IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7):769-775, 1997.

Object Recognition Robust Under Translations, Deformations, and Changes in Background

Rolf P. Würtz

Abstract— **Recognition systems based on model matching using low level features often fail due to a variation in background. As a solution I present a system for the recognition of human faces independent of hairstyle. Correspondence maps between an image and a model are established by coarse-fine matching in a Gabor pyramid. These are used for hierarchical recognition.**

Keywords— **Object recognition, face recognition, coarse-to-fine strategy, parallel processing, Gabor function, wavelet transform, image pyramid, correspondence problem, background independence.**

I. Introduction

The recognition of familiar objects in an unrestricted environment is an easy task for humans and an exorbitantly difficult one for computer vision. Images of "the same thing" can vary considerably by movements in space, changes in lighting, internal distortions and a different background. Thus, the abstract concept of an object can be defined as a huge *equivalence class* of the possible images showing that object, leaving the problem of an appropriate representation of such a class. It can be solved by storing examples of objects (which are called *models*) and comparing them with a given image. In the following I will present an algorithm which adapts this *model matching* strategy.

The problem of object recognition can be traced back to the *visual correspondence problem: Given two images of the same object, decide which pairs of points correspond to the same point on the physical object. It must also be decided which points have no corresponding partners in the other image.* Once correspondences are established for sufficiently many points between the image and the models representing different objects, the similarities of local features at corresponding points can be combined into a global similarity function. Its optimum over all models reveals the recognized object.

The algorithm is demonstrated here on face recognition. Nevertheless, the problems for which I am proposing a solution are of general nature, and it can be expected that at least parts of this algorithm will be useful for wider domains. Also, good solutions to the correspondence problem in the case of human faces will open the door to many applications beyond mere person identification. Examples include the analysis of facial expression or video manipulation.

II. Local features

For matching, local features must be as similar as possible, and their global arrangement must be preserved. If features are very local, they are also very ambiguous and unstable. If many pixels are are combined into features they become more sensitive to local distortions and many of them will be severely influenced by the background. A good compromise between those extremes is presented by Gabor features.

Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany. E-mail: Rolf.Wuertz@neuroinformatik.ruhr-uni-bochum.de

A. Gabor Wavelets

The features that have been used are extracted using 2-D Gabor functions [1] manipulated such that their Fourier transform vanishes at zero [2], [3]:

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2} \exp\left(-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}\right) \left[\exp\left(-i\vec{k}\vec{x}\right) - \exp\left(\sigma^{-2}\right)\right] . \tag{1}$$

The normalization of the kernels is — in contrast to wavelet theory — done such that all kernels roughly pick up equal amounts of "energy" from the image, see [2], [4] for details.

B. Sampling

Convolving the image I with all kernels $\psi_{\vec{k}}$ yields a continuous wavelet transform $\mathcal{W}\left(\vec{x}, \vec{k}\right)$ of the image, a function of four (two spatial and two frequency) coordinates. For implementation discrete sampling has been applied. Analogous to [2], [3] the frequency space is sampled in polar coordinates, with uniform sampling for the D directions of the center frequencies and geometric sampling for their L lengths. Spatial sampling is dependent on $\vec{k}$, and determined by the support of the thresholded kernel in frequency space.

A single complex value of $\mathcal{W}\left(\vec{x}, \vec{k}\right)$ will be referred to as a *unit* throughout this article. In order to avoid confusion with several notions of absolute value I am using the function $\mathcal{A}(\cdot)$ for the *modulus* of a complex number, and $\mathcal{P}(\cdot)$ for its *phase*. The vector of all units at a given location $\vec{x}$ and all D spatial frequencies with length $\left|\vec{k}\right|$ will be called a *feature vector* and denoted by $\vec{h}$, the single units making up $\vec{h}$ as h_i. Finally, the set of *all* responses with a fixed $\left|\vec{k}\right|$ is called a *frequency level*, for which the symbol $\mathcal{K}$ is used. Fig. 1 shows the spatial sampling on each level.

C. Amplitude Thresholding

The phases of units with low response amplitudes are ill-defined and numerically unstable [5]. Keeping such phases would diminish the reliability of the phase matching described in section III-D. For amplitude matching there is

Fig. 1. The spatial sampling points on the three frequency levels in image (above) and model (below). Some points are missing from the regular grid due to low amplitudes or background influence.

no big difference between a low amplitude and a zero amplitude. Therefore, zero amplitude and phase is assigned to all units with amplitudes smaller than a threshold t_a times the maximal amplitude.

