Perceptually-weighted error-resilient coding for MPEG-4

Simon N. Fabri, Ahmet M. Kondoz

Centre for Communications Systems Research, University of Surrey

Abstract: *Current video coding standards such as H.263++ and MPEG-4 can efficiently compress video sequences down to bitrates suitable for transmission over the forthcoming mobile networks and current low-speed Internet dial-up connections. These schemes are however generic, in that they aim to encode sequences describing any type of scene, from video conferencing to general streaming or archived video. This paper proposes a set of schemes by which the encoding criteria used will be dependent upon the nature of the scene and its intended application. This approach does not only achieve better visual quality for the represented sequences, but is also fully compatible with current MPEG-4 V1 decoders. Approaches for dealing with video conferencing sequences and transmission of sports footage are described.*

1. Introduction.

A number of different video compression schemes suitable for use at low bitrates are currently available, including a number of proprietary codecs, as well as other standardised codecs. The standardised codecs are dominated by two families, the ITU-T's H.263 family of video coding standards and the ISO MPEG-4 audio-visual standard. [1],[2] Both codecs are based on similar technology, namely the use of motion-predictive Discrete Cosine Transform to perform lossy video compression. This transform compacts the high-energy components into a few coefficients, which allows for the removal of 'visually-redundant' information with little perceptible difference to the image quality.

2. MPEG-4

When coding video sequences, a standard MPEG-4 encoder assumes that all areas of a given scene are equally important. This allows the codec to be 'generic' and be able to operate for a wide range of sequences. It is however evident that in most sequences, there are regions that are subjectively more important than others. For example, in video conferencing applications, the area of the scene containing the participant in the conference is clearly more important than whatever is happening in the background. The relative importance between different regions becomes critical when operating at the lowest practical bitrates allowed by the codec, as the correct allocation of throughput of output bits becomes extremely essential for obtaining a decoded output of acceptable quality.

The MPEG-4 standard divides a scene into macroblocks 16×16 pixels in size for carrying out motion estimation. It is also possible to dictate the quantisation parameter, i.e. the level of spatial detail, for each macroblock. This feature of the coding syntax was exploited by our encoding scheme which introduces a technique referred to as Differential Quantisation (DQ) to apply different levels of spatial detail to different areas of the scene. The criteria for determining the priority of each area for the purposes of throughput allocation are dependent upon the application the codec is being used for. The use of Differential Encoding in two different scenarios is outlined in the next sections.

3. Video Conferencing

Differential encoding requires the use of an automatic segmentation and classification algorithm to split the scene into regions with differing priorities. Traditionally, image

segmentation has been considered as being excessively computationally-intensive for real-time applications as it is extremely complex to extract a logically-contiguous region of interest from a scene containing much spatial and temporal detail. However, if classification needs only to be carried out on a macroblock-by-macroblock basis, the task becomes considerably simpler. The technique used in this scheme exploits the semantic information embedded in the MPEG-4 bit stream representation of the video sequence. It is assumed that the foreground, or region of interest, can be characterised by two main features, namely it is close to the centre of the scene, and has on average, greater temporal activity than the rest of the scene. The Sum of the Absolute Differences (SAD) is a metric used to determine the level of activity between the same macroblock in successive frames. The SAD is available to the source encoder during motion estimation, but may easily be extracted by the segmentation algorithm. The centre macroblock is selected by the algorithm as being the seed block, namely the first block for inclusion in the foreground. A metric is then computed for all neighbouring MBs consisting of the normalised SAD of the macroblock added to a factor representing its distance from the centre of the scene. The macroblock with the highest metric is then selected and added to the foreground. The subsequent macroblock is then selected from all the neighbours of the included blocks. This process continues until the number of MBs included in the foreground reaches a predetermined threshold.

3.1 Performance

Experiments were carried out on the *Foreman* sequence, which is 176×144 pixels in size and is specified by the ITU-T as one of the standard sequences to be used for low-bitrate coding. The segmentation algorithm consistently identified correctly the face region, which was found to contain approximately 24 macroblocks out of the 99 MBs which constitute the scene. In order to allow for a high quantiser differential between foreground and background MBs, an adaptation of the MPEG-4 Adaptive Intro-Refresh scheme was used. This involves coding a fixed number of Intra blocks in each frame, which eliminates the need for non-predictively-encoded Intra frames. The frame rate was set to 5fps, while the target bitrate was 40 kbit/s. Each frame was set to contain 8 Intra MBs within the foreground of each scene, and another 2 in the background. The Intra MBs are set to have a finer quantiser than the Inter MBs, which make use of predictive coding, (QP set to 10 rather than 16) so that the spatial quality of the foreground is considerably higher than the background, even though there is more motion. This can be seen in Figure 1, where the average PSNR for the foreground is 30.7dB and 31.1dB for the cases without and with DQ respectively.

