Journal of ELECTRICAL ENGINEERING, VOL. 59, NO. 5, 2008, 266–271

HUMAN FACE AND FACIAL FEATURE TRACKING BY USING GEOMETRIC AND TEXTURE MODELS

Ján Mihalík — Miroslav Kasár *

This paper deals with the tracking system of the human face and facial feature based on the human face texture model combined with an algorithm for adaptation of the wireframe 3D model Candide-3 to the human face images. The human face texture model is represented by a set of eigenfaces which are obtained by means of the principal component analysis of the training set of completely preprocessed textures. The algorithm for adaptation of 3D model needs a reasonable starting approximation and an update matrix calculated from the training set by manual deviation of 3D model for single components of the parameter vector. The designed tracking system was tested on a real videosequence with various conditions and for adaptation of 3D model both global motion parameters and animation parameters of the mouth were used.

Keywords: tracking, 3D model, Candide-3, human head, global motion parameters, animation parameters, update matrix, estimation

1 INTRODUCTION

The human face and facial feature tracking is one of the basic tasks in model-based video coding [1, 2]. Model-based video coding gives a high efficiency for coding of videosequences at very low bit rates. At the present time a number of tracking systems exist which are divided into two basic classes: attribute-based (or feature-based) [3] and template- based (or example-based) [4] systems. The presented tracking system is ranked among the template-based system, which is robust but computationally demanding.

The designed tracking system is based on the human face texture model (HFTM) combined with an algorithm for adaptation of the wireframe 3D model Candide-3 to the human face images. HFTM is obtained by applying the principal component analysis (PCA) to the training set of completely preprocessed textures which are geometrically and energically normalized. The adaptation algorithm is based on an automatic updating of the parameter vector by the residual image and the update matrix calculated from the gradient matrix.

The wireframe 3D model Candide-3 [8] provides several animation units (AU) to control not only the global but also local motion of the features (mouth, eyes, nose) of the human face. In this paper the global motion controlled by 6 global motion parameters (GMP) and the local motion of the mouth by 4 animation parameters (AP) of corresponding AU are shown. In chapter 2, obtaining of a completely preprocessed texture and construction of HFTM are described. The adaptation algorithm of 3D model to the human face in the frames of the input videosequence is presented in chapter 3 and in chapter 4 calculation of the update matrix is described. Finally, in chapter 5 tracking experiments are presented.

2 HUMAN FACE TEXTURE MODEL

Using HFTM [5] it is possible to represent the human faces inside or outside of the training set and for its construction the human face images with the adapted 3D model Candide-3 (Fig. 1b) are required. Afterwards, geometrical and energical normalization of the human face

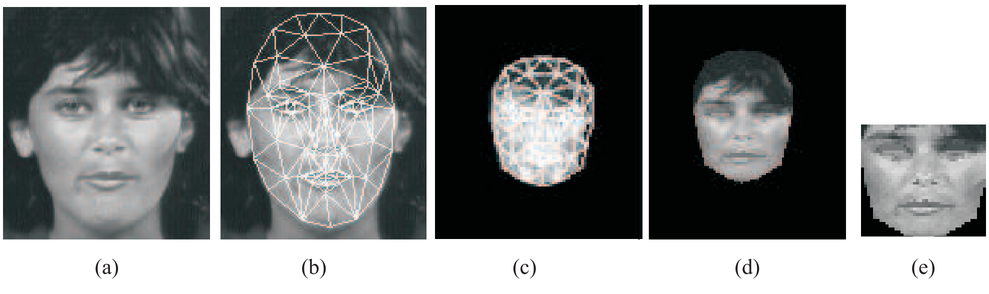
(a) (b) (c) (d) (e)

Fig. 1. a) Original image of human face, b) adapted 3D model, c) standard shape with texture, d) geometrically normalized texture, e) relevant part of the geometrically normalized texture

[*] Laboratory of Digital Image Processing and Videocommunications, Dept. of Electronics and Multimedia Telecommunications, FEI TU Košice, Park Komenského 13, 041 20 Košice, Slovakia; jan.mihalik@tuke.sk, miroslav.kasar@tuke.sk

