In Proceedings of ICPR, Hong Kong, August 2006

Performance driven facial animation using illumination independent appearance-based tracking

José M. Buenaposada
ESCET, Universidad Rey Juan Carlos
C Tulipán s/n, 28933 Móstoles, Spain
josemiguel.buenaposada@urjc.es

Enrique Muñoz, Luis Baumela
Facultad Informática, Universidad Politécnica Madrid,
Campus de montegancedo s/n, 28660 Madrid, Spain
{kike,lbaumela}@dia.fi.upm.es
http://www.dia.fi.upm.es/~pcr

Abstract

We introduce a procedure to estimate human face high level animation parameters from a marker-less image sequence in presence of strong illumination changes. We use an efficient appearance-based tracker to stabilise face images and estimate illumination variation. This is achieved by using an appearance model composed by two independent linear subspaces modelling face deformation and illumination changes respectively. The system is very simple to train and is able to re-animate a 3D face model in real-time.

1. Introduction

Realistically modelling and animating a human face is a challenging problem because of the complexity of human facial anatomy and our natural sensitivity to facial appearance. That is why current animation systems require extensive human intervention, being this one of their major drawbacks. One solution to this problem comes from the so-called *performance driven animation* approach, in which the performance of a human actor is used to control the animation of the graphical character. Now, the main difficulty is the development of motion capture techniques which can accurately describe the deformation of the actor's face. Here, computer vision has emerged as the most promising technology to achieve this goal.

Various approaches have been used for performance-driven facial animation using computer vision. Most simple ones use coloured markers painted on the face and lips of the actors, to simplify and aid the face tracker [12]. However, markings on the face are intrusive and often impractical. Feature-based procedures were the first tracking algorithms to be introduced. They are based on tracking a discrete set of texture elements such as eye or nose corners and the contours of expressive regions (eyes, eyelids or mouth) [2, 14]. They can only estimate the motion of textured regions and, therefore, they provide sparse information about the deformation of the face. Later, changes in identity, facial expressions and illumination were modelled by using linear subspace representations of shape+texture [5, 3]. The main drawback of shape+texture approaches is that they have complex training procedures which often require manual intervention [4, 1, 7]. The third approach is based on using linear subspace representations of facial appearance. These representations are gaining popularity, since there are various procedures for automatically learning linear [6, 9] and non-linear [10] subspace models and for probabilistically representing the dynamics of appearance variation [15, 8]. The major limitation for using appearance-based techniques for facial animation is the impossibility of separating some of the sources of appearance variation, for example facial deformation from illumination.

In this paper we present a face re-animation system based on an efficient appearance-based tracker, which can separate changes in appearance caused by the deformation of the face from variations in the illumination. In our system, a face is modelled by the addition of two linear subspaces, one modelling the deformation of the face (facial expressions) and the second one modelling the illumination. Both subspaces can be independently and automatically trained by processing two video sequences of the actor, one with fixed illumination and varying facial expressions and another with fixed expression and varying illumination. By using this model we will be able to track a human face and re-target the rigid and non-rigid motion of the face onto a graphical model in real-time. The main advantage of this system compared to previous approaches is the remarkable simplicity of the model, which can be easily trained, and the efficiency of the tracker, which can follow a deforming face at frequencies higher than video frame rate on an average personal computer.

2. Appearance-based tracking

In this section we introduce an appearance-based model representing the variations in the appearance of a face caused by changes in the facial expressions and the illuminations of the scene. Then we present the algorithm used for efficiently fitting the previous model to a target image.

2.1. The model

Let $I(\mathbf{x}, t)$ be the image acquired at time t, where $\mathbf{x}$ is a vector representing the co-ordinates of a point in the image, and let $\mathbf{I}(\mathbf{x}, t)$ be a vector storing the brightness values of $I(\mathbf{x}, t)$. The warping function $f(\mathbf{x}, \boldsymbol{\mu})$ models the rigid motion of the face, being $\boldsymbol{\mu}$ the vector of rigid motion parameters. Matrices $\mathtt{B}_d$ and $\mathtt{B}_i$ are linear subspace basis modelling respectively the modes of non-rigid deformation of the face and the changes in appearance caused by variations in the illumination. Our appearance-based model is represented by the brightness constancy equation [11]

