Human Visual System Based Scalable Video Coding and Communications

Ligang Lu, Zhou Wang[a], Jack Kouloheris, Alan C. Bovik[b]

Multimedia Technologies, IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
[a]ImageQuant Inc. Troy, NY 12180 UAS
[b]Laboratory for Image and Video Engineering, Dept. of ECE,
The Univ. of Texas at Austin, Austin, TX 78703, USA

ABSTRACT

This paper introduces our recent research work on the development of a scalable foveated visual information coding and communication system, which follows two emerging trends in visual communication research. One is to design rate scalable image and video codecs, which allow the extraction of coded visual information at continuously varying bit rates from a single compressed bitstream. The other is to incorporate human visual system models to improve the state-of-the-art of image and video coding techniques by better exploiting the properties of the intended receiver. The central idea of the proposed system is to organize the encoded bitstream to provide the best decoded visual information at an arbitrary bit rate in terms of foveated visual quality measurement. Such a scalable foveated visual information processing system has many potential applications in the field of visual communications. Significant examples include network image browsing, network videoconferencing, robust visual communication over noisy channels, and visual communication over active networks.

Keywords: visual communication, image coding, scalable video coding, human visual system, foveation

1. INTRODUCTION

It has been envisioned that network visual services, such as network video broadcasting, video-on-demand, videoconferencing and telemedicine, will become ubiquitous in the twenty-first century. As a result, network visual communication has become an active research area in recent years. One of the most challenging problems for the development of a visual communication system is that the available bandwidth of the networks is usually insufficient for the delivery of the voluminous amount of the image and video data. Designing a visual coding and communication system is a complicated task. Depending on the application, there are many issues related to the performance of the image/video codecs, such as quality-compression performance computational complexity, memory requirement, parallelizability, scalability, robustness, security and interactivity. Although the image/video coding standards (e.g., JBIG, JPEG, H.26X and MPEG) exhibit acceptable quality-compression performance in many visual communication applications, further improvements are desired and more features need to be added, especially for some specific applications.

Recently, two interesting research trends have emerged that are very promising and may lead to significantly improved image/video codecs. The first trend is to develop continuously rate scalable coding algorithms,[1–3] which allow the extraction of coded visual information at continuously varying bit rates from a single compressed bitstream. An example is shown in Fig. 1, where the original video sequence is encoded with a rate scalable coder and the encoded bitstream is stored frame by frame. During the transmission of the coded data on the network, we can scale, or truncate, the bitstream at any place and send the most important bits of the bitstream. The second research trend is to incorporate Human Visual System (HVS) models into the coding/communication system. It is well accepted that perceived video quality does not correlate well with Peak Signal-to-Noise Ratio

Send correspondence to Ligang Lu, E-mail: lul@ibm.us.com.

Electronic Imaging and Multimedia Technology III, Liwei Zhou, Chung-Sheng Li, Yoshiji Suzuki,
Editors, Proceedings of SPIE Vol. 4925 (2002) © 2002 SPIE · 0277-786X/02/$15.00

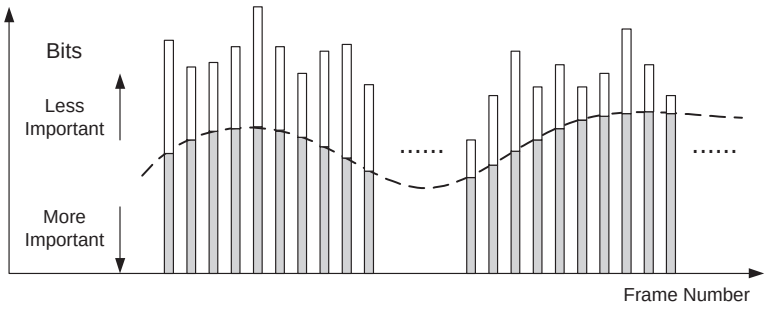

Figure 1. Bitstream scaling in rate scalable video communications. Each bar represents the bitstream for one frame in the video sequence.

