Filtered Component Analysis to Increase Robustness to Local Minima in Appearance Models

Fernando De la Torre† Alvaro Collet† Manuel Quero† Jeffrey F. Cohn‡ Takeo Kanade†
†, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.
ftorre@cs.cmu.edu acollet@cs.cmu.edu mquero@andrew.cmu.edu tk@cs.cmu.edu
‡, University of Pittsburgh. Department of Psychology. Pittsburgh, Pennsylvania 15260.
jeffcohn@pitt.edu

Abstract

Appearance Models (AM) are commonly used to model appearance and shape variation of objects in images. In particular, they have proven useful to detection, tracking, and synthesis of people's faces from video. While AM have numerous advantages relative to alternative approaches, they have at least two important drawbacks. First, they are especially prone to local minima in fitting; this problem becomes increasingly problematic as the number of parameters to estimate grows. Second, often few if any of the local minima correspond to the correct location of the model error. To address these problems, we propose Filtered Component Analysis (FCA), an extension of traditional Principal Component Analysis (PCA). FCA learns an optimal set of filters with which to build a multi-band representation of the object. FCA representations were found to be more robust than either grayscale or Gabor filters to problems of local minima. The effectiveness and robustness of the proposed algorithm is demonstrated in both synthetic and real data.

1. Introduction

Component Analysis (CA) methods such as Principal Component Analysis (PCA) have been widely applied in visual, graphics, and signal processing tasks over the last two decades. PCA is a key learning component of Appearance Models (AM). AM have proven especially powerful for face tracking and synthesis relative to alternative approaches (e.g. optical flow) [4, 15, 25, 1, 6, 8, 3].

In applications such as face detection and tracking, the goal is to search for a minimum residual between the image and the model across rigid (e.g. rotation and translation) and non-rigid parameters. For instance, consider fig. (1), in which a face has been placed in an arbitrary image. In fig. (1.a), we plot the normalized correlation surface error between the ideal template (face) and the image in a 101×101 patch centered in the middle of the face. This surface error has nice local properties: it has just one well defined global minimum that corresponds to the expected location of the face. However, if we learn a generic PCA model of the facial appearance variation from training data and try to locate the face again, two undesirable effects may occur. First, the location of the optimal parameter (translation) fails to correspond to the location of the face (delineated by the the black dot in the figure), see fig. (1.b). Second, many local minima may be found. Even if a gradient descent algorithm begins close to the correct solution, the occurrence of local minima is likely to divert convergence from the desired solution.

The aim of this paper is to explore the use of a new technique, Filtered Component Analysis (FCA). FCA learns a multiband representation of the image that reduces the number of local minima and improves generalization relative to using PCA on grayscale. Fig. (1.c) shows the main point of the paper. By building a multiband representation with FCA, we are able to locate the minimum in the right location (black dot) and reduce the number of local minima close to the optimal one.

2. Previous Work

This section reviews previous work on subspace tracking and the role of representation in subspace analysis.

2.1. Subspace detection and tracking

Subspace trackers build the object's appearance/shape representation from the PCA of a set of training samples. Let $\mathbf{d}_i \in \Re^{d \times 1}$ (see notation [1]) be the i^{th} sample of a

[1] Bold capital letters denote a matrix $\mathbf{D}$, bold lower-case letters a column vector $\mathbf{d}$. $\mathbf{d}_j$ represents the j column of the matrix $\mathbf{D}$. d_{ij} denotes the scalar in the row i and column j of the matrix $\mathbf{D}$ and the scalar i-th element of a column vector $\mathbf{d}_j$. All non-bold letters will represent variables of scalar nature. $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$ designates Euclidean norm of $\mathbf{x}$. The $vec(\mathbf{D})$ operator transforms $\mathbf{D} \in \Re^{d \times n}$ into an dn-dimensional vector by stacking the columns. $\circ$ denotes the Hadamard or point-wise product.