ISSN 1335-3632 © 2008 FEI STU

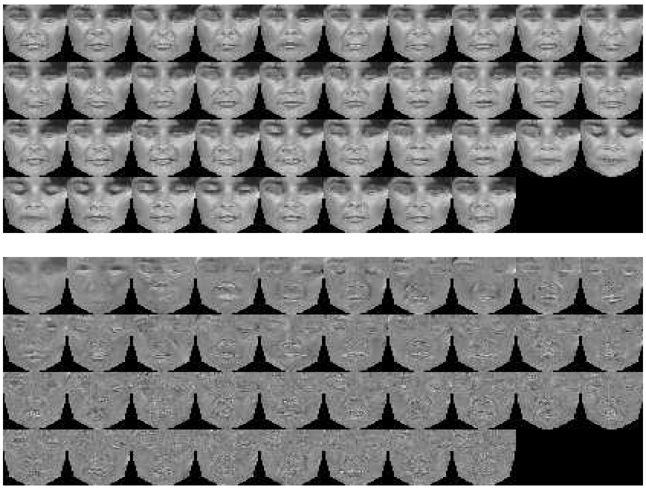

Fig. 2. (top) Set of 38 completely preprocessed textures, (bottom) HFTM with 38 eigenfaces

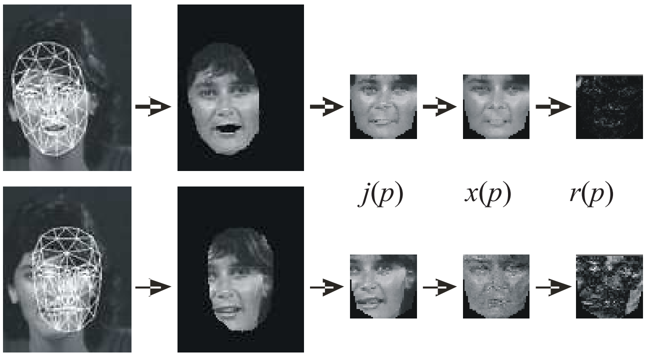

Fig. 3. (from left) Input image with well (top) and wrong (bottom) adapted 3D model, the image mapped onto 3D model, texture $j(p)$ from the frame, synthesized texture $x(p)$, residual image $r(p)$

images are carried out to remove the texture variations caused by their global and local motion and the geometrical differences between individuals. Geometrical normalization is a nonlinear transformation [6] which warps the source human face image with the adapted 3D model Candide-3 to the standard shape of 3D model with a given scale (Fig. 1c,d). After obtaining the geometrically normalized texture it is important to select its relevant part which contains the most important features of the human face like eyes, lip, nose, etc. Figure 1e shows the relevant part of the normalized texture of the human face.

Before applying PCA to the relevant parts of geometrically normalized textures it is necessary to minimize the effect of global lighting variation by their centring and energical normalization. In such a way, completely preprocessed textures are obtained and next, by applying PCA [7] to the training set of them, the HFTM is designed.

For illustration, Fig. 2 shows the completely preprocessed textures from the training set of 38 human faces. The training set was obtained by decimation with a factor of 4 of the frames of the videosequence “Miss America” In the same figure HFTM with 38 eigenfaces calculated by the above procedure is shown.

3 ADAPTATION OF 3D MODEL TO HUMAN FACE

The designed tracking system of the human face and of its features is based on HFTM combined with an algorithm for adaptation of the wireframe 3D model Candide-3 [8] to the human face in the frames of the input videosequence. The geometry and motion of 3D model are parametrized by the parameter vector

$$p = [g, z, w]^\top = [\Theta_h, \Theta_v, \Theta_r, s, t_h, t_v, z, w]^\top\,, \tag{1}$$

where g is the vector of the global motion parameters, z and w are the vectors of the shape and animation parameters, respectively.

Adaptation of 3D model to the human face assumes that its reasonable starting approximation is known. Then an optimal adaptation of 3D model in the frames with the human face means to find the parameter vector p that minimizes the distance measure. As an initial vector for the actual frame the parameter vector p is used that adapts 3D model in the previous frame of the videosequence assuming that the motion of the human face from a frame to the successive one is small enough. The summed squared error (SSE) between the synthesized texture and the texture of the human face from the frame is chosen as a distance measure. The texture from the frame is acquired in the same way as at the construction of HFTM [5]. It will be denoted as $j(p)$ because it responds to the parameter vector p. The synthesized texture will be denoted as $x(p)$ and it is acquired from the texture $j(p)$ by HFTM. The residual image is given as

$$r(p) = j(p) - x(p) \tag{2}$$

and SSE

$$e(p) = \|r(p)\|^2, \tag{3}$$

which means that it is the square value of the Euclidean norm of $r(p)$. Then corresponding peak signal/noise ratio (PSNR) is defined as follows

$$PSNR = 10 \log \frac{255^2}{e(p)/n}\,, \tag{4}$$

where n is the number of pels of the residual image $r(p)$.

