Image Statistics based on Diffeomorphic Matching

Guillaume Charpiat
Odyssee Laboratory
Ecole Normale Superieure
45 rue d'Ulm
75005 Paris, France
Guillaume.Charpiat@ens.fr

Olivier Faugeras
Odyssee Laboratory
INRIA Sophia Antipolis
2004 route des Lucioles
BP 93, 06902
Sophia-Antipolis Cedex, France
Olivier.Faugeras@sophia.inria.fr

Renaud Keriven
Odyssee Laboratory
ENPC
6 av Blaise Pascal
77455 Marne la Valle, France
Renaud.Keriven@ens.fr

Abstract

We propose a new approach to deal with the first and second order statistics of a set of images. These statistics take into account the images characteristic deformations and their variations in intensity. The central algorithm is based on non-supervised diffeomorphic image matching (without landmarks or human intervention). As they convey the notion of the mean shape and colors of an object and the one of its common variations, such statistics of sets of images may be relevant in the context of object recognition, both in the segmentation of any of its representations and in the classification of them. The proposed approach has been tested on a small database of face images to compute a mean face and second order statistics. The results are very encouraging since, wheras the algorithm does not need any human intervention and is not specific to face image databases, the mean image looks like a real face and the characteristic modes of variation (deformation and intensity changes) are sensible.

1. Introduction

How to find or recognize an object in an image? This is one of the most outstanding open problems in computer vision. Its solution will require a better understanding of the various possible visual aspects of a given object or a class of objects. For example, in the case of faces the description should include variations due to viewpoint, illumination, expression (happiness, surprise, ...), or the identity of the person. Like [3, 4] we think that statistics on images are necessary in order to tackle this problem. What we propose in this article is in a sense an extension to the set of images of an object of the work done on the statistics of 2D or 3D shapes [7, 1, 6]: by computing, from a set of images of a class of objects, the various ways these images can be warped onto one another we define and compute a mean image for that class and its second order statistics. Note that unlike previous approaches, e.g., [4] our approach does not require any manual intervention to identify landmarks or regions of interest. We work directly on the deformation fields which establish the correspondences between the whole images, since these fields are the fundamental elements of the problem. In order to do this we build upon previous work on non-supervised algorithms that build such correspondence fields between images, e.g., [7, 8, 5, 2]

In Section 2 we model the matching problem between two images and describe a variation of a matching algorithm proposed in [5] and analyzed in [2]. In Section 3 we use it to define and compute the mean image of a set of images and in Section 4 to define and compute its second order statistics.

2. Image matching

The main difficulty when defining the mean of several images is that this mean is supposed to *look like* each one of the images. This implies that the images have been registered and supposes the knowledge of a way to estimate the similarity of any couple of images. This is why we first consider the matching problem between only two images.

2.1. Basic framework

Let A and B be two images. We think of them as positive real functions functions defined in a rectangular subset Ω of the plane $\mathbb{R}^2$. We search for a deformation field $\mathbf{f}$ from Ω to Ω such that the warped image $A \circ \mathbf{f}$ resembles B. More precisely, we would like the field $\mathbf{f}$ to be smooth enough and invertible, i.e. it should be a diffeomorphism from the rectangular subset Ω to itself, which leads us to assume that the

diffeomorphism $\mathbf{f}$ equals the identity on the image boundary $\partial\Omega$. Other possibilities are offered by extending the images to a larger subset Ω_1.

In order to keep $\mathbf{f}$ continuous, we have to consider a regularizing term $R(\mathbf{f})$ on $\mathbf{f}$, for example

$$R(\mathbf{f}) = \|\mathbf{f} - Id\|_{\Omega}^{H^1}$$

where Id is the identity function on Ω and the H^1-norm is, for any field from Ω to $\mathbb{R}^2$,

$$\|a\|_{\Omega}^{H^1} = \int_{x\in\Omega} \|a(x)\|^2 + \|Da(x)\|^2 \, dx.$$

If we prefer to be sure $\mathbf{f}$ is a diffeomoprhism and remains invertible, we can consider $\|\mathbf{f} - Id\|_{\Omega}^{H^1} + \|\mathbf{f}^{-1} - Id\|_{\Omega}^{H^1}$, where $\mathbf{f}^{-1}$ is the inverse of $\mathbf{f}$.