$$\begin{aligned} \mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t) &= [\mathtt{B}_i \mathbf{c}_{i,t}](\mathbf{x}) + [\mathtt{B}_d \mathbf{c}_{d,t}](\mathbf{x}) \\ &= [\mathtt{B}\mathbf{c}_t](\mathbf{x}) \;\; \forall \mathbf{x} \in \mathcal{F}, \end{aligned} \tag{1}$$

where $\mathbf{c}_{d,t}$ and $\mathbf{c}_{i,t}$ are respectively the deformation and illumination appearance parameters, $\mathtt{B} = [\,\mathtt{B}_i | \,\mathtt{B}_d\,]$, $\mathbf{c}_t^\top = (\mathbf{c}_{i,t}^\top, \mathbf{c}_{d,t}^\top)^\top$, and $\mathcal{F}$ represents the set of pixels of the face used for tracking. By $[\mathtt{B}\mathbf{c}](\mathbf{x})$ we denote the value of $\mathtt{B}\mathbf{c}$ for the pixel with position $\mathbf{x}$. Intuitively (1) states that the rigidly rectified image $\mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t)$ can be expressed as a linear combination of the deformation subspace basis vectors, $\mathtt{B}_d$. The illumination subspace $\mathtt{B}_i \mathbf{c}_i$ corrects the deformation to take into account the illumination.

2.2. Efficient model alignment

Tracking a face consists of estimating, for each image in the sequence, the values of the motion, $\boldsymbol{\mu}$, and appearance, $\mathbf{c}$, parameters which minimise the error function

$$E(\boldsymbol{\mu}, \mathbf{c}) = ||\mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t) - [\mathtt{B}\mathbf{c}_t](\mathbf{x})||^2. \tag{2}$$

Minimising (2) can be a difficult task as it defines a non-convex cost function. We solve it using a Gauss-Newton minimisation approach. In order to make Gauss-Newton iterations, a Taylor series expansion of $\mathbf{I}$ at $(\boldsymbol{\mu}_t, \mathbf{c}_t, t)$ is performed, producing a new error function

$$E(\delta\boldsymbol{\mu}, \delta\mathbf{c}) = ||\mathtt{M}\delta\boldsymbol{\mu} + \mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t + \delta t) - \mathtt{B}(\mathbf{c}_t + \delta\mathbf{c})||^2, \tag{3}$$

where $\mathtt{M} = \left[\frac{\partial \mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}), t)}{\partial \boldsymbol{\mu}} \Big|_{\boldsymbol{\mu} = \boldsymbol{\mu}_t} \right]$ is the $N \times n$ ($n = \dim(\boldsymbol{\mu})$) Jacobian matrix of $\mathbf{I}$.

The minimum of (3) can be estimated by least-squares

$$\begin{bmatrix} \delta\boldsymbol{\mu} \\ \delta\mathbf{c} \end{bmatrix} = -(\mathtt{M}_J^\top \mathtt{M}_J)^{-1} \mathtt{M}_J \mathcal{E}, \tag{4}$$

where $\mathtt{M}_J = (\mathtt{M} | -\mathtt{B})$ and $\mathcal{E} = \mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t + \delta t) - \mathtt{B}\mathbf{c}_t$.

Solving (4) using the matrix inversion lemma, we get the solution for $\delta\boldsymbol{\mu}$ [11]

$$\delta\boldsymbol{\mu} = -(\Sigma^{-1} \Lambda_{M1} \Sigma)^{-1} \Sigma^\top \Lambda_{M2} \mathcal{E}$$

where $\Lambda_{M1} = \mathtt{M}_0^\top (\mathtt{I} - \mathtt{B}(\mathtt{B}^\top\mathtt{B})^{-1}\mathtt{B}^\top)\mathtt{M}_0$ and $\Lambda_{M2} = \mathtt{M}_0^\top (\mathtt{I} - \mathtt{B}(\mathtt{B}^\top\mathtt{B})^{-1}\mathtt{B}^\top)$ are constant matrices which can be precomputed off-line and $\mathtt{M} = \mathtt{M}_0\Sigma$. An efficient solution for $\delta\mathbf{c}$ can be obtained from (3) by least-squares, considering that $\delta\boldsymbol{\mu}$ is known

$$\delta\mathbf{c} = \Lambda_B[\mathtt{M}\delta\boldsymbol{\mu} + \mathcal{E}], \tag{5}$$

where $\Lambda_B = (\mathtt{B}^\top\mathtt{B})^{-1}\mathtt{B}^\top$ is also constant and can be precomputed off-line. This is the key for the efficiency of this algorithm.

