14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

RATE-DISTORTION PERFORMANCE OF DUAL-LAYER MOTION COMPENSATION EMPLOYING DIFFERENT MESH DENSITIES

Andy C. Yu, Heechan Park, Graham R. Martin

Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom
Email: {andycyu, heechan, grm}@dcs.warwick.ac.uk

ABSTRACT

We describe a dual layer algorithm for mesh-based motion estimation. The proposed algorithm, employing meshes at different scales, aims to improve the rate-distortion performance at fixed bit rates. Without the need for time-consuming evaluation, the proposed algorithm identifies the motion-active regions in a picture. By enhancing the mesh structure in the aforementioned regions, a significant improvement in motion estimation is evident. Furthermore, the scheme to construct two different significance maps at both encoder and decoder provides a reduction in transmission overhead. Simulations on test sequences possessing high motion activity show that the proposed algorithm results in a better PSNR-rate performance compared to single layer motion estimation using 16x16 triangular meshes. Improvements of up to 1.6dB are obtained.

1. INTRODUCTION

The use of mesh-based motion prediction provides an alternative to conventional block matching. The procedure is to divide the current frame into a number of triangular or rectangular patches, and to find the best matching corresponding patch in the reference frame when deformed by affine transformation. The mean absolute difference is commonly used as the matching criterion, and the motion is defined by the change in position of the corresponding vertices [2]. Compared to the conventional block-matching algorithm (BMA), mesh-based motion models provide more visually acceptable results in the predicted frame. This is credited to the fact that connectivity of the patches is maintained, compared with a disjointed collection of blocks in BMA. Furthermore, motion is estimated more accurately due to the utilisation of an affine transformation, which supports various spatial deformations, such as translation, rotation, and zoom [3]. The deformed patches are described by the motion displacement of the grid points (the points shared by the vertices of the patches) for motion coding purposes.

Intuitively, a better rate-distortion performance is expected when finer meshes are utilised. Fig. 1 provides evidence of this by showing the coding results for the *Football* sequence when triangular patches of size 32x32, 16x16, and 8x8 pixels are employed. It is observed that 16x16 meshes achieve a better rate-distortion performance than 32x32 meshes. However, a decrease in mesh size does not necessarily lead to a better identification of motion-active regions. This is due to the connectivity constraint of the fine mesh that reduces the effective area for the motion search. Fig. 1 further demonstrates the decrease in PSNR obtained by the employment of 8x8 meshes at low bit rates. This is attributed to the overhead of an increased number of motion vectors which is certainly a disadvantage with regard to compression.

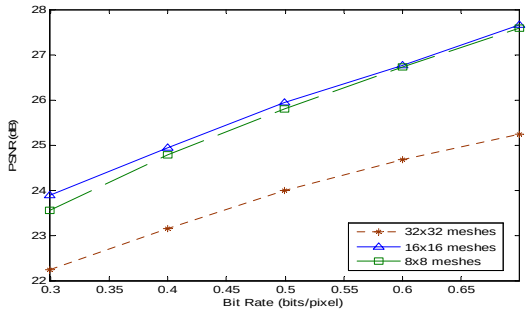

Fig. 1. Rate-distortion performance of *Football* sequence employing mesh-coding with different patch sizes.

In a previous study [4], we proposed an algorithm incorporating partial refinement by the employment of a layered mesh topology without overhead. The technique is based on the utilisation of meshes featuring multi-scale to cope with different prediction demands in a picture. In this paper, a modified algorithm is presented to improve the rate-distortion performance for video sequences that require intensive motion description. This paper is arranged as follows. A study of the relationship between motion coding and residue coding is introduced in the next section. Section 3 provides a detailed description of the proposed dual-layered motion coding. Simulation results and conclusions are presented in the last two sections.

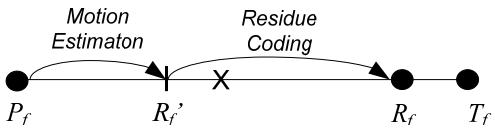

Fig. 2 Conversion process from reference frame, P_f, to target frame, T_f.

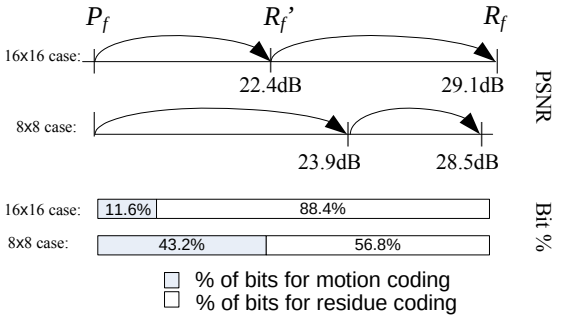

Fig. 3 PSNR improvement and bit allocation for the *Football* sequence coded at 0.30bits/pixel.