Now we have to choose a criterion $C(A, B)$ which expresses the similarity between the two images A and B. The simplest one is

$$C(A, B) = \|A - B\|_{\Omega}^{L^2} = \int_{x\in\Omega} (A(x) - B(x))^2 \, dx,$$

but we prefer the following one, which has the advantage of being based on intensity variations and consequently the one of being contrast-invariant.

2.2. Local Cross-Correlation

Given a scale σ, the cross-correlation of two images A and B at point x is defined by:

$$CC(A, B, x) = \frac{v_{AB}(x)^2}{v_A(x)\ v_B(x)}$$

where $v_A(x)$ is the local spatial variance of A in a gaussian neighborhood of size σ centered on x, and $v_{AB}(x)$ the local covariance of A and B on the same neighborhood, i.e. we define:

$$g(x, y) = e^{\frac{\|x-y\|^2}{2\sigma^2}}$$

$$\mu(x) = \int_{y\in\Omega} g(x, y)\ dy$$

$$\bar{A}(x) = \frac{1}{\mu(x)} \int_{y\in\Omega} A(y)\ g(x, y)\ dy$$

$$v_A(x) = \epsilon + \frac{1}{\mu(x)} \int_{y\in\Omega} (A(y) - \bar{A}(x))^2\ g(x, y)\ dy$$

$$v_{AB}(x) = \frac{1}{\mu(x)} \int_{\Omega} (A(y) - \bar{A}(x))(B(y) - \bar{B}(x))\ g(x, y)\ dy$$

The positive constant ϵ is added only not to have a null divider in the expression of $CC(A, B, x)$. Given this, the local cross-correlation on the whole images are defined by [5]:

$$LCC(A, B) = \int_{x\in\Omega} CC(A, B, x)\ dx$$

2.3. The Image Matching Algorithm

The two-image matching algorithm consists in minimizing with respect to the deformation field $\mathbf{f}$ (initialized to the identity) through a multi-scale gradient descent the following energy (see [2] for details)

$$E(A, B, \mathbf{f}) = LCC(A \circ \mathbf{f}, B) + R(\mathbf{f})$$

Thus we obtain a field $\mathbf{f}$ which establishes the correspondences between the two images A and B.

3. The mean of a set of images

Now that we know how to compute a diffeomorphic matching between two images, we can try to infer from this a new algorithm for the computation of the mean of n images A_i indexed by $i \in \{1, \ldots, n\}$. This is not as easy as one could guess. We present here three different methods, from the simplest, naive one, to a less intuitive but far better one.

3.1. An intuitive algorithm: find the mean

We can first define the mean as the image M which looks the most like all the warped images, i.e., if we introduce n diffeomorphisms $\mathbf{f}_i$ in order to warp each image A_i on the mean M, we could minimize

$$\sum_i E(A_i \circ \mathbf{f}_i, M, \mathbf{f}_i)$$

with respect to both M and the fields $\mathbf{f}_i$. But how do we choose the initial image M? Besides, here is the main problem: we should not minimize the energy E with respect to an image. Indeed, if we consider the case where $n = 2$ and the two images are the same one translated by a few pixels, the gradient term due to the diffeomorphisms should move them so as to find the translation, but this is prevented by the minimization with respect to the mean image M, which, by averaging the intensities, introduces new contours induced by those in the two images. Consequently, contours from both images appear in the image M, and each of the two images "sees" its contours appear in M at the same location, and the diffeomorphisms will not evolve from the identity.

