Fovea Based Coding for Video Streaming

Çağatay Dikici[1], H. Işıl Bozma[1], and Reha Civanlar[2]

[1] Intelligent Systems Laboratory,
Electrical and Electronics Engineering Department
Boğaziçi University, 34342, Bebek, Istanbul, Turkey
{dikicica, bozma}@boun.edu.tr
http://www.isl.ee.boun.edu.tr
[2] Computer Engineering Department
Koç University, 34450, Istanbul, Turkey
rcivanlar@ku.edu.tr

Abstract. Attentive robots, inspired by human-like vision – are required to have visual systems with fovea-periphery distinction and saccadic motion capability. Thus, each frame in the incoming image sequence has nonuniform sampling and consecutive saccadic images have temporal redundancy. In this paper, we propose a novel video coding and streaming algorithm for low bandwidth networks that exploits these two features simultaneously. Our experimental results reveal improved video streaming in applications like robotic teleoperation. Furthermore, since the algorithm employs the Gaussian-like resolution of human visual system and is extremely simple to integrate with the standard coding schemes, it can also be used in applications such as cellular phones with video.

1 Introduction

Motivated by biological vision systems, there has been a growing trend to have robots explore their environment and look around in an attentive manner – thereby minimizing the amount of collected information and thus reducing the required computation considerably [1–3]. In order to remotely communicate with such a robot and see its visual surroundings, one needs real-time video transmission – preferably using standard coding schemes. Interestingly, video broadcasted from a robot with such attention capabilities has two intrinsic properties which are not exploited in general by the standard compression algorithms: i.) Fovea-periphery distinction, which leads to varying spatial resolution within each frame; and ii.) Saccadic motion between foveae, as a result of which there is much overlapping information between consecutive frames. For example, in H.263 [4] all blocks in a frame and between frames are coded with the same priority. In this paper, we present a real-time video algorithm that exploits both the spatial and temporal redundancy that exist in the video sequences and thus can be used for attentive robot-video streaming over internet. The novelty of this work is twofold: i)Simultaneous foveal and temporal compression; and ii) Easy integration with any of the standard compression technology – without requiring any alterations in these standards. Furthermore, this approach can be used in general in any low-bandwidth video transmission since it matches the fall-off resolution of the human visual system.

1.1 Related Literature

Attentive robots explore their surroundings in a loop of pre-attention and attention [2]. The aim of the pre-attention stage is to determine the next attention regions. This is achieved through the fovea-periphery mechanism [5]. Unlike traditional cameras, the distribution of receptor cells on the retina is Gaussian-like with a small variance, resulting in a loss of resolution as we move away from the optical axis of the eye [6]. The fovea is the small region of highest acuity around the optical axis and the rest of the retina is called periphery. Saccades - very rapid jumps of optical axis - are used to bring images of chosen objects to fovea where resolution of fine visual detail is at its best. Non-orthogonal spatial filter responses can be used for determination of saccadic movements of foveae in still images [7]. The remote access (in particular internet based) and teleoperation of such a robot requires real-time transmission of thus generated video – which has varying resolution and redundancy. Consequently, video streaming methods that exploit these properties become crucial.

The application of varying resolution to video coding and streaming is relatively new [8]. As outlined therein, this approach presents several distinct advantages such as guaranteed compression ratios and speed. Foveal and peripheral regions are coded differently in spatial domain, and the priority assignment of the ATM cells are used for transmitting the regions of video frames with varying priorities [8]. However, in the case of a network congestion, peripheral information which attracts relatively lower attention are first lost. However, since the approach depends solely on the Quality of Service (QoS) parameters, it will potentially have problems on best effort systems like internet. Furthermore, the redundancy along the temporal dimension is not utilized at all. An approach that applies both spatial & temporal processing on MPEG2 streams has been presented in [9]. DCT coefficients of the periphery are quantized in order to reduce the length of the bitstream. However, since DCT coefficients are calculated on 8×8 blocks, the foveal region definition is limited with rectangular shape. In order to minimize the blocking artifacts between the foveal and the peripheral regions, image pyramids and raised-cosine blending are used in [10]. However, such an approach requires the generated pyramids also to be transmitted through the channel then implied increased bandwidth. Space-variant approaches such as log-polar mapping of original frame can also be applied for foveation [11], but the increased computation for achieving such transformations impede real-time applications. Finally, if methods that take particular coding algorithms like H.263, MPEG-4 and and JVT into account do not offer compression independent solutions.

