Proceedings of the Seventh IASTED International Conference
COMPUTER GRAPHICS AND IMAGING
August 17-19, 2004, Kauai, Hawaii, USA

Region-of-Interest Video Compression with a Composite and a Long-Term Frame

Athanasios Leontaris and Pamela C. Cosman
Department of Electrical and Computer Engineering
University of California, San Diego, La Jolla, CA 92093-0407
{aleontar,pcosman}@code.ucsd.edu

Abstract

Multiple frame prediction has proved successful in increasing the performance of motion compensated video coders. Region-of-interest (ROI) coding in conjunction with a dual frame predictor can lead to interesting design choices on how to allocate the available bitrate among the frame regions. In this work, we tackle the problem of applying ROI coding on aerial imagery and then on standard video-conferencing sequences. For the latter, we introduce a novel composite frame scheme that takes advantage of the flexibility offered by the dual frame predictor that captures motion history from the short term past. Experimental results show that we can achieve significant bit rate and computational complexity savings for packet loss scenarios.

Key Words
ROI, composite, multiple, feedback, video compression.

1 Introduction

Region-of-interest coding has attracted a lot of research work in recent years due to the ability to attain, for low bit rates and for certain parts of the image, image quality that otherwise would require substantially higher transmission rates. Applications include rate-constrained coding for (unmanned) aerial vehicles (UAV), lossy medical image coding for telemedicine, and low-power hand-held devices for video conferencing, where coding faces at higher bit rate can have a great effect on perceptual quality.

The actual region-of-interest map can be obtained either automatically or manually. However, even an automatic technique will sometimes need some human intervention so as to narrow down the ROI selection or refine the output of the detection algorithm. Thus, most regions-of-interest are actually obtained in a hybrid manual-automatic manner. In our work, we assume a priori ROI knowledge,

Images were an important application of ROI coding. One medical image scheme where the ROI is compressed losslessly and the rest of the image is treated in a lossy manner is presented in [1]. Wavelet methods for ROI coding are addressed in [2], where JPEG 2000 ROI functionality is described. JPEG 2000 is the first image coding standard to facilitate ROI coding. The implementations can be either tiling, coefficient scaling (one of the most widely used approaches), or code-block selection. An interesting question was addressed in [3]: whether ROI coding improves overall perceived image quality. The author came to the conlusion that perceived image quality, with ROI coding, will only be increased if the regions are small and the bitrate is low enough to produce visible compression artifacts.

ROI coding was naturally extended to image sequence compression. An early subband-based approach can be found in [4], where Rate-Distortion (RD) optimization is applied, and distortion is adaptively weighted according to the significance of each image region. A motion-compensated DCT (MC-DCT) approach was presented in [5]. A neural network architecture was used to extract ROIs. Additional bits are allocated to ROI areas, by modifying the quantization tables. Another RD optimized wavelet video coder for regions is presented in [6]. The problem of optimal bit rate allocation for multiple regions of interest is solved by using a weighted Lagrangian Multiplier technique. We partially adopted this approach in our scheme.

In this work, we consider two different scenarios; one of an unmanned aerial vehicle (UAV) that transmits aerial imagery back to its base station; and one of video-conferencing with relatively static backgrounds and small regions-of-interest. UAVs rely on a virtually uninterruptible full-duplex communications link with their ground base station for receiving navigational and command data. Similarly, video-conferencing is a full-duplex communication scenario, where the sender and the receiver continously exchange data packets. Thus, we can assume that for both scenarios, feedback information in the form of ACK and NACK signals will be available to the transmitter, although with some delay d. This feedback information can be used to enhance performance as in [7].

Throughout this paper we assume that we have a priori knowledge of the ROI. Apart from using weighted Lagrangian multipliers for ROI bit allocation we selectively enabled and disabled the use of the long-term frame buffer. In addition, we devised a novel scheme for constructing a composite long-term frame. The scheme is based on keeping many versions of the same ROI in a buffer, unlike the scheme in [8] which constructs a composite frame that still resembles a regular one. The paper is organized as follows: Section 2 describes explicit allocation of more resources to the ROI. Section 3 describes implicit allocation, while Section 4 discusses the implementation of a composite frame scheme. In Section 5 we discuss error concealment. Section 6 deals with the experimental results, and the paper is concluded in Section 7.

