. The sequence has a frame size of 256 x 256 pixels and lasts for approximately 20 seconds (15 fps). Four reconstructed frames are shown in Figure 6. The box-shaped segmentation mask encloses the car from the front. As a

result, two video objects are generated, the frontal car and the background.

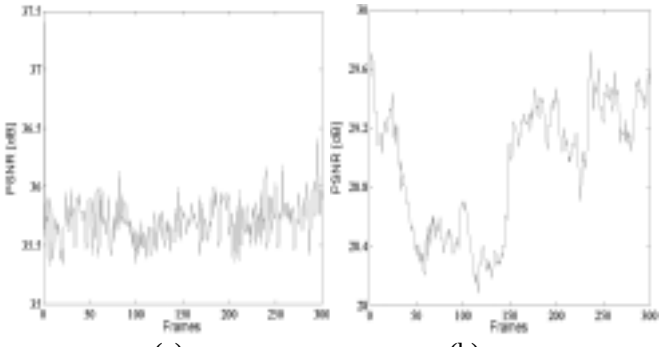
The PSNR values of the car video object shown in Figure 7a present an impressive regularity. The reason is that the segmentation mask encloses almost exactly the car object. Hence, the movements and changes in the background structure of the camera do not affect the compression quality of the car video object. By comparison, in the face encoding examples, a rectangular mask was employed to enclose the elliptical shape of the face. As a result of this approximation, some elements of the background were included in the face video object leading to a greater variability in the encoder performance.



#22 #100

#135 #219

Figure 6. Reconstructed *Pursuit* sequence. Compression ratio is 68.94.



(a) (b)

Figure 7. PSNR of the reconstructed data for *Pursuit* sequence. (a) Car. (b) Background.

6. CONCLUSIONS

This paper presented a smart camera with real-time video object creation and encoding based on the MPEG-4 standard. Our system has applications in surveillance, security, mobile videophones, and automotive industry. It combines powerful algorithms for object detection, tracking, and compression. The system performance has been demonstrated and discussed on various sequences taken with a fixed camera with pan and tilt, and with a moving camera. We showed that it is possible to obtain very good and relatively constant reconstructed quality for the object of interest even in the conditions of large camera/object movements. This work represents a step forward towards the DSP implementation of smart cameras, capable of programmable intelligent vision tasks.

Acknowledgements. We would like to thank Dr. Alok Gupta, the former Head of the Imaging and Visualization Department of Siemens Corporate Research for sponsoring A. Del Bue's internship in Princeton. We thank Professor von Seelen of the Institute for Neuro-informatik, Ruhr-Universitaet, Bochum, Germany for the pursuit sequence.

7. REFERENCES

[1] J. Batista, P. Peixoto, H. Araujo, "Real-Time Active Visual Surveillance by Integrating Peripheral Motion Detection with Foveated Tracking", IEEE Workshop on Visual Surveillance, Bombay, India, 18—25, 1998.

[2] S. Battista, F. Casalino, C. Lande, "MPEG-4: The Third Millennium Standard", IEEE Multimedia, vol. 6, no. 4, pp. 74-83, 1999.

[3] R.T. Collins, A.J. Lipton, T. Kanade, "A System for Video Surveillance and Monitoring", American Nuclear Society Eight Intern. Meeting on Robotics and Remote Systems, 1999.

[4] D. Comaniciu, V. Ramesh, P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift", IEEE CVPR, Hilton Head Island, South Carolina, 2:142—149, 2000.

[5] D. Comaniciu, V. Ramesh, "Robust Detection and Tracking of Human Faces with An Active Camera", IEEE Int'l Workshop on Visual Surveillance, Dublin, Ireland, 11-18, 2000.

[6] J.L. Crowley, F. Berard, "Multi-Modal Tracking of Faces for Video Communications", IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, 640—645, 1997.

[7] A. Eleftheriadis, A. Jacquin, "Automatic Face Location Detection and Tracking for Model-Assisted Coding of Video Teleconference Sequences at Low Bit Rates", Signal Processing - Image Comm., 7(3): 231—248, 1995.

[8] M. Greiffenhagen, V. Ramesh, D. Comaniciu, H. Niemann, "Statistical Modeling and Performance Characterization of a Real-Time Dual Camera Surveillance System", IEEE CVPR, Hilton Head Island, South Carolina, 2:335—342, 2000.

[9] R. Koenen, "Profiles and Levels in MPEG-4: Approach and Overview", Signal Processing: Image Communication, (4-5):463-478, 2000.

[10] P. Kuhn, "Algorithms, Complexity Analysis and VLSI Architectures for MPEG-4 Motion Estimation", Kluwer, Boston, 1999.

[11] W.W. Lu (Editor), "Technologies on Broadband Wireless Mobile: 3G Wireless and Beyond", IEEE Communications Magazine, 38(10):57—19, 2000.

[12] Moving Picture Experts Group, "Overview of the MPEG-4 Standard", ISO/IEC JTC1/SC29/WG11, 2000.

[13] W. Zheng, I. Ahmad, M. L. Liou, "Real-Time Software Based MPEG-4 Video Encoding", Workshop on MPEG-4, San Jose, 2001.