1.2 Problem Statement

Suppose that the visual task involves a robot looking at a scene in an attentive manner and a video is generated meanwhile. The objective can be defined as real-time transmission of this data over the Internet so that users can see precisely what the robot is seeing. Moreover, the system should also: i.) allow real-time streaming and ii.) be usable with any particular video compression algorithm.

2 Approach

Consider an incoming image sequence. Let I_v^t denote the visual field image at time t. The function $c^t : I_v^t \rightarrow C$ maps each pixel in this region to a value from the color space C. The fovea is represented by $I_f^t \subset I_v^t$. The next fovea I_f^{t+1} at time $t+1$ is chosen from the the set of candidate foveae $C(I_v^t)$ – as determined from the visual field. For each candidate fovea $I_c \in C(I_v^t)$, an attention criteria $a : I_c \rightarrow R^+$ is computed. The attention criteria is a scalar valued function of interest based on the presence of simple features with low computational requirements. This function is determined by the measure of interest on that frame and its definition will vary depending on the measure of interest such as color, intensity or human facial features. The candidate fovea that maximizes this measure is then designated to be the next fovea as:

$$I_f^{t+1} = arg \max_{\forall I_c \in C(I_f^t)} a(I_c) \tag{1}$$

The robot then moves its camera as to fixate on this newly determined fovea. Such camera movements correspond to saccadic eye movements in humans. As a result, a raw image sequence $\{I_v^0, ..., I_v^t, I_v^{t+1}, ...\}$ is generated.

In general, most video coding algorithms achieve compression by applying transforms on the original sequence that exploit spatial redundancies such as discrete cosine transform (DCT). In attentive vision, foveal and peripheral regions of the incoming visual field each incoming visual field I_v^t is processed differently. The visual data as defined by the color map c in the incoming fovea is preserved in the fovea while that of the periphery $I_p^t = I_v^t - I_f^t$ is transmitted with lower resolution. Consequently, a new color map $\tilde{c}_S^t : I_v^t \rightarrow C$ is defined as:

$$\tilde{c}_S^t(x) = \begin{cases} c^t(x) & \text{if } x \in I_f^t \\ f_S * c^t(x) & \text{if } x \in I_p^t \end{cases} \tag{2}$$

where $f_S : I_p^t \rightarrow \tilde{I}_p^t$ is a spatial filter function. The main idea in choosing this filter is that since the peripheral pixels do not attract our attention, the high frequency contained therein is not important and thus can be removed from the data. Consequently, the corresponding image areas can be coded with fewer DCT coefficients. For example, low pass filters such as Gaussian and blurring filters can be utilized.

However, with such a definition, spatial edge artifacts will appear in the reconstructed image after transmission. In order to minimize this, the color map $\tilde{c}_S$ is modified to include α-blending:

$$c_S^t(x) = \alpha^t(x)c^t(x) + (1 - \alpha^t(x))f_S * c^t(x) \tag{3}$$

The blending function $\alpha^t : I_v^t \rightarrow [0, 1]$ is time dependent function whose value at time t varies between 0 and 1. The values of α^t are 1 or close to 1 on the fovea and converge towards 0 for peripheral pixels as a function of their proximity to the fovea. Consequently, a gradual color transition is introduced in the fovea periphery neighborhood. Furthermore, introducing such a blending enables us to define non-rectangular shaped foveal regions.