Figure 3 shows the textures $j(p)$, $x(p)$ and the residual images $r(p)$ for the human faces images (frames) with the well and wrong adapted 3D model. There is visible from the residual image $r(p)$ that textures $j(p)$ and $x(p)$ are more different for the wrong adapted 3D model, when SSE is bigger in this case.

The goal is to find the parameter vector p that minimizes $r(p)$ and $e(p)$. After calculating $r(p)$ and $e(p)$ for the given p, the update vector Δp is found by multiplying the residual image with an update matrix A

$$\Delta p = A r(p)\,. \tag{5}$$

Table 1. The selected step h and the minimum and maximum values of j^{th} column of R for the single parameters

	Θ_h	Θ_v	Θ_r	s	t_h	t_v	AP1	AP2	AP3	AP4
h	0.0026	0.0026	0.0174	2.5	0.5	0.5	0.05	0.02	0.05	0.05
Min R_j	−61.8	−45.3	−8.7	−0.05	−0.3	−0.3	−0.8	−2.7	−1.6	−0.9
Max R_j	91.7	56.6	13.6	0.05	0.2	0.5	1.9	3.8	1.6	1.3

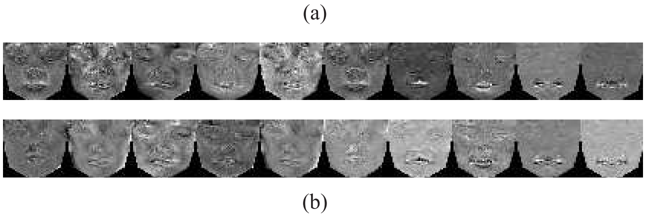

Fig. 4. a) The gradient matrix R, b) the transposed update matrix A, after back converting of their columns into 2D images of the size 40×42 pels

The update vector Δp is a vector of little changes that gives a probable direction in the search space and the new SSE is calculated as follows

$$e' = e(p + \Delta p)\,. \tag{6}$$

If $e' < e$, the parameter vector p is updated accordingly

$$p + \Delta p \to p \tag{7}$$

and the algorithm for adaptation of 3D model is based on the iteration procedure of (5) to (7) until the convergence, when $e' > e$. The magic of this is the update matrix A [9] that is created in advance by the training human face images with 3D model correctly adapted.

4 ESTIMATION OF DATA GRADIENT MATRIX FROM TRAINING DATA

Assuming that $r(p)$ is linear in p, that is

$$\frac{\partial r(p)}{\partial p} = R\,, \tag{8}$$

where R is the constant gradient matrix, it can be written

$$r(p + \Delta p) = r(p) + R\Delta p\,. \tag{9}$$

Given a p, the optimal Δp is found, that minimizes

$$e(p + \Delta p) = \|r(p) + R\Delta p\|^2. \tag{10}$$

The least square solution [10] is

$$\Delta p = -\big(R^\top R\big)^{-1} R^\top r(p)\,, \tag{11}$$

which gives the update matrix A as the negative pseudoinverse of the gradient matrix R, *ie*

$$A = -R^* = -\big(R^\top R\big)^{-1} R^\top. \tag{12}$$

Estimation of the gradient matrix R is obtained in such a way that single components of the parameter vector p are manually perturbed for the adapted 3D model in the human faces images of the training set. The j^{th} column of R denoted as

$$R_j = \frac{\partial r(p)}{\partial p_j} \tag{13}$$

can be estimated using differences

$$R_j \approx \frac{r(p + hq_j) - r(p - hq_j)}{2h}\,, \tag{14}$$

where h is a suitable step and q_j is the vector with all components equal zero expect for j^{th} component that equals one. By averaging over all N human face images with the adapted 3D model from the training set and for the suitable number K of the step multipliers, the final estimate of R_j is

$$R_j \approx \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{r(p_n + khq_j) - r(p_n - khq_j)}{2kh}\,. \tag{15}$$

Thereby R_j is estimated for all components of the vector p and together create the gradient matrix R from which the update matrix A is calculated by eq. (12).