(PSNR), which is still the most widely used method for image/video quality evaluation. HVS characteristics must be considered to provide better visual quality measurements.[4, 5]

Our work stands at the intersection of the two promising research trends. Specifically, wavelet-based embedded bitplane coding techniques are used for rate scalable coding. Further, we exploit the foveation feature of the HVS, which refers to the fact that the HVS is a highly space-variant system, where the spatial resolution is highest at the point of fixation (foveation point) and decreases dramatically with increasing eccentricity. By taking advantage of this fact, considerable high frequency information redundancy can be removed from the peripheral regions without significant loss in the reconstructed image and video quality. An example of foveated image is shown in Fig.4. If attention is focused at the foveated region, then the foveated and the original images have almost identical appearance (depending on the viewing distance). The foveation factor has been employed in previous work to improve image and video coding efficiency.[6–10] However, most of the algorithms used fixed foveation models. These methods lack the flexibility to adapt to different foveation depths and are not convenient to be implemented in a rate scalable manner. Chang[11] proposed to develop a wavelet-based scalable foveated image compression and progressive transmission system. However, human visual characteristics were not considered in depth, and no efficient coding algorithms were implemented.

2. HUMAN VISUAL SYSTEM BASED SCALABLE IMAGE AND VIDEO CODING TECHNIQUES

Psychophysical experiments have been conducted to measure the visual sensitivity as a function of spatial frequency and retinal eccentricity.[8] Based on the contrast sensitivity model introduced in [8], we developed a new foveated visual sensitivity model, which is the first foveation model[12] that explicitly distinguishes two different foveation factors – the spatial variance of the visual contrast sensitivity and the spatial variance of the local visual cut-off frequency. The model is converted into the image pixel domain. In Fig.2, we show the normalized contrast sensitivity as a function of pixel position, where the image width is 512 pixels and the viewing distance is 3 times of the image width. The cut-off frequency as a function of pixel position is also given. It can be observed that the cut-off frequency drops quickly with increasing eccentricity and the contrast sensitivity decreases even faster. The foveation model is also converted into the wavelet domain and then combined with a visual importance model for wavelet coefficients.[13] A novel structural distortion based image quality indexing approach[14] is applied to the wavelet foveation model, leading to a Foveated Wavelet Image Quality Index (FWQI).[5, 12]

The foveated visual model and its corresponding quality measure provide us with useful tools to develop and optimize foveated image/video coding and communication systems. Our research has been focused on combining the foveation model with embedded bitplane coding techniques, which are rate scalable and have achieved great success for uniform resolution image/video coding.[1–3] The central idea of our work is to organize the encoded bitstream to provide the best decoded visual information at an arbitrary bit rate in terms of foveated visual quality measurement.

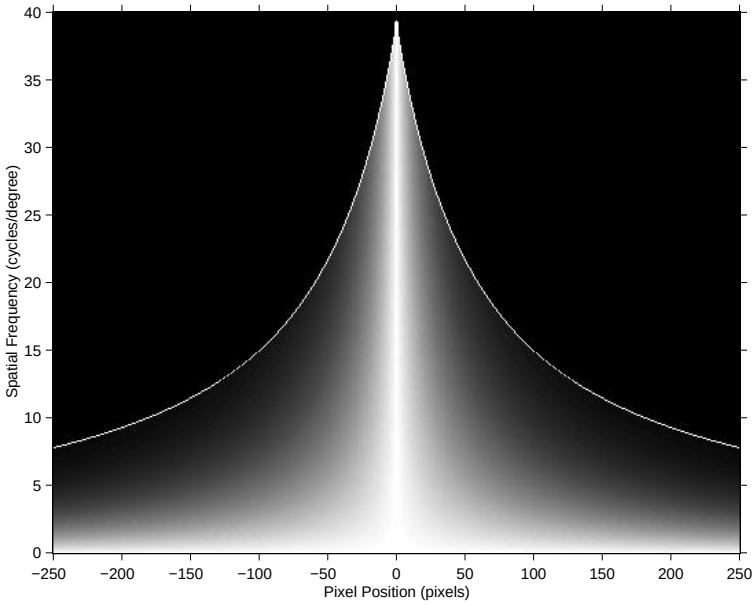

Figure 2. Normalized contrast sensitivity (Brightness indicates the strength of contrast sensitivity) and cutoff frequency (white curve).

