Bilinear Decomposition of 3-D Face Images: An Application to Facial Expression Recognition

Iordanis Mpiperis, Sotiris Malassiotis and Michael G. Strintzis

Informatics and Telematics Institute
Centre for Research and Technology Hellas
6^{th} Km Charilaou-Thermi Rd, Thessaloniki, Greece 57001
{iordanis, malasiot, michael}@iti.gr

Abstract

This paper describes a novel technique for decoupling two of the main sources of variation in 3-D facial structure, subject's identity and expression. Decoupling and controlling independently these factors is a key step in many practical applications and in this work it is achieved by modeling the face manifold with a bilinear model. Bilinear modeling, however, can only be applied to vectors, and therefore a vector representation for each face is established first. To this end, we use a generic face model that is fitted to each face under the constraint that anatomical points get aligned. The effectiveness and applicability of the proposed method is demonstrated with an application to facial expression recognition.

1. Introduction

Visual inspection of the face allows humans to recognize multiple characteristics of the individual, such as his/her identity, cognitive and psychological state. Although this seems trivial for humans, it is quite challenging for computers, both in the 2-D and 3-D case, due to the integration of unwanted variation in facial appearance and 3-D structure (*e.g.* changes in lighting conditions, head pose e.t.c).

To deal with this problem, many researchers try to devise descriptors robust to unwanted variation. For instance, geodesic polar representations [7], spherical canonical forms [1] and adapting geometric attributes [4] are descriptors that allow expression-invariant face recognition. An alternative approach is modeling the face manifold with appropriate models, most famous of which (e.g. Active Appearance Models [2]) are based on Principal Components Analysis. The advantage of this approach is that the model may be used not only for the recognition of face attributes but also in a variety of applications such as data compression and so on. The disadvantage however of the models used so far is that usually they are linear and they cannot separate the sources of variation. To this end, a few researchers have recently proposed non-linear models able to handle multiple sources of variation. For instance, Vasilescu *et al.* [11] proposed using the N-mode SVD tensor decomposition to separate the influence of identity, pose, illumination and expression in face appearance, while Wang *et al.* [12] proposed the Higher-Order SVD (Singular Value Decomposition) in order to recognize and synthesize facial expressions in 2-D images.

Motivated by the successful application of the aforementioned models to 2-D images, this paper describes a novel technique for modeling 3-D facial geometry and decoupling two of the main sources of variation, subject's identity and expression, by means of a bilinear model. Bilinear models were introduced by Tenenbaum and Freeman [10] to describe two-factor observations, where concepts like the "content" and "style" of observations should be analyzed and manipulated separately from each other. They are linear in either factor when the other is held constant and therefore they are simple in structure, computation and implementation; they can be trained with well known algorithms and they can model subtle interactions between factors. However, the bilinear decomposition may only be applied to vector representations of faces. Thus a deformable face model is fitted to each face and then the parameters of this model are used for the bilinear decomposition.

In the following, we describe the technique for fitting the deformable face model to a face surface using their geodesic polar representations [7], while in Section 3, we present the bilinear model and its training. Its use is finally demonstrated in Section 4 with an application to facial expression recognition.

2. Vector representation of faces

Vector representation of faces is achieved by fitting a parameterized deformable 3-D face model to each sample surface. Its parameters define bijectively its configuration and therefore they can be used for its vector representation. However, this is true only if the fit of the model satisfies the constraint of anatomical correspondence.

Fitting begins by defining a 3-D mesh M_0 with N vertices $\mathbf{v}_i$ (see Fig. 1). This base-mesh is subdivided using the Loop subdivision scheme [5] to give a more smooth and dense mesh. At each subdivision step, the vertices of the resulting 3-D mesh may be written as a linear combination of the vertices of the previous level mesh and eventually of the initial mesh M_0. After a few levels of subdivision (3 in our experiments) we result in a dense mesh, called subdivision-mesh, that serves as the deformable model, while the vertices of the base-mesh that define its form comprise the vector representation of the face.