3.2. Another intuitive algorithm

We can then try to substitute in E an expression for M as a function of the diffeomorphisms and images, thus effectively eliminating the unknown M, in order not to have to take the derivative of E with respect to an image. For example, we can choose

$$M = \frac{1}{n} \sum_i A_i \circ \mathbf{f}_i$$

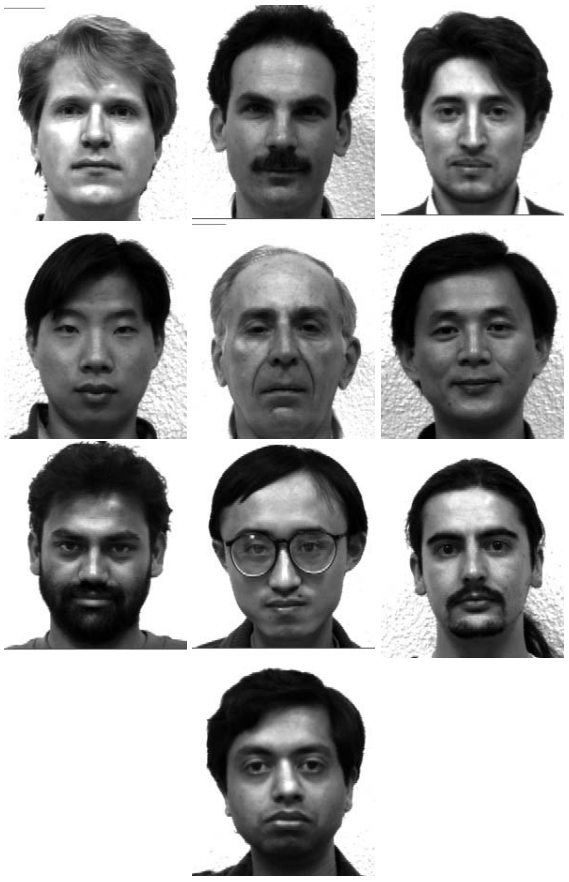

Figure 1. The ten face images.

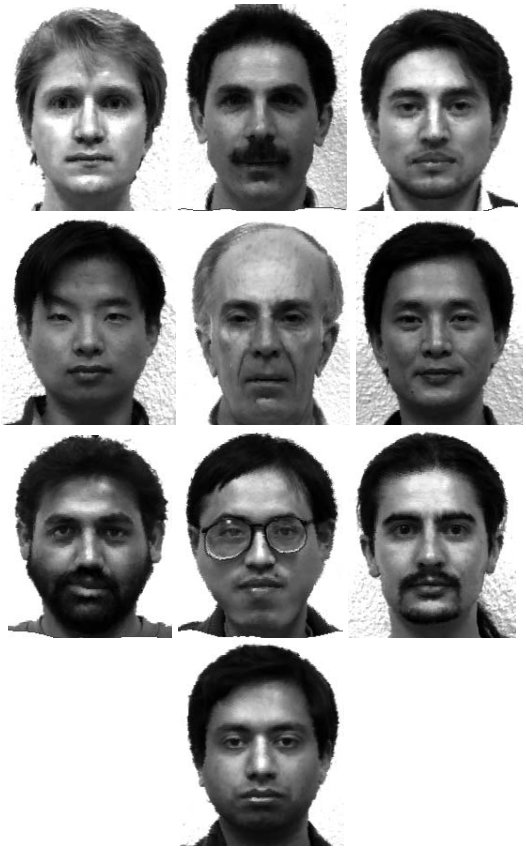

Figure 2. The ten warped images $A_i \circ \mathbf{f}_i$.

Figure 3. The mean of the previous ten faces.

and minimize with respect to the $\mathbf{f}_i$ the following criterium:

$$\sum_i E(A_i \circ f_i, \frac{1}{n}\sum_k A_k \circ \mathbf{f}_k, \mathbf{f}_i)$$

We then encounter another problem: we do not take the derivative of E with respect to an image, but we try to match for each i the warped image $A_i \circ \mathbf{f}_i$ and the mean $\frac{1}{n}\sum_i A_i \circ \mathbf{f}_i$. As $\frac{1}{n}\sum_i A_i \circ \mathbf{f}_i$ is the sum of the warped images, it contains in particular all the contours of $A_i \circ \mathbf{f}_i$, which means that we still have the same problem as before: the diffeomorphisms are immediatly stuck in a local minimum.

3.3. The final word: eliminating the mean

The problem comes mostly from the fact that we are trying to work directly on the mean of the images, whereas we should work only with the fields $\mathbf{f}_i$, which carry all the information about the problem. Indeed, the mean M contains much less information than the diffeomorphisms $\mathbf{f}_i$: for example the mean of a white disk on a black background and a black disk on a white background is uniformly grey and consequently has not a large LCC-correlation with the initial images. Therefore we should rather deal with pairs of warped images than with pairs of a warped image and the mean. The mean then becomes an auxiliary quantity, just computed at the end when the diffeomorphisms are known.