As the camera saccades from the current fovea to the next, an image sequence $\{I_v^0, ..., I_v^t, I_v^{t+1}, ...\}$ is generated. If temporal sampling is fast enough, there is much temporal redundancy between saccadic image frames. In general, most video coding algorithms also exploit temporal redundancies in their compression schemes. In motion compensation based coding, each frame I_v^{t+1} can be represented as the difference of the current frame and the previous one I_v^t, and thus be coded using Motion Vectors (MV). As expected, the efficiency of this type of coding goes up with increased temporal redundancy. In attentive vision, we can increase the temporal redundancy between frames depending on whether foveal or peripheral regions are under consideration. Since the fovea is subject to close scrutiny, all minute detail changes between frames should be preserved and no new temporal redundancy can be introduced. However, this is not the case for periphery. Since changes between frames in the peripheral regions are ignorable to some extent, temporal redundancy can be increased by applying filters across temporal dimension. For example, color maps in the periphery can be updated with every K saccades.In doing so, since the current peripheral region can be estimated from the previous one, the length of the bitstream using MVs is reduced considerably. In this case, the temporal color map $c_T^t : I_v^t \rightarrow C$ is defined as:

$$c_T^t(x) = \begin{cases} c^t(x) & \text{if } x \in I_f^t \\ c_S^{t-tmodK}(x) & \text{if } x \in I_p^t \end{cases} \tag{4}$$

If both spatial and temporal redundancies are taken into account, the resulting color map becomes a composition of the two functions c_T^t and c_S^t as $c_T^t \circ c_S^t : I_v^t \rightarrow C$.

3 Experiments

Our approach is implemented on video streaming from APES - an attentive robot developed in our laboratory [12,13]. APES can be remotely controlled and teleoperated and streams its acquired image sequence over the internet. Hence, any registered user can connect and watch the visual field of the APES robot as it explores its current surroundings. One authorized user can also control the APES remotely while watching its captured video in real time. In this setup, Helix^{TM} Project [14] is used as video coding and streaming framework. RTP/UDP/IP is used for transmission of real-time data, and Real Time Session Protocol(RTSP) is used for session initiation and control of the video stream. Furthermore, the remote control of the APES (pan & tilt controls) is sent via TCP/IP for lossless data transmission.

In order to quantify the visual quality of the foveal system, two metrics are used: The first metric is Foveal Mean Square Error(FMSE) which is similar to Mean Square Error (MSE), but is defined only on the fovea[3]

$$FMSE = \frac{1}{\mid I_f^t \mid} \sum_{n=1}^{x \in I_f^t} [c_T^t \circ c_S^t - c^t]^2 \tag{5}$$

However, it is well known that MSE value has a clear physical meaning in statistical sense, but it may not always reflect perceived visual quality [15].

As an alternative, Structural Similarity Measure (SSIM) - a metric capturing the amount of structural degradation between two images has been proposed in [15]. In this metric, luminance, contrast, and structural components of the two images are weighted and a quality index is generated. Physical meaning of SSIM can be explained such that it is a metric based on the comparison of mean, standard deviation and correlation coefficient of the normalized variance of two images. However, since the focus is on the fovea, we use a modified version Foveal Structural Similarity Measure(FSSIM) which is defined only on the fovea:

$$FSSIM = \frac{(2\mu_{c_T^t \circ c_S^t}\mu_{c^t} + C_1)(2\sigma_{(c_T^t \circ c_S^t)c^t} + C_2)}{(\mu^2_{c_T^t \circ c_S^t} + \mu^2_{c^t} + C_1)(\sigma^2_{c_T^t \circ c_S^t} + \sigma^2_{c^t} + C_2)} \tag{6}$$

[3] FMSE is a newly defined quality metric. We leave it to the vision science researchers to check its validity but for our applications, based on visual observations, it seems to be meaningful.

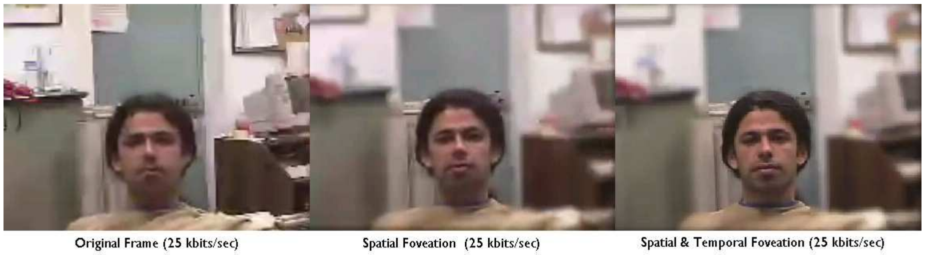
μ, and σ are the mean and variance of the respective images within the fovea. C_1 and C_2 are used in order to prevent unstable conditions if the values $\mu^2_{c^t_T \circ c^t_S} + \mu^2_{c^t}$ or $\sigma^2_{c^t_T \circ c^t_S} + \sigma^2_{c^t}$ are very close to 0. $C_1 = 0.01$ and $C_2 = 0.03$. Note that $FSSIM \leq 1$ with equality holding if and only if the source and the target images are identical.