D. Background suppression

Model and background are usually separated by a clear line. This need not be visible, but there are always points from model and background, respectively, which are direct neighbors in the image plane. When the background changes, units closer to the border than a distance $R(\vec{k}) = 2\sigma/|\vec{k}|$ will match poorly (a background can always be constructed such that these units can acquire arbitrary values). Thus, they must be discarded from the representation, which is done by convolution with a circle of radius $R(\vec{k})$.

E. Model and image representation

In summary, model representations are calculated by convolving the model image with the different Gabor kernels, followed by amplitude thresholding and discarding all units influenced by the background. The image also undergoes the convolutions and amplitude thresholding, but is not presegmented.

III. Correspondence Maps

In this section I describe four basic procedures for feature matching. Their combination leads to reliable and sufficiently dense correspondence mappings. A mapping is defined as a set of point correspondences $\mathcal{M}(M, I) = \left\{ (\vec{x}_i^M, \vec{x}_i^I) \,\middle|\, i = 1, \ldots, N \right\}$. For further evaluation its *size* $|\mathcal{M}|$ is defined as the number N of point pairs, the *average displacement* $\vec{A}(\mathcal{M})$ and the *distortion* $\vec{D}(\mathcal{M})$ with components D_1 and D_2 as follows:

$$\vec{A}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \left(\vec{x}_i^I - \vec{x}_i^M \right) , \tag{2}$$

$$D_m(\mathcal{M}) = \frac{1}{|\mathcal{M}| - 1} \sqrt{ \sum_{i=1}^{|\mathcal{M}|} \left(x_{mi}^I - x_{mi}^M - A_m(\mathcal{M}) \right)^2 } \tag{3}$$

$\vec{D}(\mathcal{M})$ is zero if and only if $\mathcal{M}$ is a simple shift or empty. Finally, given an image I, a model M, a correspondence mapping $\mathcal{M}$, and a local similarity function $\mathcal{S}_{loc}$, the *global* similarity between model and image is defined as the average over all local similarities:

$$\mathcal{S}_{glob}(M, I, \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \mathcal{S}_{loc}\left(\vec{h}\left(x_i^I\right), \vec{h}\left(\vec{x}_i^M\right) \right) . \tag{4}$$

A. Multidimensional template Matching

Template matching must be slightly modified for matching feature vectors rather than scalars. *Multidimensional template matching (MTM)* consists of finding the displacement $\vec{y}$ for a *data field* $\vec{f}(\vec{x})$ and a *template* $\vec{t}(\vec{x})$ such that the following becomes maximal:

$$\mathcal{S}\left(\vec{f}, \vec{t}\right)(\vec{y}) = \frac{\sum \mathcal{S}_{loc}\left(\vec{f}(\vec{x}), \vec{t}(\vec{x} - \vec{y})\right)}{\text{area}\left(|\vec{t}(\vec{x})|\right) \cdot \text{area}\left(|\vec{t}(\vec{x} - \vec{y})| \cdot |\vec{f}(\vec{x})|\right)} , \tag{5}$$

where the sum runs over the grid points in the support of its argument. The area-function in the denominator is the number of grid points in the support of its argument. is important not to normalize by the norm of the function but only by the number of points where the data vector $\vec{f}(\vec{x})$ and the shifted template $\vec{t}(\vec{x}-\vec{y})$ are both non-zero, because zeros are usually missing values rather than true zeros.

MTM is applied to the *amplitudes* of the units, because these are varying slowly. For the following two sections the local similarity function for two feature vectors will be:

$$\mathcal{S}_{\mathcal{A}}\left(\vec{h}^M, \vec{h}^I\right) = \frac{\sum_j \mathcal{A}(h_i^M)\mathcal{A}(h_i^I)}{\|\mathcal{A}(\vec{h}^M)\| \cdot \|\mathcal{A}(\vec{h}^I)\|}. \tag{6}$$

B. Global Matching

The first part of the mapping procedure consists in finding the part of the image where the object is located. For this it is sufficient to restrict the model and image representations to the lowest frequency level $\mathcal{K}_0$. The response amplitudes at $\mathcal{K}_0^I$ are used as the data for MTM and the amplitudes at $\mathcal{K}_0^M$ as the template. The result is a shift vector $\vec{y}_0$, which is added to every model point to yield a first estimate of the mapping:

$$\mathcal{M}_0(M, I) = \{\, (\vec{x}, \vec{x}+\vec{y}_0) \mid \vec{x} \in \mathcal{K}_0(M)\} \tag{7}$$

Although reconstruction from $\mathcal{K}_0$ would not yield a recognizable picture of the model the information suffices to find the correct location. On the lowest level the image information is smeared out so far spatially that the local distortions between model and image (which will be measured in the refinement steps) do not impede the template matching. This can only work if the spatial extent of the model is smaller than the one of the image, which is guaranteed by the background suppression.