From the training set of 38 human face images their completely preprocessed textures ($N = 38$) of the size 40×42 pels were obtained (Fig. 2). These were converted into 1D column vectors of the size 1441×1 (without black fields). Next 6 global motion parameters $(\Theta_h, \Theta_v, \Theta_r, s, t_h, t_v)$ [11] and 4 animation parameters of the mouth (AP1 — upper lip raiser, AP2 — jaw drop, AP3 — lip stretcher, AP4 — lip corner depressor) [12] were used. For each human face image with the adapted 3D model in the training set every parameter has been perturbed, in the range $< -Kh, Kh >$, where $K = 10$ and h is a suitable step experimentally selected for them as can be seen in Tab. 1. For all perturbed parameters the corresponding completely preprocessed textures have been obtained and consequently they were approximated by HFTM. Afterwards $20 \times 10 \times 38 = 7600$ residual images have been calculated and used for estimation of the gradient matrix R according to eq. (15). Finally from R by using eq. (12) the update matrix A was calculated.

When the parameter vector contains all 10 parameters (GMP and AP), the gradient matrix R has the size 1441×10 and the update matrix A 10×1441. The minimum and maximum values of j^{th} column of the gradient

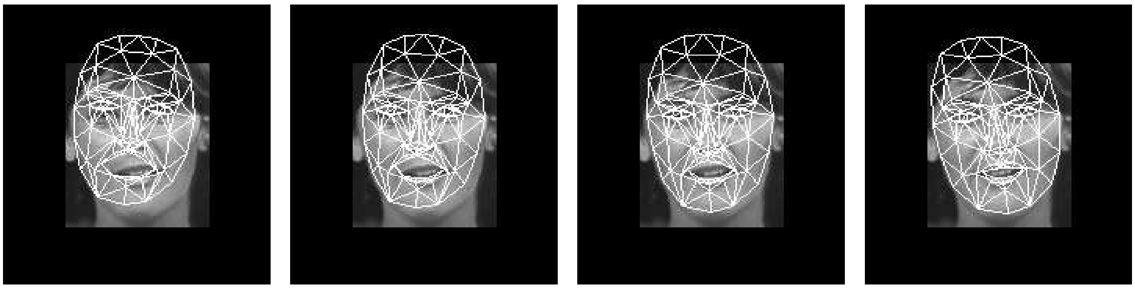

Fig. 5. Results of the adapted 3D model for the perturbed parameters GMP + AP and iterations 1, 2, 3, 8

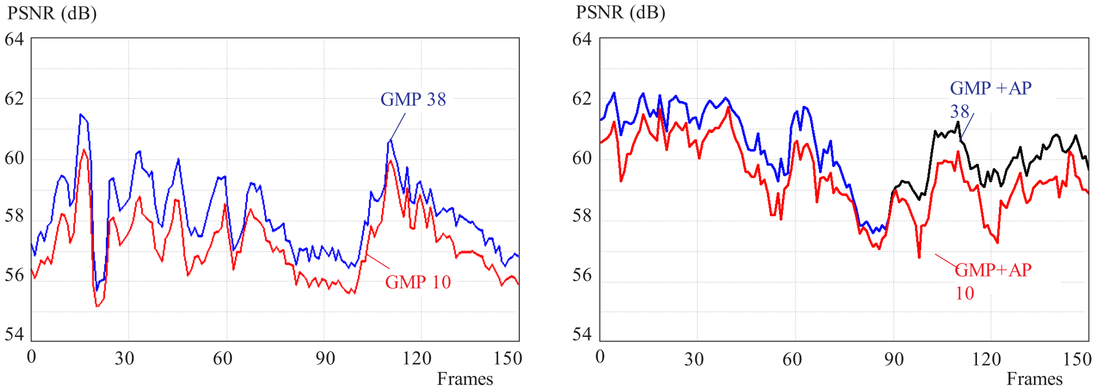

Fig. 6. PSNR of the final texture for all 4 cases

Table 2. Number of iterations, PSNR and SSE in dependence on the perturbed parameters

Perturbed parameters	Iterations	PSNR	SSE
GMP	8	70.36	0.764
AP	7	80.45	0.002
GMP + AP	8	70.66	7.17

Table 3. PSNR and the minimum and maximum values of the residual image $r(p)$ for the perturbed parameters GMP + AP and iterations 1, 3, 8

Iteration	PSNR	$\min r(p)$	$\max r(p)$
1	52.21	-2.314	2.4049
3	57.86	-2.181	1.5595
8	70.66	-0.6186	0.6813

matrix R can be seen in Tab. 1. Then the update matrix A has the minimum value -1.6158 and the maximum one 1.1233.