Finally, note that FMSE value indicates the mean square error of the fovea region. If the FMSE value increases, so does the quantity of error increase. On the other hand, FSSIM value indicates the similarity measure of the fovea region by using HVS properties. The more similar the input and the reference frames are, its value gets closer to '1'; and similarly this value being '0' indicates the contrary case.

In order to evaluate the performance of the algorithm, an extensive statistical comparative study was conducted. First, the APES Video Database was created by making the robot look at scenes consisting primarily of a person – talking and mimicking at 3 different distances (long, intermediate, short distances) and 3 different poses (left, right, frontal views)[16]. Each incoming video was recorded in 384 × 288 resolution, 20 second long, $25 frames/sec$ RGB video in uncompressed AVI format without any preprocessing. Next, the videos in this database were subjected to the following preprocessing:

1. Twelve video sequences are selected randomly from the APES Database - with 4 of long, intermediate and short distance category respectively.
2. Next, for each video sequence, the foveal area is determined after visual examination of the video sequence and ensuring that the fovea overlaps with the image area containing the person's face. For each category, 2 different fovea sizes are considered:
   - For *long-distance ($l > 4m$)* sequences, these are taken to be 100 × 100 and 130 × 130 pixels;
   - For *intermediate distance ($l = 4m$)* sequences, they are 130 × 130 and 160 × 160 pixels;
   - For *short distance sequences ($l = 2m$)*, they are taken to be 160 × 160 190 × 190 pixels respectively.
3. Each sequence is first only spatially processed - using the 2 different fovea sizes. In spatial processing, $5x5$ box blur filter is selected as spatial filter function f_S. For α-blending the transition width is chosen as 5.
4. Each sequence is next spatio-temporally processed for the two different fovea sizes. K is selected as three in this process.
5. The original raw video and the two preprocessed videos are then are then encoded with Real-Media [14] codec with bit-rates $25k/sec$ and $35k/sec$. A sample frame is shown in Fig. 1

In analysis part, all the encoded video frames are compared with the original input raw frames. The comparison is performed as follows: First, for each encoded frame, MSE,FMSE,SSIM,FSSIM values are calculated. Since the number of encoded frames and the number of input raw frames may vary because of the encoding process, minimum FMSE values within 5 frame neighborhood of input raw frames are selected as reference frames. For each sequence, the first 450 frame statistics are stored. Figure 2 presents FMSE and FSSIM values of a video sequence with a 130 × 130 sized fovea with classical coding, with spatial and spatio-temporal coding. As expected, compared to classical coding, spatial and spatio-temporal coding improve the FMSE and FSSIM values considerably. However, there is not much added performance between just spatial and spatio-temporal coding schemes – possibly due to the temporal filtering



Original Frame (25 kbits/sec)

Spatial Foveation (25 kbits/sec)

Spatial & Temporal Foveation (25 kbits/sec)

Fig. 1. Sample frames - Left to right: No preprocessing ,spatial only and spatio-temporal preprocessing.

selected for our particular application. The performance should improve with a more appropriately selected filter.

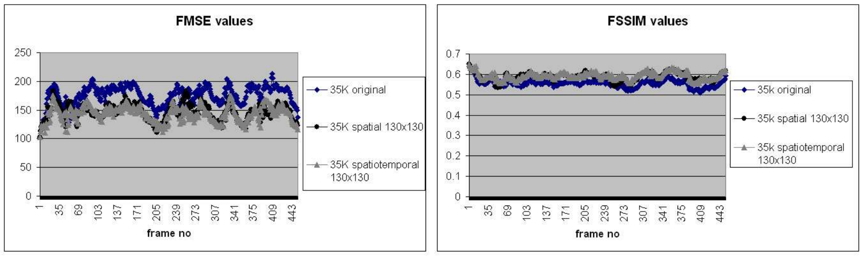


Fig. 2. FMSE &FSSIM Values.