The algorithm proceeds as follows: initialize all deformation fields $\mathbf{f}_i$ to the identity, and let them evolve in a multiscale framework in order to minimize

$$\frac{1}{n-1}\sum_{i\neq j} LCC(A_i \circ \mathbf{f}_i, A_j \circ \mathbf{f}_j) \quad + \quad \sum_k R(\mathbf{f}_k)$$

Thus, at the end of the evolution, each $A_i \circ \mathbf{f}_i$ is supposed to look like each of the others, and the mean is naturally computed as $M = \frac{1}{n}\sum_i A_i \circ \mathbf{f}_i$. The regularizing term $\sum_i R(\mathbf{f}_i)$ implies that if several sets of fields $\mathbf{f}_i$ conduct to approximatively the same energy $\sum_{i\neq j} LCC(A_i \circ \mathbf{f}_i,\ A_j \circ \mathbf{f}_j)$ (for example by adding a common diffeomorphism $\mathbf{f}_c$ to every field and replacing $\mathbf{f}_i$ with $\mathbf{f}_i \circ \mathbf{f}_c$), then the most intuitive one is chosen (the one of least regularizing cost). In order to accelerate the process in pratice, we also impose the condition $\sum_i \mathbf{f}_i = 0$ at each time step by substracting the mean of the fields $\frac{1}{n}\sum_i \mathbf{f}_i$ to each of them.

3.4. Example

We have tested this algorithm on a face database from Yale[1]. More precisely, we have computed the mean face out of photographs of ten different people with similar expressions, approximatively the same illumination and position conditions, and the same size (195 * 231 pixels). The ten image A_i are shown in figure 1, the ten warped images $A_i \circ \mathbf{f}_i$ in figure 2, and their mean in figure 3.

Note the accuracy of the mean: it looks like a real face, its features are very sharp, not blurred at all (except the ears), thanks to the simultaneous accurate matching of all images. If we had used one of the algorithm centered on the mean image instead of the diffeomorphisms themselves, we would have obtained a completely blurred image because of non-corresponding edges of different images (the fields being stuck in local minima before starting to evolve), not far better than a bad simple average of every pixel of all initial images without warping.

The strange white curved line below the eyes of the mean comes from the reflects of the light into the eighth man's glasses, which the algorithm probably confused (and matched) with the brightness of the top of the ohter cheeks.

This computation took about 10 minutes on a standard workstation. Note once again the good job done by the diffeomorphisms $\mathbf{f}_i$, on figure 2, with in mind the fact that there is no human intervention to help the algorithm find the good correspondences, that the algorithm is absolutely not specific to face databases, and that there is no use of any kind of prior on the images.

4. Second order statistics of a set of images

Now that we are able to compute the mean image of a set of images, we would like to study its characteristic modes of variation. Indeed, the knowledge of only the mean may be not sufficient to have a good idea of the whole set of images. For example, there may exist some relevant typical kinds of changes in the shape of the intensity of an object, without the knowledge of which you may not be able to discuss the belonging of a new image to the class defined by the set of images you studied before.

As the information about the shape variations in the set of images A_i lies in the diffeomorphisms $\mathbf{f}_i$, we compute statistics on these warping fields. The same way, as the information about the intensity variations (changes of skin texture, of hair color...) lies in the intensity of the warped images $A_i \circ \mathbf{f}_i$, since when they are warped their pixels are corresponding, we also compute statistics on the intensity of the warped images. Finally, as there could be links between shape variations and intensity variations, we compute combined statistics.