2 Explicit Resource Allocation to the ROI

ROI information is assumed known by some external means. The ROI map is fed to the encoder as a binary map (only ROI and Non-ROI; we do not support multiple ROIs). The map contains a “1” for each macroblock (MB) that is considered interesting, and “0” otherwise. Hence our ROI coder has MB-level granularity. Coding the ROI at a higher quality can be accomplished by allocating more bit rate to the interesting MBs.

To increase the bit rate allocated to the ROI we will relax the bit rate constraint in the Rate-Distortion Optimization cost calculation function. This is realized as in [6], by weighting the Lagrangian multiplier accordingly. Let J denote the Lagrangian cost, D the distortion and R the respective bit rate. Let λ denote the Lagrangian multiplier chosen adaptively to achieve our target bit rate, and θ denote the relaxation factor that ranges from 0 to 1 for MBs belonging to the ROI. The following Lagrangian Cost Function will be employed for the ROI:

$$\min_{(mode,QP)} J_{MB} = \min_{(mode,QP)} (D_{MB} + \theta \times \lambda \times R_{MB})$$

For macroblocks outside the ROI we set $\theta = 1$. The minimization is taken over all combinations of $modes$ and quantization parameters (QPs).

3 Implicit Resource Allocation to the ROI

Traditionally, the problem of ROI coding was relegated to allocating additional rate to portions of the image or frame that were deemed to be more interesting. Here, we consider more general resource allocation. The term resource can include memory, computational resources, and increased bit rate. Here we consider the allocation of increased computational and memory resources for the interesting MBs. To this end, the region of interest is encoded using an additional frame buffer (the long-term one) as in [9], while the rest of the frame is encoded in a normal manner with only one prediction reference. This is illustrated in Fig. 1, where only MBs belonging to the ROI can reference the long-term frame buffer. For this scheme there is no need to transmit the ROI information to the receiver. The receiver can decode the stream without any ROI knowledge.

The gain obtained by using a dual frame scheme for certain regions can not only be attributed to occlusion effects and compression performance gains such as in [9]. In the case of feedback, the additional long-term reference frame can also provide error resilience as described in [7].

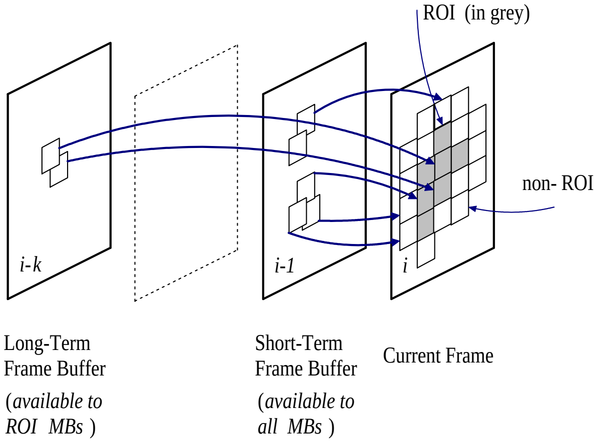

Figure 1. ROI resource allocation scheme.

In the case of UAVs the data link connecting them to the ground base station can also be utilized for sending back acknowledgement signals (ACK/NACK) for successful/unsuccessful reception of compressed video data. The portion of the link bandwidth that will be used for these signals is negligible. This information can be used by the on-board video encoder to compensate for the lost video stream.

Feedback will be applied in conjunction with a multiple frame predictor. Taking into account the tight power constraints, that translate into low memory and computational complexity, we initially limit the multiple frame predictor to a dual frame buffer and at the same time we apply object segmentation and use the additional frame buffer only for the ROI.

No explicit rate allocation was applied, but the combination of the compression efficiency and error resilience characteristics of the long-term frame ensure a higher reconstruction quality, especially in the face of packet losses, which can be very common for radio data links. Assuming that the ratio of ROI MBs over the total number is $\alpha \ll 1$ then we incur lower complexity for increased quality for ROI MBs. Results can be found in Section 7, where implicit, explicit and combinations of those allocations are compared. We will refer to this technique as selective dual frame.