2. RELATIONSHIP BETWEEN MOTION CODING AND RESIDUE CODING

Fig 2 illustrates a video coding process where T_f and P_f are the target and reference/previous frames, respectively. The goal is to convert P_f to R_f, a reconstructed frame resembling T_f, by employing motion estimation and residue coding. The overhead in the process is the encoded motion vectors and frequency spectra derived from the residue data. The limit of effective motion prediction (marked as X) is dependent on the similarity between reference frame and current frame. Once the limit is reached, the prediction cannot be improved by the expense of additional bits. In the case of a fixed bit rate, an over-allocation of bits to the motion description has consequences for the residue coding. Fig. 3 examines the *Football* sequence coded at 0.30bits/pixel, as included in Fig. 1. The diagram shows the PSNR improvement and percentage bit allocation for motion and residue coding. One can see that a significant improvement of 1.5dB in motion prediction is achieved by motion estimation utilising finer meshes. In contrast, the four-times increase in bits assigned to the motion information results in a poor performance for the residue coding. This degradation is especially serious at low bit rates which is reflected in Fig. 1. In the example, the overall PSNR is boosted by certain regions having a need for more detailed motion description. However the additional motion vectors describing static objects are essentially redundant. An attempt to deliver additional bits to predict the movement of new occluded/discovered objects ought to be avoided, as residue coding is a more efficient method of encoding them.

Meshes incorporating different sized patches provide a good solution to the aforementioned problem, since finer meshes need only be used where required. The question that

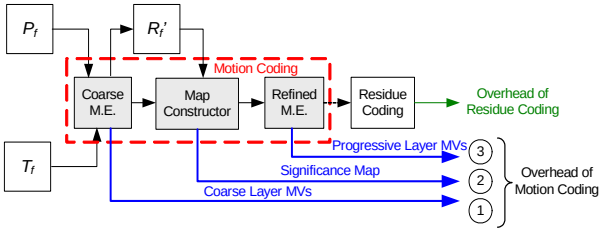

Fig. 4 The proposed motion coding at encoder side.

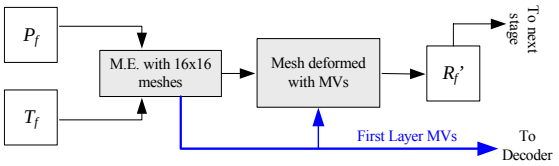

Fig. 5 Block diagram of the coarse layer at encoder side.

arises is how to apply finer meshes efficiently in order to overcome the connectivity constraint. Furthermore, the overhead in specifying the location of regions employing the finer meshes should be reduced to the lowest degree. In next section, we present a dual-layer approach to fulfil the two stated objectives.

3. PROPOSED DUAL-LAYER ALGORITHM

The proposed dual-layer algorithm predicts the motion of objects with two mesh topologies of different densities. A triangular mesh of patch size 16x16 pixels is initially applied to obtain an intermediate predicted frame, R_f'. A significance map is constructed to determine the location of motion-active regions, to which are applied second layer fine meshes of patch size 8x8 pixels. The encoder for the proposed algorithm is illustrated in Fig. 4. The motion coding can be elaborated as three stages, each of which is described in the following subsections.

3.1. Motion Estimation in the Coarse Layer

Fig. 5 shows the block diagram of this layer. The motion estimation in the coarse layer utilises the hexagonal matching algorithm (HMA) proposed in [1]. The HMA performs motion search by forming a hexagon with a grid point and its six neighbours. The motion vector for the grid point inside the hexagon is acquired by deforming the mesh with an affine transform. Estimation of the motion for the other vertices forming the hexagon is performed in a similar manner. Subsequently, a refinement process is applied to each grid point if the displacement of any grid point in the hexagon has been modified. The iterative refinement process is implemented for all grid points until convergence to local or global minima. An intermediate frame, R_f', is consequently constructed prior to

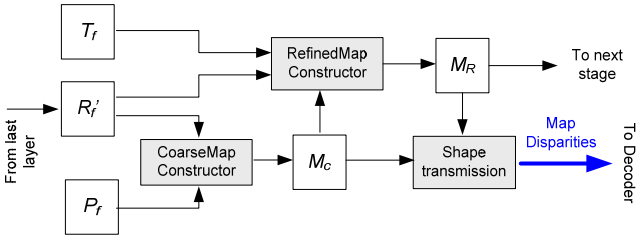

Fig. 6 Diagram showing the map construction.