C. Mapping Refinement

This section presents a method to refine a mapping $\mathcal{M}_n$ that has been established using the frequency levels $\mathcal{K}_0 \ldots \mathcal{K}_n$ to a mapping $\mathcal{M}_{n+1}$ using the information from level $\mathcal{K}_{n+1}$. The refinement is achieved by *local* MTM of the response amplitudes in the levels $\mathcal{K}_{n+1}(M)$ and $\mathcal{K}_{n+1}(I)$. Both levels consist of a rectangular matrix of feature vectors, missing feature vectors are replaced by zeros. The matrix of the model is divided up into small non-overlapping squares, whose *size* $\vec{s}^M$ depends on the level resolution and is chosen such that they contain, in general, 2×2 feature vectors. On the model borders or at possible holes resulting from amplitude thresholding some of the rectangles may contain only zero to three feature vectors. In the first case, they are dropped from processing, otherwise they are filled up with zeros. Each little square serves as a template for a local MTM.

The choice of the data field is more sophisticated. First, the point pair from the mapping $\mathcal{M}_n$ is chosen whose model point lies closest to the center of the template. If it is precisely at the center of the template, the corresponding image point becomes the center of the data field, which attains a fixed size $\vec{s}^I$, such that it contains (in general, like above) 3×3 feature vectors. If the mapping is not known yet at the center of the template, some heuristic must be applied in order to determine the center of the data field, and its size must be larger to account for the uncertainty in the correspondence. Let $\vec{c}^M$ be the center of the model template, $\vec{s}^M$ its size, $\vec{x}^M$ the model point closest to $\vec{c}^M$ which is part of the mapping $\mathcal{M}^n$, and $\vec{x}^I$ the corresponding image point. Then center $\vec{c}^I$and sizes $\vec{s}^I$ of the data area are defined as follows:

$$\vec{c}^I = \vec{x}^I + (\vec{c}^M - \vec{x}^M), \tag{8}$$
$$s_i^I = s_M^I \cdot s_i^M + 2 \cdot \left|c_i^M - x_i^M\right|, \ i = 1, 2. \tag{9}$$

This reflects the idea that where the correspondence is not known the best one can do is assume a constant deviation from the closest known mapping location. The poor knowledge is accounted for by searching over a larger data field in Eq. 9.

Now the data field is defined as all the feature vectors that fall inside the above rectangle. If it is empty, no correspondence is assigned to the points in the template. Otherwise, the existing feature vectors are arranged into a rectangular matrix, missing locations are assigned zero amplitude and phase. Then MTM is applied to the pair of template and data, yielding local shift vectors $\vec{y}_1$, which are relative to the mapping already known. For each $\vec{x}$ in $\mathcal{K}_{n+1}(M)$ *and* inside the current template the pair $\left(\vec{x}, (\vec{c}^I + \vec{y}_1) + (\vec{x} - \vec{c}^M)\right)$ is included in the mapping $\mathcal{M}_{n+1}$.

This procedure is executed for all the templates that make up $\mathcal{K}_{n+1}(M)$. It is worth noting that the single MTMs of the non-overlapping templates are independent of each other and can be executed in parallel. The data fields may overlap, which can lead to mappings that are unique but not invertible. Furthermore, the mapping need not be strictly neighborhood preserving. Both problems will be greatly alleviated by the removal of poor matches described in section III-E.

D. Phase adjustment

The two matching modules described so far have used only the *amplitudes* of the complex units. For most locations, their phases rotate with a frequency close to the center frequency of the generating wavelet. For phase matching it is assumed that the phase frequency is equal to the center frequency, except for points with small amplitude. Fleet [5] reports that it has proven sufficient to exclude points with amplitudes smaller than 5% of the maximal amplitude to reliably avoid these phase instabilities. I have adopted this value for the threshold t_a in section II-C.