Let $\widetilde{M}$ denote the subdivision-mesh and $\widetilde{\mathbf{v}}_i$, $i = 1 \dots S$ its vertices. If also h_{ij} are the coefficients of the linear combinations between the vertices of the base-mesh and the subdivision-mesh, then altogether we have

$$\widetilde{\mathbf{v}}_i = \sum_{j=1}^{N} h_{ij} \mathbf{v}_j . \tag{1}$$

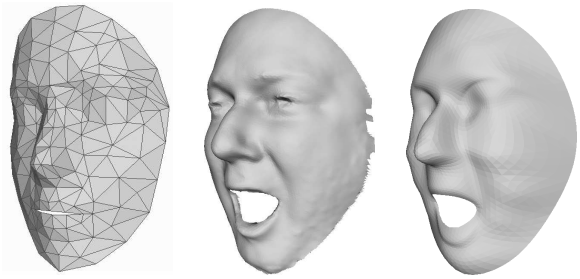
A common approach to model fitting (*e.g.* [8, 6]) is formulating it as an energy minimization problem. The energy is comprised of the Euclidean distances between the points of the model and their nearest counterparts on the surface. Considering (1) and defining the function $mc(i)$ that returns the index k of the facial point $\mathbf{p}_k$ nearest to the $\widetilde{M}$ vertex i, this energy term can be written as

$$E_{mc} = \sum_{i=1}^{S} \left(\sum_{j=1}^{N} h_{ij} \mathbf{v}_j - \mathbf{p}_{mc(i)} \right)^2 . \tag{2}$$

However, this energy term alone leads to an under-constrained problem whose solution may not represent a plausible human face (*e.g.* vertices may be set to disparate points and fold the triangles of the mesh). Therefore, a smoothness term that tries to prevent the model from distorting is also added to the energy formulation. Smoothness is defined as a measure of the elastic energy of the base-mesh that penalizes non-parallel displacements of the edges and is given by

$$E_e = \sum_{i=1}^{N} \frac{1}{N_i} \sum_{j \in \mathcal{N}_i} \left(\mathbf{v}_i - \mathbf{v}_j - \mathbf{v}_i^0 + \mathbf{v}_j^0 \right)^2 \tag{3}$$

where $\mathcal{N}_i$ is the set of $\mathbf{v}_i$'s neighbors, N_i is its cardinality and $\mathbf{v}_i^0$, $\mathbf{v}_j^0$ are the initial positions of the vertices.



a b c

Figure 1. Fitting the deformable model to a surface. a: base-mesh, b: target surface, c: subdivision-mesh fitted to the surface.

E_{mc} gives rise to forces that attract model vertices towards their nearest points on the surface instead of the anatomically corresponding points. This is not a problem if the model is relatively close to the surface, since nearest points and anatomically corresponding points almost coincide. But if the model is relatively far from the surface, vertices may be displaced so that anatomically erroneous correspondence is established. To overcome this problem, first we define a dense correspondence field between the model and the face taking into account their geodesic polar parameterizations [7]. That is, we add an extra energy term

$$E_c = \sum_{i=1}^{S} \left(\widetilde{\mathbf{v}}_i - \mathbf{p}_{c(i)} \right)^2 \tag{4}$$

where we assume that surface points $\mathbf{p}_{c(i)}$ correspond to model vertices $\widetilde{\mathbf{v}}_i$, if they have the same geodesic polar coordinates. (Correspondence is given by function $c(\cdot)$.) Again, this may result to poor anatomical correspondence if the model is too different from the face. To handle both this problem and the similar one regarding nearest points above, we adopt an iterative approach to perform fitting.

In each iteration, we compute the geodesic polar coordinates of the surface points and the subdivision-mesh vertices and we form the couples with the same coordinates. We also form the couples of vertices and their nearest points on the surface. Model parameters are then estimated by minimizing the energy function

$$E_{def} = \lambda_1 E_c + \lambda_2 E_{mc} + \lambda_4 E_e \ . \tag{5}$$

E_{def} is quadratic with respect to the unknown model parameters and therefore its minimization can be achieved easily by solving a simple linear system. Let $\hat{\mathbf{v}}[k] = \left[\hat{\mathbf{v}}_1[k]^T \dots \hat{\mathbf{v}}_N[k]^T \right]^T$ stand for the base-mesh vertices that minimize (5) in the k-th iteration and η be a step chosen in $(0, 1)$. Then, the base-mesh vertices $\mathbf{v} = [\mathbf{v}_1^T \dots \mathbf{v}_N^T]^T$ that are used for the vector representation of the surface may be

found using the following update rule

$$\mathbf{v}[k] = (1-\eta)\mathbf{v}[k-1] + \eta\hat{\mathbf{v}}[k] \tag{6}$$

upon convergence to a final position.

3. Modeling expression and identity variation

Using the vector representation described above, we may now model the face manifold by means of a bilinear model.