4.1. Definition and computation of shape variations

The deformation fields are functions from a subset Ω of the plane $\mathbb{R}^2$ to itself, therefore the natural way to express correlation between two fields $\mathbf{a}$ and $\mathbf{b}$ is to compute their scalar product for the usual norm $L^2(\Omega \to \mathbb{R}^2)$:

$$\langle \mathbf{a} \,|\, \mathbf{b} \rangle_{L^2(\Omega\to\mathbb{R}^2)} = \frac{1}{|\Omega|}\int_\Omega \mathbf{a}(x)\cdot\mathbf{b}(x)\ dx$$

Since the mean $\bar{\mathbf{f}}$ of the fields $\mathbf{f}_i$ is 0 (see above), the (shape-)correlation matrix SCM defined by

$$SCM_{i,j} = \left\langle \mathbf{f}_i - \bar{\mathbf{f}} \,\middle|\, \mathbf{f}_j - \bar{\mathbf{f}} \right\rangle_{L^2(\Omega\to\mathbb{R}^2)}$$

can be simplified in

$$SCM_{i,j} = \langle \mathbf{f}_i \,|\, \mathbf{f}_j \rangle_{L^2(\Omega\to\mathbb{R}^2)}\,.$$

Then we diagonalize the correlation matrix SCM (its size $n \times n$ depends on the number of images, not on the number of pixels), and extract its eigenvalues σ_k and normalized eigenvectors $\mathbf{v}_k$. We obtain $n-1$ modes of deformation (one being null because of the linear constraint $\sum_i \mathbf{f}_i = 0$), and the k^{th} mode $\mathbf{m}_k$ is given by the coefficients of $\mathbf{v}_k$:

$$\mathbf{m}_k = \sum_i (\mathbf{v}_k)_i\ \mathbf{f}_i.$$

As statistics were made in the linear space $L^2(\mathbb{R}^2 \to \mathbb{R}^2)$, we can continuously apply a mode $\mathbf{m}_k$ to the mean image M with an amplitude α ($\in \mathbb{R}$) by computing the image $M \circ (Id + \alpha(\mathbf{m}_k - Id))$, and then produce animations of the deformations.

4.2. Example

These modes are illustrated in figure 4. Each column represents a mode, starting from the main one (leftmost column) to the one with the smallest eigenvalue, which is actually 0 since one mode is null (rightmost column). Each

[1] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

column is divided in five images: in the central image of each column, we represent the mean we computed before; in the images just above and underneath the mean, we represent the mode applied to the mean with amplitude $+\sigma_k$ and σ_k; and then with amplitude $+2\sigma_k$ and $-2\sigma_k$ in first and last image of each column, in order to exaggerate and better visualize the deformations.

Note that the images on the second and fourth lines still look like normal faces of various people; it is a very good point since they are supposed to be characteristic examples of what shape variations the mean face can undergo without getting out of the class of face images.

On the contrary, images on the first and last lines are stranger: even if we still recognize they look human a bit, we see immediately that there are not real; which is not the case of the other lines. This is also a very good point, since these images have been obtained by applying the characteristic modes two times too far (with amplitude $2\sigma_k$ instead of σ_k), which shows that the amplitudes of the deformations (the values of σ_k) are right, and shows that a set of images is well described by its characteristical shape variations.

4.3. Intensity variations

In order to take all the face variations into account, we should not only consider the shape variations (i.e. the diffeomorphisms) but also the intensity variations. As before, we can define an intensity-correlation matrix ICM on the intensity variations I_i:

$$I_i = A_i \circ \mathbf{f}_i - M$$

for the $L^2(\mathbb{R}^2 \rightarrow \mathbb{R})$ scalar product. Thus, we can compute the principal modes of intensity variations, which correspond to skin or hair changes for a shape-fixed head.

We can also combine shape and intensity variations. If we note $\sigma_S^2 = \frac{1}{n}\sum_i \|\mathbf{f}_i\|^2$ and $\sigma_I^2 = \frac{1}{n}\sum_i \|I_i\|^2$ the standard deviations of shapes and intensities, we can define a combined correlation matrix CCM

$$CCM = \frac{1}{\sigma_S^2} SCM + \frac{1}{\sigma_I^2} ICM$$

and proceed as before, compute and display principal modes of variations. This matrix can be considered as resulting from a inner product on the set of variations (shape and intensity): from two elements $(\mathbf{f}_i, I_i)$ and $(\mathbf{f}_j, I_j)$ in this set, we can compute their correlation:

$$\langle \mathbf{f}_i, I_i \,|\mathbf{f}_j, I_j \rangle = \frac{1}{\sigma_S^2} \langle \mathbf{f}_i \,|\mathbf{f}_j \rangle + \frac{1}{\sigma_I^2} \langle I_i \,|I_j \rangle$$

where the two coefficients stand for the relative variability of each component.