Figure 4 shows the back converted columns of the gradient matrix R and of the transposed update matrix A into 2D images of the size 40×42 pels. In the case, when the parameter vector contains only 6 global motion parameters, the gradient matrix R will have the size 1441×6 and the update matrix 6×1441.

5 TRACKING EXPERIMENTS

The tracking experiments were carried out, if the designed tracking system was applied to an image from the training set with a good adapted 3D model. Then 3D model was perturbed manually from the optimal position and consequently it was adapted using HFTM with 38 eigenfaces and the described adaptation algorithm (chapter 3). In the first case only global motion parameters were perturbed, in the second one only animation parameters of the mouth, and in the third one both global motion and animation parameters were perturbed. Table 2 shows the number of iterations which are required for repeated adaptation of 3D model, PSNR and SSE of the texture calculated for the finally adapted 3D model compared to that one for its optimal position. In Tab. 3 there are PSNR, the minimum and maximum values of the residual image $r(p)$ computed according to eq. (2) for the third case. Figure 5 shows the results of adaptation of 3D model for the same case and some iterations.

As it can be seen from Fig. 5, deviation of 3D model was big enough and therefore the number of iterations for its final adaptation is big, too. On the other side in the real videosequences the distinctions between successive frames are far less and therefore the number of iterations reduces (mostly less than 4). The final precision of the adaptation was satisfactory in all cases.

Next the designed tracking system was applied to the videosequence “Miss America” with the frame rate

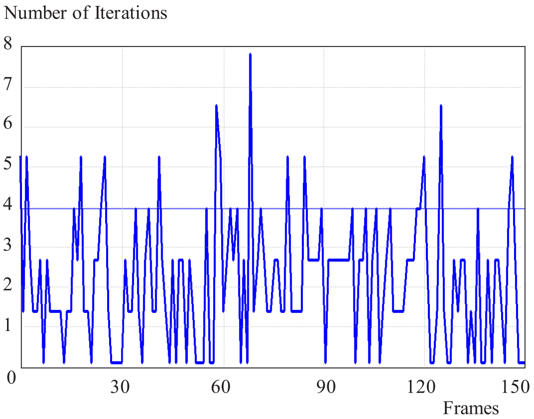

Fig. 7. The number of iterations for the second case (38 eigenfaces, GMP + AP)

30 Hz and 150 frames of the size 288 × 352 pels. Afterwards, totally 4 cases of tracking with various conditions were realized. In the first case HFTM with all 38 eigenfaces was used and the parameter vector p contained only the global motion parameters $(\Theta_h, \Theta_v, \Theta_r, s, t_h, t_v)$. 3D model Candide-3 was manually adapted in the first frame of the videosequence by changing the global motion parameters and the static shape parameters [13]. Then the parameter vector p containing GMP was iteratively optimized for each frame. In the second case HFTM with all 38 eigenfaces was used but the parameter vector p contained both the global motion parameters $(\Theta_h, \Theta_v, \Theta_r, s, t_h, t_v)$ and the animation parameters (AP1, AP2, AP3, AP4) of the mouth. 3D model was again manually adapted in the first frame of the videosequence by changing of the global motion and animation parameters of the mouth and the static shape parameters. The parameter vector p containing GMP and AP was then iteratively optimized for each frame. The third and fourth cases are similar to the previous ones except for that HFTM does not use all 38 but only 10 eigenfaces corresponding to the largest eigenvalues.

In Fig. 6 there are graphs of PSNR of the final textures from the single frames for all 4 cases. As can be seen, PSNR increases by addition AP to GMP, because the accuracy of adaptation increases. In the cases when only 10 eigenfaces are used, PSNR decreases opposite the cases with 38 eigenfaces. This is caused by an error which occurs at obtaining of the synthesized texture by using not all eigenfaces. Figure 7 shows the number of iterations which are necessary for adaptation of 3D model consequently in the following frames for the second case (38 eigenfaces, GMP + AP). Mostly less than 4 iterations are necessary but when the distinction between successive frames is bigger, the number of iterations increases and it is valid for all 4 cases. The achieved results were satisfactory and the resulting head model adaptation for the second case is shown in Fig. 8. From this figure it follows that the global adaptation (rotation, scale, translation) is very good and the animation parameters of the mouth affect the local adaptation very well.