The phase difference of two units found corresponding by amplitude matching is assumed to be caused by a local shift on the scale of the discretization. With this heuristic it is clear how the phases of two *units* have to be matched. The phase difference must be equal to the product of the displacement vector and the center frequency. For each

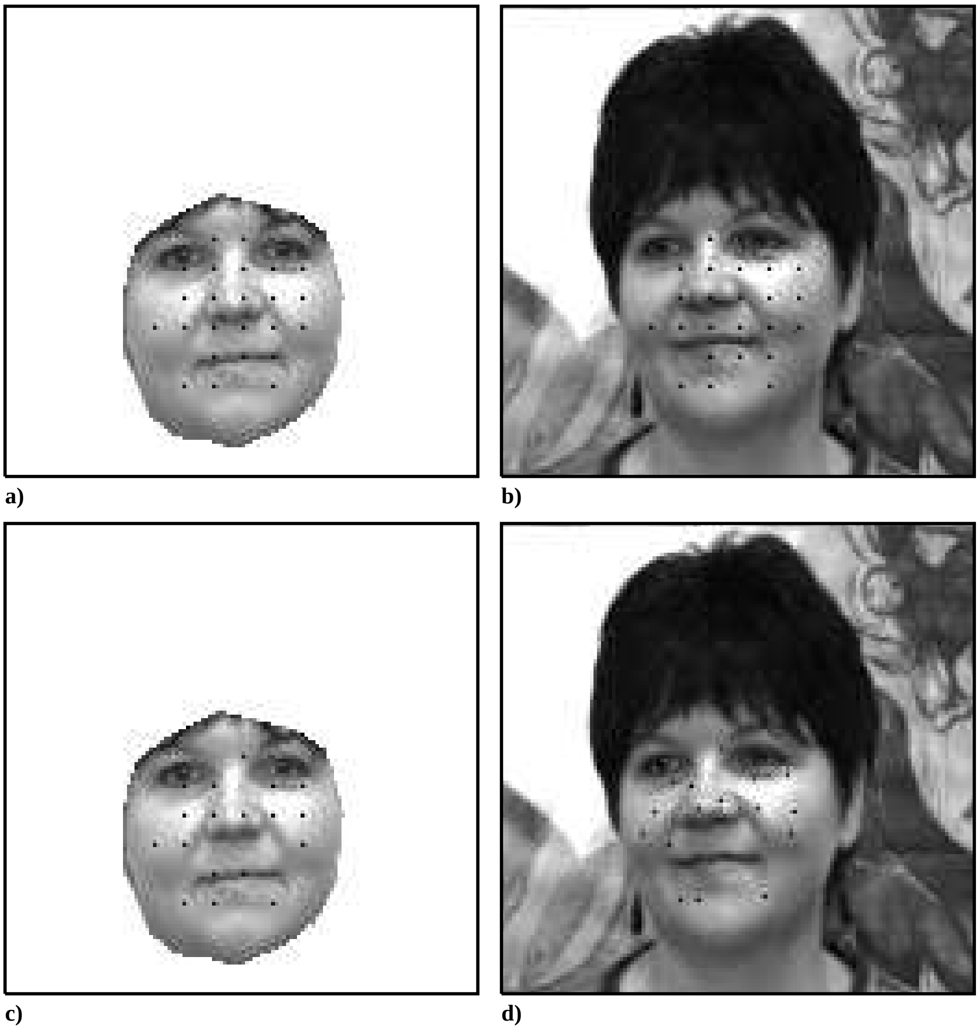

Fig. 2. Mappings on the lowest level ($|\vec{k}| = 0.4$). **a)** and **b)** show the mapping points of $\mathcal{M}_0^{\mathcal{A}}$, which results from multidimensional template matching. In **c)** and **d)** the phases have been matched and point pairs with poor similarity have been discarded resulting in $\mathcal{M}_0^{\mathcal{F}}$.

orientation this yields a displacement in the direction of $\vec{k}$. For matching a whole feature vector each of its D units votes for its displacement, and an agreement is reached by choosing the displacement $\vec{X}$ which gives the least squared deviation from the single ones. Basic trigonometric operations yield the following formula:

$$\vec{X} = \frac{2}{D} \sum_{j=0}^{D-1} \frac{\Delta\mathcal{P}\left(h_j^I, h_j^M\right)}{|\vec{k}|} \cdot \frac{\vec{k}_j}{|\vec{k}|}, \tag{10}$$

where the phase difference $\Delta\mathcal{P}$ is defined as $\mathcal{P}\left(h_j^I\right) - \mathcal{P}\left(h_j^M\right)$ if the amplitudes of both units are non-zero, and as zero otherwise. After this matching step, the similarity of the local feature vectors must be reevaluated, i.e., the remaining phase difference must become part of a new local similarity function $\mathcal{S}_{\mathcal{P}}$:

$$\mathcal{S}_{\mathcal{P}}\left(\vec{h}^M, \vec{h}^I\right) = \frac{\sum_j \mathcal{A}\big(h_j^M\big)\,\mathcal{A}\big(h_j^I\big)\cos\left(\Delta\mathcal{P}\big(h_j^M, h_j^I\big) - \vec{k}_j\vec{X}\right)}{\left\|\mathcal{A}\left(\vec{h}^M\right)\right\|\left\|\mathcal{A}\left(\vec{h}^I\right)\right\|}. \tag{11}$$

This is identical to $\mathcal{S}_{\mathcal{A}}$ if the phase differences after the applied local shift are zero, remaining phase differences leading to a penalty.