1-4244-1180-7/07/$25.00 ©2007 IEEE

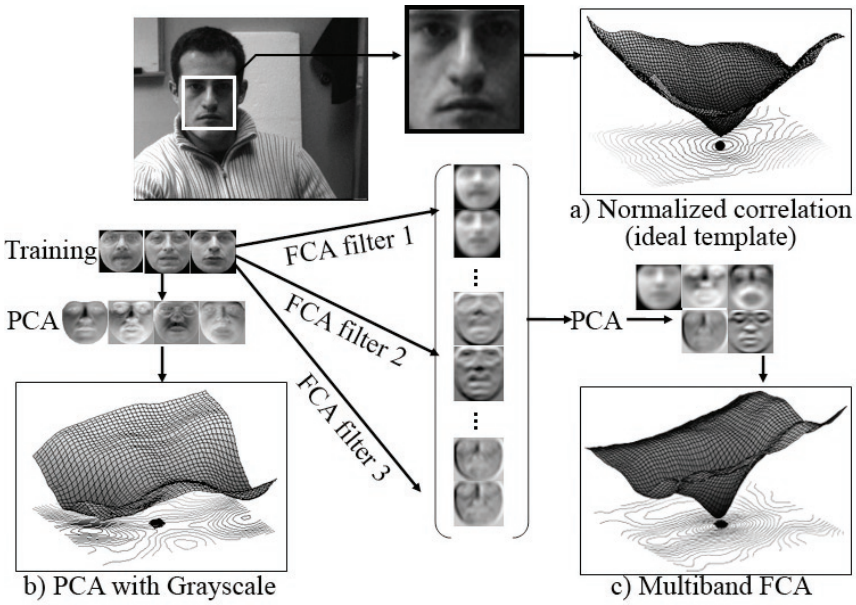

Figure 1. a). Normalized correlation error surface of the image with the face (94×97) in a 101×101 patch. b) Error function with a generic grayscale appearance model. The black dot denotes the optimal position of the face. c) Error function of a multiband learned representation using FCA. The location of the face corresponds to the minimum of the function.

training set $\mathbf{D} \in \Re^{d \times n}$ and $\mathbf{B} \in \Re^{d \times k}$ the first k principal components. The k principal components $\mathbf{B}$ maximize $\max_{\mathbf{B}} \sum_{i=1}^{n} ||\mathbf{B}^T \mathbf{d}_i||_2^2 = ||\mathbf{B}^T \mathbf{\Gamma} \mathbf{B}||_F$ under the constraint $\mathbf{B}^T \mathbf{B} = \mathbf{I}$, where $\mathbf{\Gamma} = \mathbf{D}\mathbf{D}^T = \sum_i \mathbf{d}_i \mathbf{d}_i^T$ is the covariance matrix (zero mean data). The columns of $\mathbf{B}$ form an orthonormal basis that spans the principal subspace. If the effective rank of $\mathbf{D}$ is much less than d, we can approximate the column space of $\mathbf{D}$ with $k << d$ principal components. The sample $\mathbf{d}_i$ can be approximated as a linear combination of the principal components as $\mathbf{d}_i \approx \mathbf{B}\mathbf{c}_i$ where $\mathbf{c}_i = \mathbf{B}^T \mathbf{d}_i$.

Once the model has been learned (i.e. $\mathbf{B}$ is known), tracking is achieved by finding the parameters $\mathbf{a}$ of the geometric transformation $\mathbf{f}(\mathbf{x}, \mathbf{a})$ that aligns the data w.r.t. the subspace. In the case of an affine transformation, $\mathbf{f}(\mathbf{x}, \mathbf{a}) = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} a_3 & a_4 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x - x_c \\ y - y_c \end{pmatrix}$ where $\mathbf{a} = (a_1, a_2, a_3, a_4, a_5, a_6)$ are the affine parameters and $\mathbf{x} = (x_1, y_1, \cdots, x_n, y_n)$ is a vector containing the coordinates of the pixels to track. Given an image $\mathbf{d}_i$, subspace trackers or detectors find $\mathbf{a}$ and $\mathbf{c}_i$ that minimize: $min_{\mathbf{c}_i, \mathbf{a}} ||\mathbf{d}_i(\mathbf{f}(\mathbf{x}, \mathbf{a})) - \mathbf{B}\mathbf{c}_i||_2^2$ (or other normalized error). If $\mathbf{a} = (a_1, a_2)$, i.e. just translation, the search can be done efficiently over the whole image using the Fast Fourier Transform (FFT). Searching for $\mathbf{a} = (a_3 = a_6, a_5 = a_4)$, that is, rotation and scale, can also be done efficiently in the log-polar representation of the image with the FFT [13].

$\otimes$ denotes convolution. $\mathbf{1}_k \in \Re^{k \times 1}$ is a vector of ones. $\mathbf{I}_k \in \Re^{k \times k}$ is the identity. $||\mathbf{A}||_F = tr(\mathbf{A}^T \mathbf{A}) = tr(\mathbf{A}\mathbf{A}^T)$ designates the Frobenious norm of a matrix. $tr(\mathbf{A}) = \sum_i a_{ii}$ is the trace of the matrix $\mathbf{A}$.

It is important to notice that $\mathbf{f}$ can also model non-rigid motion. For instance, consider $\mathbf{f}(\mathbf{B}^S \mathbf{c}^s, \mathbf{a}) = \sum_{i=1}^{k} c_i^s \mathbf{f}(\mathbf{b}_i^S, \mathbf{a})$, where $\mathbf{B}^S$ is a non-rigid shape model learned by computing PCA on a set of registered shapes [6] and $\mathbf{c}^s$ the non-rigid parameters. In this case, $\mathbf{f}(\mathbf{B}^S \mathbf{c}^s, \mathbf{a})$ will account for rigid and non-rigid motion. A standard approach to efficiently search over the rigid $\mathbf{a}$ and non-rigid $\mathbf{c}^s$ parameters, is to use gradient descent methods [1, 6, 15, 3, 8].

2.2. Representation in subspace analysis

Most work on AM uses some sort of *normalized* grayscale to build the representation. However, regions of graylevel values can suffer from large ambiguities, camera noise, and changes in illumination. More robust representation can be achieved by local combination of pixels through filtering. Filtering of the visual array is a key element of the primate visual system [19].