In [12], we proposed a scalable foveated wavelet image coding algorithm termed Embedded Foveation Image Coding (EFIC), which seamlessly combines foveation filtering with foveated image compression and provides very good coding performance in terms of foveated visual quality measurement.

In [15], the scalable foveated coding method is extended for video coding, resulting in a Foveation Scalable Video Coding (FSVC) system. FSVC first divides the input video sequence into Groups of Pictures (GOPs). Each GOP has one intra-coding frame (I frame) at the beginning and the rest are predictive coding frames (P frames). The general framework of the FSVC encoding system is shown in Fig. 3. The prototype of FSVC allows us to select multiple foveation points. It also limits the search space of the foveation points to save computation power. The FSVC framework is very flexible such that different foveation point selection schemes can be applied to a single framework because the best way of foveation point(s) selection is highly application dependant. We implemented the FSVC prototype in a specific application environment, where an automated foveation point selection scheme and an adaptive frame prediction algorithm are employed as the key techniques.

The methods to choose foveation points for I frames and P frames are different. For I frames, a skin color detection and template matching based face detection technique[16] is implemented to select foveation points in the face areas. A different strategy is used for P frames, where we focus on the regions in the current P frame that provide us with new information from its previous frame, in which the prediction errors are usually larger than other regions. The potential problem of this method is that the face regions may lose fixation. To solve this problem, we use an unequal error thresholding method to determine foveation regions in P frames, where a much smaller prediction error threshold value is used to capture the changes occurring in the face regions.[15] In Fig.7, we show 4 consecutive frames in the "Silence" sequence and the corresponding selected foveation points, in which the first frame is an I frame and the rest are P frames.

In fixed rate motion compensation based video coding algorithms, a common choice is to use the feedback decoded previous frame as the reference frame for the prediction of the current frame. This choice is infeasible for continuously scalable coding because the decoding bit rate may be different from the encoding bit rate and is unavailable to the encoder. In [3], a low base rate is defined and the decoded and motion compensated frame at the base rate is used as the prediction. This solution avoids the significant error propagation problems, but when the decoding bit rate is much higher than the base rate, large prediction errors may occur and the overall coding efficiency may be seriously affected. We proposed a new solution to this problem,[17] where the

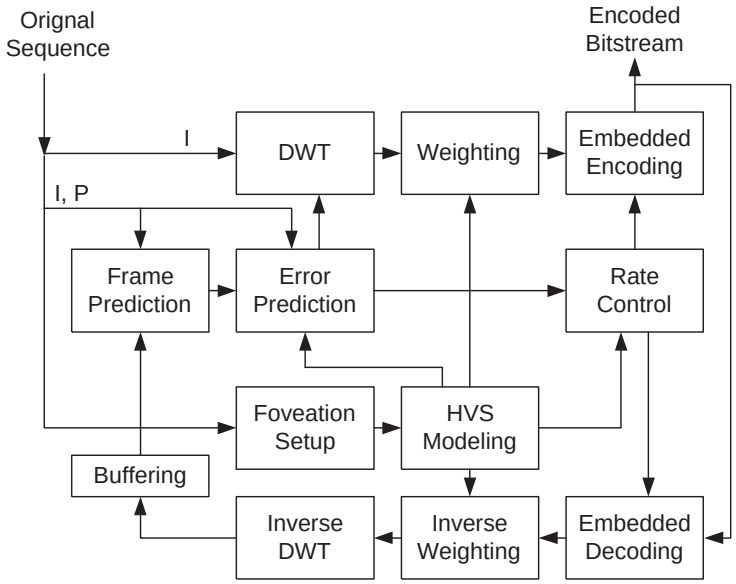

Figure 3. General framework of the FSVC encoding system.

original motion compensated frame and the base bit rate decoded and motion compensated frame are adaptively combined using the foveation model. By using the new method, error propagation becomes a small problem, while at the same time, better frame prediction is achieved, which leads to smaller prediction errors and better compression performance.