4 Composite Frame Scheme

For videoconferencing that is characterized by relatively low motion and a static background, we propose a dual frame scheme (previous and long-term) where the long-term frame does not comprise a previously buffered one, but rather one composed by parts of previous frames. These parts will be the ones corresponding to the ROI which is assumed static for the duration of the video sequence. Once again we assume a priori ROI knowledge. Here, we transmit the ROI information to the receiver, though the rate overhead is negligible for small and static ROIs.

To constrain memory usage, no more memory than the one required to buffer an entire frame is utilized. ROIs from previous frames with a buffer of one MB circling the ROI are simply cut and pasted into the composite frame buffer. Thus, a 1×2-macroblock ROI, when buffered, will take up an area of 3×4 macroblocks. The composite frame is initialized with the first frame. The ROI and its perimeter are copied to the next available position in the composite frame buffer. Once the composite buffer has been filled, then the pointer to the next available position moves to the start of the buffer, in raster-scan order and older ROI parts are overwritten.

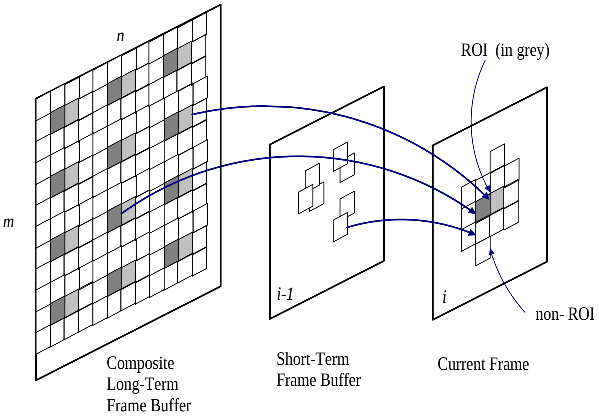

Figure 2. Composite ROI frame scheme.

Motion estimation (ME) is done in a non-traditional way. Let us assume that we are encoding frame i. If n ROIs (together with their ME perimeter) fit horizontally, and m ROIs fit vertically, then up to $n \times m$ past ROIs can be buffered in the composite frame, dating from as close as $i - 2$ to as far as $i - 1 - n \times m$. The motion estimation module will now search on a regular $[-16, 16]$ window over all those $n \times m$ stored ROIs. Thus, while for single frame ME we search once per MB, and for dual frame ME twice, now, non-ROI MBs will conduct one ME search, while MBs belonging to the ROI will search $n \times m$ times, at each one of those stored ROIs. In practice, since the ROI is very small (2 MBs for our simulated case), and only 12 ROIs fit in the buffer, we obtain $2 \times 12 = 24$ searches in addition to 99 on the previous frame, which is a 24% increase in computational complexity. Memory complexity is increased two-fold as with dual frame.

The composite frame model we described above is depicted in Fig. 2. The composite frame itself is depicted with larger size for illustration purposes alone. Its actual size in pixel area is exactly the same as any other frame. Non-ROIs reference only the short term frame buffer, while MBs belonging to the ROI conduct a search over the short term frame buffer and the composite ROI, at $n \times m$ locations. For this case of two MBs, the left one will conduct a search centered on the correspoding left one of the stored ROI window, the right one's search will be centered on the corresponding right MB, and so on. We do not search over all possible positions of the ROI window. Hence, dark grey MBs search centered on dark grey composite MBs and light grey ones in the current frame search light grey MBs in the composite frame (apart from one search on the previous frame).

5 Error Concealment

During our experiments with aerial imagery, we noticed that in the presence of errors, and if the error concealment (EC) algorithm simply consists of using all the neighboring motion vectors, this can have detrimental effects for dual frame coders. Aerial imagery is characterized by heavy motion and large valued motion vectors (MVs). This has the unwanted effect of having uncorrelated MVs that point to different references (previous frame or long-term one). EC performance is thus jeopardized, making the scheme underperform a single-frame implementation.