Using different representations for subspace recognition were explored by Bischof et al. [2]. In the training stage, the authors built a subspace by filtering the PCA-grayscale basis with steerable filters [10]. In the recognition phase, they filtered the test images and performed robust matching, obtaining improved recognition performance over grayscale. Yilmaz et al [27] show how to improve face recognition under illumination changes using PCA filtered images. On the other hand, multiband representations (e.g. Gabor) have been typically used as features for many visual classification tasks [19]. In related work on component analysis,

several tensor factorization of image ensembles have been proposed over the past few years [18, 22, 26]. These approaches avoid the vectorization effect of the image and find a reduced rank multi-linear approximation of the graylevel images.

In the context of AM, Mckenna et al. [16] proposed a facial feature tracker based on Gabor wavelets and shape models, showing improved tracking performance over grayscale approaches. Cootes et. al [5] found that a non-linear representation of edge structure could improve subspace matching. Stegmann and Larsen [24] report that building subspaces for AM in an augmented space of intensity, hue and edges performed better in the task of localizing faces. In similar fashion, [7] make use of wedgelet regression trees to reduce the computational complexity of standard Active AM. De la Torre et al. [9] found that subspace tracking was improved by using a multiband representation created by filtering the images with a set of Gaussian filters and its derivatives.

This work differs in several aspects from previous work. First, we explicitly learn a set of optimal spatial filters adapted to the object of interest, rather than using hand-picked ones. Once the filters are learned, we build a multiband representation of the image that has improved error surfaces with which to fit AM. We evaluate quantitatively the properties of the error surfaces and show how FCA outperforms current methods in appearance based detection and tracking applications.

3. Filtered Component Analysis

Many component analysis methods (e.g. PCA, LDA) build data models based on the second order statistics (covariance matrices) of the signal. In particular, PCA finds a linear transformation that decorrelates the data by exploiting the correlation across samples. PCA models the correlation across pixels of different images, but not the spatial statistics within each of the images. In this section, we propose Filtered Component Analysis (FCA) that learns a bank of orthogonal filters that decorrelate the spatial statistics of a set of images. Once the FCA filters are learned, we build a multi-band representation that provides more robust matching and generalizes better than grayscale.

3.1. Learning spatial correlation

Previous research [9, 2, 5] has shown the importance of representation in AM. However, researchers have used hand-picked filters to represent the signal. Instead, FCA will learn a set of orthogonal spatial filters optimal for variance preservation. Variance preservation of image spatial statistics is a realistic assumption to build a generative model for detection or tracking appearance. For instance, active AM [5, 15] build a model of shape/appearance based on variance preservation of the training images.

Given a set of training images, $\mathbf{D} \in \Re^{d \times n}$, our aim is to model the spatial statistics of the signal by learning the filter $\mathbf{F}$ that minimizes:

$$E_1(\mathbf{F}, \boldsymbol{\mu}) = \min_{\mathbf{F}, \boldsymbol{\mu}} \sum_{i=1}^{n} ||\mathbf{d}_i \otimes \mathbf{F} - \boldsymbol{\mu}||_2^2 \quad (1)$$

Recall that $\otimes$ denotes convolution, and $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_i \otimes \mathbf{F}$ is the mean of the filtered signal. If $\boldsymbol{\mu}$ is known, the optimal $\mathbf{F}$ can be achieved by solving:

$$\mathbf{A}vec(\mathbf{F}) = \mathbf{b} \qquad \mathbf{A} = \sum_{i=1}^{n} \sum_{(x,y)} \mathbf{d}_i^{(x,y)} \mathbf{d}_i^{(x,y)^T}$$
$$\mathbf{b} = \sum_{i=1}^{n} \sum_{(x,y)} \boldsymbol{\mu}^{(x,y)} \circ \mathbf{d}_i^{(x,y)} \quad (2)$$

where (x, y) is the domain where the convolution is valid and $\mathbf{d}_i^{(x,y)}$ is a patch of the filter size (f_x, f_y) centered at the coordinates (x, y). The matrix $\mathbf{A}$ can be computed efficiently in space or frequency from the autocorrelation function of $\mathbf{d}_i$. Analogously, $\mathbf{b}$ is estimated from the cross-correlation between $\mathbf{d}_i$ and $\boldsymbol{\mu}$. Alternatively, one could use the integral image [12] to efficiently compute eq. 2.

Without imposing any constraints on the filter coefficients, the optimal solution of eq. 1 is given by $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{F} = \mathbf{0}$ (although an iterative algorithm will rarely converge to this solution). To avoid this trivial solution, we impose that the sum of squared coefficients is 1, i.e. $vec(\mathbf{F})^T vec(\mathbf{F}) = 1$. The latter constraint can be elegantly solved by noticing that the convolution is a linear operator, and eq. 2 can be rewritten as:

$$E_2(\mathbf{F}) = \min_{\mathbf{F}} \sum_{i=1}^{n} ||(\mathbf{d}_i - \boldsymbol{\mu}') \otimes \mathbf{F}||_2^2 \quad (3)$$

where $\boldsymbol{\mu}' = \frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_i$ is the sample mean. Now eq. 3 can be solved by finding the eigenvector with smallest eigenvalue of $\mathbf{A} = \sum_{i=1}^{n} \sum_{(x,y)} (\mathbf{d}_i - \boldsymbol{\mu}')^{(x,y)} (\mathbf{d}_i - \boldsymbol{\mu}')^{(x,y)^T}$.