The term $\mathtt{M}\delta\boldsymbol{\mu}$ represents the brightness variation in $\mathbf{I}$ due to a motion of magnitude $\delta\boldsymbol{\mu}$. Intuitively equation (5) states that the appearance parameters are computed by projecting onto the subspace $\mathtt{B}$ the rectified image corrected to take into account the incremental motion $\delta\boldsymbol{\mu}$ and the already known appearance $\mathtt{B}\mathbf{c}_t$.

3. Re-animation

With the tracker introduced in section 2 we can extract stabilised images of the face in each frame of the sequence and, what is more important, its deformation appearance parameters. In this section we show how to estimate the facial graphical animation parameters from the deformation appearance parameters provided by the tracker.

In order to estimate the animation parameters for a given face region we will use n_e sample images. Let $\mathtt{D}$ and $\mathtt{A}$ be respectively the $n_d \times n_e$ and $n_a \times n_e$ matrices obtained by storing column-wise the appearance deformation and graphical animation parameters of the sample sequence[1]. Then $\mathtt{E}$ is the $(n_d + n_a) \times n_e$ matrix

$$\mathtt{E} = \begin{bmatrix} \mathtt{D} \\ \mathtt{W}_A\mathtt{A} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_{d,1} & \cdots & \mathbf{c}_{d,n_e} \\ \mathtt{W}_A[\mathbf{a}_1 & \cdots & \mathbf{a}_{n_e}] \end{bmatrix},$$

where $\mathtt{W}_A$ is a diagonal matrix of weights to compensate the difference in scale between the graphical animation and the deformation appearance parameters. In our case $\mathtt{W}_A = r\mathtt{I}$, where r^2 is the ratio of the variances of deformation appearance parameters and the total variability in the animation parameters.

[1] We assume, that all samples, $\mathbf{c}_{d,j}$, and animation parameters, $\mathbf{a}_j$, are mean centred.

Using PCA on $\mathtt{E}$, we get $\mathtt{B}_l$, the subspace basis expanded by the l eigenvectors corresponding to the largest eigenvalues of the covariance matrix $(\mathtt{E}\mathtt{E}^\top)$, which can be written as

$$\mathtt{B}_l = \begin{bmatrix} \mathtt{B}_{d2} \\ \mathtt{B}_a \end{bmatrix}.$$

Note here that we are using three eigenspaces, two for tracking and another one for re-animation.

Now we can estimate $\mathbf{c}_l$ using the $(n_e + n_a) \times l$ matrix, $\mathtt{B}_l$, that represents the relation between the deformation appearance parameters in D and the animation parameters in $\mathtt{A}$. Once $\mathbf{c}_l$ is known, we can approximate each pair $(\mathbf{c}_d, \mathbf{a})$ by $(\mathbf{c}_d{}^*, \mathbf{a}^*)$ such that:

$$\begin{bmatrix} \mathbf{c}_d{}^* \\ \mathtt{W}_A\mathbf{a}^* \end{bmatrix} = \mathtt{B}_l\mathbf{c}_l, \quad \mathbf{c}_l = \mathtt{B}_l^\top \begin{bmatrix} \mathbf{c}_d \\ \mathtt{W}_A\mathbf{a} \end{bmatrix}.$$

