A similarity-based method for the generalization of face recognition over pose and expression

Sharon Duvdevani-Bar, Shimon Edelman,
A. Jonathan Howell, Hilary Buxton

CSRP 480

January 1998

ISSN 1350–3162

UNIVERSITY OF

SUSSEX
AT BRIGHTON

Cognitive Science
Research Papers

A similarity-based method for the generalization of face recognition over pose and expression *

Sharon Duvdevani-Bar

Department of Applied Mathematics
Weizmann Institute
Rehovot 76100, Israel
sharon@wisdom.weizmann.ac.il

Shimon Edelman, A. Jonathan Howell, Hilary Buxton

School of Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton BN1 9QH, UK
[shimone,jonh,hilaryb]@cogs.susx.ac.uk

Abstract

Human observers are capable of recognizing a face seen only once before when confronted with it subsequently under different viewing conditions. We constructed a working computational model of such generalization from a single view, and tested it on a homogeneous database of face images obtained under tightly controlled viewing conditions. The model effectively constructs a view space for novel faces by interpolating view spaces of familiar ones [5]. Its performance – 30% error rate in one out of 18 recognition, and 8% in one out of three discrimination – is encouraging, given that it reflects generalization from a single view/expression to a range of $\pm 34^\circ$ rotation in depth and to two additional expressions. For comparison, human subjects in the one out of three task involving only viewpoint changes exhibit a 3% error rate [11].

1 Introduction

In the past decade, developments in the theory of object recognition have elucidated the process whereby recognition can be generalized to novel views of objects, provided that these are contained in the linear span of the familiar views [16]. Although the linear combination technique works only for the case of representation by coordinates of fiducial features and under orthographic projection, in the more general case of arbitrary features and an unconstrained imaging process generalization is still feasible, if the novel views interpolate the familiar ones [13]. In the present paper, we show that recognition can be generalized from *a single given view*, if the notion of interpolation is applied to entire *view spaces*, rather than to individual views, of faces. The theoretical exposition below follows the ideas described in [5], which are, in turn, related to the notions of class-based processing [10] and recognition by prototypes [3, 17, 4].

Consider the multidimensional space of measurements (reflected, e.g., in the pixel values of an image) performed by a visual system upon the world. A scene such as a view of an object corresponds to a single point in the measurement space, and a smoothly changing scene such as a sequence of views of an object rotating in front of the observer — to a smooth manifold that we call the *viewspace* of the object. The dimensionality of the viewspace depends on the number of degrees of freedom of the object; a rigid object rotating around a fixed axis gives rise to a one-dimensional viewspace (see the curve labeled $\mathcal{V}_1$ in Figure 1).

By continuity, the viewspaces of two nearly identical shapes will be very close to each other; a smooth deformation of the object will result in a concomitant smooth evolution of its viewspace (if the measurement functions are themselves smooth). This observation can be turned into a computational basis for the treatment of novel objects [5]. Suppose that a system has internalized the viewspaces of a number of object classes; it can then process a novel view of a novel object intelligently, to the extent that it resembles the familiar objects (see Figure 1). For this to work, the concept of similarity must be given a concrete interpretation in terms of the measurement