E. Elimination of poor correspondences

The mapping procedures described so far still have a serious drawback: Every point in the model always finds a correspondence in the image, which is not acceptable in the case of occlusion. Unless further knowledge of the object classes or three-dimensional models are employed the only grounds on which this can be detected is the actual similarity between features of the corresponding points. In order to exclude mismatches a *relative similarity threshold* q_n is introduced for every frequency level. All point pairs $\left(\vec{x}^M, \vec{x}^I\right)$ are excluded from the mapping $\mathcal{M}$ for which the condition $\mathcal{S}_{loc}\left(\vec{h}(\vec{x}^M), \vec{h}(\vec{x}^I)\right) \geq q_n$ is violated, with

$$q_0 = \mathcal{S}_{glob}\left(M, I, \mathcal{M}_0\right), \tag{12}$$

$$q_{n+1} = \mathcal{S}_{glob}\left(M, I, \mathcal{M}_n\right), \; n > 0\,. \tag{13}$$

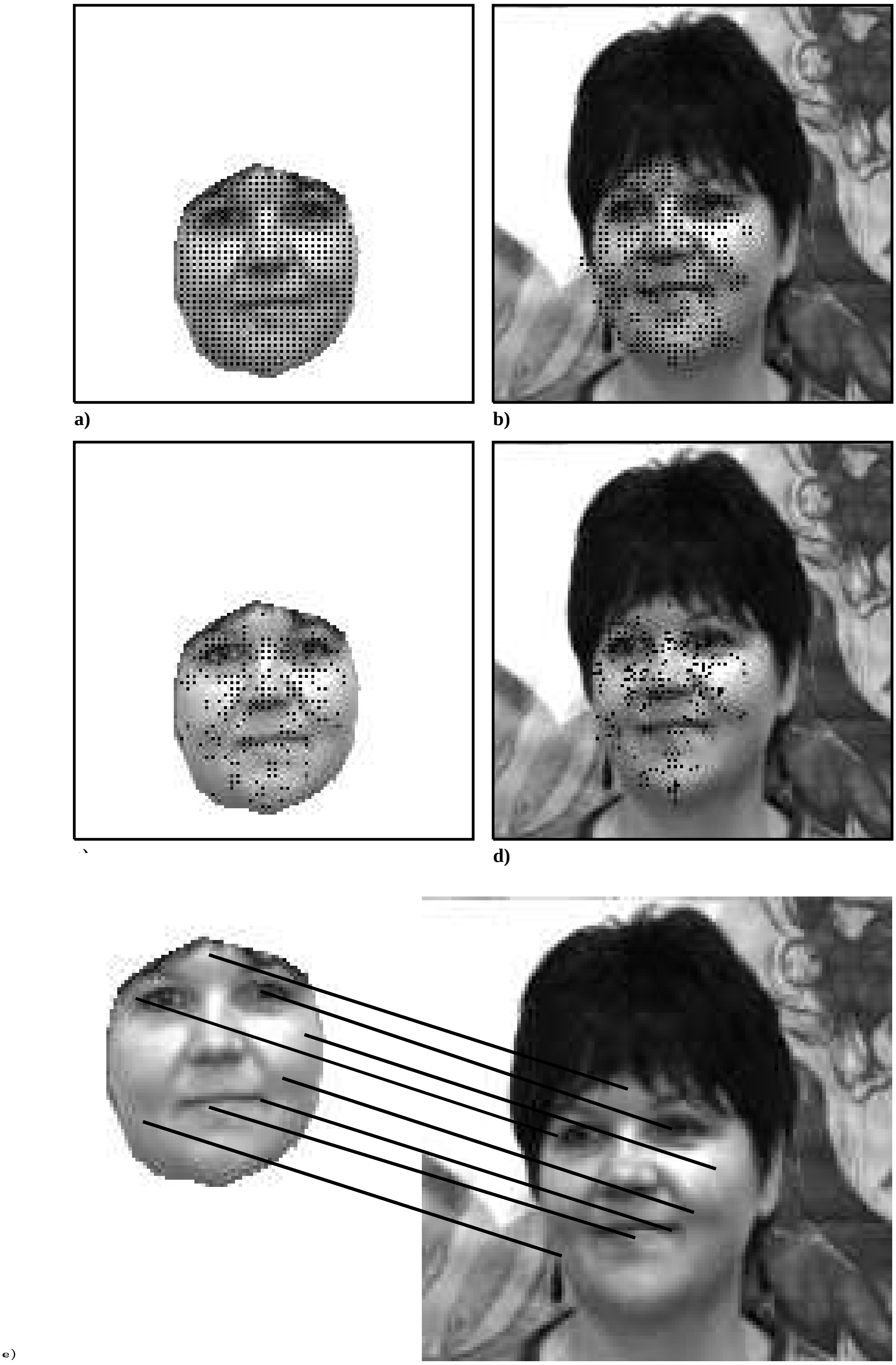

Fig. 3. Mappings on the highest level ($|\vec{k}| = 1.5$). **a)** and **b)** show the mapping points of $\mathcal{M}_2^{\mathcal{A}}$ obtained by local template matching using information from the middle level, which was in turn created from the one on the lowest level shown in Fig. 2. **c)** and **d)** show the mapping $\mathcal{M}_2^{\mathcal{F}}$ after phase matching and discarding poor matches, **e)** selected correspondences from $\mathcal{M}_2^{\mathcal{F}}$.

$\mathcal{S}_{glob}(\mathcal{M}_n)$ is the global similarity from Eq. 4, $\mathcal{S}_{loc}$ is the local similarity function from Eq. 6 or Eq. 11, respectively. For all levels except the lowest one this threshold is calculated using only information from mappings already known. Thus, no global information about the current mapping is required, and the refinement steps can be executed in parallel over all model locations.

F. Overall mapping procedure

