

# Eigenfiltering for Flexible Eigentracking (EFE)

Fernando De la Torre †, Jordi Vitria ‡, Petia Radeva ‡, Javier Melenchon †

†Departament of Communications and Signal Theory. La Salle School of Engineering.

Universitat Ramon LLull. Passeig Bonanova 8. Barcelona 08022. Spain

‡Computer Vision Center. Computer science department.

Universitat Autònoma de Barcelona. 08193 Bellaterra (Barcelona) . Spain

**e-mail:** *ftorre@salleURL.edu jordi@cvc.uab.es petia@cvc.uab.es jmelen@salleURL.edu*

## Abstract

*Traditional techniques for tracking non-rigid objects such as optical flow, correlation, active contours or color, can not deal with situations where image changes are not due to motion but appearance (e.g. tracking the lips when the teeth appear). Two main contributions for appearance tracking of flexible objects are proposed. The first one is a flexible generalization of Eigentracking within the same robust continuous optimization framework. The second one is a generalization of traditional graylevel eigenspaces, constructing a multiple channel "eigenspace" using filter responses to give robustness against variations in the training conditions such as illumination changes. Additionally, 3D geometric transformations are incorporated, a regularization term is added for numerical stability reasons and the optimization problem is solved in closed form. Experiments on lip tracking are reported.*

## 1 Introduction

Visual tracking has emerged as an important component for systems using vision as feedback for surveillance, continuous control or human computer interaction [1, 7, 5, 13]. Faces have received considerable attention from both the computer vision and signal processing communities, where ability to track motion is useful in applications such as face based biometrics person authentication, facial expression analysis, animation or teleconferencing.

Since *snakes* appear, a lot of work has been done to recover non-rigid motion [1, 4, 6, 5, 9]. While *snakes* enforce smoothness in their shapes, they do not take into account restrictions for a particular non-rigid motion and they are sensitive to occlusion and noise. Other trackers have emerged to solve these problems, such as Active Contours [1], all techniques relying on template matching in image or Eigenspace [4, 12, 10], the ones using optical flow

[2], deformable templates, feature trackers [9] and Active Shape Models [6]. Active contours [1] are not well suited for tracking complex objects without contextual information or special signals. All optical flow techniques [2] have reported good results for recognition purposes, but are not valid ones for accurate tracking, since the optical flow constraint is violated in situations where changes are due to appearance. All the techniques based on normalized correlation need a template which is not invariant to affine changes or takes into account complex appearance representations. Active Shape Models [6] just use the mean of the normal to the curve to track, being not very accurate when several texture changes are possible. Moreover, since it is posed as least square optimization problem it is not robust to outliers and not 3D motion is taken into account. In sum, all the previous techniques are not well suited for an accurate tracking of non-rigid motion when changes in image are not due to motion but texture. To take into account changes in view or texture, Black and Jepson [3] have proposed Eigentracking, where all possible configurations of the object are registered in an Eigenspace. However, the object to track must be inside a box and it is not very practical for tracking flexible objects such as mouth. Active Appearance Models(AAM) [5] have arisen to incorporate the shape constructing a couple shape/texture eigenspace. Although similar in spirit to AAM, this work differs from it in several aspects; we derive the whole tracking problem as a robust continuous optimization problem providing the covariance matrix of the estimated parameters. We do not model a coupled shape-texture graylevel eigenspace unlike AAM. We construct a distributed eigenspace with shape constraints in a hierarchical manner, modeling the whole object at the lowest multiresolutive level and in higher levels just the peripheral regions. Therefore, our work extends Eigentracking to deal with non-rigid motion, constructs an appearance representation from filter responses, closed form and more stable numerical results are achieved, and 3D geometric changes are taken into account.

## 2 Previous Work

### 2.1 Encoding Appearance and Shape

Since the early work of Kirby and Sirovich [12] parameterizing the face using Principal Component Analysis, many computer vision people have used this technique to parameterize shape, appearance or motion [3, 6, 5, 10].

Let us consider a set of  $m$  images/filters/shapes with  $N$  pixels/coefficients/points, given by the columns of a matrix  $\mathbf{A} \in \mathbb{R}^{N \times m}$ . These columns form the training set with all possible configurations of the texture, filter responses or shapes of an object. The matrix  $\mathbf{A}^1$  can be factorized using Singular Value Decomposition (SVD),  $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ . The  $k$  first columns of the orthogonal matrix  $\mathbf{U} \in \mathbb{R}^{N \times k}$ <sup>2</sup>, the eigenvectors of the covariance matrix  $\mathbf{A}\mathbf{A}^T$ , will expand the principal subspace of the columns of  $\mathbf{A}$ ,  $\mathbf{\Sigma} \in \mathbb{R}^{k \times k}$  will contain the singular values and the orthogonal matrix  $\mathbf{V} \in \mathbb{R}^{m \times k}$  will expand the row space of  $\mathbf{A}$ . An image, filter or shape  $\mathbf{I}$  will be represented by projection onto eigenvectors  $\mathbf{u}_j$  (columns of  $\mathbf{U}$ ), i.e  $\mathbf{I} \simeq \sum_{j=1}^k c_j \mathbf{u}_j = \mathbf{U}\mathbf{c}$  where  $\mathbf{c}$  are the projection coefficients.