The results are shown on figure 5. Note again how these faces are realistic and diversified (hair, skin, illumination, mustache). We can see more various attitude than before, when we considered only shape variations. This is partly due to the fact that illumination and shadow carry information on the 3D shape (for example, the shape of the cheeks) which is not directly retrievable from the sharpest edges of 2D images.

5. Summary and Conclusions

We have defined and computed first and second order statistics of a set of images with a diffeomorphic matching approach (without landmarks or human intervention). We have tested this general approach, which is non specific to any set of images, on a face database, and the results are very encouraging: the mean face really looks like that of a real human being, with sharp contours, and the characteristical modes of variations (shape and intensity) are very sensible, which proves the quality of this approach. We insist on the fact that our methods are not specific to the particular case of face databases and do not use any prior on the kind of images. We are in the process of including these relevant statistics to segmentation and classification algorithms.

References

[1] G. Charpiat, O. Faugeras, and R. Keriven. Approximations of shape metrics and application to shape warping and empirical shape statistics. *Foundations of Computational Mathematics*, 2004. Accepted for publication.

[2] O. Faugeras and G. Hermosillo. Well-posedness of two non-rigid multimodal image registration methods. *Siam Journal of Applied Mathematics*, 64(5):1550–1587, 2004.

[3] U. Grenander. *General Pattern Theory*. Oxford University Press, 1993.

[4] P.L. Hallinan, G.G. Gordon, A.L. Yuille, P. Giblin, and D. Mumford. *Two- and Three-Dimensional Patterns of the Face*. A K Peters, 1999.

[5] Gerardo Hermosillo. *Variational Methods for Multimodal Image Matching*. PhD thesis, INRIA, The document is accessible at ftp://ftp-sop.inria.fr /robotvis/html/Papers /hermosillo:02.ps.gz, 2002.

[6] E. Klassen, A. Srivastava, W. Mio, and S.H. Joshi. Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):372–383, 2004.

[7] M. Miller and L. Younes. Group actions, homeomorphisms, and matching : A general framework. *International Journal of Computer Vision*, 41(1/2):61–84, 2001.