Next, statistical metrics are gathered for each processed sequence. FSSIM and FMSE values are first normalized by dividing each frame's FSSIM and FMSE values by their originally coded frames FSSIM and FMSE values respectively and then averaged over the video. Finally, average normalized FSSIM and normalized FMSE values for each fovea size and encoding bit-rate is computed. For each group, the mean, minimum and maximum values are calculated – using the 2700 frames in each group. Fig. 3 presents the FMSE results. The FMSE values for both the spatial and spatial-temporal processing are all less than one. As expected, the quality goes down as the fovea size is increased. Furthermore, these values are improved for higher bit rates as can be seen by comparing the results for the 25k/sec and 35k/sec video streaming respectively. Similarly, the FSSIM values are shown in Fig. 4. Averaged FSSIM values are greater than one, which is consistent with the FMSE outputs.

We also computed the required processing overhead for the spatial and spatial-temporal coding schemes in order to check its suitability for real-time applications (on a Pentium II/1000 Mhz with 224 MB RAM). Each fovea size was considered seperately using randomly selected 200 frames. The results are as shown in Table 1. First of all, it is observed that the worst is about 10msec – which is quite acceptable with the frame rate of 25 frames/sec. As expected, the overhead is reduced considerably with the spatial-temporal coding.

4 Conclusion

In this paper, we present a fovea based coding scheme for video streaming through low bandwidth networks that exploits two important aspects of human vision: fovea-

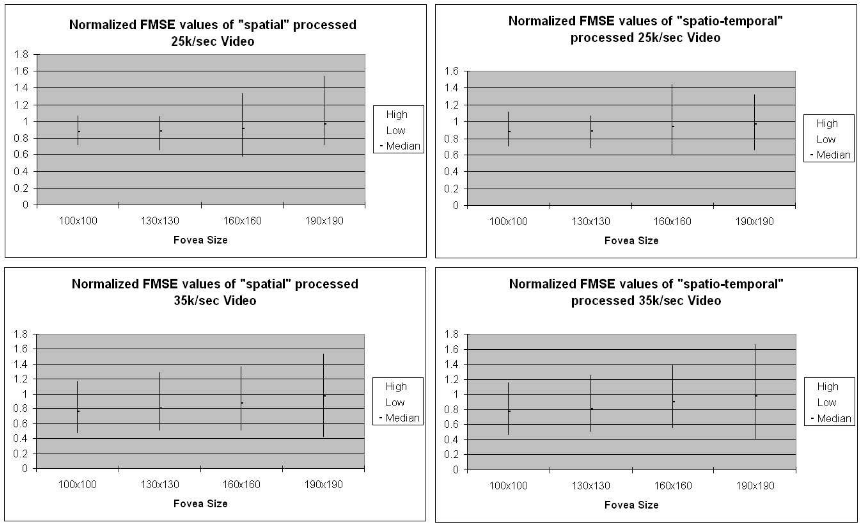


Fig. 3. Average FMSE values.

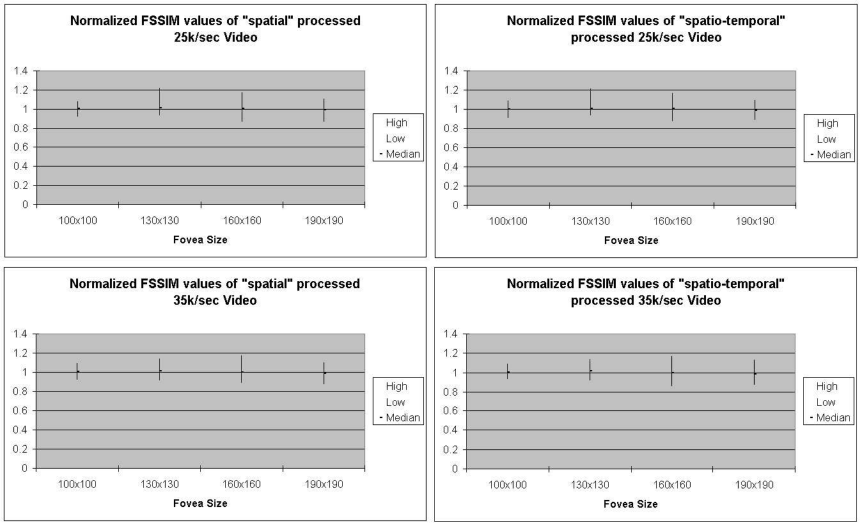


Fig. 4. Average FSSIM values.