At the construction of HFTM, 3D model Candide-3 is adapted to the human face image with semi-automatic method, where at the end some vertices of 3D model are manually adjusted and so the optimal position of 3D model is obtained. After manual perturbation of 3D model from the optimal position and the following adaptation by the designed tracking system closest to this optimal position, the relevant part of the geometrically normalized texture is similar to those ones, from which HFTM was obtained.

At the adaptation of 3D model Candide-3 to the human face from the videosequence, in the first frame 3D model was adapted automatically by the shape units without further manual correction [13] and therefore the adapted 3D model is not fully optimal. The adaptation of 3D model in next frames is obtained from the adapted 3D model in the previous frame by the animation units. Then the relevant parts of the geometrically normalized textures from the adapted 3D models are less similar to those ones, from which HFTM was obtained.

Therefore PSNR of the final textures for all 4 cases (Fig. 6) obtained at the adaptation of 3D model to the human face in the frames of the videosequence by the designed tracking system is smaller than PSNR of the final texture obtained at the perturbation of 3D model from the optimal position and consequently after its adaptation by the designed tracking system (Tab. 2). Even though the adaptation of 3D model to the human face in the frames of the videosequence is good enough and the designed tracking system is suitable for using in the model-based videocoding.

6 CONCLUSION

The designed tracking system worked correctly what is seen from the achieved results. The deviation of 3D model was set up bigger for the human face image from the training set compared to that one between successive frames of the videosequence. Even though adaptation of 3D model was satisfactory for the image, as can be seen from Tab. 2 (PSNR is big, SSE is small), the big deviation needs using more iterations (8 iterations).

If the tracking system was applied on the videosequence, the adaptation of 3D model is done consequently in its frames by using less number of iterations (Fig. 7). Almost never all 38 eigenfaces are necessary, 10 eigenfaces are enough in HFTM (PSNR decreases about 1 dB). The tracking with parameter vector p, containing both GMP and AP of the mouth, is more accurately than the tracking without AP of the mouth (Fig. 6) what proves the significance of the animation units of 3D model. The number of iterations was similar for all cases and seldom was bigger than 4.

The designed tracking system based on the HFTM combined with the algorithm for adaptation of the wireframe 3D model Candide-3 can track the human face and facial features in the head-and-shoulder videosequences. In standard videocodec MPEG-4-SNHC [14] the designed

