3D Motion Estimation and Texturing of Human Head Model

Ján MIHALÍK, Viktor MICHALČIN

Lab. of Digital Image Processing and Videocommunications, Dept. of Electronics and Multimedia Telecommunications,
Park Komenského 13, 041 20 Košice, Slovak Republic

Jan.Mihalik@tuke.sk, viki07@pobox.sk

Abstract. *This paper deals with 3D motion estimation of the wire frame head model on the basis of the analysis of the parameters of 3D global motion of the real human head for each frame of videosequence. The proposed algorithm of 3D global motion estimation is given by solution of 6 linear equations for three extracted feature points of the real human head in each frame. Next there is presented an algorithm of texturing of 3D wire frame model of human head after its estimated global motion. Texturing is carried out by two dimensional affine transform directly in synthesized frames. Both proposed algorithms can achieve very low bit rate in model based image coding.*

Keywords

3D motion estimation, texturing, human head model, perspective projection, affine transform, model based image coding.

1. Introduction

The standard videocodecs H.261, H.263, MPEG-1, MPEG-2 [1] achieve data compression on the basis of the reduction of the intra and inter frame correlation of videosignals. The core of the standard videocodecs is the interframe hybrid coding system [2] with motion compensation that uses block matching motion estimation for interframe prediction and two dimensional discrete cosine transform for coding of the prediction error. The standard videcodecs employ the statistical properties of videosignals without of knowledge of the content of its frames therefore can be used for coding any visual scene.

If semantic information about the content of the frames is known, very effective coding of the videosignal by model based image coding [3],[4] is possible. The coding is based on modeling of videoobjects inside of a visual scene by using three dimensional (3D) models. For modeling of videoobjects general or specific 3D models [5] can by used. The general 3D models are mesh based and can model any forward unknown videoobjects in a visual scene. Afterwards the model based image coding is known as object based image coding [6]. On the other side, the specific 3D models are wire frame based, which are very often used in computer graphics. A specific 3D model is used for modeling only one videoobject like human head forward known in a visual scene. For more known videoobjects in a visual scene by the beginning partern recognition of a videoobject has to be done and afterwards its specific 3D wire frame model is applied on it. Model based image coding on the basis of the specific 3D wire frame models is known as knowledge based image coding [7]. The algorithms of model based image coding are used in the standard videocodecs MPEG-4 [1] except for the classical algorithms of the above given standard videocodecs.

The paper presents 3D motion estimation and texturing of the specific 3D wire frame model of human head for knowledge based image coding of videosignals of a visual scene with one videoobject of the real human head.

2. Basic Idea of Knowledge Based Image Coding

The block scheme of the knowledge based image coding and decoding system is in Fig. 1.

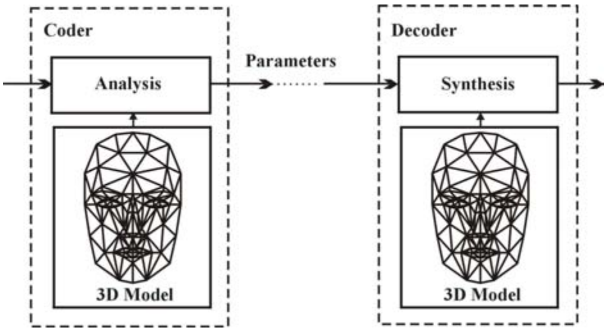


Fig. 1. Block scheme of the knowledge based image coding and decoding system.

The frames of input videosignal are analyzed to get the parameters of 3D wire frame model. For example, information about the shape, size and location of the human head in a visual scene belong among the parameters. Next the parameters of its global and local 3D motion but for real view of the 3D wire frame model very important information about the texture of real human head. Compared to the classical methods of image coding [8], [9] in this case only the parameters are coded and transmitted instead of all picture elements of the frames for the classical methods. The result is very low bit rate in output of the coder. On the basis of the received parameters and the same 3D wire frame model in the decoder, a synthesis of the human head is carried out. The parameters of 3D motion are coded and transmitted for each frame but the parameters of the shape, size and local position only one times. The texture of the human head may be coded by a classical method of image coding but again only one times.