To demonstrate the effectiveness and scalable feature of our human visual system based scalable coding technique and the automatic foveation points selection algorithm, Fig. 5 shows the decoded luminance frame number 18 of the *News* sequences at 600Kbps, 400Kbps and 200Kbps, respectively, where the base bit rate is 200Kbps from the same bitstream encoded at 2Mbps by our human visual system based scalable coding technique and then decoded at different bit rates. Fig. 6 shows the reconstructions of the 32nd frame of the *Salesman* sequence at 200Kbps, 400Kbps, and 800Kbps. Fig.7 shows the FSVC compression results of the "Silence" sequence coded at 200 Kbits/sec. It is clear that the important regions (the face and moving areas) are captured very well using our algorithms.

3. HUMAN VISUAL SYSTEM BASED SCALABLE VISUA COMMUNICATIONS

The major purpose of the proposed scalable foveated visual information processing system is to facilitate visual communications over heterogeneous, time-varying, multi-user and interactive networks, where variable bandwidth video streams need to be created to meet different user requirements. The traditional solutions, such as layered video,[18] video transcoding,[19] and simply repeated encoding, require more resources in terms of computation, storage space and/or data management. More importantly, they lack the flexibility to adapt to time-varying network conditions and user requirements. By contrast, with a continuously rate scalable codec, the data rate of the video being delivered can exactly match the available bandwidth of the network. The foveation technique provides useful tradeoff between foveation depth, frame rate, and resolution. For example, if the available bandwidth drops dramatically, a fixed data rate coding system has to stop transmission. A uniform resolution scalable coding system can still work properly but might transmit unacceptable quality video to the client. A foveation-based scalable coding system, however, may still deliver useful information to the client, who might be specifically interested in certain areas in the video frame during each time period.

One direct application of the proposed system is network image browsing. There are two significant examples. In the first example, prior to using the encoding algorithm, the foveation point(s) are predetermined. The coding system then encodes the image with high bit rate and high quality. One copy of the encoded bitstream

is stored at the server side. When the image is required by a client, the server sends the bitstream to the client progressively. The client can stop the transmission at any time once the reconstructed image quality is satisfactory. In the second example, the foveation points are unknown to the server end when the transmission starts. Instead of a fully encoded bitstream, a uniform resolution coarse quality version of the image is pre-computed and stored at the server side. The client first sees the coarse version of the image and clicks on a point of interest in the image. The selected point of interest is sent back to the server and activates the scalable foveated encoding algorithm. The encoded bitstream that has a foveation emphasis on the selected point of interest is then transmitted progressively to the client.

Another application is network videoconferencing. Compared with traditional videoconferencing systems, a foveated system can deliver much lower data rate video streams since much of the high frequency information redundancy can be removed in the foveated encoding process. Interactive information such as the locations of the mouse, touch screen and eye-tracker can be sent back to the other side of the network and used to define the foveation points. Face detection and tracking algorithm may also help to find and adjust the foveation points. Furthermore, in a highly heterogeneous network, the available bandwidth can change dramatically among the end users. A fixed bit-rate video communication stream would either be terminated suddenly (when the available bandwidth drops below the required bit-rate) or suffer from the inefficient use of the bandwidth (when the fixed bit-rate is lower than the available bandwidth). By contrast, a rate scalable foveated videoconferencing system can deal with these problems smoothly and efficiently.