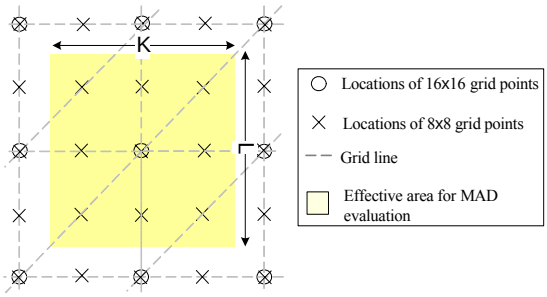

Fig. 7 Effective area for MAD evaluation.

the motion estimation. It is restored by deforming the meshes of the reference frame with updated displacements of all grid points.

The grid points with non-zero displacement show the existence of motion activity within an effective area. Intuitively one would expect a better prediction to be achieved by a further estimation with a finer mesh. However, this is not always the case. If a rigid movement exists (simply panning camera action), the motion estimation at this stage is sufficient. Another case is the movement of the surrounding grid points resulting from the sudden appearance of an occluded object. As mentioned, the allocation of additional bits in this case does not necessarily result in any significant improvement in the prediction. Ideally, a finer mesh should be applied to those parts of an object that appear to have disparate motion, for instance at the boundaries of a moving object. In the next subsection, we describe how these areas are detected through the construction of a significance map.

3.2. Construction of the Significance Map

The significance map contains binary decisions to determine whether the grid points require further motion examination. To reduce transmission overhead, maps of different precision are constructed at each side of the codec. In the decoder, a less accurate map, M_C, is built; while the encoder has a more refined map (denoted as M_R). Fig. 6 illustrates map construction at the encoder.

Since the coarse map is developed at both sides of the codec, the construction has to take into account the constraints of the decoder. A feasible assumption is made that both encoder and decoder restore the same intermediate frame by sending the motion vectors of the previous layer in advance of other motion information. An evaluation scheme based on the Mean Absolute Difference (MAD) is developed for the construction of the coarse map. The proposed scheme calculates a MAD value for a region (of size *K*-by-*L*) centred at each grid point in the reference and intermediate frames, as shown in Fig.7. Note that the positions of the grid points have been redefined for the intermediate frame, so that an identical coordinate system is established in both frames. The proposed MAD evaluation is recorded as follows:

$$\mathrm{MAD}_{G(i,j)} = \frac{1}{K \times L} \sum_i \sum_j \left| \widetilde{B}_P(G(i,j)) - \widetilde{B}_{R'}(G(i,j)) \right| \quad (1)$$

where $G(i, j)$ is the grid point at location (i, j); $\widetilde{B}_P(G(i,j))$ and $\widetilde{B}_{R'}(G(i,j))$ are rectangular areas of size *K*-by-*L* centred at $G(i, j)$ in the previous frame and intermediate frame, respectively.

Subsequently, a decision process is used to provide a binary outcome for each grid point with a threshold, Δ_{Th1}:

$$M_c(G(i,j)) = \begin{cases} 1 & \text{if } \mathrm{MAD}_{G(i,j)} \geq \Delta_{Th1} \\ 0 & \text{if } \mathrm{MAD}_{G(i,j)} < \Delta_{Th1} \end{cases} \quad (2)$$

A positive outcome in (2) indicates a significant intensity difference in the corresponding regions of each frame. One explanation of a positive outcome could be the deformed mesh which indicates that the region contains non-homogenous content. Thus, a more advanced examination of the target frame is required to determine the content. This is performed in a similar fashion to that of (1) except that the reference frame is replaced by the target frame, T_f:

$$\mathrm{MAD}'_{M_c(G(i,j))=1} = \frac{1}{K \times L} \sum_i \sum_j \left| \widetilde{B}_T(G(i,j)) - \widetilde{B}_{R'}(G(i,j)) \right| \quad (3)$$

Note that (3) is performed selectively on the valid grid points resulting from (2). The decision process for the refined map, M_R, is then:

$$M_R(G(i,j)) = \begin{cases} 1 & \text{if } \Delta_{Th1} \leq \mathrm{MAD}'_{M_c(G(i,j))=1} \leq \Delta_{Th2} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

One can see that a positive outcome is generated if the MAD' value in (3) is found to be between the two thresholds, Δ_{Th1} and Δ_{Th2}. The application of the first threshold removes the regions which have similar content to the target frame. In contrast, the application of Δ_{Th2} detects an inaccurate prediction in a region resulting from the previous layer. One possible scenario is the appearance of a new occluded object. Thus the regions are not considered for further improvement.

