Model design for scalable two-dimensional model-based video coding

M. Hu, S. Worrall, A.H. Sadka and A.M. Kondoz

Scalable and very low bit rate video coding is vital for audio-visual conversational services over narrow bandwidth channels. A novel model design scheme is proposed in order to make the points of an object model represent the motion more accurately, which will in turn enable better compression. Experimental results demonstrate the performance of the proposed scheme.

Introduction: To improve compression efficiency, much research has focused on model-based video coding, resulting in many 2-D and 3-D model-based video-coding schemes being proposed [1–2]. Scalable video coding is also very important for video transmission over channels with varying bandwidth, intended for terminals with different capabilities.

In this Letter, a new model design scheme is proposed for scalable model-based video codes, based on 2-D mesh-based motion compensation. Previously, much research has been conducted on the representation of video objects using hierarchical 2-D models in order to achieve scalable coding of objects [1–2]. Most of these methods are derived from 3-D model construction schemes used in computer graphics, which cannot represent the motion of objects precisely although motion information is used during the model design. In addition, as a fine-to-coarse strategy was employed in these methods, the computational complexity is usually heavy. To overcome these problems, two steps are used in the proposed scheme as follows.

Scheme description: Fig. 1 shows the flowchart of the proposed scheme. First, the object is segmented into patches with different motion pattern, colour and texture. Particular patch meshes are then constructed for every patch individually and hierarchically, and then combined as a scalable model of the object. In our research, a face is considered as a special object and its model is built separately. The reason is that, for head-shoulder sequences, people concentrate on the face more than other objects, and therefore, errors on the face are very obvious. During experiments, we have found that large errors occur on the eye and mouth if we do not process the face separately. A threshold is used on the ratio of face area to frame area in order to decide whether or not the sequence is a head-shoulder.

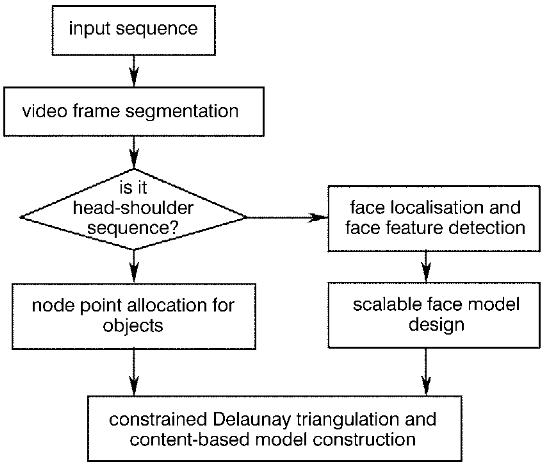
input sequence
video frame segmentation
is it head-shoulder sequence?
face localisation and face feature detection
node point allocation for objects
scalable face model design
constrained Delaunay triangulation and content-based model construction

Fig. 1 *Flowchart of model design scheme*

For node point allocation during model design, the following criteria should be satisfied. First, the motion of the node points should be estimated precisely. To overcome the aperture problem commonly encountered in motion estimation, node points should be allocated on the pixels that contain sufficient grey-level variation, such as corners. Next, we should reduce the area of patches that contain motion discontinuities. In this way, we can reduce the number of pixels with erroneous motion estimation. The best way is to ensure that the edge of the mesh triangles conform to the object boundary.

As a coarse-to-fine strategy is used, the most important points, which can represent the motion of model more precisely, are included in the coarser level. The less important points are included in the finer levels to refine the model. The scheme is described in the following four parts.

(i) Video frame segmentation: In our research, a new video segmentation scheme is employed. First, a watershed transformation is performed on the gradient of the prefiltered image. A Fisher linear discriminant (FLD) [3] is then applied iteratively on the colour, luminance and motion information to merge patches with similar properties. Statistical hypothesis testing [4] is incorporated to detect the intensity change and to merge the patches into foreground and background. Optical flow estimation is used to obtain the motion information [5]. In this way, the image is represented as objects with different colour and motion properties.