To counter this problem, we ensured EC MV prediction only from MVs pointing to the same references. We used the ROPE [10] algorithm for efficient mode selection and assume that every Group-Of-Blocks (GOB) is packetized into a single and independently decodable variable-length packet. Traditionally, the median of the three closest MVs from the upper GOB is employed for predicting the lost MVs and applying error concealment, in what we call median-3. Instead of just using the three closest MVs from the upper GOB, we used the three MVs from the bottom GOB as well. We use the median of as many of the 6 MVs as are received. We will refer to this scheme as median-6. The modification of the EC algorithm has the implication that the ROPE recursive estimations will have to be modified. The required changes can be inferred as in [9].

6 Experimental Results

In this work we used the ROPE estimator, but our framework could use any reliable distortion estimation metric. The objective of this paper is to point to the flexibility and performance that can be obtained when ROI coding is done using both rate allocation of memory and computational resources with the selective dual frame coder.

Three markedly different video sequences were used for our experiments. One is the aerial surveillance video sequence “escondido3”. It tracks a car travelling on a freeway, and was taken by a gyroscopically stabilized camera mounted on a helicopter. The video sequence is captured at 30fps and 320×240 pixels. An example of this sequence can be seen in Fig. 3(a)-(b). Their corresponding ROI maps (the lighter color denotes the ROI) are in Fig. 3(c)-(d). It is characterized by continuous panning and some light zooming in and out. 100 frames were encoded with a modified H.263+ video coder. Feedback was available with a delay d assumed equal to 3.

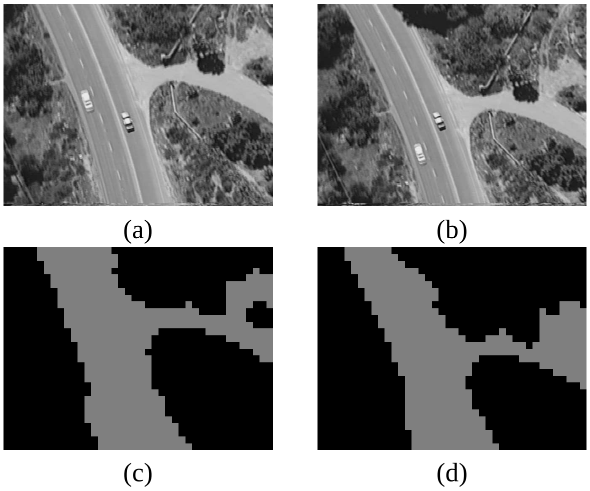
(a) (b)
(c) (d)

Figure 3. Aerial Surveillance Video. (a) Frame 0. (b) Frame 9. (c) ROI Map 0. (d) ROI Map 9.

The second image sequence used was “escondido6” which was also an aerial imagery video. However, this one shows the horizon. The top-half shows the sky and clouds, while the bottom half shows terrain which is set to be our ROI. The third image sequence used was the video-conferencing sequence “akiyo” that is very static and easy to classify into ROI and non-ROI.

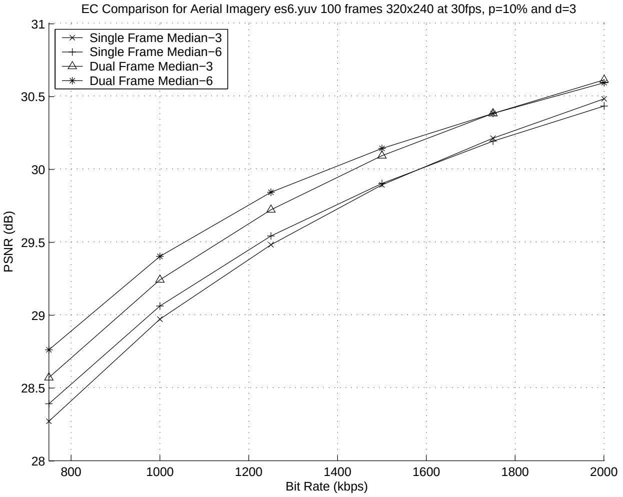

Figure 4. EC schemes comparison for “escondido6”.

Four distinct coding approaches were studied: single and dual frame as in [9], selective dual frame, where only ROI MBs have access to the long-term frame, and composite frame, where only ROI MBs have access to the composite frame buffer. ROI PSNR is averaged per-pixel over all ROI pels throughout the sequence, and *not* per-frame. Likewise, non-ROI PSNR is calculated per-pixel for all non-ROI pixels in the sequence.