The most commonly used methods for robust visual communications on noisy channels are error resilience coding at the source and channel encoders and error concealment processing at the decoders.[20] Scalable foveated image and video stream gives us the opportunity to do a better job by taking advantage of its optimized ordering of visual information in terms of HVS measurement. It has been shown that significant improvement can be achieved by unequal error protection for scalable foveated image coding and communications.[21]

Active network is a hot research topic in recent years.[22] It allows the customers to send not only static data but also programs that are executable at the routers or switches within the network. An active network becomes useful and effective for visual communications only if an intelligent scheme is employed to modify the visual contents being delivered in a smart and efficient way. The properties of scalable foveated image/video stream matches the features of active networks very well because the bit rate of the video stream can be adjusted according to the network conditions monitored at certain routers/switches inside the network (instead of at the sender side), and the feedback foveation information (points and depth) at the receiver side may also be dealt with at the routers/switches. This may result in much quicker responses that benefit real-time communications.

4. CONCLUSIONS

We presented our recent work on human visual system based scalable video coding system. Our system can provide very good scalable feature and coding effectiveness based on a foveation model. The automatic foveation point select algorithm can effectively capture or preserve the important regions (foveated point areas) of interests. We have also demonstrated that our system can have a wide range of potential applications in visual information communications.

REFERENCES

1. J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
2. A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits & Systems for Video Tech.*, vol. 6, pp. 243–250, June 1996.
3. K. S. Shen and E. J. Delp, "Wavelet based rate scalable video compression," *IEEE Trans. Circuits & Systems for Video Tech.*, vol. 9, pp. 109–122, Feb. 1999.
4. T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image & Video Proc.* (A. Bovik, ed.), Academic Press, 2000.
5. Z. Wang, A. C. Bovik, and L. Lu, "Wavelet-based foveated image quality measurement for region of interest image coding," in *Proc. IEEE Int. Conf. Image Proc.*, Oct. 2001.

Figure 4. Left: original image; Right: foveated image.

6. E. L. Schwartz, "Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding," *Vision Research*, vol. 20, pp. 645–669, 1980.
7. P. L. Silsbee, A. C. Bovik, and D. Chen, "Visual pattern image sequence coding," *IEEE Trans. Circuits & Systems for Video Tech.*, vol. 3, pp. 291–301, Aug. 1993.
8. W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," in *Proc. SPIE*, vol. 3299, pp. 294–305, July 1998.
9. S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Processing*, vol. 10, pp. 977–992, July 2001.
10. H. R. Sheikh, B. L. Evans, and A. C. Bovik, "Real-time foveation techniques for low bit rate video coding," *Real Time Imaging*, 2002. accepted.
11. E.-C. Chang, *Foveation techniques and scheduling issues in thinwire visualization*. PhD thesis, Dept. of CS, New York University, 1998.
12. Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Processing*, vol. 10, pp. 1397–1410, Oct. 2001.
13. A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, pp. 1164–1175, Aug. 1997.
14. Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, Mar. 2002.
15. Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection." submitted to *IEEE Trans. Image Processing*, 2001.
16. H. Wang and S.-F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. Circuits & Systems for Video Tech.*, vol. 7, pp. 615–628, Aug. 1997.
17. L. Lu, Z. Wang, and A. C. Bovik, "Adaptive frame prediction for foveation scalable video coding," in *Proc. IEEE Int. Conf. Multimedia and Expo*, (Tokyo, Japan), Aug. 2001.
18. S. Arawvith and M.-T. Sun, "MPEG-1 and MPEG-2 video standards," in *Handbook of Image and Video Processing* (A. Bovik, ed.), Academic Press, May 2000.
19. H. Sun, W. Kwok, and J. W. Zdepski, "Architectures for MPEG compressed bitstream scaling," *IEEE Trans. Circuits & Systems for Video Tech.*, vol. 6, pp. 191–199, Apr. 1996.
20. J. D. Villasenor, Y.-Q. Zhang, and J. Wen, "Robust video coding algorithms and systems," *Proceedings of the IEEE*, vol. 87, pp. 1724–1733, Oct. 1999.