Given $\mathbf{c}_d$, $\mathtt{B}_{d2}$ and $\mathtt{B}_a$, the re-animation problem is to estimate the corresponding animation parameters, $\mathbf{a}^*$. From the structure of $\mathtt{B}_l$ we can write $\mathtt{B}_{d2}\mathbf{c}_l = \mathbf{c}_d$, where $\mathbf{c}_l$ is the only unknown. So, the solution for $\mathbf{c}_l$ will be given by

$$\mathbf{c}_l^* = \arg\min_{\mathbf{c}_l} ||\mathtt{B}_{d2}\mathbf{c}_l - \mathbf{c}_d||^2 = \text{pinv}(\mathtt{B}_{d2})\mathbf{c}_d,$$

where the $l \times n_d$ matrix $\text{pinv}(\mathtt{B}_{d2})$, is the pseudo-inverse of $\mathtt{B}_{d2}$. Then, the graphical animation parameters of $\mathbf{c}_d$ are given by $\mathbf{a}^* = \mathtt{W}_A^{-1}\mathtt{B}_a\text{pinv}(\mathtt{B}_{d2})\mathbf{c}_d = \mathtt{R}_d^a\,\mathbf{c}_d$, where the $n_a \times n_d$ matrix $\mathtt{R}_d^a$ is constant and can be precomputed off-line.

4. Experiments

In this section we will first describe how to train the model introduced in section 2.1. Then, using synthetic sequences, we will evaluate the quality of the estimation of the graphical animation parameters and the separation of appearance deformation and illumination parameters. Finally, we will show some results of a real experiment in order to qualitatively validate of our approach with a live video sequence.

4.1. Model training

One of the advantages of the appearance model introduced in section 2.1 is that deformation and illumination subspaces are decoupled, and so, they can be independently trained. Each subspace is trained with one video sequence. For the illumination subspace we use a sequence in which a light orbits in front of the target face with a neutral expression. For the deformation subspace we use a sequence captured with a non-saturating frontal illumination in which the target face performs different facial expressions. The face is located and aligned in the first frame of both sequences, then, with a procedure similar to the one described in [9], both sequences are independently tracked and both linear subspace models independently built.

4.2. Synthetic experiments

In order to have ground truth animation parameters we generate synthetic image sequences. Using a modified version of Parke and Waters' face model [13] we have rendered two test sequences (1100 frames each) of a face performing the same expressions. In the first sequence the illumination is constant and is produced by a distant light in front of the virtual head. In the second, illumination changes are introduced by moving the light in a plane parallel to the head. In the first and second rows of Fig. 1 are respectively displayed some images from these sequences.

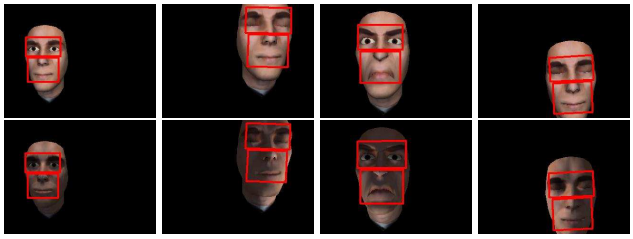

Figure 1. Some images from the synthetic sequences.

We reanimate the synthetic sequences with two trackers, one with the model introduced in section 2.1 and another without $\mathtt{B}_i$, the subspace modelling changes in illumination. In Fig. 2 are shown the results of these experiments for the mouth (left) and eyes (right) graphical animation parameters. In the horizontal axis we represent the graphical animation parameter number (e.g. for the mouth 1:jawopeness, 2:lipscontraction, etc.; for the eye 1:lefteyelid, 2:lefteyevertrotation, 3:lefteyehorizrotation, ..., 8:righteyelid, 9:righteyevertrotation,...) and in the vertical the ratio between average rms estimation error and maximum parameter range. With labels `SDI_TDI` and `SD_TDI` are plotted the results obtained with the complete tracker for the sequences with and without illumination changes respectively. With labels `SDI_TD` and `SD_TD` are plotted the results for the tracker without $\mathtt{B}_i$ for the sequences with and without illumination changes respectively. From these plots we can see that, when tracking a sequence with fixed illumination, there are no significative differences in tracking it either with or without the illumination subspace. On the contrary, if the same sequence has illumination changes, then the tracker using the full model performs significantly better. Moreover, the tracker without the illumination subspace looses track in this sequence. From these experiments we can conclude that: a) the tracker successfully separates changes in appearance caused by face deformation and illumination; b) it accurately reanimates the graphical model, even with the simple linear mapping introduced in section 3.