Valid candidates in the refined map indicate that the corresponding regions require motion refinement in the next stage. A post-process is still needed to remove solitary candidates from the map. The reason is to prevent 'expensive' bits being consumed by isolated regions which do not necessarily lead to a significant improvement in prediction. Eventually the contents of the refined map are coded for transmission. Since the coarse map has been constructed at the decoder, it is only

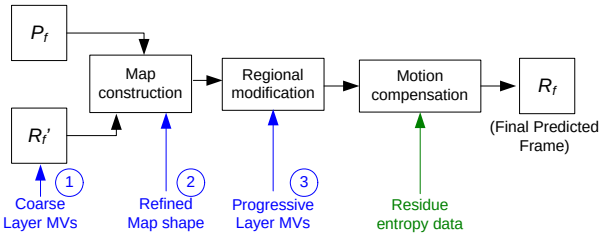

Fig. 8 Diagram of the proposed decoder.

necessary to transmit the corresponding valid locations recorded in the coarse map.

3.3. Refinement in Progressive Layer

The progressive layer aims to improve the regional prediction in the intermediate frame with the enhancement of the finer mesh. The motion refinement is performed selectively at the valid grid points outlined in the refined map. However, the significance map is built to a different scale to that of the coordinate system. Scale conversion is performed by activating the eight adjacent 8x8 grid points surrounding each valid candidate in the map. The corresponding location of grid points at the different scale is illustrated in Fig. 7. The grid points in the map indicated as 'background' are not permitted to change the displacement during the motion refinement. This restriction does not violate the application of the hexagon matching algorithm. Subsequent to the refinement process, the updated motion vector in the progressive layer is transmitted to decoder in raster order without specifying the location.

The proposed algorithm provides a solution to the mesh-connectivity problem described earlier. This is due to the strategy of employing two different scales of motion search. The first layer corrects overall motion displacement in the reference frame. The moving details in the intermediate frame are further predicted with the sections of finer mesh. The scheme results in a more detailed search compared to the application of finer meshes on the reference frame directly. Thus, a better regional prediction is sometimes expected.

3.4. Decoder of the proposed algorithm

The decoder for the proposed algorithm undergoes the procedures described in Fig. 8. The received bit-stream comprising both motion information and residue entropy-coded data is decoded on a frame-wise basis. The motion information for each frame can be further decomposed into three layers. The first layer specifies the motion vector differences (MVDs) for the grid points in the reference frame. By deforming the triangular meshes with the updated grid point displacements, an intermediate predicted frame is restored. Subsequently, the decoder repeats the same procedures described in the encoder to create the coarse map. The shape of the map is further refined with the second layer motion information. Eventually, the decoder modifies the details of the moving object with the third layer motion information prior to the residue compensation process.

Table 1
Average PSNR and bits spent per frame prior to residue coding.

Sequence	Average **PSNR(dB)** / bit spent (bits) per frame		
	16x16 mesh	8x8 mesh	The Proposed
Football	**18.46**/918.90	**19.60**/2710.07	**20.68**/3072.62
Fun Fair	**24.10**/359.17	**25.42**/1414.76	**25.43**/1192.79
Ice Skate	**22.35**/441.31	**23.90**/1642.28	**25.23**/1283.72
Mobile	**25.85**/215.52	**26.11**/ 963.38	**26.01**/ 310.55
Stefan	**25.00**/284.07	**26.10**/1070.97	**26.45**/ 895.79

4. SIMULATION RESULTS

This section discusses the performance of the proposed algorithm in a complete codec. Results are presented as improvements over the motion estimation employing 16x16 triangular meshes. The selected video sequences were of QCIF resolution (176x144) and 30 frames of each sequence were processed. Except for the first frame, all the frames are encoded with the hexagonal matching algorithm (HMA) [1]. The search range for the motion estimation is set to ± 8 pixels for both layers. The dimensions MAD evaluation regions in (1) and (3) are of size 16x16 pixels. Δ_{Th1} and Δ_{Th2} in (4) are defined as 5.0 and 40.0, respectively. Finally, the motion information was encoded using the Exp-Golomb method and the residue data using JPEG 2000 at fixed bit rates.