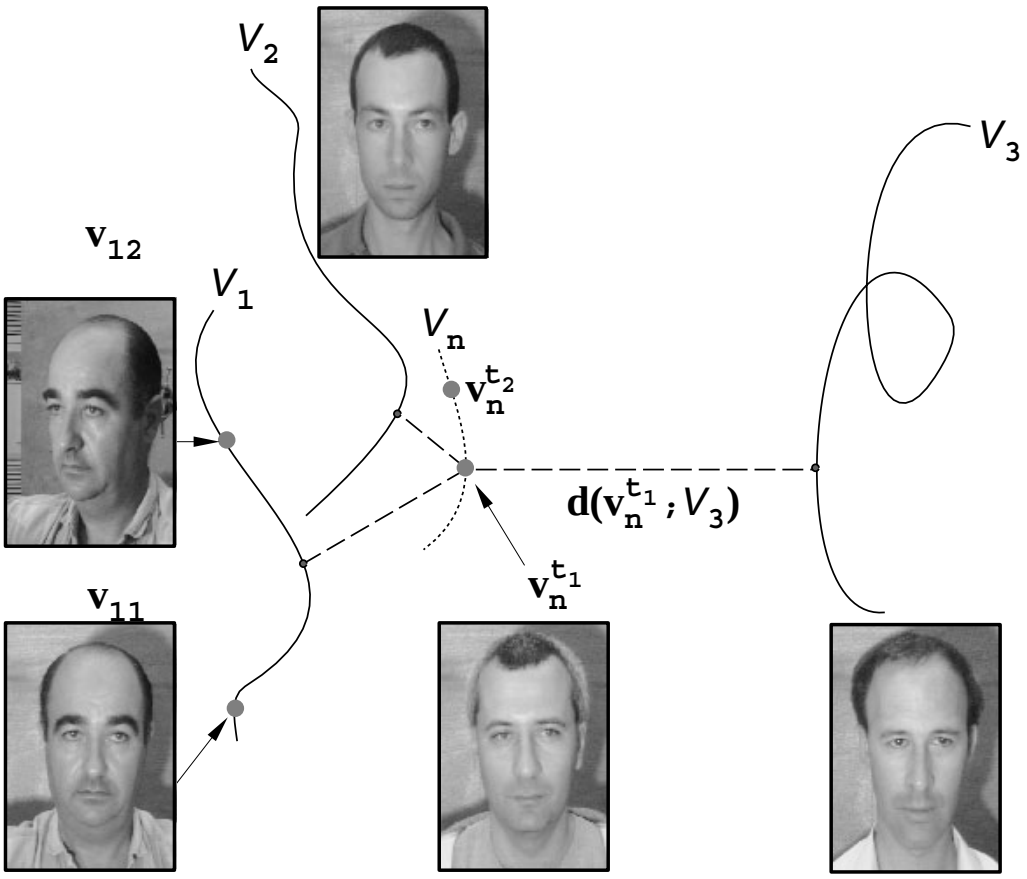

Figure 1: Interpolation of prototypical viewspaces (after [5]). The change in the view (appearance) of a person unfamiliar to the system (that is, previously seen from only one viewpoint) can be estimated by interpolating corresponding changes in the appearance of reference (prototype) faces (seen before from many viewpoints).

space. A computational mechanism suitable for this purpose is *interpolation*.

2 Viewspace interpolation mechanism

The interpolation of viewspaces involves irregularly spaced data. Among the many interpolation methods that can treat such data [1], we chose inverse-distance weighting [15, 7] — a simple algorithm in which the contribution of a known data point to the interpolated value at the test point is inversely proportional to the distance between the two. Note that our data "points" are actually entire manifolds – the viewspaces of the reference faces. Accordingly, the success of the interpolation approach here depends on the availability of a mechanism for dealing with viewspaces of individual familiar faces. Because such a mechanism has been discussed extensively elsewhere [13, 6], we treat it here as a "black box" module (cf. Figure 2) that can be trained to output a constant for different views of some target object and lower values for views of other objects. Because the output of such a module for a given image covaries monotonically with its dissimilarity (i.e., distance) from the viewspace of the object on which the module had been trained [6], it is precisely the quantity suitable for the inverse-distance weighted interpolation.

Consider a system composed of k modules, each trained to output 1 for a number of representative views of some reference object. As observed above, the output of the i'th module for a given test view $\mathbf{v}_n^t$ of a novel object, $x_i(\mathbf{v}_n^t)$, can serve as an indicator of the relevance of the i'th prototypical viewspace $\mathcal{V}_i$ to estimating the structure of the viewspace of the novel object $\mathcal{V}_n$. Consequently, the weight of $\mathcal{V}_i$ in determining the shape of $\mathcal{V}_n$ should be set to $x_i(\mathbf{v}_n^t)$.