3.2. Learning a multiband representation

In this section, we show how to find a set of filters $\mathbf{F}^{1, \cdots, F}$ that decorrelates the spatial statistics of the image and are orthogonal to each other. Observe that FCA is analogous to PCA but now rather than decorrelating the signal with the covariance of the data, we decorrelate the spatial statistics over a set of images.

In our particular tracking application, we are interested in finding a set of filters that preserve the spatial statistics of the object of interest and has minimal response to background. This filter set can be obtained by maximizing $E_{FCA}(\mathbf{F}^{1, \cdots, F})$:

$$E_{FCA} = \sum_{f=1}^{F} \sum_{i=1}^{n} ||\mathbf{d}_i \otimes \mathbf{F}^f||_2^2 - \lambda \sum_{j=1}^{n_2} ||\mathbf{d}_j^b \otimes \mathbf{F}^f||_2^2 \quad (4)$$

where $\mathbf{d}_j^b$ denotes the j^{th} sample of the background. Let $\mathbf{T} = [vec(\mathbf{F}^1)\ vec(\mathbf{F}^2)\ \cdots\ vec(\mathbf{F}^F)]$ be a matrix of all the vectorized filters, the filters should satisfy $\mathbf{T}^T\mathbf{T} = \mathbf{I}_{F\times F}$. After taking the derivatives with respect to $\mathbf{F}^f$, it can be shown that the optimal solution satisfies the following eigenvalue problem:

$$(\mathbf{A} - \lambda\mathbf{U}\alpha)\mathbf{T} = \mathbf{T}\Omega \tag{5}$$

$$\mathbf{A} = \sum_{i=1}^{n}\sum_{(x,y)} \mathbf{d}_i^{(x,y)}{\mathbf{d}_i^{(x,y)}}^T \quad \alpha = \frac{max(\mathbf{A})}{max(\mathbf{U})}$$

$$\mathbf{U} = \sum_{j=1}^{n_2}\sum_{(x,y)} \mathbf{d}_j^{b(x,y)}{\mathbf{d}_j^{b(x,y)}}^T$$

If λ is large, the set of filters will predominantly cancel the background. If λ is small the filters will be adapted to the object.With λ close to one the filters will achieve trade-off between modeling the signal (i.e object) and removing the background. Typically $0 \leq \lambda \leq 2$. α is an artificially introduced parameter to normalize the energies of $\mathbf{A}$ and $\mathbf{U}$.

The solution to eq. 5 is given by the leading eigenvectors of $(\mathbf{A} - \lambda\alpha\mathbf{U})$. At this point, it is interesting to consider again the analogy with PCA. PCA will find the leading eigenvectors of $\sum_{i=1}^{n}\mathbf{d}_i\mathbf{d}_i^T$ whereas FCA will find the leading eigenvectors (assuming $\lambda = 0$) of $\mathbf{A} = \sum_{i=1}^{n}\sum_{(x,y)}\mathbf{d}_i^{(x,y)}{\mathbf{d}_i^{(x,y)}}^T$. While PCA finds the directions of maximum variation of the covariance matrix, FCA finds the directions of maximum variation of the sum of all overlapping patches.

Also recall that FCA is different from previous tensor factorization approaches [18, 22, 26] in several aspects. First, our goal is to build a multi-band signal representation by concatenating filtered versions of images and computing PCA after that, rather than performing tensor factorization on graylevel images. Tensor approaches explore the correlation between all rows and columns, but do not explore the correlation between overlapping patches. Also, note that our particular filters are not separable.

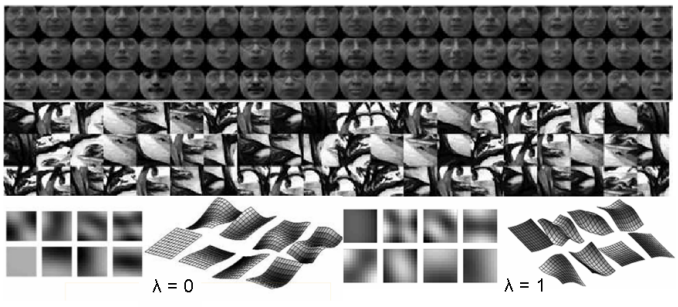

Figure 2. a) Training images of faces and background (top image). b) FCA filters for $\lambda = 0$, $\lambda = 1$ and size 11×11.

Fig. (2.a) shows many examples of faces and background patches. Fig. (2.b) shows the set of FCA filters for $\lambda = 0$ and $\lambda = 1$ for size 11×11. Observe that the first FCA filter is an average filter (left corner), and the other filters are differential filters at different orientations and scales.