6.1 Error Concealment Comparison

The error concealment method that was developed in Section 5 is evaluated in Fig. 4. It can be observed that the gain from using top and bottom GOBs (median-6) versus only the top ones (median-3) is close to 0.25dB, and is more pronounced for the single frame scheme. The median-6 EC algorithm was found to be particularly effective when coding aerial imagery due to the heavy motion involved. No ROI implicit or explicit resource allocation was used. Total sequence PSNR is reported.

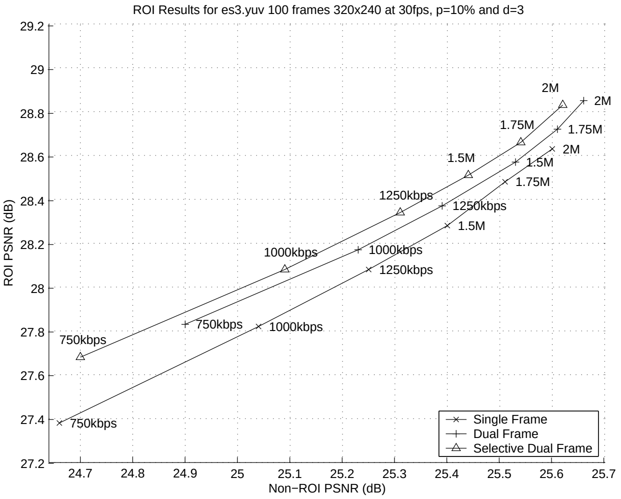

Figure 5. Comparison of coding schemes. ROI vs. Non-ROI PSNR for varying bit rate.

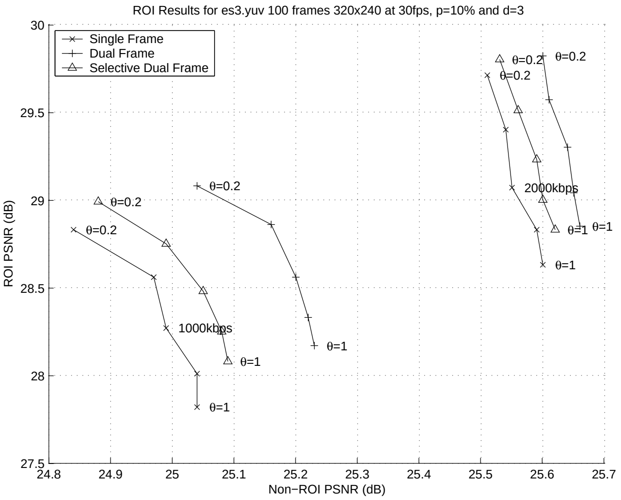

Figure 6. Comparison of coding schemes. ROI vs. Non-ROI PSNR for varying bit rate and θ.

6.2 ROI Resource Allocation

Results for “escondido3” at a resolution of 320×240 are presented in Fig. 5 and Fig. 6. In Fig. 5 we examine the performance of single-frame, dual-frame and selective dual-frame. The y axis shows the Peak Signal-to-Noise Ratio (PSNR) of the region of interest, while axis x shows the PSNR of the non-interesting parts of the image. Hence we can now judge how the ROI quality is improved with respect to the rest of the image. When comparing dual frame to single frame, it is observed that dual outperforms single for both metrics, ROI and non-ROI. If we now compare the y axes we observe that selective dual frame is slightly better than single, but is outperformed by dual frame, which was expected since non-ROI is encoded as single-frame. Comparing the y axes, for equal total bit rates, we see that the result for selective dual frame is very close to the one for dual frame, and captures most of the gain from single to dual, at a fraction of the computational complexity. Unfortunately it is not feasible to capture the entire gain due to ROI boundary effects.