(ii) Face detection and 2-D face model construction: The scheme for face feature detection and scalable face model construction is conducted on the segmented patches. First, the skin colour model on YUV space is used to locate the likely position of the face. Face features are then detected and used to verify the position of the face, based on their symmetry. Therefore, the corners of eyes, mouth and nose are detected. The chin is also detected using active snake and deformable template methods. Nineteen points are allocated to model the eyes, mouth, nose and chin. As these points are very important to represent the motion of face features, they belong to level 1. More points are allocated in higher levels based on the muscle distribution and anatomical models in order to build a scalable face model [6]. These points are less important for motion representation, but they are useful for representing the variance of texture due to different face expressions.

(iii) Node point allocation on boundary of object: Two criteria are used for the approximation of contour: one is the curvature of the boundary; the other is point distance. Using curvature as a criterion has some advantages. First, the number of points can be easily controlled. Points are allocated on the position with highest curvature first. Secondly, it can guarantee that the points are allocated to the corner of the boundary. As the boundary is almost a line in some positions, no point is allocated if we only use curvature as a criterion. This is not the optimal mesh structure based on the results in [7]. Therefore, point distance has been adopted as another criterion.

The curvature is calculated in two steps, which are listed as follows:

1. For every point on the boundary, determine the region of support:

   1.1 Define the length of the chord joining the points p_{i-k} and p_{i+k} as $l_{ik} = |\overline{p_{i-k}p_{i+k}}|$. Let d_{ik} be the perpendicular distance of the point p_i to the chord $\overline{p_{i-k}p_{i+k}}$

   1.2 Start with $k = 1$. Compute l_{ik} and d_{ik} until either (a) $l_{ik} > l_{i,k+1}$ or (b) $d_{ik}/l_{ik} > d_{i,k+1}/l_{i,k+1}$

   Then region of support p_i is the set of points which satisfy either condition (a) or (b).

2. Estimate the curvature based on the following formula:

$$\vec{u}_{i-k} = (x_{i-k} - x_i, y_{i-k} - y_i) \text{ and } \vec{u}_{i+k} = (x_{i+k} - x_i, y_{i+k} - y_i)$$

Curvature square:

$$\kappa^2 = \frac{1}{k}\sum_{j=1}^{k}\left|\frac{\vec{u}_{i-k}}{|\vec{u}_{i-k}|} - \frac{\vec{u}_{i+k}}{|\vec{u}_{i+k}|}\right|^2$$

During our research, two levels are used to approximate the boundary (more levels can be used, which depends on the complexity of boundary). The first level is based on the curvature. Level 2 is based on the distance between the node points in level 1.

(iv) Node point allocation in interior of objects: The node point allocation in the interior of the object is related to video frame segmentation, which has been discussed previously. The node points of the coarser levels are allocated on the intersection points and corners of patch boundaries in order to guarantee that the edges of triangles conform to object boundaries. They are important for the motion estimation of the mesh. For the finer levels, more control points are allocated in the interior of patches with larger gradient. The distance between the control points is also considered in order to avoid a skinny mesh. During our experiments, four levels are used. The number of points in every level is decided by the size of object, texture and motion of the frame.

After node point allocation is completed, the constrained Delaunay triangulation algorithm is used to build the mesh structure of the object model [7]. As the node points in level 1 and level 3 are located on the object contour and patches contours, they are considered as the constraint in order to guarantee that the edges of mesh triangles conform to the contour of motion discontinuity.

Results: In our experiments, several test sequences, such as Carphone (CIF), Mother-daughter (QCIF) and Akiyo (QCIF) are used. Carphone and Akiyo are considered as head-shoulder sequences and face feature detection is conducted. Fig. 2 shows the results of model design for the Carphone sequence. For the boundary, 25 and 15 points are used to represent the contour hierarchically in level 1 and 2, respectively. For the interior of the object (except face), 15, 15, 10 and 30 points are used. The face is represented using four levels with 25, 5, 5 and 7 points, respectively.

