Vision Research, *in press*

Matching Expression Variant Faces

Aleix M. Martínez
Department of Electrical Engineering
The Ohio State University
aleix@ee.eng.ohio-state.edu

Abstract

Several models have been proposed that attempt to explain how the brain identifies people by looking at their faces. However, to date, it is still not clear by which mechanism the brain successfully accomplishes the matching of two or more face images when differences in facial expression make the (local and global) appearance of these images different from one another. There seems to be a consensus that faces are processed holistically rather than locally, but there is not yet consensus on whether information on facial expression is passed to the identification process to aid recognition of individuals or not. Models have been proposed that exploit each of these two views, and psychophysical data exist in favor of and against each view. In this article, we show how the experimental data of these two opposite views can be explained by incorporating a key process of motion estimation in the classical feedforward model of face processing. This new model will then lead us to hypothesize that to successfully match expression variant faces, it is convenient to use the information supplied by this motion estimation process within the matching task. We will show experimental results in favor of this hypothesis. Finally, we will show how we can also use the same motion estimator to recognize facial expressions.

Keywords: face recognition, matching, model, motion, expression recognition, emotions.

1 Introduction

Faces are objects capable of large deformations. Through these deformations we can communicate emotions, interest (disinterest), language (e.g. gesturing in speech, and gesturing and meaning in sign languages), etc. (Bruce & Young, 1998; Hill & Johnston, 2001; Haxby et al., 2000; Messing & Campbell, 1999). Our vision, however, seems to have little problem in identifying individuals (or matching faces) even when these differences in facial expression are present. How our vision solves this problem is a fundamental question of cognitive science. This problem has also faced computer scientist as they attempt to build face recognition algorithms invariant to these deformations.

To answer the above stated question, several models have been proposed. However, to date, there is no consensus on the procedure employed by us when matching two (or more) faces bearing different facial expressions. By "matching", we mean that our cognitive system is attempting to determine whether the current images belong to the same class (person) or not. Fig. 1 depicts two examples that should help to clarify this point. In these examples, we generally have little difficulty in establishing that the images shown in (a) belong to the same subject whereas the images shown in (b) correspond to different subjects. Although in these two examples, there is a large variation in facial appearance, it is still possible to judge to what extent the images may or may not belong to the same class.

Some models proposed in the past suggest that to recognize people's identity we use a process that is completely decoupled from that of recognizing the facial expression (Bruce & Young, 1986). Others propose that a connection must exist between the two processes (Hansch & Pirozzolo, 1980). Psychophysical data exist in favor of and

(a)

(b)

Figure 1. People can usually determine whether the two images shown in (a) and the two in (b) are from the same person or not.

against each view (Bruyer et al., 983; Kurucz & Feldmar, 1979; Endo et al., 1992; Etcoff, 1984; Young et al., 1996; Schweinberger & Soukup, 1998; Baudouin et al. 2000).

The strongest evidence for a possible dissociation between the identification of faces and that of facial expression recognition comes from agnosic patients with impaired ability to identify individual faces (even those of their family and themselves), but preserved ability to recognize facial expressions (Bruyer et al., 983; Farah, 1990). The opposite can also occur; i.e., cases have been described where agnosic patients were unable to classify facial expressions, but were found to have a normal ability to identify faces (Kurucz & Feldmar, 1979).

One of the most notable experiments against the independence of identity and facial expression recognition is the finding that people are slower in identifying happy and angry faces than they are in identifying faces with neutral expression (Endo et al., 1992). Similarly, subjects are slower in identifying pictures of familiar faces when those are shown with uncommon facial expressions (Hay et al., 1991) or when some artificial deformations are added to the images.

The existence of experimental data in favor and against the two face recognition strategies described above makes a deeper understanding of many aspects of face processing difficult. In this article, we propose a new model that is consistent with the experimental data discussed above.

The proposed model does not have a direct connection between the processes of identity and facial expression recognition. However, both processes get information from the common modules of dynamic and static processing cues, as depicted in Fig. 2. Within the dynamic cues, we have only focused on the key process of motion estimation named "deformation of the face," whose task is to calculate the apparent physical deformation between the faces to be matched (for simplicity we shall refer to this process as *DF*). The *DF* process computes the motion field between images (i.e., the implied motion between images). As we will show below, this is key to explaining the psychophysical data discussed above.

The model shown in Fig. 2 is consistent with the idea advanced by Hubel and Wiesel (Hubel & Wiesel, 1965) of a sequence of more and more complex and invariant features. Early processing of facial features may be carried out in the inferior occipital gyri, which (hierarchically) sends information to the lateral fusiform sulcus and the superior temporal sulcus (STS). Cells have been found in the STS that are selective to biological motion (Servos et al., 2002; Grossman et al., 2000) and to *social* behaviors of moving facial features (Allison et al., 2000; Haxby et al., 2000). These cells may or may not use the information of our *DF* process, but their physical proximity to the other processes seems to make this plausible. The model depicted in Fig. 2 assumes, however, a dissociation

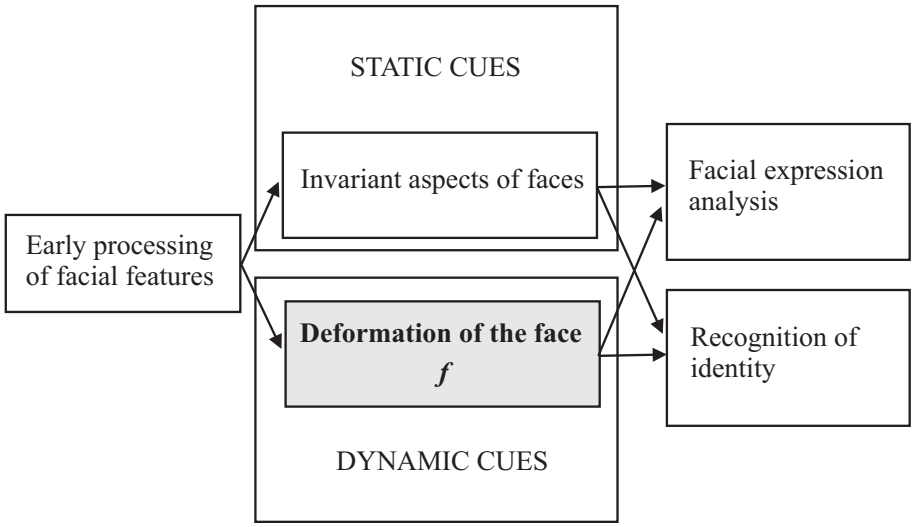


Figure 2. Depiction of the different processes of the model proposed in this paper. The key process is the one responsible for estimating the *deformation of the face*, which is connected in a feedforward manner to the processes of identification and facial expression recognition. These two last modules are dissociated, although they both obtain information from the (new) common process *DF* and from the processes that compute static cues. This is key to explaining the psychophysical data described in the past. (See text.)

between the modules responsible for computing dynamic cues and those responsible for computing static cues. Evidence exists to support such a claim. For example, Humphreys et al. (1993) describes an agnosic patient impaired to recognition of identity and with poor recognition of facial expressions from static images, but with normal classification capacities of facial expressions from video sequences.

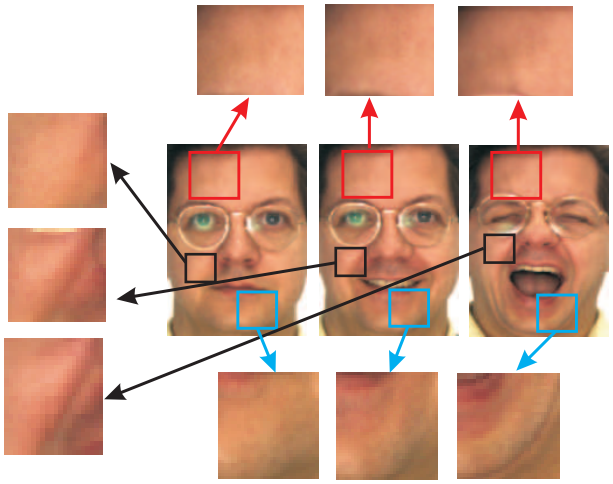
The hierarchy of the proposed model starts diverging toward the specialized areas of identity and expression recognition only after the *DF* area, while in previous models this separation was assumed to occur in a much earlier stage – probably as early as in the inferior occipital gyri. Finally, once both models become independent, the recognition of identity continues toward anterior temporal areas, while the recognition of facial expressions of emotions proceeds toward the amygdala, insula, limbic system or other areas associated to the processing of emotional cues.