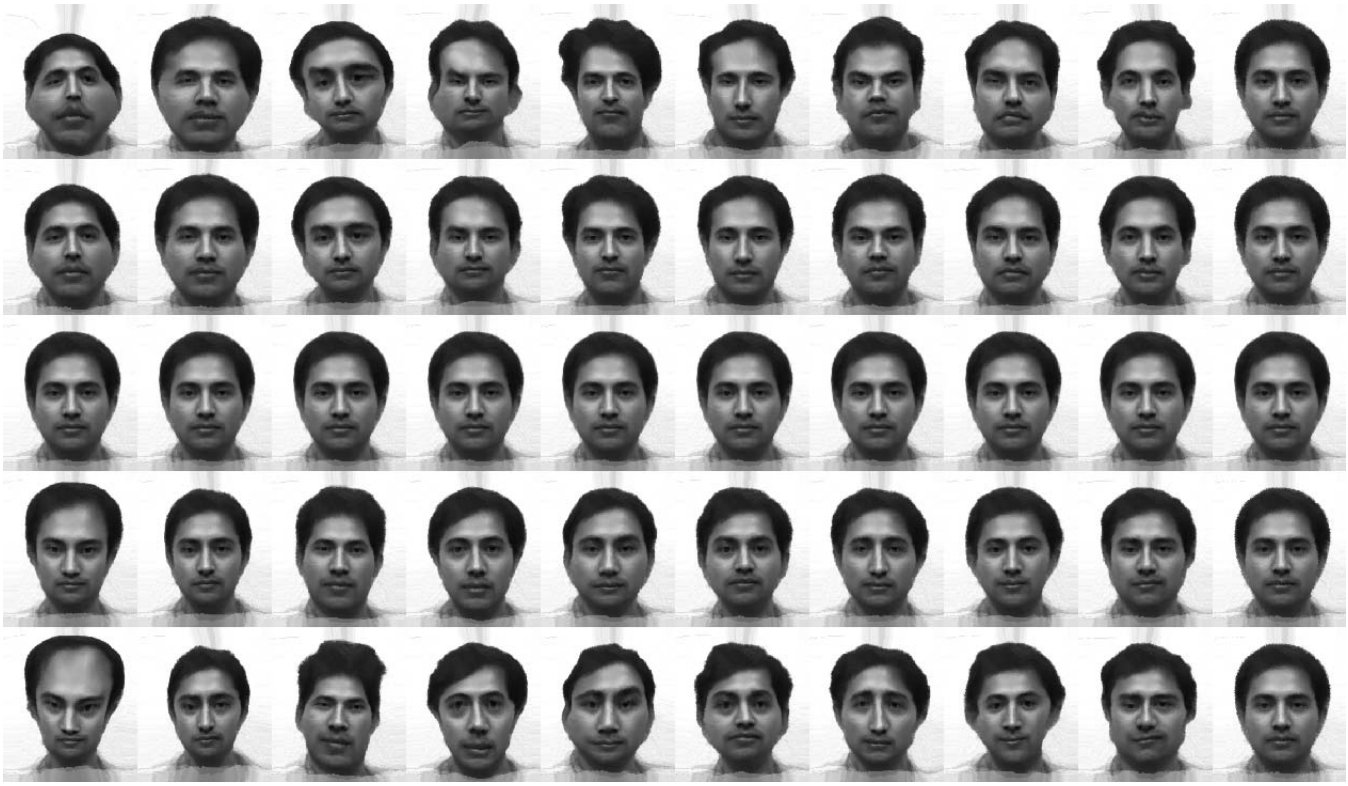

Figure 4. The shape modes of deformation of the previous set of images. Each column represents a mode, applied to the mean image with amplitude $\alpha = \{2\sigma_k, \sigma_k, 0, -\sigma_k, -2\sigma_k\}$.The (relative) values of the eigenvalues are, from left to right, 1, 0.5, 0.3, 0.1,. . . , 0.05, 0.

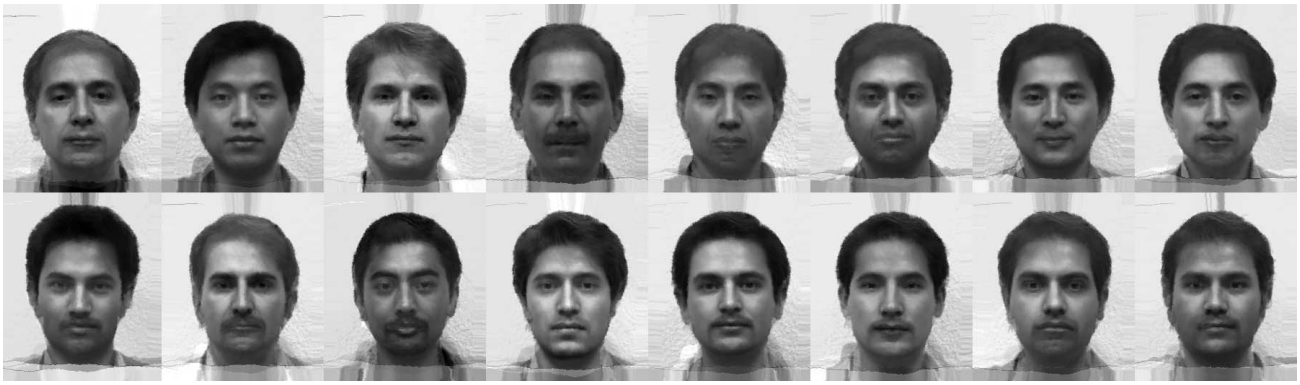

Figure 5. The eight non-zero combined modes of deformation of the same set of images without the subject with glasses. Each column represents a mode, applied to their mean image with amplitude $\alpha = \{\sigma_k, -\sigma_k\}$. The (relative) values of the eigenvalues are, from left to right, 1, 0.555, 0.505, 0.424, 0.286, 0.232, 0.162, 0.135.

[8] A. Trouvé and L. Younes. Metamorphoses through lie group action. *Foundation of Computational Mathematics*, 2005. To appear.