We now apply this principle to the computation of a quantity Y that is intended to remain constant over changes in the test view $\mathbf{v}_n^t$ of a novel object. First, we compute the vector of responses of the k modules to a test view t_1; denote it by $\mathbf{w} = \mathbf{x}(\mathbf{v}_n^{t_1})$.

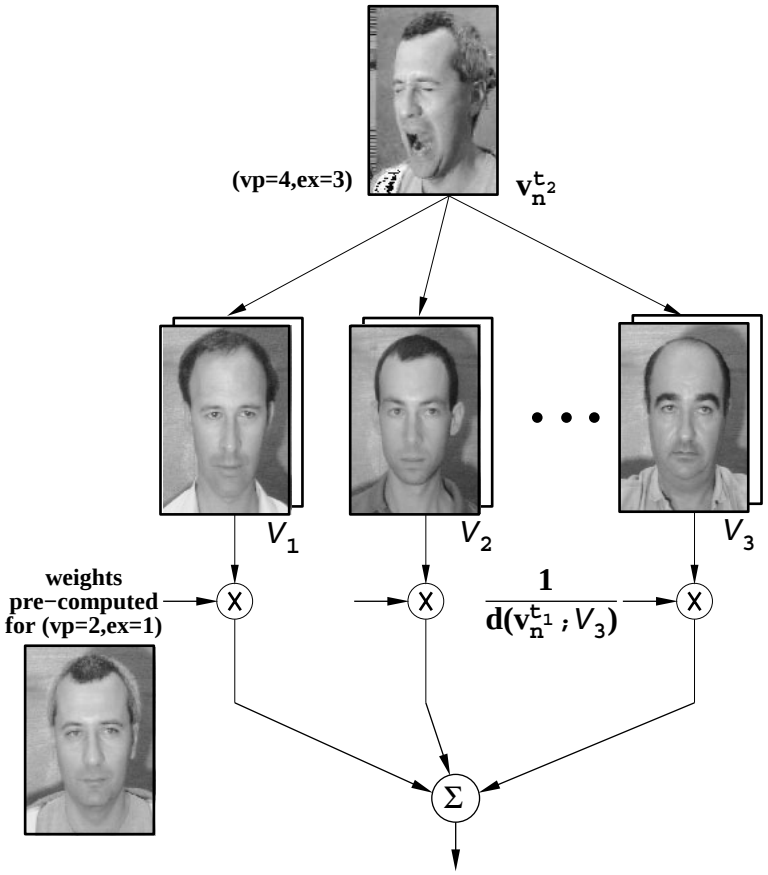

Figure 2: A mechanism for estimating the viewspace of a novel face by interpolating the viewspaces of several familiar ones. Inverse-distance weighting [15] is used to combine the outputs of a few "black boxes" — classifiers tuned to the reference faces [6]. This scheme can estimate the viewspace of a novel object by interpolation of the familiar ones.

Now, the estimate of Y for another test view t_2 is $Y(\mathbf{v}_n^{t_2}) = \mathbf{w}^T \mathbf{x}(\mathbf{v}_n^{t_2})$, where T denotes the transpose. Note that the weights are pre-computed for a certain input, then used for other inputs (i.e., in other parts of the input space). Clearly, $Y(\mathbf{v}_n^{t_2})$ will remain approximately constant, as long as the test view $\mathbf{v}_n^{t_2}$ is not too far from the view $v_n^{t_1}$ used to estimate the weights $\mathbf{w}$, and as long as the novel object is not too different from at least some of the reference ones.