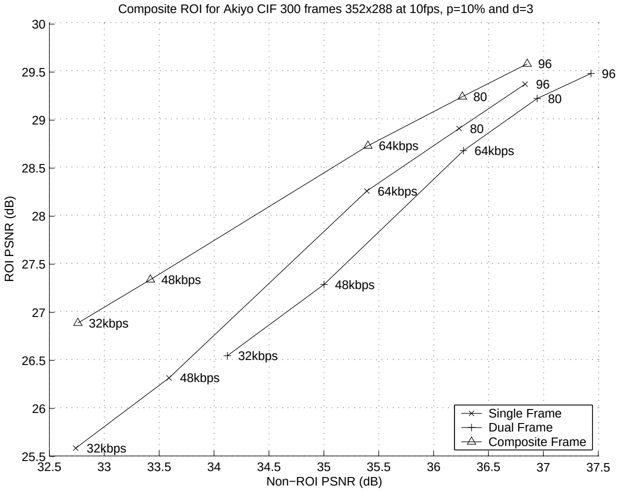

Figure 7. Composite scheme. “Akiyo” at 10fps.

In Fig. 6 we compare single, dual and selective dual frame for two discrete bit rates by varying the relaxation parameter θ. It is obvious that the lower θ is, the more bit rate is allocated to the ROI. As a result, the quality of the non-ROI regions drops. Again, we can make the argument that most of the gain from going to a dual frame scheme is captured with a selective dual frame scheme. For $\theta = 1$ and 2Mbps we observe that the ROI PSNR of selective dual frame is almost identical to dual.

6.3 Composite Frame

Results for the composite frame scheme are presented in Fig. 7 and 9. An actual composite long-term frame can be seen in Fig. 8. Fig. 7 deals with “akiyo” CIF at 352×288 and 10fps. In this case we observe that for equal bit rates, while the non-ROI PSNR of the composite scheme is equal or somewhat worse to the one for single-frame, the ROI PSNR is significantly better than single and surpasses dual frame. Hence, with a fractional increase in computational complexity we obtain ROI reconstruction quality 0.5dB better than the one of single-frame, and still better than the one for dual frame.

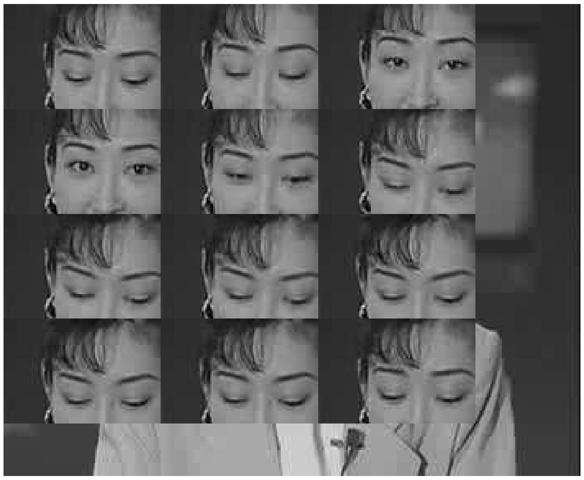

Figure 8. Composite long term frame when encoding frame 100 of “Akiyo”.

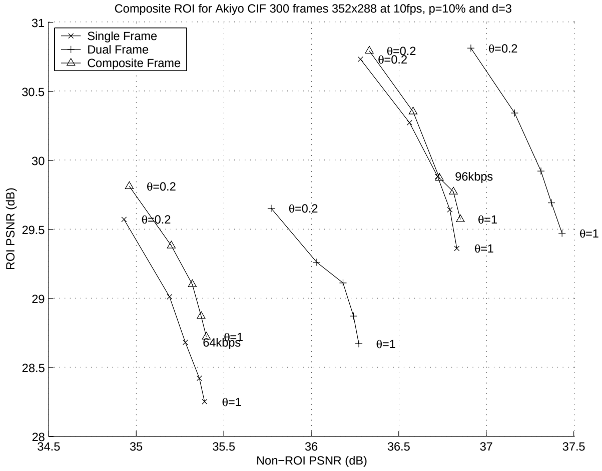

Figure 9. Composite scheme. Variation of θ.

In Fig. 9, which is analogous to Fig. 6, the composite frame scheme outperforms (in the y axis) even the dual frame scheme. The non-ROI PSNR is again fairly identical to single-frame, and much lower than dual frame. Our accomplished objective was the high quality coding of the ROI while maintaining overall low-complexity.