Note that the suppression of the *DF* area would affect the recognition of identity and facial expression but would not completely prevent recognition in either. However, if only the area dedicated to the recognition of facial expressions or the one dedicated to identification of faces was damaged, we would obtain the result demonstrated by the agnosic patients introduced earlier (i.e., those that are impaired either with regard to identity or expression, but not both).

Our model can also explain the psychophysical data against the independence. Note that it is now logical to expect the identification of faces to be slower for those cases where a larger deformation of the face exists, since we need to go through the motion estimation module *DF*. The more complex the facial expression, the more time (generally) needed to compute an approximation of the muscle activity of the face (see *Results*). For example, the *DF* module shown in Fig.2 does not need to estimate any motion for the neutral facial expression case, but requires the computation of an approximation of the motion of a smile and other facial expressions.

However, the question still remains as to why there should be a motion estimation process, which computes the deformation between the images to be matched, within our face recognition model. One reason is given by the finding that motion plays an important role in recognizing identity and facial expressions in a sequence of images (Hill & Johnston, 2001; Wallis & Bulthoff, 2001; Lander et al., 1999). Uncommon deformations or uncommon sampling times disrupt identification of individuals (Hay et al., 1991) and of facial expressions (Kamachi et al., 2001).

In this paper, we further hypothesize that this motion field is necessary (or, at least, useful) to successfully match the local and global features of a set of faces when those bear distinct facial expressions. This seems



(a)

Figure 3. As the face of a person changes facial expression, the local appearance (texture) of her/his face also changes. In this figure, we compare three local areas of the face under three different expressions (neutral, happy, and scream).

reasonable, because the local texture of a face changes considerably as the facial expression also changes. An example is shown in Fig. 3. In this figure, three local areas of three face images of the same person, but with different facial expressions, have been highlighted. It is apparent that as the facial expression changes, the local texture of each of these areas becomes quite distinct (even in the forehead). For the scream face, one could even wonder if there exists any invariant features (within these three areas) useful for identification. A classification (or identification) algorithm, however, will need to find where the invariant features are. If we compute the motion field (deformation) between the two images that one wants to match, we can then know which are the features that have changed the least between the two images, and thus, which are the best candidates for matching purposes. We will experimentally show how this is indeed useful for classification purposes.

We will compare the results of our model (accuracy and reaction times) with the results obtained with human participants in an experiment that uses the same data for both. These results will demonstrate the ability of our model to predict human responses.

We will also show that when we incorporate the motion estimation process (DF) within the model of recognition of facial expressions, we can justify several of the psychophysical data encountered in the literature (e.g. slower recognition rates for image composites of two different emotions (Young et al., 1997)). This shows that our model has implications beyond the ones discussed above, and that indeed the motion process can play an important role within the face processing system. Furthermore, we can use the extended model to make new hypotheses about our ability to identify facial expressions. We will show how our model predicts that when matching or classifying faces with facial expressions that look alike the reaction time increases, while when facial expressions are very distinct, reaction times are shorter. Similarly, we can also expect faces with large deformations to require longer reaction times and faces with small deformations to have shorter reaction times. We will confirm these hypotheses in another experiment with human participants.

2 The Model

2.1 Matching: recognition of identity

Since it is believed that faces are processed holistically rather than by parts, most computational models for face recognition have exploited such a possibility. A common way to accomplish this is by means of a unified pixel to pixel comparison of the whole face (Ullman, 1996; Fukunaga, 1990). Formally, consider two different images, I_1 and I_2, both of n pixels. We can redefine the images as vectors taking values in an n-dimensional space, i.e.

$\Re^n$. We shall denote this as V_1 and V_2. The advantage of doing this is that it allows comparisons of the images by means of vector operations such as subtraction:

$$\|V_1 - V_2\| \tag{1}$$

where $\|\cdot\|$ denotes the L_2 norm (i.e., Euclidean distance). In this definition stated here, we assume that all faces have been aligned (with respect to the main facial features) in such a way that the eyes, mouths, noses, etc. of each of the images are at roughly the same pixel coordinates, e.g. (Beymer & Poggio 1996; Martínez, 2002). The approach defined above, in Eq. (1), has proven to perform well when frontal face images with similar facial expressions are compared to each other (Brunelli & Poggio, 1993). However, when matching face images bearing different facial expressions, this comparison becomes unstable (Martínez, 2002); hence pixels can now carry information of different features. An example of this was depicted in Fig. 3.

A fundamental question in face recognition is whether or not the identification process receives (or interchanges) information from (with) the process of facial expression recognition to aid in the recognition of individuals. Psychophysical evidence exists to support a positive and a negative answer to this question. Nevertheless, as shown above, this data can now be explained by incorporating the *DF* process in our model, as depicted in Fig. 2. This new model of face processing can be expressed mathematically as:

$$\|f^{-1}(V_1 - V_2)\|, \tag{2}$$

where f is an n-dimensional vector with detailed information on how the pixels of the first image have moved so that they can represent the facial expression of the second image. Intuitively, f is a vector that keeps correspondences between the pixels of the first and second images. Eq. (2) can be interpreted as follows; pixels (or local areas) that have been deformed largely due to local musculature activity will have a low weight, whereas pixels that are less affected by those changes will gain importance. We can formally define f^{-1} as taking values linearly inverse to those of f, i.e.:

$$MAX_F - \|\mathbf{F}_i\| \tag{3}$$

where $\mathbf{F}$ is the motion flow (i.e., motion between two images), $\mathbf{F}_i$ the motion vector at the ith pixel, and $MAX_F = max_{\forall i}\|\mathbf{F}_i\|$ (the magnitude of the largest motion vector in the image).

The value of f corresponds thus to the outcome of the *DF* process. Note that f defines the face deformation (motion) between two images and, therefore, can also be used to estimate the facial expression of a new incoming face image (see Section 2.3). As mentioned earlier, experimental data supports this belief.

2.2 Motion estimation

Several plausible neurological models for the computation of visual motion between two images have been proposed, such as the Barlow-Levick's circuit. In general, this can be expressed mathematically by local deformations that occur in small intervals of time, δt, as

$$I(x, y, t) = I(x + u\delta t, y + v\delta t, t + \delta t), \tag{4}$$

where $I(x, y, t)$ is the image value at point (x, y) at time t, (u, v) are the horizontal and the vertical image velocities at (x, y) and δt is considered to be small (Horn & Schunck, 1981). We note that in our model $f = (u, v)$.

If we assume that the motion field (i.e., the pixel correspondences between the two images) is small at each pixel location, the motion estimator can be represented by the first-order Taylor series expansion as

$$E_D = \int\int \rho\left(I_x u + I_y v + I_t\right) dxdy \tag{5}$$

where (I_x, I_y) and I_t are the spatial and time derivatives of the image, and ρ is an *estimator*.

To resolve the above equation, it is necessary to add an additional constraint. The most common one is the spatial coherence constraint (Horn & Schunck, 1981), which embodies the assumption that neighboring pixels in an image are likely to belong to the same surface and, therefore, a smoothness in the flow is expected. The first-order model of this second constraint is given by

$$E_S = \int\int \rho\left(\nabla(u,v)\right) dxdy \tag{6}$$

where ∇ represents the gradient.

The goal is to minimize the regularization problem $E_D + \lambda E_S$. To gain precision at motion boundaries it is convenient to make a good choice for ρ, e.g. $\rho(x) = log\left(1 + 1/2(x/\sigma)^2\right)$ (Black & Anandan 1996).