3 Computational experiments

The experiments we describe next were intended to validate this approach on images of faces that differed along two dimensions: orientation (rotation of the head around the vertical axis) and expression. A subset of the images from the 28-person Weizmann FaceBase [11] was used in the experiments (Figure 3, left). The images were cropped and subsampled to a size of 100×100, then convolved with a bank of Gabor filters (the arrangement termed $A3$ in [8]). The filters were at four scales and three orientations, and formed a sparse, non-overlapping grid to provide 510 coefficients per image (Figure 3, right). Fifteen images (corresponding to all the combinations of five orientations and three expressions) of each of 10 faces were used to train the 10 reference-face modules; the remaining 18 faces (270 images altogether) were used to test the generalization ability of the system.

For each of the 18 test faces, the image corresponding to VP=2 (full face orientation) and EX=1 (neutral expression) was used as the single view from which generalization was to be carried out. The vectors of 10 module responses to that single view of each of the 18 test faces were pre-computed and used as the sets of weights in the generalization test stage. During testing, the system computed the weighted sum of the 10 module responses using each of the 18 sets of weights in turn. The set that yielded the highest sum (out of

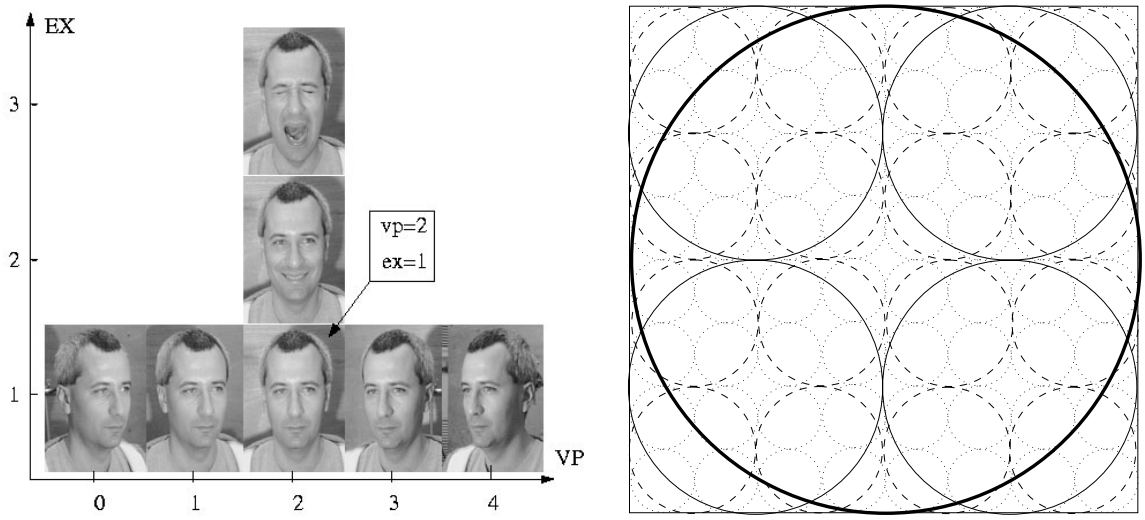

Figure 3: *Left:* the dimensions of variation in the face data used in the present study. For each of the 28 persons included in the database (a subset of the Weizmann FaceBase [11]), we used 15 face images (5 viewing positions in increments of 17° × 3 expressions). *Right:* the grid of filters that served for the preprocessing; see [8] for details.

the 18 possibilities) determined the identity of the test view.

The performance of this method is illustrated in Figure 4, which shows the generalization error as a 2D surface (plotted vs. VP, or viewing position, and EX, or expression), as well as the marginal means (over VP and over EX). The mean error rate over all 15 views was about 31% (details in Table 1 and Figure 5).

	1	2	3
0	0.5556	0.5000	0.5556
1	0.1667	0.1111	0.2222
2	0.0000	0.0556	0.0556
3	0.2222	0.3889	0.3333
4	0.5556	0.4444	0.4444

Table 1: Error rate vs. VP and EX (these data are also plotted in Figure 4). The mean error rate over the five viewing positions and the three expressions was 0.3074.