For modeling of the human head we have used the 3D wire frame model Candide [10] in Fig. 2b), which contains 113 vertices and 184 triangles (polygons).

By the operation of calibration we can adopt the shape and size of the model Candide to the real human head. The calibration changes the coordinates of vertices of the model Candide according to the human head in the reference frame of videosequence. We have used a simple calibration of the model Candide by scaling factors k_h, k_v, k_r for its horizontal, vertical and depth sizes, respectively. The factors k_h or k_v are calculated by the ratio of horizontal or vertical sizes of the model Candide and the real human head in the reference frame as follows

$$k_h = \frac{|BC_s|}{|BC_m|}, \tag{1}$$

$$k_v = \frac{|AD_s|}{|AD_m|} \tag{2}$$

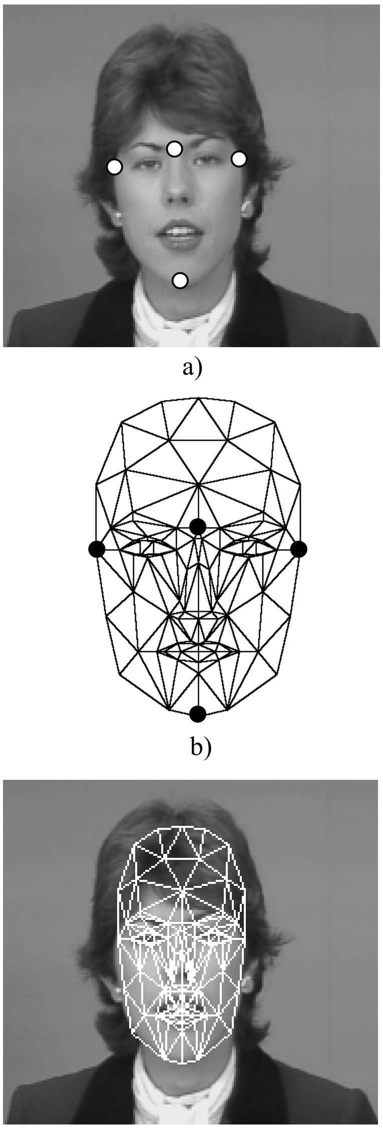
where $|BC_m|$, $|BC_s|$ are horizontal and $|AD_m|$, $|AD_s|$ - vertical sizes of the model Candide and the human head, respectively as it is seen in Fig.2. The scaling factor

$$k_r = \frac{k_h + k_v}{2} \tag{3}$$

is calculated by the average value of k_h and k_v, because from the reference frame the depth size of the human head can not be obtained. By multiplication of the coordinates of vertices of the model Candide by the corresponding scaling factors calibration of the shape and size is carried out. In Fig. 2, the calibrated model Candide is projected on the reference frame of videosequence Claire where it can be positioned and better adopted by hand manner.

3. 3D Motion Estimation

The parameters of 3D global motion of the human head [11] can be calculated by its extracted feature points [12] in the frames of videosequence. The perspective projection of the feature points from a frame on the corresponding vertices of 3D wire frame model in the space is shown in Fig. 3.


a)

b)

c)

Fig. 2. Calibration of the model Candide a) feature points of Claire in the reference frame, b) corresponding points of the model, c) calibrated model projected on the reference frame.

3D global motion tracking of the human head by its wire frame model in the model coordinate system (MCS) is given by the rotation matrix $\underline{\mathbf{R}}$ and translation vector $\overline{\mathbf{T}}$. Moving of a model vertex $\mathbf{M}=(h,v,r)^T$ in MCS from its initial position can by expressed as follows

$$\begin{pmatrix} h' \\ v' \\ r' \end{pmatrix} = \underline{\mathbf{R}} \begin{pmatrix} h \\ v \\ r \end{pmatrix} + \overline{\mathbf{T}} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} h \\ v \\ r \end{pmatrix} + \begin{pmatrix} t_h \\ t_v \\ t_r \end{pmatrix} \tag{4}$$

where $\mathbf{M}'=(h',v',r')^T$ is the rotated and translated (moved) vertex. Recalculation of the coordinates in MCS of the moved vertex $\mathbf{M}'$ on the ones in the camera coordinate system (CCS), as is seen in Fig. 3, can be done such a way