3.3. Multiband subspace detection

In traditional subspace detection, PCA is computed from a set of training images. After the training stage, the goal is to detect the object of interest over different orientation, scales and translations. If the scale and orientation is known, detection can be achieved finding the translational parameters $\mathbf{a} = (a_1, a_2)$ that minimize:

$$E_3 = min_{\mathbf{c}_i,\mathbf{a}}\frac{||\mathbf{d}_i(\mathbf{x}+\mathbf{a}) - \mathbf{B}\mathbf{c}_i||_2^2}{||\mathbf{d}_i(\mathbf{x}+\mathbf{a})||_2^2} \tag{6}$$

Evaluating eq. 6 at each location (x, y) can be computationally expensive. For a particular position (x, y) computing the coefficients (i.e. $\mathbf{c}_i$) is equivalent to correlating the image with each basis of subspace $\mathbf{B}$, and stacking all values for each pixel. For large regions, this correlation is performed efficiently in the frequency domain using the Fast Fourier Transform (FFT) (i.e. $\mathbf{C}_1 = \mathbf{b}_1^T\mathbf{I} = IFFT(FFT(\mathbf{b}_1) \circ FFT(\mathbf{I}))$). Similarly, the local energy term, $||\mathbf{d}_i(\mathbf{x}+\mathbf{a})||_2^2$, can be computed efficiently using the convolution in the space or frequency domain. Alternatively, these expressions can be computed efficiently using the integral image [12].

In multiband tracking, we represent an image as a concatenation of filtered images. For a particular image $\mathbf{d}_i$ and a set of filters $(\mathbf{F}^1, \cdots, \mathbf{F}^f)$, there are several ways to modify eq. 6:

$$E_4 = \sum_{f=1}^{F}\Omega_f\frac{||\mathbf{d}_i\otimes\mathbf{F}^f - \mathbf{B}^f\mathbf{c}_i||_2^2}{||\mathbf{d}_i\otimes\mathbf{F}^f||_2^2} \tag{7}$$

$$E_5 = \sum_{f=1}^{F}\Omega_f\frac{||\mathbf{d}_i\otimes\mathbf{F}^f - \mathbf{B}^f\mathbf{c}_i^f||_2^2}{||\mathbf{d}_i\otimes\mathbf{F}^f||_2^2} \tag{8}$$

Parameters Ω_f are the eigenvalues of $(\mathbf{A} - \lambda\alpha\mathbf{U})$, obtained by FCA. E_4 filters the training images and builds PCA based on the set of stacked filtered images. On the other hand, E_5 computes an independent PCA for each representation such that the coefficients for each filtered image are uncoupled (i.e. $\mathbf{c}_i^f$ differs for each filtered image).

4. Experiments

To test the validity of our approach, we have performed several sets of experiments in face detection and facial feature tracking. The first set of experiments consists on detecting a face embedded in an arbitrary image (see fig. 1) using a generic model. In the second set, we test the ability of FCA to improve tracking in Active AM [6, 1, 25, 4, 15, 9].

In all experiments a generic face model is built from 150 subjects from the IBM ViaVoice AV database [17] and the CMU Multi-PIE Database [11], after aligning the data with

Procrustes Analysis [6]. Once the FCA filters are learned, a multi-band representation is built for each of the 150 images, and PCA is computed retaining 80% of the total energy. For comparison purposes, multi-band PCA is also calculated for other representations (e.g. Gabor, graylevel and derivatives, oriented pair filters [14]). In the experiments, we consider Gabor Filters because of the good results reported by other researchers in the area. In addition, these filters have been shown to provide optimal localization properties in both spatial and frequency domain and thus are well suited for tracking problems.

4.1. Understanding FCA

In order to compute a FCA filter set, 400 images containing faces and 400 background patches are randomly selected from the IBM database. Using these training samples, FCA filters are computed at 5 different scales (3×3, 5×5, 7×7, 9×9 and 11×11 pixels), using eq. 5 for different λ values.

Given a new face image not present in the training set, we embedded it in a bigger background image (see fig. 3). We efficiently compute the error in all possible translations with the FFT. Fig. (3) shows an example of the resulting error surface for each FCA band, in comparison with the error surfaces given by normalized grayscale. The grayscale representation has several local minima and the global minimum is misplaced. On the other hand, the sum of the three FCA bands produces an error surface with a correctly-placed global minimum. The first band is an average filter that smoothes the error surface and decreases its variability (avoids some spurious saddle points and local minima), already giving a reasonable approximation to the desired output. The second and third bands (derivative filters in different orientations) also have the global minimum in the correct position; in addition, they cancel out other spurious local minima and widen the gap from the global minimum to the closest local minimum.

4.2. Robustness to noise and illumination

This experiment is designed to test the robustness of FCA to noise and varying illumination conditions. A subset of 100 subjects from the IBM database (not in the training set) are randomly chosen and embedded in background images. Then, random impulsional noise is added (see fig. 4.a) and the error in each location is efficiently computed (orientation and scale are known) with the FFT. To quantitatively compare each filter bank, three different surface error statistics have been computed. Given a patch of 101×101 pixels around the optimal location of the face (which is known beforehand), we compute the following statistics: 1) distance between the global minimum and the face center, 2) distance between the correct minimum and closest local minimum, and 3) Amount of local minima. The amount of local minima in an error surface is calculated by counting those pixels with sign change in x and y derivatives and positive values in the second derivatives.