### 2.2 EigenTracking

Motion of planar surfaces under orthographic projection can be described in terms of affine transform with 6 parameters  $\mathbf{a}^6 = (a_1, a_2, a_3, a_4, a_5, a_6)^T$ :

$$\mathbf{f}(\mathbf{x}, \mathbf{a}^6) = \begin{bmatrix} a_1 \\ a_4 \end{bmatrix} + \begin{bmatrix} a_2 & a_3 \\ a_5 & a_6 \end{bmatrix} \begin{bmatrix} x - x_c \\ y - y_c \end{bmatrix} \quad (1)$$

where  $\mathbf{x}_c = (x_c, y_c)^T$  is the center position of the object template to track.

Under assumption of coplanarity, appearance tracking can be achieved by treating the region to track ( $N$  pixels) as function of affine parameters,  $\mathbf{I}(\mathbf{f}(\mathbf{x}, \mathbf{a}^6)) = [I(\mathbf{f}(\mathbf{x}_1, \mathbf{a}^6)), \dots, I(\mathbf{f}(\mathbf{x}_N, \mathbf{a}^6))]^T$ . Black and Jepson [3] propose to accomplish tracking by recovering both the affine parameters  $\mathbf{a}$  and the projection coefficients  $\mathbf{c}$  by minimizing a cost function  $\min_{\mathbf{a}, \mathbf{c}} \rho[(\mathbf{I}(\mathbf{f}(\mathbf{x}, \mathbf{a})) - \mathbf{U}^G \mathbf{c}), \sigma]$  where  $\rho(x, \sigma)$  is a robust function to take into account violations of the model (e.g. non-coplanarity assumption, specular reflections, ...).  $\sigma$  is a scale factor that controls the convexity of the robust function for annealing procedure.

## 3 Flexible EigenTracking

Eigentracking [3] can not be applied to track flexible or non-rigid targets (e.g. lips), since all possible configurations of the object to track have different spatial domain.

<sup>1</sup>We assume zero mean, otherwise the mean is subtracted off.

<sup>2</sup>Throughout this paper, we will use  $\mathbf{U}^G, \mathbf{U}^F, \mathbf{U}^S$  for denoting the orthogonal basis for graylevel, filters and shape modes.

In this section we generalize previous work on Eigentracking, proposing Flexible Eigentracking (FE) which is able to track and estimate rigid and non-rigid motion.

The shape of an object is represented by an  $N$ -dimensional vector of image coordinates<sup>3</sup>, and to each point  $i$ , we will associate a graylevel appearance neighbourhood. Figure 1 shows the neighbourhood used for lip-tracking. This representation is similar to Point Distribution Model proposed by Cootes *et. al.* [6], but differs from it in several aspects. FE constructs a spatially distributed Eigenspace with shape constraints and not only the mean of the normal to the curve is taken into account. We have also posed the tracking as a *robust continuous* optimization problem achieving sub-pixel accuracy and the covariances of the estimation are provided.

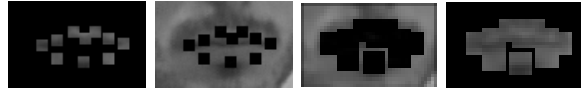


Figure 1. Neighbourhood used for each characteristic point at two different scales.

Affine transform does not take into account 3D changes in view or perspective projection. In order to compensate such model errors, we add a new basis  $\mathbf{e}_i$  to the original affine one (eq. 1). This new basis is obtained from the principal components of the residual between the 3D mouth changes and their affine approximations [1]. The rigid parameter vector will be  $\mathbf{a} = (\mathbf{a}^6, a_7, \dots, a_{6+t})^T$ , where  $t$  is the number of additional basis.