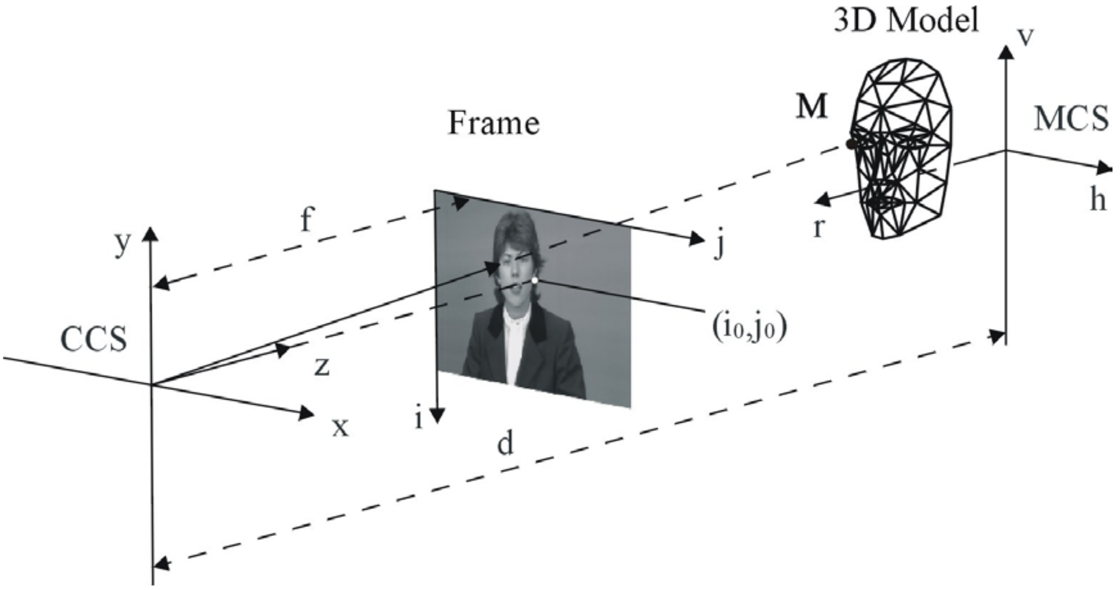


Fig. 3. Perspective projection of the human head on the 3D wire frame model

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} h' \\ v' \\ d - r' \end{pmatrix} \tag{5}$$

where d is the distance between origins of CCS and MCS. Afterwards the perspective projection of the moved vertex $\mathbf{M}'=(h',v',r')^T$ now represented in CCS gives a point (i',j') inside the frame which coordinates are given

$$i' = -f_y \frac{y'}{z'} + i_0 \tag{6}$$

$$j' = f_x \frac{x'}{z'} + j_0 \tag{7}$$

where f_x and f_y denote the focal length multiplied by scaling factors of the camera for the frame width and height, respectively and $(i_0\,,j_0)$ is a center of the frame in the direction of the optical axis z. If the rotation matrix $\underline{\mathbf{R}}$ and translation vector $\overline{\mathbf{T}}$ are known from eq. (4) to (7), we get

$$\frac{i'-i_0}{f_y} = -\frac{y'}{z'} = -\frac{\overline{\mathbf{R}}_2^T \mathbf{M} + t_v}{d - \left(\overline{\mathbf{R}}_3^T \mathbf{M} + t_r\right)} \tag{8}$$

$$\frac{j'-j_0}{f_x} = \frac{x'}{z'} = \frac{\overline{\mathbf{R}}_1^T \mathbf{M} + t_h}{d - \left(\overline{\mathbf{R}}_3^T \mathbf{M} + t_r\right)} \tag{9}$$

where $\overline{\mathbf{R}}_k$, k=1,2,3 are rows of the matrix $\underline{\mathbf{R}}$. By arrangement of eq. (8) and (9) we obtain a linear system of two equations whit 12 unknown parameters represented by vector $\overline{\mathbf{P}}$ of the global motion