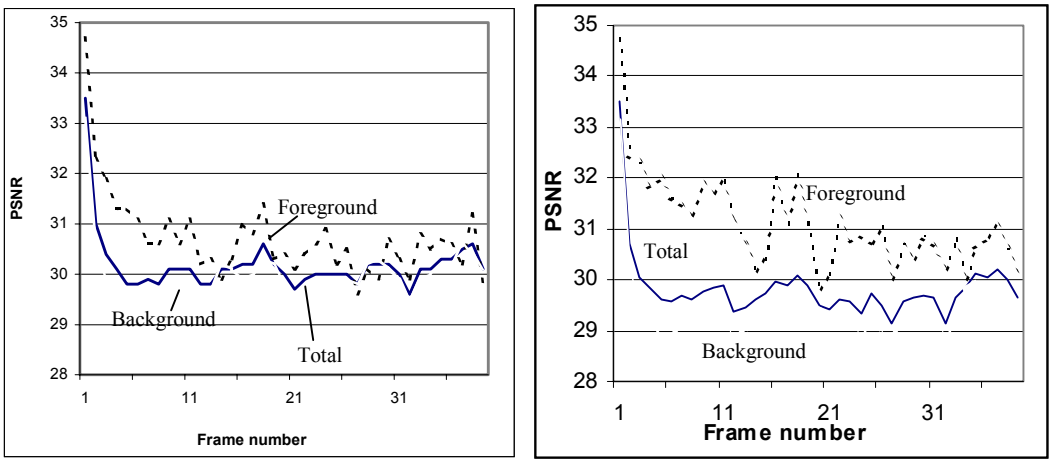


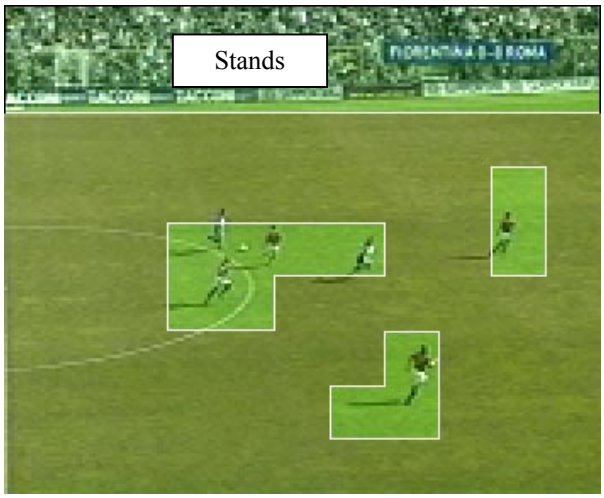
Figure 1. Comparison between standard MPEG-4 (*left)* and MPEG-4 with DQ *(right)*

4. Football Sequences

One of the most demanding applications for low-bitrate video compression is the efficient representation of sports footage. These types of video sequences are characterised by high-activity, rapid movement, camera panning, and the necessity of high-resolution display of some areas of the scene, such as the participants in the sport. All these contribute towards making the task of encoding such sequences at bitrates suitable for transmission over mobile links notoriously difficult. In order to achieve high levels of perceptual quality at the decoder, video sequences containing football footage must satisfy a number of criteria. The most important of these are that there is a high temporal resolution and sufficient spatial detail in the areas of high activity. The quality of the scene around the area where the action takes place should be sufficiently detailed for the viewer to be able to easily discern what is happening. The approach taken in improving the quality of the video sequences exploits all these features as found in footage generated for broadcast transmission. The scene was divided into four regions, namely, in descending order of importance, the football itself, the players and officials, the pitch and the stands.

In order to minimise computational complexity, the segmentation and classification algorithm exploited the characteristics of football sequences. The source video images were manipulated in the hue-saturation-value domain, in which the hue roughly represents the colour of the pixel being transmitted. The pitch area was identified as being all pixels whose hue value is within 2% of the median hue value of the image. This approach was found to be robust to variations in shadowing on the pitch, as these are represented by changes in saturation values, whereas the hue remains relatively unvaried. The histogram of the hue component of a typical scene clearly shows the contribution of the colour of the playing field, which is represented by a prominent peak. The main weakness of this technique lies in the sub-sampling of the chrominance components of the YUV format used as input to the video codec, which reduces the resolution of the hue component. The remaining areas are then deemed to belong to players and officials on the field, field markings and other paraphernalia and crowds on the terraces. The last of these were found to be the easiest to extract, as whenever a camera is aimed at the playing field lengthwise from a height above the field as is generally used in broadcast-quality footage, perspective dictates that the border of the pitch always enters the field of view from one of the four corners. This allows for all contiguous blocks of non-pitch areas at the corners of the scene to be classified as belonging to the terraces. All remaining macroblocks are classified as being either players or officials, with the exception of the ball, which is identified by carrying out a least-means-error search with a mask representing the football.