The iterative update equation for minimizing $E = E_D + \lambda E_S$ at the image pixel i at step $r+1$ can be expressed as (Black & Anandan 1996):

$$u_i^{r+1} = u_i^r - w\frac{1}{T(u_i)}\frac{\partial E}{\partial u_i} \tag{7}$$

where w is an over-relaxation parameter that is used to over-correct the estimate of u^{r+1}. It has been shown that with $0 < w < 2$ this equation converges (Varga, 1962). The partial derivative $\frac{\partial E}{\partial u_i}$ of the above equation can be approximated with:

$$\frac{\partial E}{\partial u_i} \approx \sum_{s\in R}[I_x\psi(I_x u_i + I_y v_i + I_t, \sigma) + \lambda_S \sum_{s\in\mathcal{R}_s} \psi(u_i - u_s, \sigma)] \tag{8}$$

where $\mathcal{R}_s$ represents the set of 4-neighboring pixels (left, right, up and down) and $\psi(x,\sigma) = \frac{2x}{s\sigma^2+x^2}$ is the derivative of the Lorentzian. The term $T(u_i)$ is the upper bound of the second partial derivative of E, which we can define as:

$$T(u_i) \leq \frac{\lambda_D I_x^2 + 4\lambda_S}{\sigma^2} \tag{9}$$

Here we have only shown derivations with respect to u, but exactly the same applies to v.

Although the objective function E is non-linear (and a direct solution does not exist for minimizing it), methods exist that can find a minimum E close or even equal to the global minimum. One solution corresponds to first selecting a large value for σ in such a way that a convex approximation of E is obtained (Blake & Zisserman, 1987). In this convex approximation, we can readily locate the minimum of the function. In general, this solution will not be the desired one, but it will serve as a good starting point. We can then decrease the value of σ and find a new minimum for E. When doing this, however, we will use the minimum obtained previously as an initial guess. By repeating (iterating) this process, we can converge to a good approximation of the global minimum.

The procedure detailed above will yield good results only when the object displacements between consecutive images are small (since we used a first-order model). In order to correctly detect large motions, a coarse-to-fine strategy can be employed. In our experiments, the pyramid method of (Black & Anandan 1996) was used. In this approach we begin with a reduced-resolution representation of the images so that the small-displacement assumption (made above) is satisfied. The optical flow is computed for the low-resolution images and then projected to the next level of the pyramid where the images in the sequence have a higher resolution. At each level of the pyramid, the optical flow computed from the previous level is used to warp the images in the sequence so that the small displacement assumption holds for the new resolution too. This process is repeated until the flow has been computed at the original resolution. The warping process can be formally described as:

$$I_{warped}(x,y) = I(x - u(x,y), y - v(x,y)). \tag{10}$$

The final flow field is obtained by combining the flow information of each of the levels of the pyramid. The number of levels on the pyramid will be dictated by the largest motion in the sequence of images.

2.3 Matching: expression recognition

Psychological and neurological evidence suggest that motion plays an important role in the recognition of emotions and facial expressions (Basilli 1978; Bruce & Velentine, 1988; Kamachi et al., 2001). We now show how our model can achieve this by using the outcome of the *DF* area.

Facial expressions (and specially those associated with the so called universal emotions (Darwin, 1872; Ekman & Friesen, 1978)) can be perceived in a categorical manner by humans (Etcoff & Magee, 1992; Calder et al., 1996; Young et al., 1997). Categorical perception has also been observed in other face recognition tasks, such as in the recognition of identity (Beale & Keil, 1995; Leopold et al., 2001) and in the recognition of grammatical content

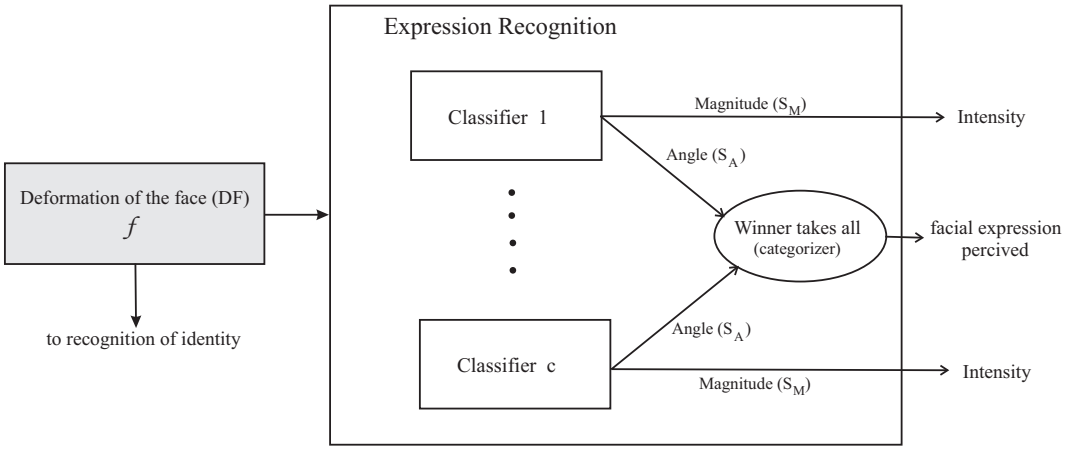


Figure 4. Depicted here are the different processes that involve the recognition of facial expressions and that use the *DF* module described in this article.

in the facial expressions of sign languages (Campbell et al., 1999; Messing & Campbell, 1999). Other visual task, such as color perception, can also be categorical (DeValois & DeValois, 1975; Bornstein, 1987). Outside vision, categorization has been observed in several tasks such as in the perception of language.

However, images are not always classified within one category or another with equal ease. For example, Young et al. (Young et al., 1997) have used a set of morphed face images which represent a continuum between pairs of facial expressions of emotions to show that while subjects prefer to classify the images in a categorical manner, their reaction times (RT) are slower for the morphed faces that combine features of two expressions. In their experiments, a face is morphed to represent $n\%$ of an emotion and $(100-n)\%$ of another emotion. Young et al. showed that human RT's get slower as the values of n and $(100-n)$ get closer. A classifier that solely uses the pixel intensities of expressive faces to categorize each facial expression in a group of emotions could not justify the changes in RT. Such a classifier would be equally fast (or equally slow) for all emotional displays.

However, these RT differences can be justified if we incorporate the *DF* module within the model of recognition of facial expressions. In our model, facial expressions are also processed using the same *DF* process described above, but the final categorization (matching) task is different. Fig. 4 depicts this part of the (extended) model. Note that while the *DF* module is common to the tasks of identity and facial expression recognition, the categorization tasks are independent processes.

In our model, each facial expression is classified in a category according to the motion field of the face, f. The direction of motion is used to determine the class (Bartlett et al. 1999), while the magnitude of the motion can be used to specify the intensity of a given expression. These two parts of the motion can be expressed mathematically as:

$$S_{M_i} = abs(\|\mathbf{F}_{\mathbf{t}i}\| - \|\mathbf{F}_{\mathbf{p}_i}\|) \quad and \quad S_{A_i} = arccos\frac{< \mathbf{F}_{\mathbf{t}i}, \mathbf{F}_{\mathbf{p}_i} >}{\|\mathbf{F}_{\mathbf{t}i}\|\|\mathbf{F}_{\mathbf{p}_i}\|} \tag{11}$$

where $\mathbf{F}_{\mathbf{t}i}$ and $\mathbf{F}_{\mathbf{p}_i}$ are the vector flows of the two expressions to be compared at the i^{th} pixel, $< \mathbf{a}, \mathbf{b} >$ represents the dot product of $\mathbf{a}$ and $\mathbf{b}$, S_{M_i} is the similarity between the magnitude of the i^{th} pixel in the two image flows, and S_{A_i} the similarity between the angles of the two vectors at pixel i.