$$\underline{\mathbf{D}}\overline{\mathbf{P}} = \overline{\mathbf{B}} \tag{10}$$

where

$$\underline{\mathbf{D}} = \begin{pmatrix} 0 & 0 & 0 & f_y h & f_y v & f_y r & -(i'-i_0)h \\ -f_x h & -f_x v & -f_x r & 0 & 0 & 0 & -(j'-j_0)h \end{pmatrix.$$
$$\begin{matrix} -(i'-i_0)v & -(i'-i_0)r & 0 & f_y & -(i'-i_0) \\ -(j'-j_0)v & -(j'-j_0)r & f_x & 0 & -(j'-j_0) \end{matrix}\Bigg) ,$$

$$\overline{\mathbf{P}} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{21} & r_{22} & r_{23} & r_{31} & r_{32} & r_{33} & t_h & t_v & t_r \end{pmatrix}^T ,$$

$$\overline{\mathbf{B}} = \begin{pmatrix} -d(i'-i_0) & -d(j'-j_0) \end{pmatrix}^T .$$

From the previous equations it follows out that for each corresponding pair of the moved point (i',j') in the frame and vertex (h,v,r) in initial position on 3D wire frame model we can compose a separate system (10). The number of unknowns in the system can be reduced from 12 to 6, because the freedom degree of ration matrix $\underline{\mathbf{R}}$ is only 3. Then the rotation matrix

$$\underline{\mathbf{R}} = \begin{pmatrix} \cos\Theta_r & -\sin\Theta_r & 0 \\ \sin\Theta_r & \cos\Theta_r & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\Theta_v & 0 & \sin\Theta_v \\ 0 & 1 & 0 \\ -\sin\Theta_v & 0 & \cos\Theta_v \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\Theta_h & -\sin\Theta_h \\ 0 & \sin\Theta_h & \cos\Theta_h \end{pmatrix} \tag{11}$$

where Θ_h, Θ_v, Θ_r are Euler's rotation angels around of axes h,v,r, respectively in clock wise direction. A solution of the nonlinear system is complex and needs a lot of operations. Assuming a small global motion of the human head between two successive frames, when for any small angle Θ<<1 the functions sinΘ≈0 and cosΘ≈1, the eq. (11) can be simplified to the next form

$$\begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} = \begin{pmatrix} 1 & -\Theta_r & \Theta_v \\ \Theta_r & 1 & -\Theta_h \\ -\Theta_v & \Theta_h & 1 \end{pmatrix} . \tag{12}$$

After the substitution of the rotation matrix elements from eq. (12) in the parameter vector $\overline{\mathbf{P}}$, eq. (10) can be written as follows

$$\begin{pmatrix} -(i-i_0)v - f_y r & (i-i_0)h & f_y h & 0 & f_y & -(i-i_0) \\ -(j-j_0)v & (j-j_0)h - f_x r & f_x v & -f_x & 0 & -(j-j_0) \end{pmatrix} \begin{pmatrix} \Theta_h \\ \Theta_v \\ \Theta_r \\ t_h \\ t_v \\ t_r \end{pmatrix} =$$

$$\ldots = \begin{pmatrix} -f_y v + (i-i_0)r - d(i-i_0) \\ f_x h + (j-j_0)r - d(j-j_0) \end{pmatrix}. \qquad (13)$$

For exact calculations of the parameters of 3D global motion for the wire frame model tracking of the human head in a frame we need coordinates only of three pairs of feature points in the frame and their corresponding vertices on the 3D wire frame model in initial positions. Afterward the 3D motion tracking of the human head by its wire frame model is given by rotation and translation of the model vertices according to the next equation

$$\begin{pmatrix} h' \\ v' \\ r' \end{pmatrix} = \underline{\mathbf{R}} \begin{pmatrix} h \\ v \\ r \end{pmatrix} + \overline{\mathbf{T}} = \begin{pmatrix} 1 & -\Theta_r & \Theta_v \\ \Theta_r & 1 & -\Theta_h \\ -\Theta_v & \Theta_h & 1 \end{pmatrix} \begin{pmatrix} h \\ v \\ r \end{pmatrix} + \begin{pmatrix} t_h \\ t_v \\ t_r \end{pmatrix} \qquad (14)$$

in dependence on the calculated 3D global motion parameters for every frames of videosequence.