FE tracks the object under rigid (a), non-rigid (b) motion and changes in appearance (c), minimizing:

$$\min_{\mathbf{a}, \mathbf{c}, \mathbf{b}} \rho \left[ \mathbf{I} \left( \mathbf{f}(\bar{\mathbf{x}}, \mathbf{a}^6) + \sum_{i=1}^k b_i \mathbf{f}(\mathbf{u}_i^S, \mathbf{a}^6) + \sum_{i=1}^t a_{i+6} \mathbf{e}_i \right) - \mathbf{U}^G \mathbf{c}, \sigma \right] \quad (2)$$

where  $\bar{\mathbf{x}}$  is the mean shape of the object and  $\mathbf{b}$  are the non-rigid coefficients. Coefficients  $\mathbf{c}$  reconstruct the graylevel appearance of the object to track. Eq. 2 just differs from Eigentracking equation in that we warp the image with rigid and non-rigid motion. Observe that the warping function inside the Cartesian coordinates of the image is a bilinear mapping of rigid (a) and non-rigid parameters (b). In our scheme, we use the robust German-McClure error function given by  $\rho(x, \sigma) = \frac{x^2}{x^2 + \sigma^2}$ .

FE does not warp the whole tracked region, since only warping of the representative points of the curve is necessary. We also use a coarse-to-fine strategy [3], and the number of samples for each multiresolution level (see fig.1) is

<sup>3</sup>We use a B-spline with  $N$  control points but we omit it for notational convenience.

the same. In this way, at lower resolution levels we can process the entire object and in higher levels we will have a coarse estimation, but just in the regions of interest. These two facts make the algorithm much faster.

Figure 2 shows a sequence of lip tracking with FE when a person is smiling. Observe that an accurate tracking is difficult due to the appearance of the teeth and previous correlation, color or active contour could easily fail in such task.



**Figure 2. Mouth Tracking when a person is smiling and speaking**

#### 4 EigenFiltering for Flexible EigenTracking

Small regions of graylevel values, can suffer from huge ambiguities, as well as sensitivity to camera noise and slight changes in illumination. More robust representation can be achieved by local combination of these pixels using filtering [9, 11]. In fact, there are biological evidences that show that human visual system uses the response of some filters to represent the visual information. Following this biological evidence, we will represent each image point  $i$  by the response of some filters. Gaussian filters and its derivatives are a well-known basis for natural images [11] and are computationally inexpensive because of its steerability property [8].

Each characteristic point  $i$  will become represented by a mean and a covariance  $(\mu_i, \Sigma_i)$ . These statistics will result from filtering the point  $i$  (of a training set of images with different mouth configurations) with a bank of filters at different orientations and scales. To reduce the number of filters, SVD is applied to the set of filters in order to eliminate the spatial correlation between them. Each component of the vector  $\mathbf{I}_F^i$  will be the result of filtering the image at point  $i$  by one of the  $M$  Eigenfilters (columns of  $\mathbf{U}^F$ ), that is  $\mathbf{I}_F^i = [I_{F_1}^i, I_{F_2}^i, \dots, I_{F_M}^i]^T$ .

Given an image  $\mathbf{I}$  and considering that the error over the  $N$  shape points is I.I.D, tracking will be achieved maximizing the following likelihood function  $L(\mathbf{I}|\mathbf{a}, \mathbf{b})$ :

$$L = \prod_{i=1}^N \frac{1}{(2\pi)^{M/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2} (\mathbf{I}_F^i(\mathbf{x}'_i) - \mu_i)^T \Sigma_i^{-1} (\mathbf{I}_F^i(\mathbf{x}'_i) - \mu_i)}$$

$$\mathbf{x}'_i = \mathbf{f}(\bar{\mathbf{x}}_i, \mathbf{a}^6) + \sum_{i=1}^k b_i \mathbf{f}(\mathbf{u}_i^S, \mathbf{a}^6) + \sum_{i=1}^t a_{i+6} \mathbf{e}_i \quad (3)$$

Observe that, unlike eq. (2), eq. (3) does not have coefficients  $\mathbf{c}$  to reconstruct any subspace. We just calculate the rigid and non-rigid parameters of the geometrical transformation, which makes the tracker be closer to the original gaussian distribution of filter responses ( in the Mahalanobis distance sense). This is the same measurement equation used by [10]. As we are working with low dimensional spaces, there is no need to reconstruct a subspace of filter responses, since the covariance matrix can be robustly estimated. In fact, similar results are obtained adding new parameters to reconstruct the filter responses, constructing in such a way an eigenspace. Although more accurate tracker can be achieved with the eigenspace reconstruction, adding the eigenspace makes it more computationally intensive and for low dimensional spaces makes no sense, since a worse model constraint is achieved. For minimizing eq. 2 and 3 we use Iteratively Recursively Least Squares, unlike [3], but we omit the details for lacking space.

#### 5 Experimental Results

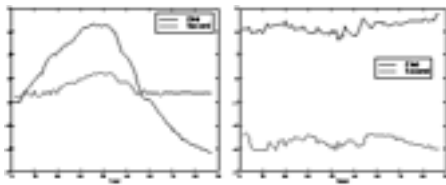
The algorithms are implemented in unoptimized Matlab code in a standard 450M Hz PC at 2 frames/sec. The first step is the construction of graylevel and filter responses models from training data with people. For each image in the training set, the whole lip contour is followed by hand and a B-spline with  $N$  control points is fitted to it in order to construct the shape space. The tracker is automatically initialized in the first image frame with a stochastic search over the parameter space of rigid and non-rigid motion [7], after which tracking is fully automatic and deterministic.