The method described in the preceding paragraph is normally used to compare two images (i.e. matching), but we can also used this to classify (or identify) facial expressions within a group of pre-learned categories. These categories can either be universal and may even be wired from birth (Darwin, 1872; Ekman & Friesen, 1978), may be associated with language categories (Davidoff, 2001) or may be learned after birth.

This categorical comparison can be carried out at each pixel location or at specific areas that are known to be most discriminant for a given expression. We can formally express this as:

$$S_M = \sum_{i=1}^{m} S_{M_i} \quad and \quad S_A = \sum_{i=1}^{m} \frac{S_{A_i}}{m_o} \tag{12}$$

where m is the number of pixels where comparison takes place, $m \leq n$, and m_o is the total number of vectors in m with magnitude greater than zero. Note that since the angle similarity can only be computed between actual vectors (of magnitude greater than zero), it is necessary to normalize S_A by the number of comparisons to prevent biases towards images with associated small motions. Similar measurements have been successfully used to identify the active units of Ekman and Friesen (Ekman & Friesen, 1978) by other authors (Bartlett et al. 1999; Donat et al., 1999).

In order to appropriately select the value of m, it is convenient to search for those features (i.e. pixels) that best discriminate between categories and those that are most stable within categories. A common way to do this is by means of the between-class and within-class measurement of Fisher's Linear Discriminant Analysis (LDA) (Fisher, 1938; Fukunaga, 1990). The use of discriminant features will help us to be more precise in our classifications, more robust to image changes and will speed up computation.

Formally, we define the within and between class scatter matrices of LDA as (Fisher, 1938):

$$S_W = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\mathbf{v}_{i,j} - \mu_j)(\mathbf{v}_{i,j} - \mu_j)^T \quad and \quad S_B = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T \tag{13}$$

where S_W is the within-class scatter matrix, S_B is the between-class scatter matrix, c is the number of classes, N_j is the number of samples for class j, $\mathbf{v}_{i,j}$ is the i^{th} sample of class j, μ_j is the mean vector of class j, and μ is the mean of all classes.

However, due to singularity problems, we generally find it difficult to compute the LDA transformation of large face images. Additionally, S_B limits us to a maximum of $c-1$ dimensions (where c is the number of classes). Since we usually deal with small values of c and it is known that LDA may perform poorly if the dimensionality of the space is small (Martínez & Kak, 2001), it is convenient to only use that information which directly specifies the usefulness of each pixel. This is represented in the variances of each feature (pixel) within S_W and S_B, which is given by the values at the diagonal of each of these matrices:

$$\hat{S}_W = diag(S_W) \quad and \quad \hat{S}_B = diag(S_B). \tag{14}$$

By first finding those pixels (areas) of the face that are most different among classes ($\hat{S}_B$) and, then selecting those that are most similar across samples of the same class ($\hat{S}_W$), we can build a classifier that computes the values of S_A in a smaller set of pixels. This classifier is also generally more robust and efficient than the one that uses all the pixels of the image.

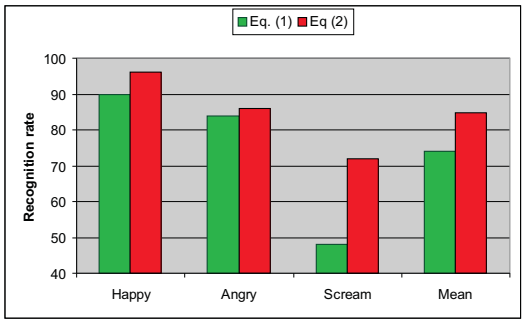
We can predict, using the model described above, that when faces are to be classified within very distinct categories (e.g. happy and neutral), the task will result easier than when the two facial expressions are alike (e.g. angry and neutral). As a consequence, it is logical to expect faster responses (RT) when we attempt to classify two very distinct classes (easier task), than while attempting to classify very similar classes. Since the model uses the *DF* procedure described above, we can also predict that when classifying faces within two distinct groups, those that involve larger motions will usually have longer RT. Similarly, those facial expressions that are more difficult to be classified or are more alike, will require the analysis of additional local parts – resulting in longer RT. According to this discussion, when classifying or identifying facial expressions of emotions, the RT should be slower for those faces that carry expressions that are most difficult to classify (identify) – as those will require the comparison of extra local areas. When a face cannot be reliably classified within one of the categories by looking at the most discriminant areas, we will need to extend our comparison to other areas of the face. This latest prediction is consistent with the findings of Young et al. discussed earlier. When images are compositions of two expressions, two classifiers will successfully classify the image. In order to choose one of them a closer analysis will be necessary, i.e. more pixels (or local areas) will be added to the comparison. For composites of 60% and 40%, a pixel-to-pixel comparison may be necessary.

neutral (a)

happy

angry (b)

scream



Recognition Rates (c)

Figure 5. (a) An example of a neutral (no-expression) face image. (b) An example of a happy, an angry and a scream face. (c) The recognition rates obtained when matching: *i)* happy and neutral faces, *ii)* angry and neutral faces, and *iii)* scream and neutral faces. Note that when we incorporate the information of the *DF* process in our model (i.e. f), the results improve and the matching process becomes less sensitive to the differences in facial expression.

3 Results

In this section we will first show how the model presented in this article is consistent with the psychophysical findings described in the introduction – dependent versus independent processes. We will compare our results with a group of human subjects. Then, we will show how we can recognize facial expressions by means of the extended model introduced above, and we will confirm our hypothesis made above with another experiment with human subjects.

3.1 Identity recognition

We can now test two important points advanced in the previous section: *a)* how the suppression of the *DF* process would affect the identification of known individuals, and *b)* how the identification of happy and angry faces is now slower than the recognition of neutral expression faces. In these experiments, we will use the face images of 100 individuals of the AR database (Martínez & Benavente, 1998). The images of this database for one of the subjects are shown in Fig. 5(a-b).

3.1.1 Recognition

As sample images we will use the neutral faces, Fig. 5(a). As test images (i.e., images to be matched with the sample ones), we will use the happy, angry and screaming faces, Fig. 5(b). For each of the test images, we will select the sample image that best matches it, as given by Eq. (2). If the retrieved image belongs to the same person (class) as the one in the testing image, we will say that our recognition was successful. Fig. 5(c) shows the percentage of successful identifications. We have detailed the recognition rates for each of the facial expression images to show the dependency between the recognition of identity and facial expression. We have also shown, in this figure, what would happen if the *DF* process was damaged or absent. This is represented by omitting the value of f in Eq. (2) (as given by Eq. (1)). This latest result makes us hypothesize that if there exist an agnosic patients with localized brain damage that exclusively involves the *DF* process (area), this patient's ability to recognize facial expressions will be quite deteriorated whereas his/her ability to identify faces will mostly be altered for those images with large muscular activity only (such as the "scream" face). The results of such a patient as predicted by our model are obtained with Eq. (1) in Fig. 5.

3.1.2 Delays

We can also analyze how the recognition of identity slows down as the expressions on the images diverge more from each other. The reader may have already noted that this is quite obvious, because to estimate large motions

we will need extra iterations (as advanced in Section 2.2). It is important to note that as the differences between facial expressions increase, it is necessary to be more careful in estimating the motion field. Since most methods of motion estimation carry the assumption of small displacements between images, it is necessary to use another approach for larger motions. We have shown above that we can solve this using a coarse-to-fine strategy (we will refer to this as the pyramid or the hierarchical approach). When small motions are present, this approach will require fewer samplings of the original image. For larger deformation, more levels of the hierarchy will be necessary.

Furthermore, at each level of the hierarchy we will need to find the global minimum of a non-convex function. In general, when the motions of different parts of the image is more distinct (i.e., moving in different directions), the function to minimize is more complex (i.e., far from non-convex). As mentioned earlier, this can be solved by means of an iterative approach. We do that by first searching the minimum in a convex approximation and then finding the solution to progressively less convex functions. At each iteration though, we use the previous estimate as a starting guess. We stop when we finally find the minimum of the original non-convex function. As defined in Section 2, in our computational model we can control the convexity of the function to be minimized by varying the value of σ. For complex motions (functions), many iterations will be necessary. For simple motions, few iterations will suffice.