The efficacy of the Differential Quantisation technique used can be seen by examining some of the obtained results. Figure 2 shows the result of the segmentation as performed upon a frame extracted from one of the sequences used. It can be seen, that the players are correctly identified, as is the crowd. The sequence was then encoded at a rate of 100 kbit/s at 10 frames/sec both with and without the Differential quantisation scheme enabled. A close-up view of one of the players clearly shows the improvement in the spatial quality of the player. Examination of the PSNR traces of the four regions, clearly emphasises the performance advantages. At the same rate, when using a single quantisation parameter, it can be seen that the spatial quality of the players and the football is consistently below the average PSNR for the whole sequence and well, below that of the pitch itself, which is fairly featureless. When DiffQuant is enabled, the highest quality macroblocks are those containing the highest-priority information, namely the players and the football.



Quality of player using standard MPEG-4 *(top)*, and MPEG-4 with Differential Quantisation *(bottom)*

Figure 2. Performance of Segmentation and Classification Algorithm

	Pitch	Players	Crowd	Total
Diff. Quant.	26.6 dB	27.3 dB	20.9 dB	24.8 dB
Standard MPEG-4	27.5 dB	24.7 dB	21.9 dB	25.3 dB

Table 1. Peak Signal-to-Noise Ratios for Football Sequence

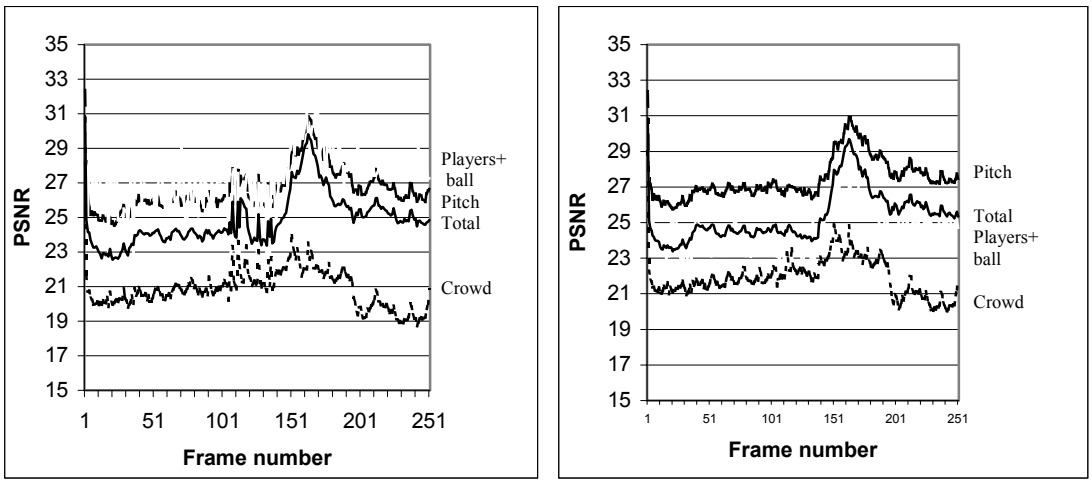


Figure 3. Comparison between standard MPEG-4 (*right)* and MPEG-4 with DQ *(left)*

5. Conclusion

This paper has shown that the MPEG-4 coding syntax can be adapted to apply different levels of spatial detail to different areas of a video scene. This feature was used to enhance the perceptual quality of low-bitrate video sequences as used in video-conferencing applications and streaming video applications. The Differential Quantisation technique introduced will facilitate the implementation and enhance the quality of low bitrate video services for use in mobile and fixed low-bitrate environments as it may be used with any standard MPEG-4-compliant decoder.

6. References

1. Schafer R., "MPEG-4: A Multimedia Compression Standard for Interactive Applications and Services", *IEE Electronics & Communication Engineering Journal,* Dec. 1998

2. Talluri R., "Error-Resilient Video Coding in the ISO MPEG-4 Video Standard", *IEEE Communications Magazine*, June 1998, Vol. 36, No. 6, pp. 112-119.