Let $\mathbf{v}^{xp}$ be the K-dimensional stacked column vector of the N base-mesh vertices of the facial surface of person p with expression x ($K = 3N$). Then each component v_k^{xp} is given by the general bilinear form [10]

$$\mathrm{v}_k^{xp} = \sum_{i=1}^{I}\sum_{j=1}^{J} w_{ijk} a_i^x b_j^p \tag{7}$$

where a_i^x and b_j^p are the control parameters that control expression and identity respectively, while w_{ijk} are the coefficients that model the interaction of the factors. Equation (7) shows that coefficients w_{ijk} are weighted symmetrically by a_i^x and b_j^p and thus this model is called *symmetric* in the literature.

The symmetric model is able to generalize equally on both directions, identity and expression. In practice however, this is true only if the model is trained with approximately equal number of samples with respect to identity and expression. But this is not valid for the majority of face databases, since most of them contain hundreds of individuals displaying few expressions, usually the 6 prototypical expressions proposed by Ekman [3]. To overcome this problem and achieve better generalization with respect to expressions, we may use an asymmetric model by letting mixing coefficients w_{ijk} vary with the expression control parameters a_i^x, that is

$$a_{kj}^x = \sum_{i=1}^{I} w_{ijk} a_i^x \ . \tag{8}$$

Using the above definition and (7), the vector representation of the face is now given by

$$\mathrm{v}_k^{xp} = \sum_{j=1}^{J} a_{kj}^x b_j^p \qquad\qquad \mathbf{v}^{xp} = \mathbf{A}^x \mathbf{b}^p \tag{9}$$

where now matrix $\mathbf{A}^x$ controls expression. The identity of the face is still controlled by vector $\mathbf{b}^p$.

Let us assume that there exist T faces in our database belonging to T_p individuals and depicting one of T_x possible expressions. Let also $h_{xp}[t]$ be a zero-one function that is one if the t-th face $\mathbf{v}[t]$ belongs to individual p with expression x. Unknown coefficients arise from the minimization of the total squared error [9]

$$E_a = \sum_{t=1}^{T}\sum_{x=1}^{T_x}\sum_{p=1}^{T_p} h_{xp}(t)\left(\mathbf{v}(t) - \mathbf{A}^x\mathbf{b}^p\right)^2 \ . \tag{10}$$

By differentiating E_a with respect to $\mathbf{A}^x$ and $\mathbf{b}^p$ and setting the partial derivatives equal to zero, we end up with equations

$$\mathbf{A}^x = \left(\sum_{p=1}^{T_p} \mathbf{m}_{xp}\mathbf{b}^{pT}\right)\left(\sum_{p=1}^{T_p} n_{xp}\mathbf{b}^p\mathbf{b}^{pT}\right)^{-1} \tag{11}$$

$$\mathbf{b}^p = \left(\sum_{x=1}^{T_x} n_{xp}\mathbf{A}^{xT}\mathbf{A}^x\right)^{-1}\left(\sum_{x=1}^{T_x} \mathbf{A}^{xT}\mathbf{m}_{xp}\right) \tag{12}$$

where $n_{xp} = \sum_{t=1}^{T} h_{xp}[t]$ and $\mathbf{m}^{xp} = \sum_{t=1}^{T} h_{xp}[t]\mathbf{v}[t]$.

Matrices $\mathbf{A}^x$ and vectors $\mathbf{b}^p$ may now be found by iterating equations (11) and (12) according to an update rule similar to (6) [1].

4. Facial expression recognition

In this section, we present an application of the deformable and bilinear model to facial expression recognition using the BU-3DFE face database [13]. BU-3DFE contains 2,500 face scans of 100 subjects, who display the 6 prototypical expressions of *anger*, *fear*, *disgust*, *happiness*, *sadness* and *surprise*.

First, the base-mesh M_0 is built by selecting $N = 169$ vertices lying on an average facial surface and then the vector representation of every face is established by fitting the deformable model as described in Section 2.

Then, in order to obtain statistically safe experimental results, we follow the 10-fold cross-validation approach. In each experiment, BU-3DFE subjects are divided randomly in two sets, a training set consisting of 90 subjects and a test set consisting of the rest 10 subjects. The training set is used to train the asymmetric bilinear model. That is, we estimate the 6 matrices $\mathbf{A}^x$ corresponding to the $T_x = 6$ possible expressions and the $T_p = 90$ identity control vectors $\mathbf{b}^p$ corresponding to the subjects of the training set. The number of columns of $\mathbf{A}^x$ and the dimension of $\mathbf{b}^p$ is set to 80 while the number of rows K is equal to the triple of vertices number, $K = 507$. Entries of $\mathbf{A}^x$ and $\mathbf{b}^p$ are initialized randomly and then they are computed by iterating (11) and (12) until the relative change in their Frobenius norm gets below a threshold (0.01).