To calculate the computational time required to compute the motion field for each of the expressions, we need to determine: *i)* the number of levels of the pyramid required to compute the largest motions of the image, and *ii)* the number of iterations necessary to correctly calculate the minimum of the non-convex function at each level of the pyramid.

For each of the facial expressions in the AR database (i.e., happy, angry and scream) as well as for the neutral expression image, we have calculated the minimum number of iterations and levels of the pyramid required as follows. First, we computed the motion fields, f, using levels of the pyramid that range from 1 to 4 – for each of the expressions independently. (This was done using a randomly selected group of 30 people, 15 male and 15 female, of that database.) We then compared the results obtained when using $h+1$ levels of the pyramid and when only using h levels. If the similarity in magnitude (as computed by S_M/m_o) and angle (S_A) between the two (h and $h+1$) was below a threshold, we determined that h levels suffice for the computation of the motion in that image; otherwise $h+1$ levels were necessary. We will refer to this chosen value as H. The threshold used in this experiment was equal to one pixel; i.e., the motion estimated at $h+1$ had to be at least 1 pixel longer than that calculated with only h levels to be sufficiently large to justify $h+1$ levels. Note that less than one pixel would reflect a zooming effect rather than an actual difference in motion estimation.

To determine the number of iterations required at each level of the pyramid, we compared the results obtained when using $g+1$ and g iterations. Again, if the comparison was below a threshold, we selected g, otherwise we selected $g+1$. We will refer to this value as G. In this case, the threshold was 0.1 and g was tested for the range of values from 10 to 50.

Finally, we combined the two selected values into a single one as $CT = G*H$ (computational time = the number of iterations necessary at each level multiplied by the number of levels needed). The results (mean across samples) were: *Neutral faces*: $H=1$, $G=10$ and $CT=10$, *Happy faces*: $H=3$, $G=26$ and $CT=78$, *Angry faces*: $H=2.4$, $G=20$ and $CT=48$, *Scream faces*: $H=4$, $G=33$ and $CT=152$. These results are plotted in the graphical representation of Fig. 6. These results do not include the time necessary to compute Eq. (2), but since in our current implementation of the model this time is always constant, we have omitted it for simplicity.

3.1.3 Comparison to human subjects

Human participants: Ten people (5 males and 5 females) from different backgrounds voluntarily participated in this experiment. The age of the participants varied from 20 to 61 (mean=34.7). All had normal or corrected-to-normal vision. None of the participants had previous experience with the face stimuli shown in the experiments and all were naive as to the research questions under study and to what variables were recorded during the experiment.

Stimuli: Eighty (80) images corresponding to the neutral expression, happy, angry and scream faces of twenty (20) people of the AR face database were selected for this experiment. To prevent the results from being affected by other image cues than those under investigation, all twenty selected individual were males with no glasses. All images had been recorded under strictly controlled lighting conditions to guarantee uniformity across samples. The images were warped to a standard image size as described earlier. This prevents configurational recognition,

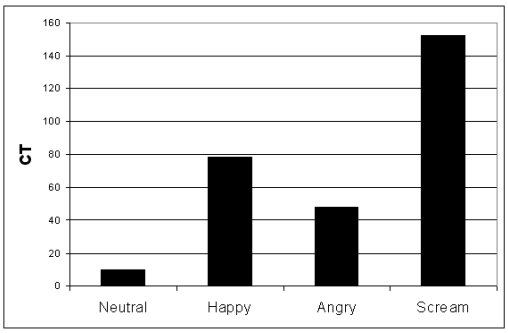


Figure 6. Shown here are the mean delays (Computational Time, CT) required to compute the motion fields, f, for each facial expression group.

making the identification task (in general) more difficult (as configurational recognition is thought to play a crucial role in adult perception of faces). The images shown to the participants were of 165 by 120 pixels, which in a 21 inch monitor corresponded to approximately 15 by 11 cm. When viewed from a normal distance of approximately 60 cm this size corresponds to 14 by 10.4 degrees of visual angle.

Design and procedure: The experiment consisted of two blocks of images, each with the images of ten of the selected people from the AR database of face images. In each block, pairs of images were shown in sequence: first a neutral image of a randomly selected person was displayed for 800 ms (prime face), an interstimulus interval of 300 ms followed, then a neutral, happy, angry or screaming face (target face) of another randomly selected person was displayed. Participants had to decide whether the two images shown in sequence correspond to the same individual or not. Participants were instructed to respond as soon as they knew the answer, but not sooner. Reaction times (RT) as well as correct and incorrect choices were recorded.

The identity of the prime and target face images as well as the facial expression of the target face were randomly selected. Each participant saw 80 pairs of images in each of the two blocks of images, which adds up to a total of 160 pairs.

Results: Fig. 7(a) shows the mean RT values of all participants when deciding whether the prime and target face images belong to the same person or not. That corresponds to neutral-neutral, neutral-angry, neutral-happy and neutral-scream pairs. As expected, the more the target face diverged (in muscle activity) from the prime face, the slower it was to reach a decision. In Fig. 7(b), we show the percentage in recognition rate achieved by the participants for each possible sequence pair; i.e., the prime image being a neutral expression face and the target as shown.

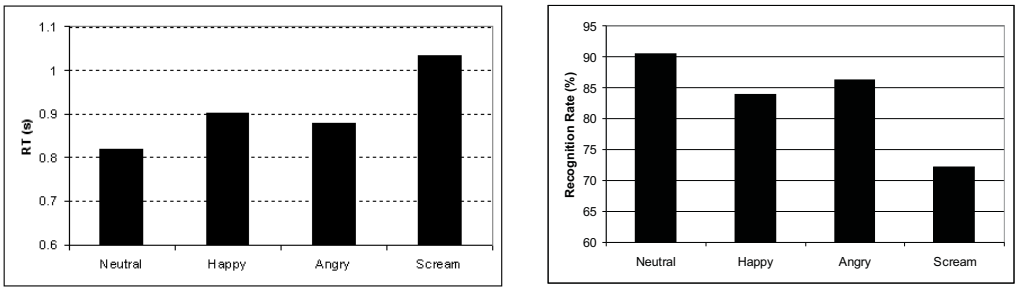
It is important to note that our model successfully predicted the responses of our human subjects. Unfortunately, a numerical comparison would be difficult, because while the RT include the matching time (which is not necessarily constant for all expressions), the CT correspond only to the time necessary to compute the deformation of the face (i.e. *DF* process).

3.2 Facial expression recognition

3.2.1 Matching facial expressions

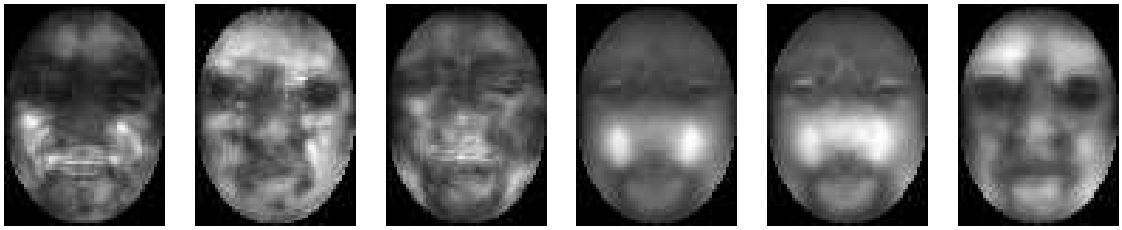
Similar to the matching experiments of identity recognition described above, we now show how the motion vectors can be used to recognize facial expressions. We will calculate the similarity between pairs of images by using the value of S_A described earlier in Eq. (12).