In addition to the calibration of the model Candide we need to calibrate the camera to get its parameters f_x, f_y and d. They have to be calculated such as calibration parameters of the model Candide before 3D motion estimation. For the calibrated model Candide according to the human head in reference frame the 3D motion parameters $\Theta_h = \Theta_v = \Theta_r = 0$ and $t_h = t_v = t_r = 0$. After substitution of the zero values to eq. (13) we get

$$\begin{pmatrix} f_y v \\ f_x r \end{pmatrix} - \begin{pmatrix} (i'-i_0)r - (i'-i_0)d \\ -(j'-j_0)r + (j'-j_0)d \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \qquad (15)$$

Assuming the distance d between CCS and MCS to be much larger in comparison with the dimensions of frames, scaled focal lengths f_x and f_y of the camera can be calculated from eq. (15) for a feature point. More precise values of the camera parameters f_x and f_y can be achieved for several feature points of human head in the reference frame on the basis of the least square method of solution of eq. (15) by minimizing the mean square error

$$\left\| \begin{pmatrix} f_y v \\ f_x r \end{pmatrix} - \begin{pmatrix} (i'-i_0)r - (i'-i_0)d \\ -(j'-j_0)r + (j'-j_0)d \end{pmatrix} \right\|^2. \qquad (16)$$

4. Texturing of Human Head Model

The texture of human head gives to the 3D model the final appearance. Assuming constant luminance conditions, the texture can be represented by the reference frame of videosequence. Then texturing of the calibrated model Candide projected on synthesized frames is done after its 3D motion estimation. The coordinates of all vertices of the model Candide are changed according to estimated 3D motion parameters for each frame of the input videosequence Claire. Synchronously the coordinates of the perspective projected vertices on to the plane of synthesized frames are changed, too. The moving projected model Candide is textured polygon by polygon in synthesized frames. The relationship between vertices in the reference frame (RF) and the ones in synthesized frames (SF), as it is seen in Fig. 4, is given by two dimensional affine transform [13]

$$\mathbf{S} = \underline{\mathbf{A}}\mathbf{V} + \overline{\mathbf{T}} \qquad (17)$$

where $\underline{\mathbf{A}}$ is affine matrix, $\overline{\mathbf{T}} = (t_i, t_j)$ translation vector, $\mathbf{S}=(s_i,\ s_j)^T$ and $\mathbf{V}=(v_i,\ v_j)^T$ are vertices in RF and SF, respectively. After substitution we get

$$\begin{pmatrix} s_i \\ s_j \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} v_i \\ v_j \end{pmatrix} + \begin{pmatrix} t_i \\ t_j \end{pmatrix}. \qquad (18)$$

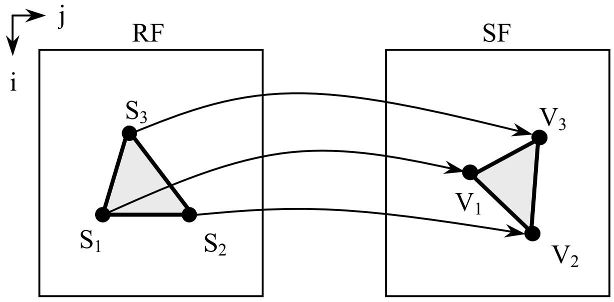


Fig. 4. Affine transform of the referred polygons in the reference and synthesized frames.