To assess the significance of this result, we conducted two additional experiments. First, we trained 18 radial basis function networks [13] on the same single given view/expression (VP=2, EX=1) of each of the test faces used before (because there was just one example, the networks had one center or basis function each). The identity of each test image was then decided by an 18-way winner take all procedure, resulting in a mean error rate of 47%. Compared to the performance of state of the art face recognition systems, this is a high error rate, which would have been even higher if more faces were used for testing. It must be noted, however, that significantly lower error rates have been achieved so far only by systems that rely on additional information (multiple views of faces, or a 3D model of an average face), and use complicated computational procedures (correspondence based on optic flow, or elastic matching). Furthermore, human subjects (unlike those systems) perform relatively poorly on generalization from a single view, if the task involves discrimination among more than a few faces [12].

To clarify this latter point, we compared the performance of the present system with that of human subjects, as reported in [11]. In that study, subjects carried out a 3-way discrimination of faces drawn from the same database we used here, achieving about 3% error rate for generalization over viewing position. To facilitate the direct comparison of the error rate of the system to that of the human subjects, we re-run the experiment using 3-way (instead of 18-way) discrimination. The mean error rate exhibited by the system (over the 816 triplets, or all possible combinations of three out of 18 faces) was about 8% (Figure 6).

4 Discussion and summary

We described a computationally viable method for the generalization of face recognition over differences in pose and facial expression, from a single given view of novel faces. According to this method, prior ex-

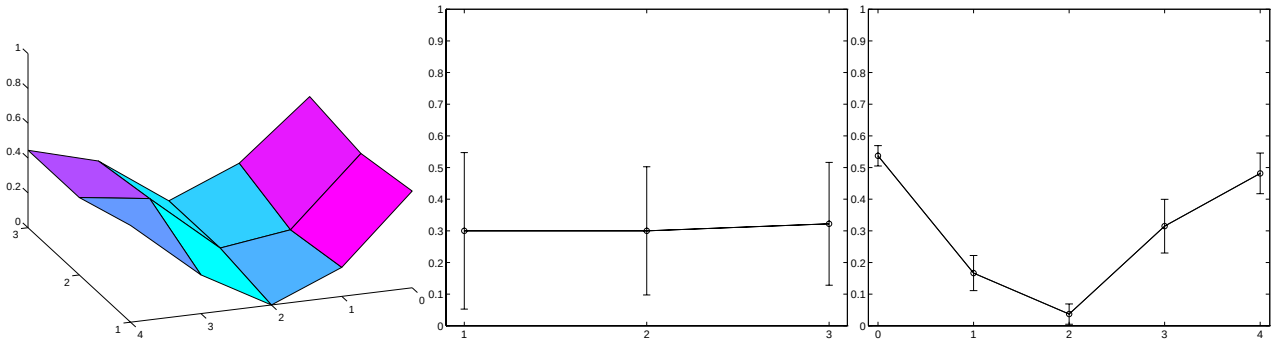

Figure 4: *Left:* a surface plot of the error rate vs. VP and EX (the numbers are listed in Table 1). *Middle:* error rate vs. VP, averaged over the three different values of EX. *Right:* error rate vs. EX, averaged over the five different values of VP. The mean error rate over the five viewing positions (spanning a range of $\pm 34^\circ$ in orientation), and the three expressions was 0.3074. The error bars correspond to ± 1 standard error of the mean computed over the 18 test faces.

perience with similar objects (i.e., other faces seen in a variety of conditions) serves to guide the system in its treatment of the stimulus. Since the introduction of this concept of so-called class-based processing [10, 14, 2, 11], several applications to face recognition and related problems have been published [17, 4, 3]. Typically, these methods rely on the establishment of a dense correspondence field, before any recognition or generalization is attempted. Approaches that gave up this constraint showed a certain promise [9], but could not compete, performance-wise, either with the human subjects, or with the more sophisticated correspondence-based methods.