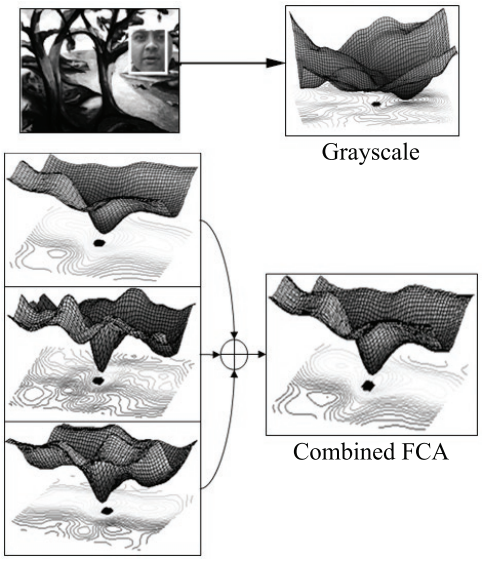

Figure 3. Error surfaces for grayscale and for each FCA band

Figure 4. a) (left) Original image and test image with added impulsional noise. b) (right) FCA(11,4) and Gabor(8,4) .

Table 1 shows the average results for the described error statistics for three representations: a set of four 11×11 pixels FCA filters (see fig. 4.a (top)), the best-performing Gabor filter set (see fig. 4.b (bottom)) and the normalized grayscale. In all our experiments, we report the results of the set of Gabor filters with the same spatial domain than the corresponding FCA filter set. A global minimum is said to be correct if it falls within a region of 3×3 pixels around the theoretical minimum. All the representations have similar accuracy; however, the amount of local minima is very high in the grayscale, and both grayscale and Gabor fail to provide a sufficiently high global-closest minimum margin in comparison with FCA filters. These results are quite stable across spatial domains of the FCA filter sets and have therefore been omitted in the interest of space.

The second experiment tests the robustness of FCA to illumination changes. A total of 120 faces (30 subjects, 4 images each) under varying illumination conditions (see fig. 5) are taken from the CMU PIE database [23]. Using the same approach as in the previous experiment, each face is embedded in a background image and the error surfaces are

	gray	$FCA_{\lambda=0}$	$FCA_{\lambda=0.5}$	Gabor(8,4)
(1)	98	**99**	**99**	99
(2)	9.73	**24.36**	**24.03**	19.01
(3)	30.06	**1.45**	**1.49**	2.46

Table 1. Experiments on noisy data. Statistics: (1) Percentage of correct global minimum. (2) distance between correct and closest local minimum. (3) Average number of local minima.

computed for each filter set. Results from this experiment are shown in table 2. In this case, FCA clearly outperforms any other technique in all three statistics of the error functions. Accuracy is higher than grayscale and Gabor by 33% and 12% respectively, while keeping the closest minimum at least 25.37% further away and having the lowest density of local minima. It is worth noting that the best-performing filter set has been $FCA_{\lambda=0}$ due to the different background training and testing statistical properties. Fig. (6) shows the error surface for a particular subject; as we can observe, the properties of FCA are more desirable than grayscale or Gabor filters in terms of location and density of local minima.

Figure 5. Changes in illumination on the PIE database.

	gray	$FCA_{\lambda=0}$	$FCA_{\lambda=0.5}$	Gabor(8,4)
(1)	41	**74**	**73**	62
(2)	14.59	**26.37**	**26.04**	19.68
(3)	3.28	**1.4**	**1.41**	1.92

Table 2. Experiments on illumination. (1),(2),(3) see table1.

The last experiment of this section explores FCA performance on real images. 10 images have been collected in the lab (see Fig. 7) with an inexpensive webcam, and roughly manually-selecting the same scale in the faces as in the training images. Table 3 shows the detection results of this experiment. As we can see FCA consistently outperforms other representations that included Gabor and grayscale in all metrics.

	gray	$FCA_{\lambda=0}$	$FCA_{\lambda=0.5}$	Gabor(8,4)
(1)	20	**80**	**80**	70
(2)	15.71	**18.05**	**25.52**	13.53
(3)	2	**2**	**1.2**	2.4

Table 3. Experiments on images taken in the lab.(1), (2), (3) see table 1

4.3. Tracking with Active Appearance Models

In this experiment, we test the ability of FCA to overcome local minima problems in Active Appearance Models

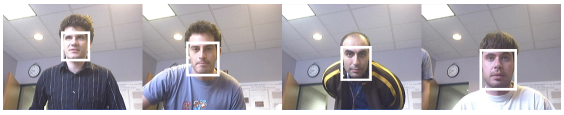

Figure 7. Some test images.