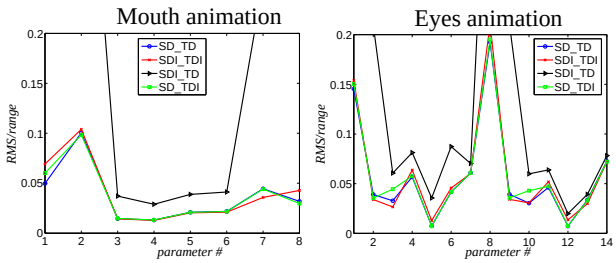

Figure 2. Re-animation results for the synthetic sequences.

4.3. Real experiments

In this case, matrix E was built with parameters obtained from a set of manually selected pairs of facial expressions in the graphical model (key frames) and the corresponding images of the human actor. We used 21 key frames for the eyes and 18 for the mouth. Then we acquired a live sequence using an Apple iSigth camera. We changed the illumination conditions by moving a light source in front of the user, with roof lights on. A C++ implementation of the system described in this paper was able to track the sequence and reanimate the graphical model in real-time running in a PentiumIV 3.2GHz computer. In Fig. 3 are shown some tracking and re-animation frames from this sequence.

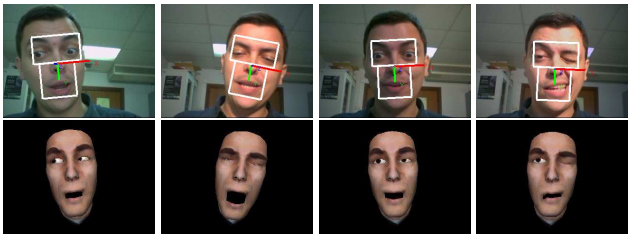

Figure 3. Some images from the real live sequence.

5. Conclusions

We have presented a computer vision system which can estimate human face animation parameters (muscle deformations) from a marker-less image sequence under strong illumination variations. It is based on a model of facial appearance composed of two independent linear subspaces modelling face deformations and illumination. With an efficient image alignment procedure we estimate the parameters of the model and quantify facial motion and deformation.

In the synthetic experiments performed we have shown that the tracker correctly separates changes in the appearance of the face caused by deformations and changes in illumination. Using a simple model representing the linear correlations between the appearance parameters estimated by the tracker and the animation parameters of the graphical model we have been able to reanimate a face graphical model. A C++ implementation of this system is able to track a live sequence of an actor and reanimate a graphical model in real-time.

References

[1] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *Trans. on PAMI*, 26(10):1380–1384, October 2004.

[2] B. Bascle and A. Blake. Separability of pose and expression in facial tracing and animation. In *Proc. of ICCV*, pages 323–328. IEEE, 1998.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. of SIGGRAPH*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[4] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Trans. on PAMI*, 25(9):1–12, September 2003.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Proc. of ECCV*. Springer-Verlag, 1998.

[6] F. de la Torre and M. J. Black. Robust parameterized component analysis: Applications to 2d facial modeling. In *Proc. of ECCV (4)*, volume 2353 of *LNCS*, pages 653–669. Springer, 2002.

[7] F. Dornaika and J. Ahlberg. Fast and reliable active appearance model search for 3d face tracking. *Trans. on SMC-B*, 34(4):1838–1853, 2004.

[8] H. Fei and I. Reid. Joint bayes filter: A hybrid tracker for non-rigid hand motion recognition. In *Proc. of ECCV*, volume 3023 of *LNCS*, pages 497–508, 2004.

[9] L. Jongwoo, D. Ross, L. Ruei-Sung, and Y. Ming-Hsuan. Incremental learning for visual tracking. In *Advances in Neural Information Processing Systems*, 2004.

[10] K.-C. Lee and D. Kriegman. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In *Proc. of CVPR*, 2005.

[11] No-Author. Intentionally blank to keep the anonymity of the review. 2006.

[12] J. Ohya, Y. Kitamura, H. Takemura, H. Ishi, F. Kishino, and N. Terashima. Virtual space teleconferencing: Real-time reproduction of 3d human images. *Journal of Visual Comunications and Image Representation*, 6(1):1–25, March 1996.

[13] F. I. Parke and K. Waters. *Computer Facial Animation*. AK Peters Ltd, 1996.

[14] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *Trans. on PAMI*, 15(6), June 1993.

[15] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. of ICCV*, 2001.