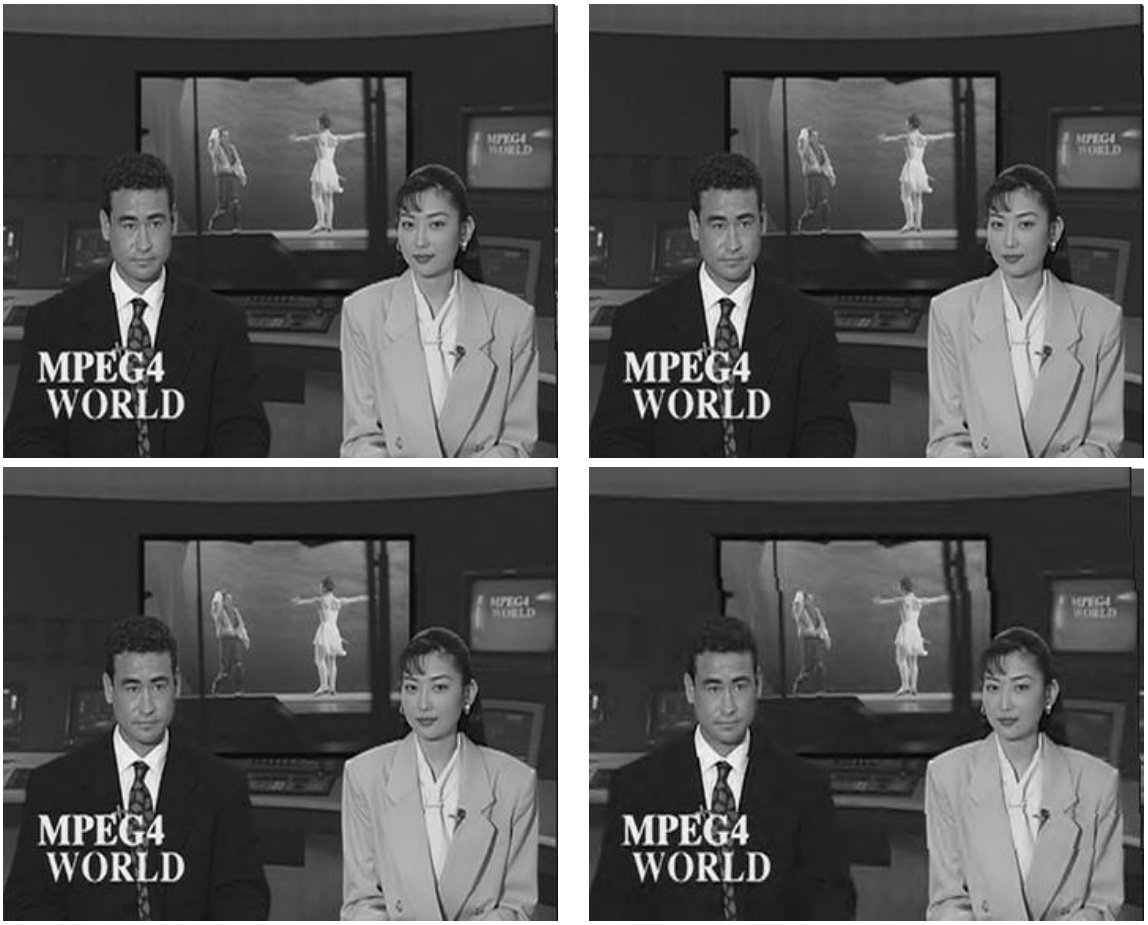

Figure 5. The 18th Frame of *News* sequence coded at 2Mbps. Top-Left: original image; Top-Right: decoded at 600Kbps; Bottom-Left: decoded at 400Kbps; Bottom-Right: decoded at 200Kbps.

Figure 6. The 32th Frame of *Salesman* sequence coded using FSVC. Top-Left: original image; Top-Right: coded at 200Kbps; Bottom-Left: coded at 400Kbps; Bottom-Right: coded at 800Kbps.