In the present work, the employment of a front end containing Gabor filters at multiple scales and orientations [8] served to reduce the need for detailed pixel-by-pixel correspondence, and allowed the viewspace interpolation method [5] to be utilized to its full potential. We conjecture that a further improvement in the front-end measurement stage, combined with a more advanced approach to interpolation (which is currently done by inverse-distance weighting), will close most of the remaining gap between the system's 3-way discrimination error (8%) and the error exhibited by human subjects (3%).

References

[1] P. Alfeld. Scattered data interpolation in three or more variables. In T. Lyche and L. Schumaker, editors, *Mathematical Methods in Computer Aided Geometric Design*, pages 1–33. Academic Press, New York, 1989.

[2] R. Basri. Recognition by prototypes. A.I. Memo No. 1391, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1992.

[3] R. Basri. Recognition by prototypes. *International Journal of Computer Vision*, 19(147-168), 1996.

[4] D. Beymer and T. Poggio. Image representations for visual learning. *Science*, 272:1905–1909, 1996.

[5] S. Edelman and S. Duvdevani-Bar. Similarity-based viewspace interpolation and the categorization of 3D objects. In *Proc. Similarity and Categorization Workshop*, pages 75–81, Dept. of AI, University of Edinburgh, 1997.

[6] S. Edelman and S. Duvdevani-Bar. Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, 9:701–720, 1997.

[7] W. J. Gordon and J. A. Wixom. Shepard's method of 'Metric Interpolation' to bivariate and multivariate interpolation. *Mathematics of Computation*, 32:253–264, 1978.

[8] A. J. Howell and H. Buxton. Receptive field functions for face recognition. In *Proc. 2nd Int. Workshop on Parallel Modelling of Neural Operators for Pattern Recognition (PAMONOP)*, pages 83–92, Faro, Portugal, 1995.

[9] M. Lando and S. Edelman. Receptive field spaces and class-based generalization from a single view in face recognition. *Network*, 6:551–576, 1995.

[10] Y. Moses. *Computational approaches in face recognition*. PhD thesis, Feinberg Graduate School of the Weizmann Institute of Science, 1993.

[11] Y. Moses, S. Ullman, and S. Edelman. Generalization to novel images in upright and inverted faces. *Perception*, 25:443–462, 1996.

[12] A. J. O'Toole, H. Bülthoff, and C. L. Walker. Face recognition across viewpoint. MPIK TR 21, Max Planck Institut für biologische Kybernetik, Tübingen, Germany, September 1995.