Now there is a mapping initialization, a method to refine a mapping using information from a higher frequency level and two methods for improving a mapping, namely phase matching and dropping correspondences with poor similarity. They are combined as follows: The initialization yields a mapping $\mathcal{M}_0^{\mathcal{A}}$. Phase matching matched yields $\mathcal{M}_0^{\mathcal{P}}$, from which q_0 is derived. Then the poor matches are removed from $\mathcal{M}_0^{\mathcal{P}}$, which leads to $\mathcal{M}_0^{\mathcal{F}}$, the final mapping on this level. This mapping together with the levels $\mathcal{K}_1(M)$ and $\mathcal{K}_1(I)$ is used for the refinement step, which results in $\mathcal{M}_1^{\mathcal{A}}$. Phase matching yields $\mathcal{M}_1^{\mathcal{P}}$ and with q_1 as derived from $\mathcal{M}_0^{\mathcal{F}}$ this is reduced to $\mathcal{M}_1^{\mathcal{F}}$. The same step is iterated until the frequency levels are exhausted, (one more time with the current parameters), and results in the mappings $\mathcal{M}_2^{\mathcal{A}}$, $\mathcal{M}_2^{\mathcal{P}}$, and $\mathcal{M}_2^{\mathcal{F}}$.

The rough global organization of the mapping is achieved at a low spatial frequency, where distortions do not matter. The low level can only yield background independent mappings for the interior of the model. The refinement steps find reliable mappings closer to the boundary, because the influence of the background is reduced to a smaller area. Also holes in the mapping can be filled in at higher frequencies. Phase matching at each level achieves high accuracy, see Figs. 2 and 3 for examples.

IV. Recognition

A. Recognition from correspondence mappings

Each correspondence mapping can be used to define a global similarity between model and image as a linear combination of the global similarity $\mathcal{S}_{glob}$ (see Eq. 4) and the length of the distortion vector (see Eq. 3).

$$\mathcal{S}_{rec}(M, I, \mathcal{M}) = \mathcal{S}_{glob}\left(\mathcal{M}(M, I)\right) - \left|\vec{D}(\mathcal{M}(M, I))\right| . \quad (14)$$

Leaving out the distortion led to poor recognition results, which shows that pure feature similarities are not enough to discriminate between the models. Evaluating the similarity $\mathcal{S}_{rec}(M_i, I, \mathcal{M}_i)$ for a database of models $\{\, M_i \mid i = 0 \ldots N\}$ yields a series of similarities, whose maximum corresponds to the recognized model.