The first test (matching) corresponds to determining for each possible combination of two facial expressions (a total of 10 combinations) if the two images shown have the same facial expression or not. To do this, we used the neutral, happy, angry and screaming face images of 50 randomly selected individuals of the AR face database which gives us a total of $12,750$ different pairs. For each of these pairs, we compute the motion field (i.e., face



(a) (b)

Figure 7. (a) Mean RT of the ten participants when deciding wether two consecutive viewed images belong to the same individual or not (the prime image with a neutral expression and the target image with the expression as shown in the x-axes). (b) Mean recognition rate (in percentage) of the ten participants when performing the task described in (a).



(a) (b) (c) (d) (e) (f)

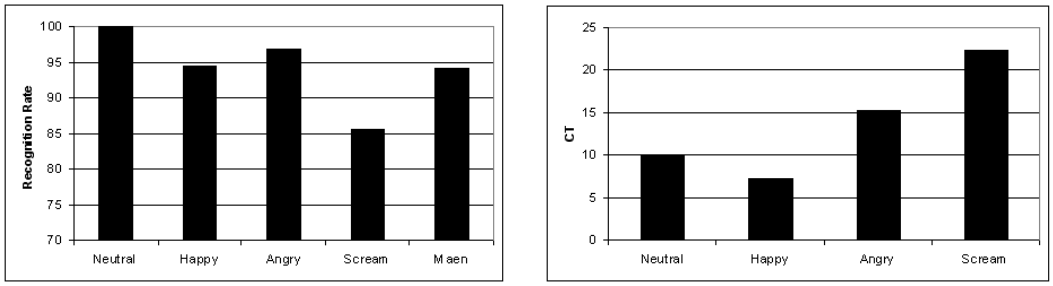
Figure 8. (a) $\hat{S}_{w_{happy}}$, (b) $\hat{S}_{w_{angry}}$, (c) $\hat{S}_{w_{scream}}$, (d) $\hat{S}_b$ including all the expressions, (e) $\hat{S}_b$ for expressions happy and scream, and (f) $\hat{S}_b$ for expressions neutral and angry.

deformation, *DF*) that exists between the neutral image and the facial expression selected. The two resulting motion fields are then compared by using the similarity measure S_A. This value is expected to be low for similar motion fields (i.e. similar expressions) and large for different ones.

Once the value of S_A has been obtained for each of the $12,750$ pairs of images, we search for the value of S_A that optimally divides the pairs with equal expression in one group and those with different expression within another group. We then use this threshold to classify the image pairs of a different set of 50 people. The correct classification in this second group (using the threshold obtained with the first group) was of 82.7%.

As advanced in section 2.3, we can improve this result by means of a discriminant function that helps us to determine which areas of the face are most discriminant within classes (i.e., same facial expression) and which are most distinct between classes (i.e., different facial expressions). One way to do that is by means of Eq. (14). For example, when comparing happy and screaming faces we can use the values of $S_{b(happy,scream)}$ (shown in Fig. 8(e)) and the values of $\hat{S}_{w_{happy}}$ and $\hat{S}_{w_{scream}}$ (shown in Fig. 8(a,c)) to determine those pixels that are most discriminant. We then order (rank) the pixels inversely proportional to the values of $\hat{S}_w$ and proportionally to the values of $\hat{S}_b$. Since most of the pixels will have an associated ranking of zero or close to zero, we can make our comparison faster by only using those pixels with a value of $\hat{S}_b/\hat{S}_w$ larger than a pre-determined threshold. This threshold can also be learned from the training data – in which case we select that value that best classifies the training data. By following this procedure, the results improved to 91.3%.

We have already discussed that in several cases categories may already exist and that classification (or identification) may be another alternative to consider. That is to say, if we see two images expressing emotions (e.g. happy and happy – same, or, happy and angry – different), we may determine that they have (or not) the same expression because they are classified within the same (different) category rather than because they look alike. That means that when a new image is to be classified, we first compute its motion field and then compare this



(a) (b)

Figure 9. (a) Recognition rates obtained by our model when classifying each of the face images in four different groups: neutral, happy, angry and scream. (b) Mean computational time (CT) required to calculate the class for those images with neutral, happy, angry and scream facial expressions.

with each sample motion field (previously) stored in memory. The stored motion that is found to be most similar to the testing one will determine the class of the testing image.

We used the neutral, happy, angry and scream face images of 10 randomly selected individuals as samples and the neutral, happy, angry and scream face images of 90 different individuals as testing images. For each of the 360 testing images, we determine the closest sample (among the 40 stored in memory) using the value of S_A. If the facial expression in the testing image and in the closest sample were the same, we recorded a successfully classified image. Again, we use the values of $\hat{S}_b$ and $\hat{S}_w$ to improve the classification results and speed up computation. These results are shown in Fig. 9(a).

3.2.2 Delays

According to our model, the delays observed when we recognize facial expressions can be due to: *i)* the time required to compute the motion field (DF) of the expression displayed on the (testing) image, or *ii)* the difficulty associated in classifying the facial expression of a test image in a set of pre-selected categories.

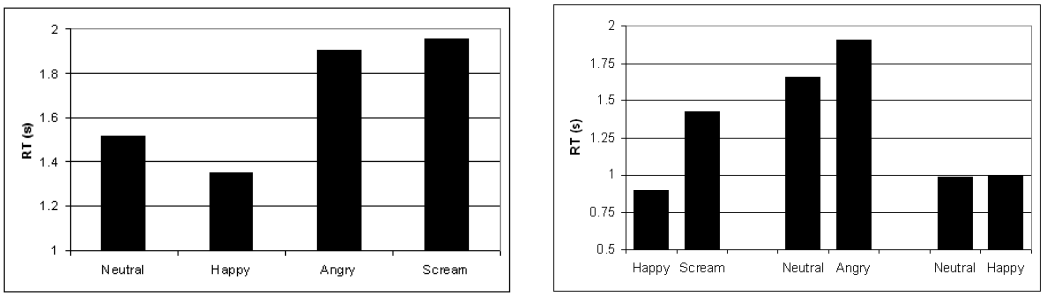
For example, when classifying images as either happy or screaming, we expect to have longer RT for those images with a scream expression because it takes longer to compute the motion field (DF) of a scream face. Moreover, we would expect longer RT when classifying images as either neutral or angry than when classifying images as either happy or screaming, because the images in the first task (group) are more alike and thus a more detailed analysis will be required. While happy and screaming faces can be easily distinguish by looking at a small number of pixels (such as the eyes or the corners of the mouth), a pixel-to-pixel comparison may be necessary to decide whether an image is a neutral expression or a not-excessively-marked angry face. [1]

In Fig. 9(b) we show the Computational Times (CT) of Fig. 6 multiplied by the percentage (range: 0 to 1) of pixels that were necessary to use in order to obtain the best classification rate when classifying the images as either neutral expressions or the expression under consideration. The pixels were selected according to the rankings given by $\hat{S}_b$.

3.2.3 Comparison to human subjects

Human Participants: A new group of ten people (5 males and 5 females) voluntarily participated in this experiment. The age of the participants varied from 24 to 33 (mean=27.2). All had normal or corrected-to-normal vision. None of the participants had previous experience with the face stimuli shown, none had participated in the experiment of Section 3.1.3, and all were naive as to the research questions under study and to what variables were recorded during the experiment.

[1] In some cases, the angry face images of the AR face database are not very prominent. This may be due to the difficulty associated with posed emotions that one does not feel.



(a) (b)

Figure 10. Mean RT when classifying the images in: (a) four different groups (neutral, happy, angry and scream), (b) two categories classification (happy-scream, neutral-angry, and neutral-happy). (Reaction time in seconds.)

Stimuli, design and procedure: The neutral, happy, angry and scream face images of twenty (20) males (with no glasses) of the AR face database were selected for this experiment. To prevent recognition by shape alone, images were warped to a standard image size of 165 by 120 pixels. Subjects participated in four different tests. The first required them to classify each of the images of the AR database within one of the four categories of that dataset. Subjects were told in advance of those categories and an image for each of the expressions was shown to participants before the experiment started. The other three tests only involved two types of facial expressions. In these two-class experiments, subjects were asked to classify images within these two categories only. The two-class experiments comprise the following facial expression images: a) happy and scream, b) neutral and angry, and c) neutral and happy.