In fig. 3 some frames of a 3D rotation of the mouth are shown. Fig. 4 shows the two last coefficients of the rigid parameters  $a_7, a_8$  ( which codify the 3D changes ) and coefficients  $\mathbf{b}$ . Observe that it is possible to infer 3D position of the face with coefficient  $a_7$ , just noting if it is positive or negative. The non-constant parameters  $\mathbf{b}$  are caused by inaccuracies of the model and linear dependency between the non-rigid subspace and the new added basis.



**Figure 3. 3D mouth motion**

In Figure 5 we can observe a sequence where a person is speaking, smiling, rotating, and zooming in/out, tracked with EFE. There are some frames where the tracker is lost due to very fast movement. In such cases, we detect that the tracker is lost and we re-initialize it in the previous frame with a prediction of the rigid parameters.



**Figure 4. Two coefficients of additional rigid motion. Two non-rigid motion parameters**



**Figure 5. Lip tracking**

Fig. 6 shows two frames of a person speaking and how EFE can track it. The illumination conditions of the training set are different from the tracking sequence. The third and fourth images in fig. 6 show that without previous normalization of gray-level values, FE tracker is lost.

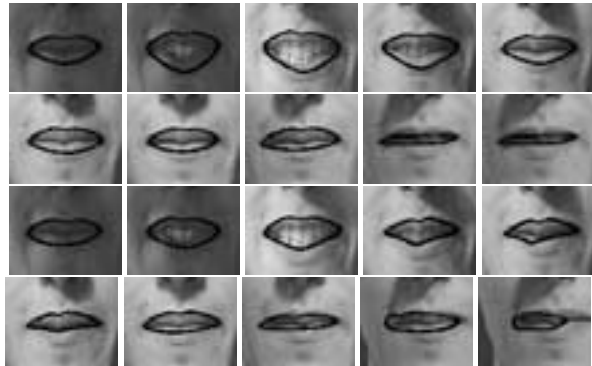


**Figure 6. EFE and FE tracking under different training conditions.**

Figure 7 shows the comparison between FE and EFE when abrupt changes in illumination are produced. We normalize locally each patch in order to give robustness against local shadows. However EFE (first two rows, fig. 7) presents better performance than FE (third and fourth rows, fig. 7), in the sense, that the adaptation to the lips is visually better. Each patch of FE is composed of  $15 \times 15$  pixels square neighborhood, whereas EFE uses 3 eigenfilters<sup>4</sup> in 10 positions. In this case we have reduced significantly the

<sup>4</sup>24 filters: a gaussian, 2 first derivatives and 3 second derivatives over 4 scales. The first 3 columns of  $U^F$ , preserve the 90% of the energy.

number of points used for each patch ( $3 \times 10$  vs.  $15 \times 15$ ).



**Figure 7. EFE and FE tracking under changeable illumination conditions.**

**Acknowledgements** The work was partially supported by TAP98-063, TIC98-1100 CICYT and 2FD97-0220 EU grants.

## References

- [1] B. Bascle and A. Blake. Separability of pose and expression in facial track and animation. In *ICCV*, 1998.
- [2] M. J. Black. Explaining optical flow events with parameterized spatio-temporal models. In *CVPR*, 1999.
- [3] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of objects using view-based representation. In *ECCV*, pages 329–342, 1996.
- [4] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. In *ICCV*, 1995.
- [5] T. Cootes, G. J. Edwards, and C. Taylor. Active appearance models. In *ECCV*, 1998.
- [6] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models- their training and application. *CVIU*, 61(1):38–59, January 1995.
- [7] F. de la Torre, Y. Yacoob, and L. Davies. A probabilistic framework for rigid and non-rigid appearance based tracking and recognition. In *FG-2000*.
- [8] W. T. Freeman and E. Adelson. The design and use of steerable filters. *PAMI*, (13), 1991.
- [9] S. McKenna, S. Gong, R. Wurtz, J. Tanner, and D. Banin. Tracking facial feature points with gabor wavelets and shape models. In *Audio-Video Based Biome. Person Auth.*, 1997.
- [10] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *ICCV*, 1995.
- [11] R. Rao and D. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 12, 1995.
- [12] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A*, 4(3):519–524, March 1987.
- [13] J. Vitria, J. Serra, E. Raso, and A. Puertas. Visual eliza: a testbed for developing visual perception processes for conversational interfaces. In *CCIA*, 1998.