For the texturing of each polygon in SF by the texture values of the corresponding polygon in RF we need to calculate the affine matrix $\underline{\mathbf{A}}$ and the translation vector $\overline{\mathbf{T}}$ belonging to them. If they are known for all points inside the polygon of SF will be determined points inside the corresponding polygon in RF. Consequently the points of the polygon in SF can take the texture values of the ones inside the corresponding polygon in RF. The elements of the matrix $\underline{\mathbf{A}}$ and the components of the translation vector $\overline{\mathbf{T}}$ for each couple of the corresponding polygons can be immediately calculated from the known coordinates of vertices $\mathbf{V}_1=(v_{i1},\ v_{j1})$, $\mathbf{V}_2=(\ v_{i2},\ v_{j2})$, $\mathbf{V}_3=(\ v_{i3},\ v_{j3})$. The coordinates follow out from the perspective projection of the vertices of moved model Candide by its estimated 3D motion parameters for a given frame of videosequence Claire. After separate substitution of the coordinates of vertices $\mathbf{V}_1$, $\mathbf{V}_2$, $\mathbf{V}_3$ and their corresponding fixed vertices $\mathbf{S}_1=(s_{i1},\ s_{j1})$, $\mathbf{S}_2=(s_{i2},\ s_{j2})$, $\mathbf{S}_3=(s_{i3},\ s_{j3})$, in reference frame to eq. (18) and after further arrangement we get two systems of linear equations

$$\begin{pmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \end{pmatrix} = \begin{pmatrix} s_{i1} & s_{j1} & 1 \\ s_{i2} & s_{j2} & 1 \\ s_{i3} & s_{j3} & 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ t_i \end{pmatrix}, \tag{19}$$

$$\begin{pmatrix} v_{j1} \\ v_{j2} \\ v_{j3} \end{pmatrix} = \begin{pmatrix} s_{i1} & s_{j1} & 1 \\ s_{i2} & s_{j2} & 1 \\ s_{i3} & s_{j3} & 1 \end{pmatrix} \begin{pmatrix} a_{21} \\ a_{22} \\ t_j \end{pmatrix}. \tag{20}$$

Finally, the parameters a_{11}, a_{12}, a_{21}, a_{22}, t_i, t_j of the affine transform follow out from the solution of the systems (19) and (20) for a couple of corresponding polygons in reference and synthesized frames.

5. Experimental Results

Experimental results of 3D motion estimation and texturing of the human head model Candide have been obtained for the videosequence Claire of 166 frames of the size 288×352 pels. From four extracted feature points [14] for the purpose of 3D motion estimation only three ones, i.e. central points of eyes and mouth have been used. Their corresponding vertices on the wire frame model Candide had the same positions. As a reference frame for the calibration of the model Candide the first one of the video sequence Claire was taken. The calibration of the camera (obtaining its scaled focal lengths f_x and f_y) was done by eq. (15) for zero values of 3D motion parameters valid for the reference frame. Assuming the distance d=10000 pels between CCS and MCS, we calculated from eq. (16) f_x=8757 and f_y=10722 pels using the three extracted feature points in the reference frame on the basis of the least square approach. After the calibration of the model Candide and the camera by the reference frame, 3D motion parameters Θ_h, Θ_v, Θ_r, t_h, t_v, t_r for next frames of the videosequence Claire were exactly calculated from eq. (13). In Fig. 5 there is a tracking of the human head in selected frames of videosequence Claire by the wire frame model Candide. The tracking was done by moving of the model Candide on the basis of 3D motion estimated parameters and its projection by eq. (6) and (7) on the selected frames.