Reaction times (in seconds) and percentage of correct choices were recorded. Fifty images were randomly selected and displayed, one at a time, until the subject pressed a key to indicate her/his classification choice. A two second pause (with blank screen) separated each of the images shown to the participants.

Results: In Fig. 10(a) we show the RT means of all the participants when classifying the images within each of the four groups. These results should be compared to the CT predicted by our model and shown in Fig. 9(b).

As discussed above our model also predicts that when comparing happy and screaming faces (i.e., when classifying images in only two clearly distinguishable classes), the latter will generally require longer RT because (as demonstrated in Section 3.1) longer time is required to estimate the *DF*. This was confirmed by our group of subjects, Fig. 10(b). We also predicted that when classifying face images within two similar classes, the RT will generally increase. This is the case for neutral and angry faces, Fig. 10(b). Another particular case is that of classifying face images as either neutral or happy. This task can be readily solved by looking at a small number of pixels (such as those around the corners of the lips and the eyes). Thus, in this case, similar RT are expected. This was indeed the case in our experiment, Fig. 10(b).

4 Discussion

Our claim is that the psychophysical data previously described and believed to be contradictory can be explained by including a process of motion estimation (whose task is to calculate the deformation between the faces we want to match), *DF*, within a hierarchical model of face processing as depicted in Fig. 2. This is key to explaining the reason why in some experiments, e.g. (Endo et al., 1992; Hay et al., 1991; Baudouin et al. 2000), slower recognition times are obtained when attempting to identify faces with distinct facial expression (e.g., smiling versus neutral faces). At the same time, the model does not require a direct interaction between the processes of face identification and facial expression recognition. This is important, because it is consistent with the observation that some agnosic patients (Kurucz & Feldmar, 1979; Bruyer et al., 983; Farah, 1990) are impaired only with regard to one of the two tasks (either identification of people or facial expression recognition).

In our model, motion (dynamic) cues are processed independently from static cues. This is based on neurophysiological evidence which supports the assumption that dynamic cues are computed separately from static ones

(Humphreys et al., 1993). Although dynamic and static cues are processed separately in our model, they are combined to accomplish the tasks of recognition of identity and facial expression at the end of the hierarchy. This is also consistent with experimental data that show disruption in recognition when one of the two cues (dynamic or static) is altered (Hay et al., 1991; Kamachi et al., 2001).

The model presented in this article, and depicted in Fig. 2, leads to the hypothesis that motion is actually useful for successful matching of face images bearing distinct facial expressions. We further hypothesize that the computed motion fields could be used to select the most invariant (textural) features between the images we want to match. We have reported results that show the usefulness of adding the information supplied by the *DF* process within the matching task. As we have seen in the preceding section, recognition of identity is reduced by discarding the outcome of the *DF* module from the similarity function (i.e., going from Eq. (2) to Eq. (1)). We do not argue that this is the only way by which we accomplish such a task, but that it is an important one.

In addition, these motion features could also be used to construct a (motion) feature-space useful for recognition. Motion may be used as an alternative, independent means for identifying people and expressions. In computer vision, reasonable results have been obtained by constructing feature-spaces based solely on motion cues (Yacoob & Davis, 1996; Beymer & Poggio 1996; Bartlett et al. 1999; Pantic & Rothkrantz, 2000). These results could ultimately be used to reinforce the recognition task, or help to make a decision where other processes are not adequate.

We have shown, in this article, how we can extend and use our model to classify faces within a set of facial expression categories. We have also experimentally shown that the *DF* carries the necessary information to successfully achieve this task. Motion fields and linear discriminant analysis have shown to be useful for classifying facial expressions of emotions and to identity the AUs of an expression in several previous studies (Yacoob & Davis, 1996; Bartlett et al. 1999; Donat et al., 1999; Lyons et al., 1999; Calder et al., 2001; Pantic & Rothkrantz, 2000).

It is important to note that by combining the *DF* and a linear classifier, we were able to predict the classification RT of each of the facial expressions of the AR database. Furthermore, we were able to make some predictions that were later confirmed by a group of human subjects.

Other algorithms, such as the dynamic link architecture (Lades et al., 1993), may also justify some of the psychophysical data described in the introduction. However, as we have shown in this article, the motion estimation process not only is supported from a large number of psychophysical studies, but also has demonstrated to be useful for many other tasks in face recognition.

Additionally, neuroimaging and neurophysiological studies as well as single-cell recordings in monkeys, reveal areas in and near the STS that respond to social moving percepts (such as eye direction (Wicker et al., 1998; Puce et al., 1998)). These areas have also been found to be activated by static images of the face, which suggests that they are also sensitive to implied motion (Allison et al., 2000). Implied motion could be detected by a process similar to the *DF* module described in this article.

The model proposed in this article is also consistent with the idea of a hierarchical organization of the visual path along the ventral path (Ungerleider, 1995). In face recognition, it has also been argued that as the process of identification gets more specific, the areas activated shift progressively toward more anterior parts (Ungerleider, 1995; Haxby et al., 1994; Sergent et al., 1992). Our model is consistent with these findings.

Finally, the model proposed in this paper predicts that there could be agnosic patients that are impaired in facial expression recognition but have preserved identity recognition in the presence of facial expression changes. This would be possible when the facial expression categorizer(s) is (are) damaged but the *DF* area remains intact. Also, it may be possible to find an agnosic patient with damage into the *DF* area but intact face recognition and identification processes. This patient should have strong difficulties in classifying facial expressions and some difficulties in identifying identity with faces displaying large deformations (e.g., the scream face as shown in Fig. 5).

A limitation of the current model is that it is primarily focused on textural changes. Configurational changes though, are also expected to vary the RT and, therefore, it would be interesting to include them in the model.

Appendix A: Face images

The AR database of faces contains color images (i.e., three channels as given by the values of R, G, B) of 768×576 pixels. For our experiments, we reduced the original size to 165×120 pixels, and we converted the images to grey-level (i.e., one channel) by using the function $I = (R + G + B)/3$.

Before computing the flow field or similarities between faces, the face images need to be aligned. This is done to prevent results that are effected by factors such as scale or orientation (Beymer & Poggio 1996; Martínez, 2002). To do this, we manually marked the coordinates of the eyes, mouth, nose, chin and ears of each face image. Once these facial features have been localized, using the differences between the x and y coordinates of the two eyes, the original image is rotated until obtaining a frontal view face where both eyes have the same y value. Mathematically, $atan(\|y_1 - y_2\|/\|x_1 - x_2\|)$, where (x_1, y_1) and (x_2, y_2) are the right and left eye coordinates. Finally, the image is warped (or re-sampled by parts) until all the facial features detailed above are aligned. The results shown in Fig. 8 were obtained by further sampling the images to a size of 82 by 60 pixels.

Acknowledgements

The author would like to thank the referees and the editor for their comments, which contributed significantly to the improvement of an earlier version of this paper. The author was partially supported by NSF grant 99-05848. This research was partially conducted in the RVL and SLLL labs at Purdue University.

References

Allison,T., Puce, A., McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Science* 4:267-278.

Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. (1999) Measuring spatial expressions by computer image analysis. *Psychophysiology* 36:253-263.

Basilli, J. (1978). Facial motion in the perception of faces and emotional expression. *J. Exp. Psychology* 4:373-379.

Baudouin, J., Gilibert, D., Sansone, S., Tiberghien, G. (2000). When the smile is a cue to familiarity. *Memory* 8:285-292.

Beale, J.M., & Keil, F.C. (1995). Categorical effects in the perception of faces. Cognition 57:217-239.

Beymer, D., Poggio, T. (1996). Face recognition from one example view. *Science* 272(5250).

Black, M.J., Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding* 63:75-104.