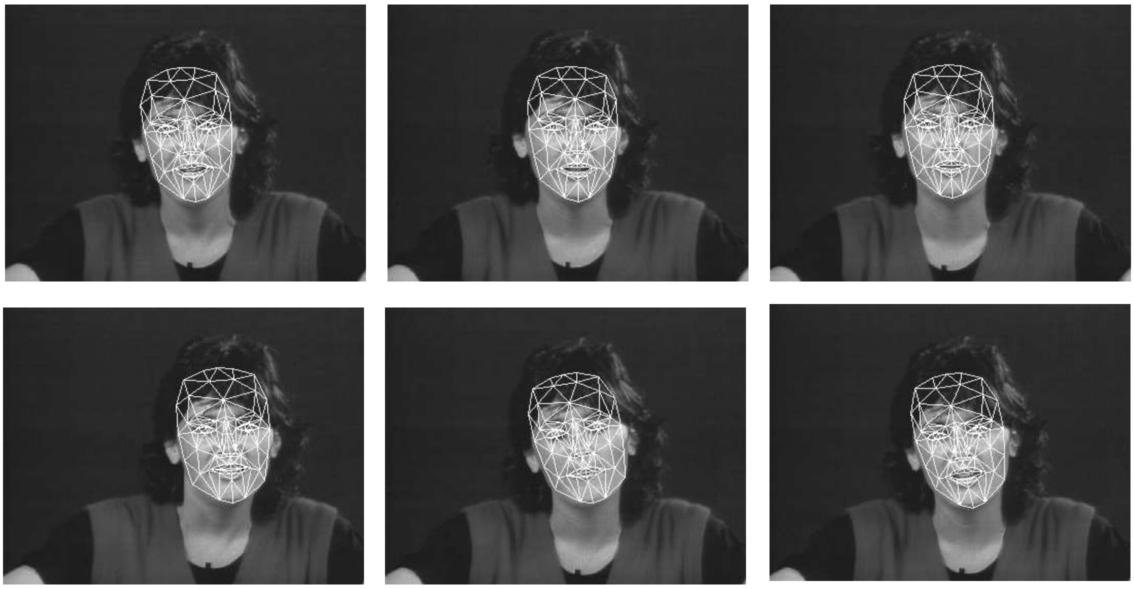

Fig. 8. The tracking results for the second case (38 eigenfaces, GMP + AP), (every 30^{th} frame is shown)

system can be an important part for the model-based video coding of the videosequence with very low bit rate.

Acknowledgements

The work was supported by the Scientific Grant Agency of Ministry of Education and Academy of Science of Slovak republic under Grant No. 1/3133/06.

References

[1] The Special Issue of the IEEE Trans.on Circuits and Systems for Video Technology on MPEG-4 SNHC, July 2004.

[2] STRINTZIS, M.—SARRIS, N. : 3D Modeling and Animation: Synthesis and Analysis Techniques for the Human Body, IRM Press, Hershey, PA, July 2004.

[3] HESS, M.—MARTINEZ, G. : Facial Feature Extraction Based on the Smallest Univalue Segment Assimilating Nucleus (SUSAN) Algorithm, Picture Coding Symposium, USA, 2004, pp. 232–238.

[4] DORNAIKA, F.—AHLBERG, J. : Fast and Reliable Active Appearance Model Search for 3D Face Tracking, Proceedings of Model-based Imaging, Rendering, Image Analysis, and Graphical Special Effects (Mirage), 2003, pp. 113–122.

[5] MIHALÍK, J.—KASÁR, M. : Basis of Eigenfaces for Tracking of Human Head, J. Electrical Engineering **58** No. 3 (2007), 134–139.

[6] GALLIER, J. : Curves and Surfaces in Geometrical Modeling – Theory and Algorithms, Morgan Kaufmann Publishers, 2000.

[7] KASÁR, M.—MIHALÍK, J.—ZAVACKÝ, J. : Design of Basis of Eigenfaces, Proc. Scientific Conference "Nové smery v spracovaní signálov VIII", Tatranské Zruby, 2006, pp. 182–186.

[8] AHLBERG, J. : Candide-3 – an Updated Parameterized Face. Report No. LiTH-ISY-R-2326, Dept. of EE, Linköping University, 2001.

[9] AHLBERG, J. : Fast Image Warping for Active Models, Tech. Report No. LiTH-ISY-R-2355, Dept. of EE, Linköping University, 2001.

[10] BJORCK, A. : Numerical Method for Least Squares Problems, SIAM, Philadelphia, 1996.

[11] MIHALÍK,J.—MICHALČIN, V. : 3D Motion Estimation of Human Head by Using Optical Flow, Radioengineering **15** No. 2 (2006), 37–44.

[12] MIHALÍK, J.—MICHALČIN, V. : Animation of 3D Model of Human Head, Radioengineering **16** No. 1 (2007), 58–66.

[13] MIHALÍK,J.;KASÁR, M. : Shaping of Geometry of 3D Human Head Model, Proc. 17^{th} International Conference "Radioelektronika 2007", Brno, Czech Republic, 2007, pp. 483-486.

[14] ZHONG, J. : Flexible Face Animation Using MPEG-4/SNHC Parameter Streams, Proc. International Conference on Image Processing, Chicago, 1998, pp. 924–928.

Received 25 June 2007

Ján Mihalík graduated from Technical University in Bratislava in 1976. Since 1979 he joined Faculty of Electrical Engineering and Informatics of Technical University of Košice, where received his PhD degree in Radioelectronics in 1985. Currently, he is Full Professor of Electronics and Telecommunications and the head of the Laboratory of Digital Image Processing and Videocommunications 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 videocommunications.

Miroslav Kasár was born in Hnúšťa, Slovakia, in 1980. He received the Ing (MSc) degree from the Technical University of Košice in 2003. At present he is a PhD student at the Department of Electronics and Multimedia Telecomunications of the Technical University of Košice. His research interest is video coding with a very low bit rate.