B. Recognition significance

This process always yields a model with highest similarity to the given image, no matter if the correct person is actually contained in the database. Thus, the *significance* of a recognition must be extracted from the the series of global similarities $\mathcal{S}_i$ ordered in ascending sequence, and M_i be the model with similarity $\mathcal{S}_i$. For the recognition to be significant $\mathcal{S}_0$, the similarity of the "candidate" model M_0 must be clearly distinct from all the other values. With s the standard deviation of the series $\mathcal{S}_i$ without $\mathcal{S}_0$ the criteria for the acceptance of a match is:

$$\kappa_1 = \left[\frac{\mathcal{S}_1 - \mathcal{S}_0}{s} > t_1\right], \kappa_2 = \left[\mathcal{S}_0 > t_2\right] . \quad (15)$$

Like in [2], both criteria are be combined with a logical OR in order to keep more significant recognitions while ruling out all incorrect ones. In the present evaluation, the model database always contained the correct person, and the following cases are possible. The best matching model M_0 can be correct (**C**) or false (**F**), and the significance criterion can accept (**A**) or reject (**R**) it. The combination of both yields a total of four cases, of which, ideally, only **CA** should occur. Any safe recognition algorithm must rule out case **FA**, while the case **CR** reveals an imperfection of the image or the algorithm. The quality of recognition can be judged by counting the number of **CA** cases once the thresholds t_1 and t_2 have been adjusted such that no **FA** cases remain.

C. Hierarchical recognition

An insignificant recognition means that no reliable decision was possible from the data available. Thus, the multilevel structure of the algorithm can be used to improve average recognition time and recognition quality. First, a recognition is attempted using only the mapping $\mathcal{M}_0^{\mathcal{F}}$ on the lowest level. Only for the **R** cases the next mapping $\mathcal{M}_1^{\mathcal{F}}$ is used and for the **R** cases on this level, recognition is again attempted using the mapping $\mathcal{M}_2^{\mathcal{F}}$.

Beside the practical advantages hierarchical recognition also models the psychophysical effect that a if low-pass filtered and subsampled image is represented by sharp squares, it can only be recognized after low-pass filtering. On the frequency levels present in the low pass filtered image a correct recognition is possible but it is not significant enough for the visual system to be satisfied. If no higher frequency information is available, this is the final result of the recognition attempt. In the presence of faulty high frequency information recognition is again tried on the next higher level, where it completely fails.

V. Experiments and results

The results have been obtained with the following parameters: The image resolution was 128×128 pixels, the number of frequency levels and directions was $L = 3$ and $D = 4$, respectively. Minimal and maximal frequency was $k_{min} = 0.4$, $k_{max} = 1.5$ in frequency space coordinates ranging from $-\pi$ to π. The ratio σ of window width and wavelength in the Gabor function was chosen to be 2.0, yielding kernels that are close to receptive fields found in the visual cortex [1], [3].

For performance measures two model databases **M1** and **M2** and three image databases have been set up. **M1** consists of 83 persons looking straight into the camera, whose images have been segmented by a simple rectangle which has the same size for all models for a fair comparison with the system described in [2]. **M2** consists of the same images as **M1**, now segmented by hand such that the hair is

Method	M1↔I1		M1↔I2		M2↔I1		M2↔I2		M1↔I3	
	C	**CA**	**C**	**CA**	**C**	**CA**	**C**	**CA**	**C**	**CA**
Hier. level 0	71	54	68	23	42	17	41	11	0	0
Hier. level 1	40	29	62	45	73	54	59	18	0	0
Hier. level 2	15	12	20	8	24	23	50	23	0	0
Hier. total	–	95	–	76	–	94	–	52	–	0
Level 2 only	94	89	86	69	94	93	78	49	0	0
FACEREC	95	93	92	81	19	1	14	1	95	93

TABLE I

Recognition results. All numbers are percentages of the size of the test database.

invisible and only the faces proper remain. There is no need for precision in this segmentation, which demonstrates the capabilities for recognition independent of the background, and more specifically, recognition of persons independent of their hairstyle.

Image database **I1** was used to test the performance under moderate conditions and consists of the same 83 persons looking 15° to their right. Database **I2** introduces hard conditions, namely 3 pictures of each person looking 15° and 30° to their right and one showing a facial expression of their choice (a total of 249 images). Experiments were conducted by comparing all images from **I** with all models from **M**. For each experiment the thresholds t_1 and t_2 have been adjusted such that false positive recognitions (**FA** cases) are reliably excluded. The results are summarized in Tab. I.