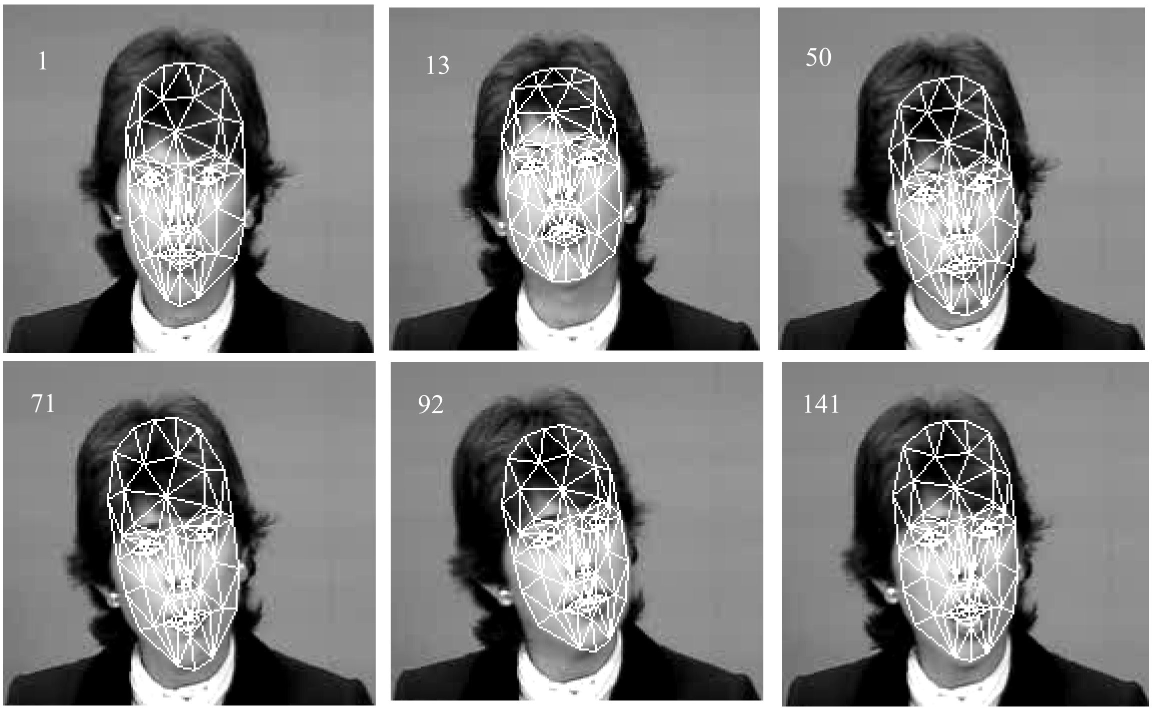


Fig. 5. Tracking of the human head Claire by estimated 3D motion of the model Candide projected on the selected frames.

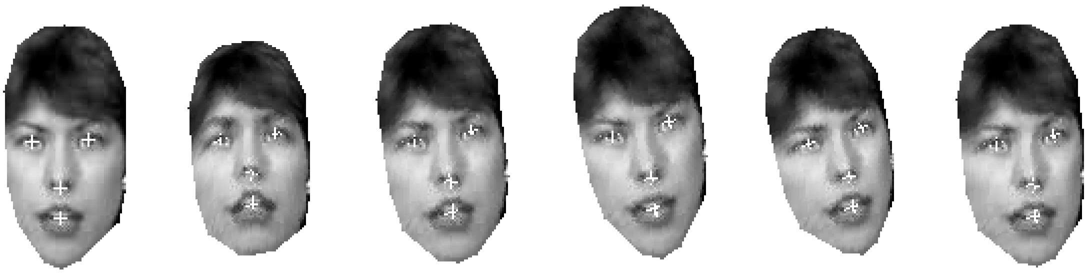


Fig. 6. Textured model Candide from the selected synthesized frames number 1, 13, 50, 71, 92, 141.

For texturing of the human head model Candide the texture of reference frame of the videosequence Claire has been used. After 3D motion estimation and moving of the model Candide in dependency on motion of the head in the videosequence Claire the model is immediately projected on the synthesized frames. Finally by using affine transform the texture from the reference frame was translated polygon by polygon to the corresponding ones inside of synthesized frames. In Fig. 6 there is the textured model Candide from the selected synthesized frames.

6. Conclusion

In this paper we presented the 3D motion estimation and texturing of the human head model Candide for model based image coding with very low bit rate. The proposed algorithm of the 3D global motion estimation used only three extracted feature points of the human head in the videosequence Claire. Its complexity is given by solution of 6 linear equations for each frame of the videosequence. The solution is independent on the size of frames and the kind of 3D wire frame model. The main advantages of the proposed algorithm of 3D global motion estimation are simplicity, low calculation requirements, and the possibility of using any kind of the 3D wire frame model.

The proposed algorithm of texturing of the human head model Candide is based on the affine transform. By the transform the texture of reference frame is directly translated polygon by polygon to the corresponding ones of the projected Candide, after its 3D moving, inside of the synthesized frames. The texture of reference frame has to be transmitted by the beginning of the video transmission and using a classical method of its coding. Afterward only 6 estimated parameters of 3D global motion are transmitted for each frame of the videosequence Claire. Texturing is carried out on the receiver side in the decoder on the basis of the received texture of reference frame and estimated 3D motion parameters.