[6, 15]. In this case, we have constructed a multiresolution model of appearance patches around each of the 68 landmarks from 150 different subjects [9], taking 3 images per subject. The image samples were randomly chosen from the CMU PIE database [11] and aligned with Procrustes Analysis [6]. Once the shape and appearance FCA models are built (see fig. 8) retaining 80% of the energy, we use standard gradient descent methods to fit a new image to the model [9], although more efficient methods could use inverse composition [15, 1].

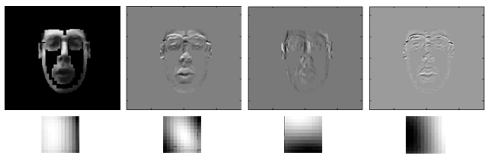

Figure 8. Multiband representation for each FCA filter.

In the case of AAM, evaluating the performance of the algorithm in terms of density of local minima is harder due to the high dimensionality of the parameter space. To evaluate the algorithm, we run two different tests: first, shape and rigid motion coefficients are randomly perturbed and the algorithm convergence ratio is measured, as well as the mean squared error between the final solution and the initial landmarks. Second, we test if the ideal solution is a local minimum of the model as follows: starting in the correct position, deviation after convergence is compared to the ground truth. In both tests, FCA have shown superior performance w.r.t. grayscale, Gabor filters, gradient combinations and oriented pair filters [14] that we omit in the interest of space.

4.3.1 Convergence analysis

In this section, we report results on convergence after perturbing the ground truth parameters with gaussian noise (up to 7 pixels/landmark). Fig. (9) shows a perturbed ground truth image (9.a) and the same image after convergence (9.b). The convergence threshold has been set to 3 pixels/landmark w.r.t. the ground truth in terms of mean squared error. Table 4 shows the average results, for 100 random faces using different filtering techniques. The testing images do not include any of the subjects used in the training stage. All results are reported after 50 iterations of the algorithm.

As it is shown in Table 4, $FCA_{\lambda=1}(11,4)$ is the best performing representation, outperforming grayscale by 35%

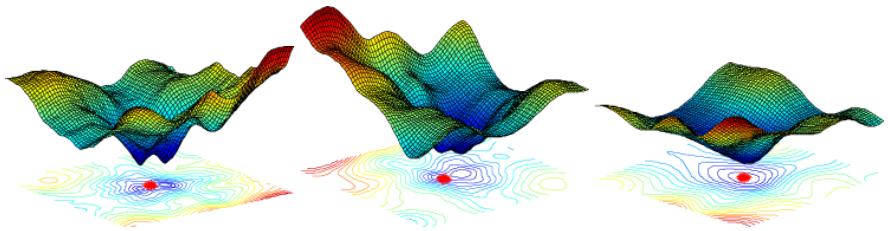

Figure 6. (1) Error surface for **grayscale**. (2) Error surface for **Gabor(8,4)**. (3) Error surface for $\boldsymbol{FCA_{\lambda=0}(11,4)}$.

Filter set	Conv(%)	Mean Error
Grayscale	36	3.52
Grayscale+Gradient(X,Y)	40	3.45
Gabor(8,4)	43	3.42
$FCA_{\lambda=0}(11,4)$	69	2.84
$FCA_{\lambda=1}(11,4)$	**71**	**2.82**

Table 4. AAM convergence tests for the CMU PIE database after random perturbation of the initial parameters. Gabor(X,Y) and FCA(X,Y) denote a set of Y filters with spatial scale X.

and Gabor(8,4) by 28% in the CMU PIE database [11]. Fig. (10) shows the corresponding error distributions for this test.

Figure 9. a)Random perturbation of the ground truth. b) Converged image.

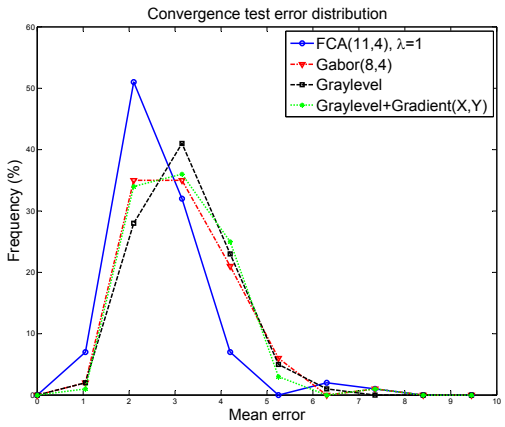

Figure 10. Error distribution for several filters in PIE database.

4.3.2 Stability of local minima

In this experiment, we test the stability of local minima. That is, we verify if there is a local minimum in the correct location (ground truth). The AAM model and fitting strategy is the same as the previous experiment [9]. We randomly select 100 subjects not present in the training set, and the fitting algorithm is initialized to the correct position (manually labeled). Table 5 shows the non-diverged tests percentage after 50 iterations.

Filter set	Conv(%)	Mean Error
Grayscale	60	3.28
Grayscale+Gradient(X,Y)	51	3.40
Gabor(8,4)	56	3.46
$FCA_{\lambda=0}(11,4)$	75	2.80
$FCA_{\lambda=1}(11,4)$	**86**	**2.69**

Table 5. AAM stability tests for the CMU PIE database. Gabor(X,Y) and FCA(X,Y) denote a set of Y filters with scale X.