7 Discussion and Conclusion

The experimental results from the previous section show that in the case of selective dual frame we can obtain ROI PSNR gains comparable to the ones for full dual frame, at a fraction of the computational resources. The full dual frame motion estimation requires a 100% increase in computational complexity over a single frame search, whereas the selective dual frame applied only to the ROI requires a 24% increase and provides quality in the ROI comparable to the full dual frame. It is safe to say that the additional complexity is proportional to the size of the ROI. The composite frame scheme proved to be particularly effective for video-conferencing applications with small and easy to define ROIs, and provided ROI performance better than the one of a full dual frame scheme at a fraction of the computational cost. Still, memory requirements remain the same for both proposed approaches. Further work could be pursued in the direction of finding a near-optimal rule that would quantify the correspondence between θ and the resulting increase in PSNR of the ROI. Selective forward error correction (FEC) with respect to the ROI could be applied by designing an Unequal Error Protection (UEP) scheme, with increased amounts of channel coding to the interesting parts of the image. For multiple ROIs this can lead to a more granular allocation of FEC protection.

Segmenting the frames into multiple regions-objects and encoding them in a different manner has already been proposed and incorporated into the MPEG-4 Video Coding standard [11]. Objects, obtained with some arbitrary segmentation scheme, are encoded and decoded independently from one another into Video Object Planes (VOP). During motion estimation and compensation these VOPs only reference pels belonging to the same VOP at a previous time instance. VOPs can be encoded with different encoding parameters such as the Quantization Parameter, and as such can be decompressed at different reconstruction qualities.

However, in our approach, object boundaries *do* overlap during motion estimation, and *more* than one frame is made available as a reference during motion estimation. This is done selectively to enhance the quality of the ROI alone. The additional reference frame (apart from the previous one), can be either a regular frame or a composite one, storing multiple instances of the same object. We investigated prioritizing frame regions not by lowering the QP for a specific object/region, but by allowing it to use more computational and memory resources.

8 Acknowledgments

We would like to thank Ben Ochoa and Prof. Serge Belongie for providing the aerial imagery used in our experiments, and the anonymous reviewers for their helpful comments. This work was supported in part by the National Science Foundation, the CoRe program of the State of California, and the Office of Naval Research.

References

[1] J. Strom and P. C. Cosman, “Medical image compression with lossless regions of interest,” *Signal Processing, Special Issue on Medical Image Compression*, vol. 59, pp. 155–171, June 1997.

[2] A. P. Bradley and F. W. Stentiford, “JPEG 2000 and region of interest coding,” in *Proc. Digital Image Computing Techniques and Applications*, Jan. 2002.

[3] A. P. Bradley, “Can region of interest coding improve overall perceived image quality?,” in *Proc. APRS Workshop on Digital Image Computing*, Feb. 2003, pp. 41–44.

[4] E. Nguyen, C. Labit, and J.-M. Odobez, “A ROI approach for hybrid image sequence coding,” in *Proc. IEEE International Conference on Image Processing*, Nov. 1994, vol. 3, pp. 245–249.

[5] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, “Low bit-rate coding of image sequences using adaptive regions of interest,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 8, pp. 928–934, Dec. 1998.

[6] Y. Yang and S. S. Hemami, “Rate-distortion optimizations for region and object based wavelet video coding,” in *Proc. 34th Asilomar Conference on Signals, Systems, and Computers*, Nov. 2000.

[7] A. Leontaris and P. C. Cosman, “Dual frame video encoding with feedback,” in *Proc. 37th Asilomar Conference on Signals, Systems and Computers*, Nov. 2003.

[8] R. Kutka, “Content-adaptive long-term prediction with reduced memory,” in *Proc. IEEE International Conference on Image Processing*, Sept. 2003, vol. 3, pp. 817–820.

[9] A. Leontaris and P. C. Cosman, “Video compression with intra/inter mode switching and a dual frame buffer,” in *Proc. IEEE Data Compression Conference*, Mar. 2003, pp. 63–72.

[10] R. Zhang, S. L. Regunathan, and K. Rose, “Video coding with optimal inter/intra-mode switching for packet loss resilience,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, June 2000.

[11] T. Sikora, “The MPEG-4 video standard verification model,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 19–31, Feb. 1997.