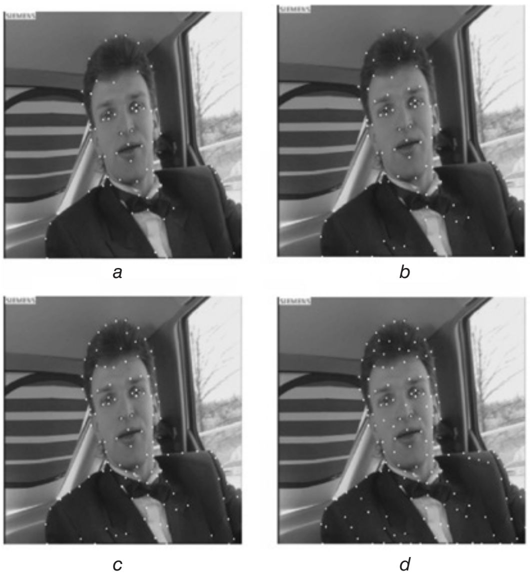

Fig. 2 Models for Carphone sequence

a Level 1 including points of boundary level 1 and interior level 1 (BL1 + CL1)
b Level 2 (BL1 + CL1 + CL2)
c Level 3 (BL1 + CL1 + CL2 + BL2 + CL3)
d Level 4 (BL1 + CL1 + CL2 + BL2 + CL3 + CL4)

The performance of the model design scheme is tested by motion estimation and warping. A hexagonal matching algorithm [8] is used to estimate the motion vector of node points with $\frac{1}{4}$ pixel resolution. Frames 2, 4, 6 and 8 are chosen. They are warped from frame 1 after motion estimation. Table 1 gives the average PSNR value (dB) for Carphone, Akiyo and Mother-daughter sequences after motion estimation and warping. Compared with the results in [1], there is about 2–5 dB increase using the proposed model scheme. This shows that the scalable model can represent the motion of the object precisely.

Table 1: Average PSNR values of Carphone (CIF), Mother-daughter (QCIF) and Akiyo sequence (QCIF) for different representing levels

	Average PSNR value [dB]			
	Level 1	Level 2	Level 3	Level 4
Carphone sequence (CIF)	32.60	34.16	34.58	35.51
Mother-daughter (QCIF)	24.08	28.98	30.73	32.94
Akiyo (QCIF)	31.25	32.91	34.56	35.91

Conclusion: A new model design scheme has been proposed to achieve scalable 2-D model-based video coding. It has good performance in representing the motion of video objects. It also has less computational complexity than other published methods.

29 August 2002

Electronics Letters Online No: 20020980
DOI: 10.1049/el:20020980

M. Hu, S. Worrall, A.H. Sadka and A.M. Kondoz (*Centre for Communication Systems Research, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom*)

References

1 VAN BEEK, P., TEKALP, A.M., NING ZHUANG, CELASUM I., and MINGHUI, XIA: ‘Hierarchical 2-D mesh representation, tracking, and compression for object-based video’, *IEEE Trans. Circuits Syst. Video Technol.*, 1999, **9**, (2), pp. 353–369

2 CELASUM, I., and TEKALP, A.M.: ‘Optimal 2-D hierarchical content-based mesh design and update for object-based video’, *IEEE Trans. Circuits Syst. Video Technol.*, 2000, **10**, (7), pp. 1135–1153

3 FUKUNAGA, K.: ‘Introduction to statistical pattern recognition’ (Academic Press, INC., New York, 1990, 2nd edn.)

4 KIM, M., JEOR, J.G., KWAK, J.S., LEE, M.H., and AHN, C.: ‘Moving object segmentation in video sequences by user interaction and automatic object tracking’, *Image Vis. Comput.*, 2001, **19**, pp. 245–260

5 HORN, B.K.P., and SCHUNCK, B.G.: ‘Determining optical flow’, *Artif. Intell.*, 1981, **17**, pp. 185–203

6 WATERS, K., and TERZOPOULOS, D.: ‘Analysis and synthesis of facial image sequence using physical and anatomical models’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1993, **15**, (6), pp. 569–579

7 DE BERG, M., VAN KREVELD, M., OVERMARS, M., and SCHWARZKOPF, O.: ‘Computational geometry: algorithms and applications’ (Springer-Verlag, Berlin, 1997)

8 NAKAYA, Y., and HARASHIMA, H.: ‘Motion compensation based on spatial transformation’, *IEEE Trans. Circuits Syst. Video Technol.*, 1994, **4**, (3), pp. 39–356