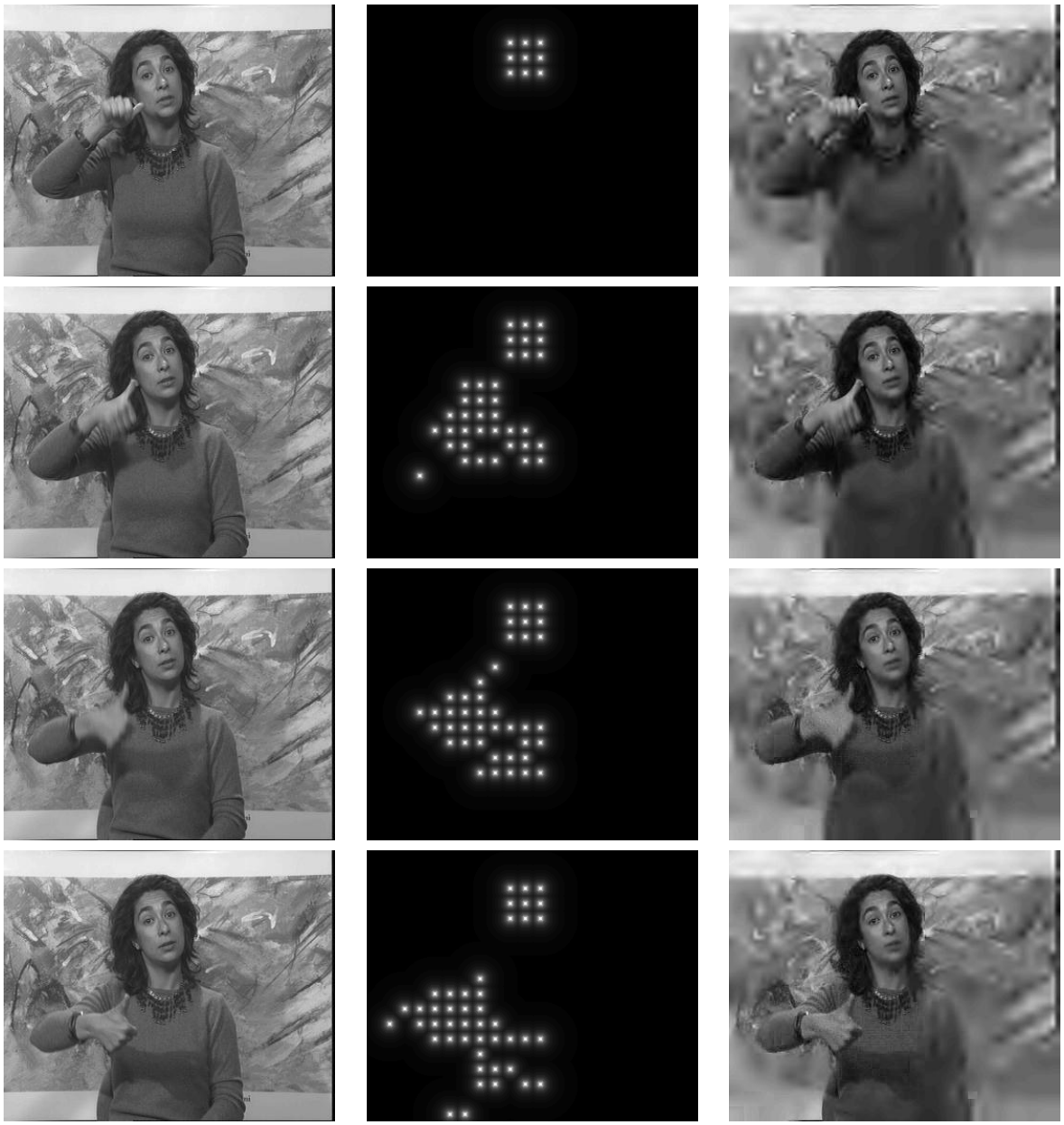

Figure 7. "Silence" sequence (left); selected foveation points (middle); and FSVC compression results at 200 Kbits/sec (right).

21. M. F. Farooq, "Unequal error protection for scalable foveated image communication," Master's thesis, Dept. of Electrical and Computer Engineering, The University of Texas at Austin, May 2002.
22. D. L. Tennenhouse, J. M. Smith, W. D. Sincoskie, D. J. Wetherall, and G. J. Mindenk, "A survey of active network research," *IEEE Comm. Magazine*, vol. 35, Jan. 1997.