Image data base **I1** did not pose problems for recognition with either model database. For database **I2** the number of correct and significant recognitions drops considerably when switching from model database **M1** to **M2**. This shows that the feature vectors inside the face are distorted strongly, and the recognition must rely more on the overall appearance or outer form. In [2], the high similarity of all silhouettes of the same persons probably made the recognition problem simpler than it really is.

A third image database **I3** has been used which consisted of the negatives of **I1** in the sense that the grey value $I(\vec{x})$ of each pixel was replaced by $255 - I(\vec{x})$. Such images of human faces are hard to recognize [6], which proves that human face recognition must make some use of the Gabor phases. The recognition system failed to recognize the face from a negative image in all cases for both segmentation schemes, because the assumption that phase differences are caused by local displacements becomes completely wrong.

The results in Tab. I show that correct recognition is indeed possible from the lowest level on and the average recognition time can be reduced by the hierarchical approach. Furthermore, hierarchical recognition was always superior to the one from the highest level alone. This gives interesting insights into the distribution of prominent recognition cues across the spatial frequency range. More details can be found in [4].

As a general observation over many examples the correspondence maps obtained by this method are very reliable. This is of importance not only for good recognition results but also for the tracking of facial points [7] and possibly for measuring emotion. As there is no objective method of finding the true correspondences, this claim has been checked on a variety of model/image pairs but not been proved quantitatively. Examples for correspondence maps are shown in Figs. 2 and 3.

VI. Discussion and relation to other work

I have presented an object recognition system which is robust under changes in background, small deformations and translations. Its capabilities on hairstyle-invariant face recognition have been demonstrated. The background invariance makes it superior to the system described in [2].

In the system from [2] (which is called FACEREC in Tab. I) the concept of a *jet* requires frequency independent sampling of Gabor responses, which makes it hard to prevent the contamination of jet components by the background because the large supports of the low frequency kernels leads to the exclusion of nearly all jet locations. This is even more serious because the optimal value of the relative bandwidth σ is 2π in the case of FACEREC. Here, $\sigma = 2$ has been used, which leads to smaller kernels while covering the same frequency range, and those kernels are closer to physiological data. Applying the FACEREC-procedures without modification to the model database **M2** yields very poor results (see Tab. I). This shows that the problem of background influence on the features may not be neglected lightly. Finally, the concept of elastic graphs hinder massively parallel implementation, because the update of one position propagates via the elasticity term to all other positions. In the hierarchical system, all refinement steps are completely independent of each other, which makes the parallel complexity proportional to the number of frequency levels involved.

I close the discussion by comparing this system to the ones in [8], which describes the state of the art in face recognition in 1995. The proposed system performs better than all of those in at least one of the following respects:

1. No manual point correspondences or pose normaliza-

tion have been used.

2. Recognition is independent of background and hairstyle.
3. No specific properties of faces have been used.
4. Neither a 3-D model nor multiple views have been used.

Acknowledgements

The author is indebted to Christoph von der Malsburg and Charles H. Anderson for a multitude of ideas which have shaped this work. Thanks also go to Tino Lourens for critical reading of the manuscript. Financial support by grants from the HCM program of the European Community and the NAMOS project by the German Minister for Science and Technology is gratefully acknowledged.

References

[1] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. ASSP*, vol. 36, no. 7, pp. 1169–1179, 1988.

[2] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 300–311, 1993.

[3] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, 1996.

[4] R. P. Würtz, *Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition*, vol. 41 of *Reihe Physik*. Thun, Frankfurt am Main: Verlag Harri Deutsch, 1995.

[5] D. J. Fleet, *Measurement of Image Velocity*. Dordrecht, Netherlands: Kluwer Academic Publishers, 1992.

[6] R. Phillips, "Why are faces hard to recognize in photographic negatives?," *Perception and Psychophysics*, vol. 12, pp. 425–426, 1972.

[7] S. J. McKenna, S. Gong, R. P. Würtz, J. Tanner, and D. Banin, "Tracking facial feature points with Gabor wavelets and shape models," in *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication Crans-Montana, Switzerland, March 1997* (J. Bigün, G. Chollet, and G. Borgefors, eds.), vol. 1206 of *LNCS*, pp. 35–42, Springer Verlag, 1997.

[8] M. Bichsel, ed., *International Workshop on Automatic Face- and Gesture-Recognition*, MultiMedia Lab, Department of Computer Science, University of Zürich, June 1995. Winterthurerstraße 190, CH-8057 Zurich, Switzerland, ebner@ifi.unizh.ch.