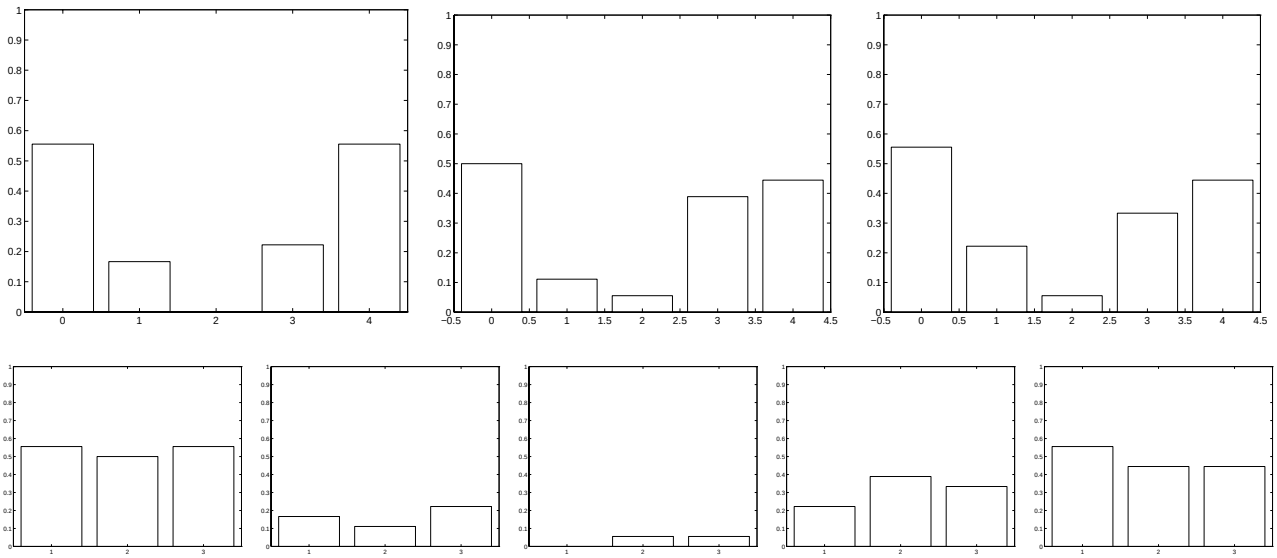

Figure 5: *Top:* error rates obtained for the different values of VP, while holding EX fixed. *Bottom:* results obtained for the different values of EX, while holding VP fixed.

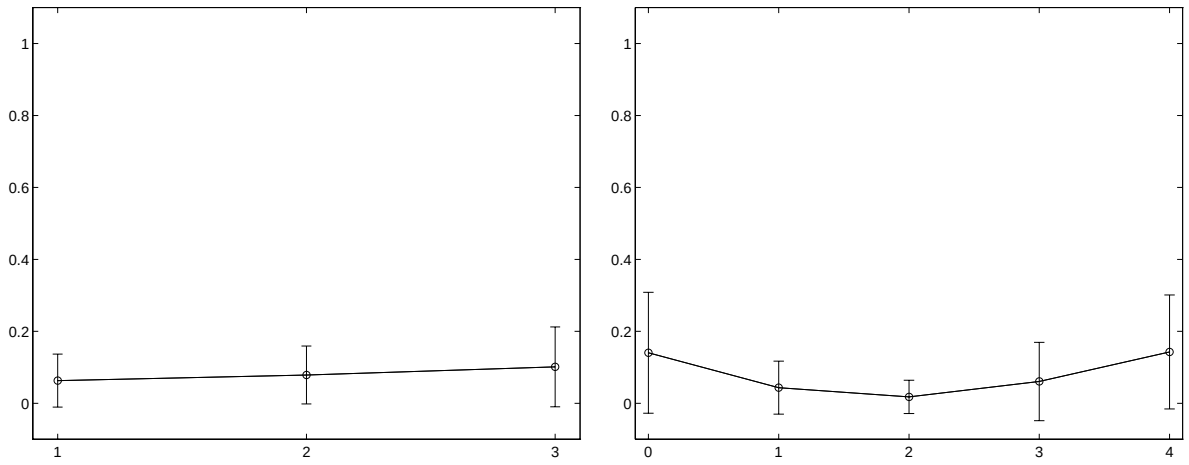

Figure 6: Performance of the system in the 3-way classification task. *Left:* error rate vs. VP, averaged over the three different values of EX. *Right:* error rate vs. EX, averaged over the five different values of VP. The mean error rate over the five viewing positions and the three expressions was 0.08. Human subjects in a comparable task exhibited an error rate of about 0.03 [11].

[13] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[14] T. Poggio and T. Vetter. Recognition and structure from one 2D model view: observations on prototypes, object classes, and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[15] D. Shepard. A two-dimensional interpolation function for irregularly spaced data. In *Proc. 23rd National Conference ACM*, pages 517–524. ACM, 1968.

[16] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005, 1991.

[17] T. Vetter and T. Poggio. Image synthesis from a single example image. In B. Buxton and R. Cipolla, editors, *Proc. ECCV-96*, number 1065 in Lecture Notes in Computer Science, pages 652–659, Berlin, 1996. Springer.