periphery distinction and saccadic motion. Thus, each frame in the acquired image sequence has nonuniform sampling and consecutive saccadic images have temporal redundancy. Such a coding scheme is suitable for applications such as video broadcasting from attentive robot or cellular phones with video where the perceiver fixates on objects in a continual manner. Our experimental results indicate that compared to classical coding, spatial and spatio-temporal coding improve the transmission quality. For our future work, we will work on methods for increasing temporal redundancy through careful generation of the saccadic movements.

Table 1. Computational overhead statistics.

Fovea size	Overhead (msec) for spatial coding	Overhead (msec) for spatial-temporal coding (K=3)
100 × 100	9.39	5.3
130 × 130	9.64	5.29
160 × 160	9.47	5.48
190 × 190	9.48	5.55

5 Acknowledgment

This work has been supported by Bogazici University Scientific Research Projects Grant #BAP02S106. We gratefully acknowledge Doğan Can for his meticulous collection of the APES video database and helping with the experiments. We also gratefully acknowledge Oksijen Teknoloji A.S. for supporting Çağatay Dikici's project and completing his masters thesis.

References

1. Akins, K. (ed.): Perception, pages 290-316. Oxford University Press, (1996)
2. Ballard, D.H., Brown, C.M.: "Principles of Animate Vision", CVIP: Image Understanding, (1992), 56
3. Koch, C, Itti, L.: "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention", Vision Research, (2000)
4. ITU-T Recommendation H.263: "Video coding for low bit rate communication", (1996)
5. Kowler, E. (ed.): "Eye Movements and Their Role in Visual and Cognitive Processes", Elsevier, (1990)
6. Gouras, P., Bailey, C.H.: "The retina and phototransduction". In J. H. Schwartz and E. R. Kandel, (eds.), Principles of Neural Science. Elsevier, (1986)
7. Takacs, D., Wechsler, H.: "A Dynamic and Multiresolution Model of Visual Attention and Its Application to Facial Landmark Detection", Computer Vision and Image Understanding, Vol. 70. No. 1 ,(1998), 63-73
8. Wiebe, K., Basu, A.: "Improving image and video transmission quality over ATM with foveal priorization and priority dithering" Pattern Recognition Letters, (2001), 22
9. Reeves, T.H., Robinson, J.A.: "Rate Control of Foveated MPEG Video", CCECE (1997)
10. Geisler, W.S., Perry, J.S.: "A Real-time Foveated Multiresolution System for Low-bandwidth Video Communication", SPIE Proceedings: Human Vision and Electronic Imaging, Vol. 3299, (1998), 294-305
11. Grosso, E, Manzotti, R., Tiso, R., Sandini, G.: "A Space-Variant Approach to Oculomotor Control", Proceedings of International Symposium on Computer Vision, (1995), 509-514
12. Soyer, C., Bozma, H.I., Istefanopulos, Y.: "Apes: Actively Perceiving Robot", Proceedings of IEEE/RSJ International Conference on Robots and Systems, Lausanne, Switzerland, (2002)
13. Soyer, C., Bozma, H.I., Istefanopulos,Y.: "Attentional Sequence Based Recognition: Markovian and Evidential Reasoning", IEEE Transactions on Systems, Man and Cybernatics, Vol. 33, No. 6, (2003), 937-950
14. Helix Encoder, http://www.helixcommunity.org
15. Whang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: "Image Quality Assessment: From Error Measurement to Structural Similarity", IEEE Transactions on Image Processing, Vol. 13, (2004), No. 1
16. "APES Video Database", http://www.isl.ee.boun.edu.tr/Apes/VideoDatabase.html