The experimental results of 3D motion estimation and texturing of human head model Candide show acceptable quality of the synthesized videosequence Claire at very low bit rate. Further increasing of the quality can be achieved by animation and using more complex wire frame model of the human head.

Acknowledgement

The work was supported by the Scientific Grant Agency of the Ministry of Education and the Academy of Science of the Slovak republic under Grant No. 1/0384/03.

References

[1] MIHALÍK, J. *Image Coding in Videocommunications*. Mercury-Smékal, ISBN 80-89061-47-8, Košice, 2001.(In Slovak)

[2] MIHALÍK, J. Adaptive Hybrid Coding of Images. *Journal of Electrical Engineering*, 1993, vol. 44, No.3, p.85-89.

[3] FORCHHEIMER, R., KROMANDER, T. Image Coding - from Waveforms to Animation. *IEEE Trans. Acoust., Speech and Signal Proc.* 1989, vol.ASSP-37, no.12, p.2008-2023.

[4] AIZAWA, K., HUANG, T. S. Model-Based Image Coding: Advanced Video Coding Techniques for Very Low Bit-Rate Applications. *Proc. IEEE*, 1995, vol.83, no.2, p.259-271.

[5] PEARSON, D. E. Development in Model-Based Video Coding. *Proc. IEEE*, 1995, vol.83, no.6, p.892-906.

[6] MUSMANN, H. G., HÄTTER, M., OSTERMAN, J. Object-Oriented Analysis-Synthesis Coding of Moving Images. *Signal Processing: Image Communication*, 1989, vol.1, no. 2, p. 117-139.

[7] WELSH, W. J. Model-Based Video Coding of Videophone Images. *Electronics & Commun. Engineering J.,* 1991, p.29-36.

[8] MIHALÍK, J. Adaptive Transform Coding of Image. *Electronic Horizon,* 1991, vol. 52, no.11-12, p.253-257.

[9] MIHALÍK, J., GLADIŠOVÁ, I., MICHALČIN, V. Two Layer Vector Quantization of Images. *Radioengineering*, 2001, vol.10, no.2, p.15-19.

[10] RYDFALK, M.: CANDIDE: A Parameterised Face. *Dep. Elec. Eng. Rep*. LiTH-ISY-I-0866, Linköping Univ., 1987.

[11] TSAI, C. J., EISERT, P., GIROD, B., KATSAGGELOS, A. K. Model-Based Synthetic View Generation from a Monocular Video Sequence. In *Int. Conf. on Image Proc*. Santa Barbara, 1997, vol.1, p.444-447.

[12] ZHANG, L. Estimation of Eye and Mouth Corner Point Position in a Knowledge-Based Coding System. *Proc. SPIE*, 1996, vol.2952, p.21-28.

[13] FOLEY, J. D, VAN DAM, A., FEINER, S. K, HUGHES, J. F. *Computer Graphics, Principles and Practicles*. Addison-Wesley, 2nd edition, 1990.

[14] ANTOSZCZYSZYN, P. M., HANNAH, J. M., GRANT, P. M. A New Approach to Wire-Frame Tracking for Semantic Model-Based Coding Moving Image Coding. *Signal Processing: Image Communication*, 2000, vol.15, p.567-580.

About Authors…

Ján MIHALÍK graduated from the Technical University in Bratislava in 1976. Since 1979 he has joined the Faculty of Electrical Engineering and Informatics of the Technical University of Košice, where he received his PhD degree in Radio electronics in 1985. Currently, he is Full Professor of Electronics and Telecommunications and the head of the Laboratory of Digital Image Processing and Videocommunication at the Department of Electronics and Multimedia Telecommunications. His research interests include information theory, image and video coding, digital image and video processing and multimedia videocommunication.

Viktor MICHALČIN was born on 1976 in Ukraine. He received the Ing. Degree from the Technical University of Košice in 2000. Currently he is PhD student at the Department of Electronics and Multimedia Telecomunications of the Technical University, Košice. His research interests are vector quantization and model based image coding.