Table 1 shows the average picture degradation and bits spent per frame for the proposed dual-layer algorithm and the single layer system featuring two sizes of mesh. Both measurements are made prior to residue coding. Except for the Football sequence, the proposed algorithm generally requires fewer bits for motion coding compared to the single layer employing a global 8x8 mesh. This is attributed to the decreased number of motion vectors transmitted in the progressive layer. Furthermore, it is observed that for some sequences, the proposed algorithm provides the best motion prediction. This is due to a better motion identification resulting from the finer mesh topology on the intermediate frame rather than directly on the reference frame. Thus, it explains the increased number of bits experienced for the Football sequence. In conclusion, the proposed algorithm obtains a better compromise between prediction accuracy and bits spent in motion coding.

Fig 9 illustrates the PSNR v. bit-rate performance for the *Football* and *Ice Skating* sequences employing the proposed algorithm and three other approaches. The results are made after residue coding with JPEG 2000 at fixed bit rates. The proposed algorithm provides significant PSNR gain over the single layer using 16x16 mesh, compared to the other two algorithms. For the Ice skating sequence, the proposed algorithm benefits from bit reductions obtained from motion information coding. A reduction in motion coding allows more bits to be utilised for residue coding. Thus, a PSNR gain of up to 1.6dB is observed compared to 1.0dB and 0.7dB by the algorithm recorded in [4] and the single layer method em-

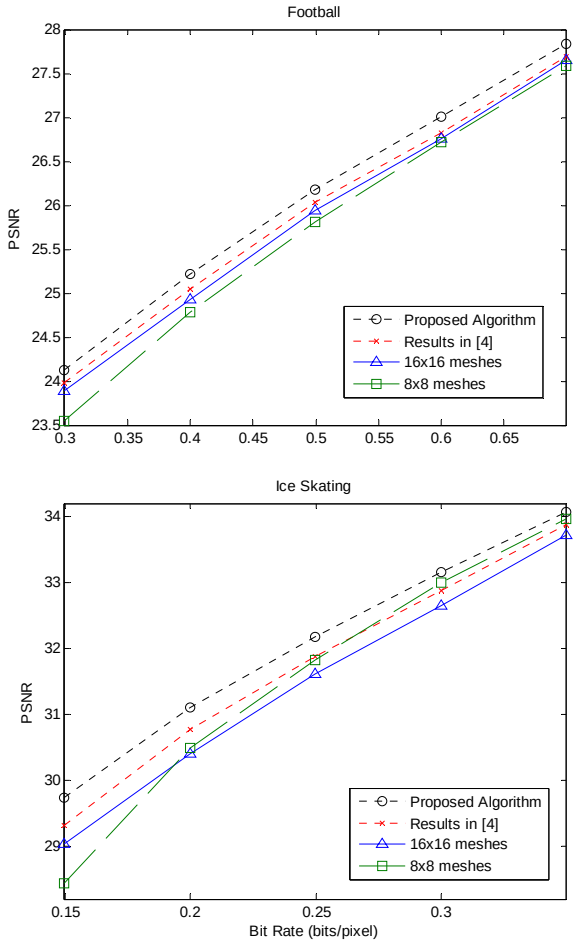

Fig. 9 The PSNR-rate diagram of Ice Skating (top) and Football (bottom) sequences at fixed bit rates.

ploying 16x16 meshes, respectively. In contrast, the PSNR gain in the Football sequence suffers from an increased overhead for motion coding. However, an average improvement of 0.6dB is still found.

5. CONCLUSIONS

In this paper, we propose a dual-layer approach to improve the rate-distortion performance of mesh-based motion compensation. The algorithm initially applies meshes of size 16x16 pixels to identify the moving objects in a picture. A finer mesh is then selectively used to compensate for any inaccurate predictions in the previous layer. By constructing a significance map at both sides of the codec, the overhead of specifying exact locations for the finer mesh is reduced. Simulation results show a significant gain in PSNR compared to other algorithms reported in the literature.

REFERENCES

[1] Y. Nakaya and H. Harashima, "Motion compensation based on spatial transformations," *IEEE trans. on Circuit and System for Video Technology*, vol. 4, no. 3, pp. 339-367, Jun 1994.

[2] A. Nosratinia, "New kernels for fast mesh-based motion estimation," *IEEE trans. on Circuit and System for Video Technology*, vol. 11, no. 1, pp. 40-51, Jan 2001.

[3] Y. Altunbasak and M. Tekalp, "A hybrid video codec with block-based and mesh-based motion compensation modes," *Special Issue of Int. Journal of Imaging System and Technology*, vol. 9, no.4, pp. 248-256, Aug. 1998.

[4] H. Park, A. Yu and G. Martin, "Progressive mesh-based motion estimation using partial refinement," in Proc. of International Workshop on Very Low Bit-rate Video (VLBV) 2005, 4 pp., Sep. 2005.