[1] Convergence is guaranteed if J, the dimensionality of vector $\mathbf{b}^p$, is less than or equal to T_p, the number of individuals

Table 1. Expression recognition based on asymmetric bilinear model.

In\Out	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**82.5**	0.0	0.0	0.0	17.5	0.0
Disgust	0.0	**100.0**	0.0	0.0	0.0	0.0
Fear	0.5	0.0	**98.0**	0.0	1.5	0.0
Happiness	0.0	1	0.0	**99.0**	0.0	0.0
Sadness	33.0	0.0	3.5	0.0	**63.5**	0.0
Surprise	0.0	0.0	0.0	0.0	0.0	**100.0**

Estimated matrices $\mathbf{A}^x$ and vectors $\mathbf{b}^p$ are then used to build a Maximum Likelihood classifier. We assume that the vertex vector $\mathbf{v}$ defining the facial surface of person p with expression x is a random vector with spherical gaussian probability density function (p.d.f.) centered at the prediction of the asymmetric bilinear model. That is

$$f(\mathbf{v}|p,x) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^K} e^{-\frac{1}{2\sigma^2}||\mathbf{v}-\mathbf{A}^x\mathbf{b}^p||^2} \qquad (13)$$

where $\sigma^2 = 10^5$ is the error variance. Using the Total Probability Theorem, the conditional p.d.f. of $\mathbf{v}$ assuming expression x may now be written as

$$f(\mathbf{v}|x) = \sum_{p=1}^{T_p} \frac{1}{T_p} f(\mathbf{v}|p,x) \ . \qquad (14)$$

Now, the expression of a novel test face may be classified simply to the prototypical expression with the greatest likelihood, that is the expression x_i for which

$$f(\mathbf{v}|x_i) > f(\mathbf{v}|x_j) \ \ \forall\, x_j \neq x_i \ . \qquad (15)$$

The above experiments are repeated on several randomly chosen subdivisions of training and test sets, under the constraint that all subjects are included at least once in the test set. The recognition results are averaged and presented in Table 1 showing a total average recognition rate of 90.5%. The highest misclassification occurs between the expressions of *anger* and *sadness*. The main difference between these expressions lies mostly on the configuration of the eyebrows, which means that our deformable face model cannot localize them accurately enough. Perhaps this problem may be resolved by the introduction of color information to the face model.

5. Conclusion

In this paper, we presented a technique for modeling face geometry with a bilinear model that allows decoupling identity and expression. First we proposed a method for establishing a vector representation of the face, which is necessary for bilinear decomposition and then we demonstrated an application of bilinear modeling to facial expression recognition. This application showed the advantage of separate control, since the same method could be used for expression-invariant recognition simply by interchanging the roles of expression and identity.

References

[1] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Expression-invariant representations of faces. *IEEE Trans. on Image Processing*, 16(1), January 2007.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[3] P. Ekman and W. Friesen. *Facial Action Coding System (FACS): Manual*. CA: Consulting Psychologists Press, Palo Alto, 1978.

[4] X. Li and H. Zhang. Adapting geometric attributes for expression-invariant 3D face recognition. In *IEEE Int. Conf. on Shape Modeling and Applications*, pages 21–32, 2007.

[5] C. Mandal, H. Qin, and B. C. Vemuri. Novel FEM-based dynamic framework for subdivision surfaces. *Computer-Aided Design*, 32(8):479–497, 2000.

[6] D. Metaxas and I. Kakadiaris. Elastically adaptive deformable models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(10):1310–1321, October 2002.

[7] I. Mpiperis, S. Malassiotis, and M. G. Strintzis. 3D face recognition with the geodesic polar representation. *IEEE Trans. on Information Forensics and Security*, 2(3):537–547, September 2007.

[8] C. Shelton. Morphable surface models. *International Journal of Computer Vision*, 38(1):75–91, 2000.

[9] J. Tenenbaum and W. Freeman. Separating style and content. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 662–668, 1997.

[10] J. Tenenbaum and W. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12:1247–1283, 2000.

[11] M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 93–99, June 2003.

[12] H. Wang and N. Ahuja. Facial expression decomposition. In *Proc. Ninth IEEE Int. Conf. on Computer Vision*, volume 2, pages 958–965, October 2003.

[13] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition (FG06)*, pages 211–216, April 2007.