Blake, A., Zisserman, A. (1987). Visual reconstruction. *Cambridge: The MIT Press.*

Bornstein, M.H. (1987). Perceptual categories in vision and audition In *Categorical Perception: the groundwork of cognition*, S. Harnad (Ed.), Cambridge University Press.

Brunelli, R., Poggio, T. (1993). Face recognition: features versus templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 15: 1042-1053.

Bruce, V., Velentine, T. (1988) When a nod's as good as a wink: The role of dynamic information in facial recognition. In *Practical Aspects of Memory: Current research and issues,* Vol. 1, Ed. M.M. Gruneberg, P.E. Morris, R.N. Sykes, U.K. Chichester, pp. 169-174, Wiley.

Bruce, V., Young, A. (1986). Understanding face recognition. *British Journal of Psychology* 77:305-327.

Bruce, V., Young, A. (1998). In the eye of the beholder, the science of face perception. *Oxford University Press.*

Bruyer R., Laterre C., Seron X., Feyereisen P., Strypstein E., Pierrard E., Rectem D. (1983). A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and Cognition* 2:257-284.

Calder, A.J., Young, A.W., Benson, P.J., Perrett, D.I. (1996). Categorical perception of morphed facial expressions. *Visual Cognition* 3:81-117.

Calder, A.J., Burton, A.M., Miller, P., Young A.W. & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research* 41:1179-1208.

Campbell, R., Woll, B., Benson, P.J. & Wallace, S.B. (1999). Categorical perception of face actions: their role in sign language and in communicative facial displays. *The Quat. J. Exp. Psych.* 52A(1):67-95.

Darwin, C. (1872). The expression of the emotions in man and animals. *London:John Murray, 1872.* (Re-printed by The University of Chicago Press 1965.)

Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cog. Sci. 5(9):382-387.*

DeValois R.L. & DeValois, K.K. (1975). Neural coding of color. In Handbook of Perception, E.C. Carterette & M.P. Friedman (Eds.), Vol. 5, Academic Press.

Donato, G., Bartlett, M.S., Hager, J.C., Ekman P. & Sejnowski, T.J. (1999) Classifying Facial Actions. IEEE Trans. Pattern Analysis and Machine Intelligence *21(10):974-989.*

Ekman P. & W. Friesen, W. (1978) Facial Action Coding System: A technique for the measurements of facial movements. Consulting Psychologists Press.

Endo, N., Endo, M., Kirita, T. & Maruyama, K. (1992). The Effects of Expression on face Recognition. Tohoku Psychologia Folia *52:37-44.*

Etcoff, N.L. (1984). Selective attention to facial identity and facial emotion. Neuropsychologia *22:281-295.*

Etcoff N.L. & Magee, J.J. (1992). Categorical perception of facial expressions. Cognition *44:227-240.*

Farah, M.J. (1990). Visual Agnosia: Disorders of object recognition and what they tell us about normal vision. Cambridge: MIT Press.

Fisher, R.A. (1938). The statistical utilization of multiple measurements. Annals of Eugenics *8:376-386.*

Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition. Academic Press (Second Edition).

Grossman, E., Donnelly, M., Price, R., Pickens, P., Morgan, V., Neighbor, G. & Blake, R. (2000). Brain areas involved in perception of biological motion, J. Cog. Neurosci. 12:711-720.

Hancock, P.J.B., Bruce, V. & Burton, A.M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Science* 4:330-337.

Hansch E.C. & Pirozzolo, F.J. (1980). Task Relevant Effects on the Assessment of Cerebral Specialization for Facial Emotions. *Brain and Language* 10:51-59.

Haxby J.V., Horwitz B., Ungerleider L.G., Maisog J.M., Pietrini P., Grady C.L., (1994). The functional-organization of human extrastriate cortex – a PER-RCBF study of selective attention to faces and locations. *J. Neuroscience* 14:6336-6353.

Haxby, J.V., Hoffman E.A. & Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Science* 4:223-233.

Hay, D.C., Young, A.W. & Ellis, A.W. (1991). Routes through the face recognition system. *Q. J. Exp. Psychol. A-Human Exp. Psy.* 43:761-791.

Hill, H. & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Curr. Biol.* 11:880-885.

Horn, B.K.P. & Schunck, B.G. (1981). Determining optical flow. *Artificial Intelligence* 17:185-203.

Hubel, D. & Wiesel, T. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophys.* 28:229-289.

Humphreys, G.W., Donnelly, N. & Riddoch, M.J. (1993). Expression is computed separately from facial identity, and it is computed separately for moving and static faces – Neuropsychological evidence. *Neuropsychologia* 31:173-181.

Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S. & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. *Perception* 30:875-887.

Kurucz, J. & Feldmar, G. (1979). Prosopo-affective agnosia as a symptom of cerebral organic-disease. *Journal of American Geriatric Society* 27:225-230.

Lades, M., Vorbrüggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz R.P. & Konen W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. on Computers* 42(3):300-311.

Lander, K., Christie, F. & Bruce, V. (1999) The role of movement in the recognition of famous faces. *Memory Cognition* 27:974-985.

Leopold, D.A., O'Toole, A.J., Vetter T. & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat. Neurosci.* 4(1):89-94.

Lyons, M.J., Budynek J. & Akamatsu, S. (1999) Automatic Classification of Single Facial Images. *IEEE Trans. Pattern Analysis and Machine Intelligence* 21(12):1357-1362.

Martínez, A.M. (2002). Recognizing Imprecisely Localized, Partially Occluded and Expression Variant Faces from a Single Sample per Class. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24(6):748-763.

Martínez A.M. & Kak A.C. (2001). PCA versus LDA. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 23(2):228-233.

Martínez A.M. & Benavente, R. (1998). The AR-Face Database. *CVC Tech. Report #24.*

Messing, L.S. & Campbell, R. (1999). Gesture, Speech, and Sign. *Oxford University Press.*

Pantic, M. & Rothkrantz, L.J.M. (2000). Automatic analysis of facial expressions: the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(12):1424-1445.

Puce A., Allison T., Bentin S., Gore J.C. & McCarthy G., (1998). Temporal cortex activation in humans viewing eye and mouth movements. *J. Neuroscience* 18:2188-2199.

Russell, J.A. (1980). A Circumplex model of affect. *J. Personality and Social Psych.* 39:1161-1178.

Schlosberg, H. (1941). A scale for the judgment of facial expressions. *J. Experimental Psych.* 29:497-510.

Schweinberger, S.R. & Soukup, G.R. (1998). Asymmetric relationship among perception of facial identity, emotion, and facial speech. *J. Exp. Psychology: Human Perception and Performance* 24(6):1748-1765.

Sergent, J. Otha, S. & MacDonald, B. (1992). Functional neuroanatomy of face and object processing – a positron emission tomography study. *Brain* 115:15-36.

Servos, P., Osu, R., Santi A. & Kawato, M. (2002). The neural substrates of biological motion perception: an fMRI study. *Cerebral Cortex* 12:772-782.

Ullman, S. (1996). High-level Vision: Object recognition and visual cognition. *MIT press*

Ungerleider, L.G. (1995). Functional brain imaging studies of cortical mechanisms for memory. *Science* 270:769-775.

Varga, R.S. (1962). Matrix iterative analysis. *Englewood Cliffs:Prentice-Hall*

Wallis, G. & Bulthoff, H.H. (2001). Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. USA* 98:4800-4804.

Wicker, B., *et al.* (1998). Brain regions involved in the perception of gaze: a PET study. *NeuroImage* 8:221-227.

Yacoob, Y. & Davis, L. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 18:636-642.

Young, A.W., Hellawell, D.J., Van De Wal, C. & Johnson, M. (1996). Facial expression processing after amygdalotomy. *Neuropsychologia* 34:31-39.

Young, A.W., Rowland, D., Calder, A.J., Etcoff, N.L., Seth, A. & Perrett, D.I. (1997). Facial expression megamix: Test of dimensional and category accounts of emotion recognition. *Cognition* 63:271-313.