In this test, FCA also outperforms any other single or multiband representation at any scale. Particularly, it is 26% better than grayscale and 30% than the best Gabor set. Fig. (11) shows the error distributions for different filters in this test.

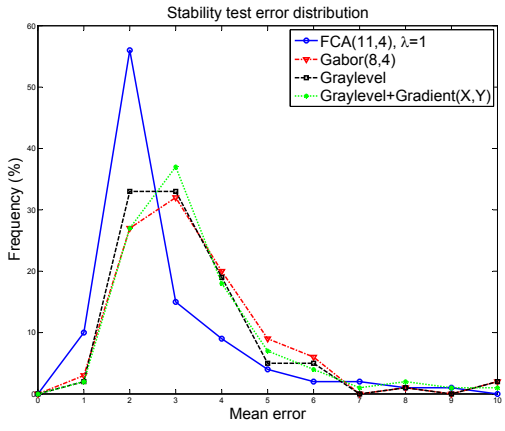

Figure 11. Stability test error distribution for several filters in PIE database.

5. Conclusions and Future Work

In this paper, we have proposed FCA to build a multi-band representation for appearance models that provides a more robust matching. FCA outperforms Gabor, oriented pair filters and grayscale representations. Additionally, we have introduced quantitative metrics for evaluating the error surface.

FCA has shown promising results, however future work should consider the use of different constraints for the filters (e.g. $vec(\mathbf{F})^T \mathbf{1}_{f_x \times f_y} = 1$). Also, it will be worth to explore the use of some recently proposed non-linear filters (e.g. [21, 20]) in the context of appearance models.

Acknowledgements This work was partially supported by the National Institute of Justice award 2005-IJ-CX-K067 and from NIH Grant R01 MH 051435. Takeo Kanade is partially supported by the National Science Foundation under Grant No. EEEC-0540865. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Thanks to Iain Matthews, Simon Baker and Simon Lucey for their helpful comments and discussions.

References

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

[2] H. Bischof, H. Wildenauer, and A. Leonardis. Illumination insensitive recognition using eigenspaces. *Computer Vision and Image Understanding*, 1(95):86 – 104, 2004.

[3] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of objects using view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999.

[5] T. Cootes and C. Taylor. On representing edge structure for model matching. In *CVPR*, 2001.

[6] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. tech. report. university of manchester. 2001.

[7] S. Darkner, R. Larsen, M. Stegmann, and B. Ersboll. Visual tracking of high dof articulated structures: An application to human hand tracking. In *2on International Workshop on Generative Model Based Vision (GMBV)*, 2004.

[8] F. de la Torre and M. J. Black. Robust parameterized component analysis: theory and applications to 2d facial appearance models. *Computer Vision and Image Understanding*, 91:53 – 71, 2003.

[9] F. de la Torre, J. Vitrià, P. Radeva, and J. Melenchón. Eigenfiltering for flexible eigentracking. In *International Conference on Pattern Recognition*, pages 1118–1121, 2000.

[10] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:891–906, 1991.

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. The cmu multi-pose, illumination, and expression (multi-pie) face database. Technical report, Carnegie Mellon University Robotics Institute.TR-07-08, 2007.

[12] J. P. Lewis. Fast normalized cross-correlation. In *Vision Interface*, 1995.

[13] H. Liu, B. Guo, and Z. Feng. Pseudo-log-polar fourier transform for image registration. *Signal Processing Letters, IEEE*, 13(1):17– 20, 2006.

[14] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 2001.

[15] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, Nov. 2004.

[16] S. J. McKenna, S. Gong, R. P. Würtz, J. Tanner, and D. Banin. Tracking facial feature points with Gabor wavelets and shape models. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication Crans-Montana, Switzerland*, pages 35–42, 1997.

[17] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. Technical Report WS00AVSR, Johns Hopkins University, CLSP, 2000.

[18] D. O'Leary and S. Peleg. Digital image compression by outer product expansion. *IEEE Trans. on Communications*, 31:441–444, 1983.

[19] R. Rao and D. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 12:441–444, 1995.

[20] S. Roth and M. Black. Fields of experts: A framework for learning image priors. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 860–867, 2005.

[21] S. Roth, L. Sigal, and M. Black. Gibbs likelihoods for bayesian tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume I, pages 886–893, 2004.

[22] A. Shashua and A. Levin. Linear image coding for regression and classification using the tensor-rank principle. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[23] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *IEEE Conference on Automatic Face and Gesture Recognition*, 2002.

[24] M. Stegmann and R. Larsen. Multi-band modelling of appearance. In *First International Workshop on Generative Model-Based Vision - GMBV*, 2002.

[25] N. F. Troje and T. Vetter. Representations of human faces. Technical Report 41, Max-Planck-Institut fr biologische kybernetik.

[26] M. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proc. European Conf. on Computer Vision*, 2002.

[27] A. Yilmaz and M. Gokmen. Eigenhill vs. eigenface and eigenedge. *Pattern Recognition*, 34(1):181184, 2001.