Estimation of Face Position and Pose with Labeled Graphs[*]

Norbert Krüger, Michael Pötzsch, Thomas Maurer, Michael Rinne

Ruhr-Universität Bochum,
Institut für Neuroinformatik,
D–44780 Bochum, Germany

Abstract

We present a new system for the automatic determination of the position, size and pose of the head of a human figure in a camera image. The system is an extension of the well–known face recognition system [WFK] to pose estimation. The pose estimation system is characterized by a certain reliability and speed. We improve this performance and speed with the help of statistical estimation methods. In order to make these applicable, we reduce the originally very high dimensionality of our system with the help of a number of *a priori* principles.

1 Introduction

In this paper we deal with two problems. Firstly, we describe a pose estimation algorithm based on Elastic Graph Matching (EGM) [LVB, WFK]. The algorithm is an extension of the face representation introduced in [WFK] to the problem of pose estimation, in [WFK] the poses of faces is assumed to be known. Secondly, we improve the performance and speed of the pose estimation algorithm by learning. This learning algorithm can be seen as an intermediate step towards the construction of an autonomous object recognition system that is based on examples and not on manually constructed world knowledge. At present, our mechanisms for comparing different poses are still in part constructed manually.

As basic local image features we use Gabor-based wavelets. As others before us, we treat the set of wavelets centered on one image point as a unit which we call it a "jet." Like many other object recognition systems (e.g., [LTC, KDN]), ours is based on object models (or rather, models for two-dimensional aspects of objects as they appear in the image). In our hands, aspect models have the form of graphs, the nodes of which are labeled with jets or bunches of jets and the links of which are labeled with distance vectors between nodes (for examples see Fig. 1b,c). This representation of objects applied to faces combined with EGM allows us to determine the pose and position of faces.

[*] Supported by EC grant ERBCHRX–CT–930097, an HFSPO grant and an US Army Research Lab grant (01/93/K–0109). For the experiments we have used the FERET database of facial images collected under the ARPA/ARL FERET program.

BMVC 1996 doi:10.5244/C.10.27

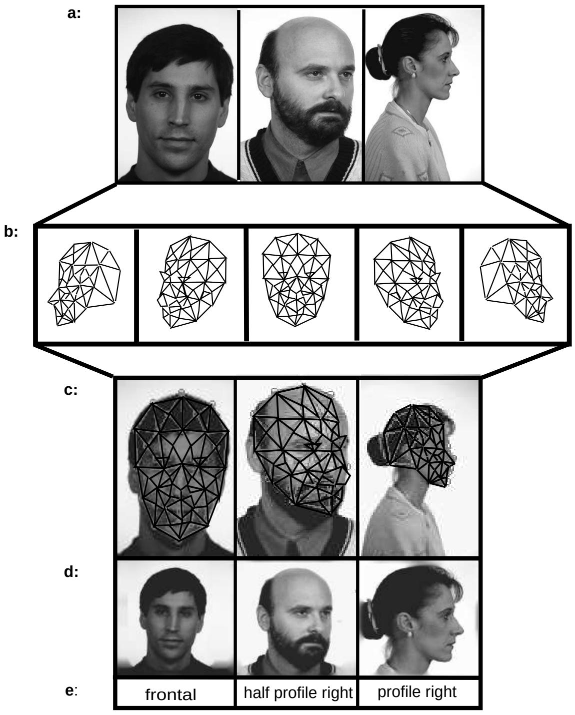

Figure 1: The head finding and pose estimation system. **a:** Input pictures with faces of different size and pose. **b:** Graphs representing the five poses: profile left, half profile left, frontal, half profile right and profile right. **c:** Input picture with the grid achieving the highest graph similarity S^G. **d, e:** The output of the system. The face normalized according to scale and the pose estimated by the system.

Starting from this extension of the original face recognition system described in [WFK] to pose estimation we optimize the representation of faces by statistical methods. We want to achieve an optimal but sparse representation of faces of different sizes and poses. The free parameters we have to determine in our model are the nodes of the graphs and the jets included in the bunches used to represent a face. The computational–time of the pose estimation system depends linearly on the number of nodes and jets in a bunch. We have formulated a number of *a priori* principles. These principles guide the selection of nodes and jets included in a bunch and allow an effective search in the huge combinatorical space of subgrids and subsets of jet bunches. The principles speak about the selection of landmarks on a face, starting with an (as yet hand-crafted) comprehensive set of landmarks, and about the selection of jets from sample images, to form bunches of jets for the landmarks. Our principles are

P0 (Locality): Features (jets, nodes) referring to different landmarks are treated as independent.

P1 (Maximal Discrimination): Features varying little within classes (poses) and varying much between classes are preferred.

P2 (Minimal Redundancy): Features should be selected for minimal redundancy of information.

Principles P0, P1 and P2 are applied in the form of several special formalizations in our algorithm. We distinguish between derived principles and formalized principles. A derived principle is a conceptual adaptation of P1 or P2 to a specific task, a formalized principle is a particular mathematical expression into which a derived principle is cast. With this hierarchy of abstraction we would like to stress the general applicability of our *a priori* principles to representations based on labeled graphs, independent of the features used as labels and of details of the matching process. In this paper we only introduce the derived principles; the formalized principles are desribed in [KPM].

Given a set of training images together with manually provided ground-truth as to the correct position of landmarks, each principle on its own would deterministically select a subset of landmarks and subsets of sample jets. However, these selections would differ for different principles and arbitration between them is required. This arbitration depends on relative weights given to the selection principles. We treat these relative weights and the number of features used during matching as free parameters and optimize them with respect to the overall performance of the system. These parameters constitute the total search space of our system of just five dimensions, a space we can afford to search exhaustively.

2 The Algorithm for Pose Estimation

In this section we describe our representation of faces of different size and pose with bunch graphes and our matching algorithm based on this representation.

Representation of faces

As models for object aspects our system employs labeled graphs (see figure 1). For estimating position, size and pose of a human head, the system uses a collection of such graphs. The edges of the graphs are labeled with distance vectors between node positions. Nodes are labeled with image information referring to landmarks, that is, local areas on a head or face such as the tip of the nose or the left eye. These labels are bunches of jets (a detailed description of a jet is given in the next paragraph). Each jet is derived from the image of a different person; a bunch thus covers a variety of forms a single landmark may take. We call a graph with nodes labeled with bunches of jets a *bunch graph*, an idea first introduced in [WFK]. The total model for heads used in our pose estiamation algorithm we call a *collection of bunch graphs*, each bunch graph representing a head in a certain pose and size.

Jets are derived from a set of linear filter operations in the form of convolutions of the image I with a set of Gabor wavelets $\psi_{\vec{k}}$ (cf. [Dau]), whose wavelength and orientation are parameterized by $\vec{k}$. The $\psi_{\vec{k}}$ take the form of plane waves restricted by a Gaussian envelope function (see Fig. 2a). Due to the spatial extent of the wavelets, jets describe a local area around their position. A bunch of jets taken at the same landmark (that is, at corresponding positions) of different faces forms a generalized representation of this landmark.

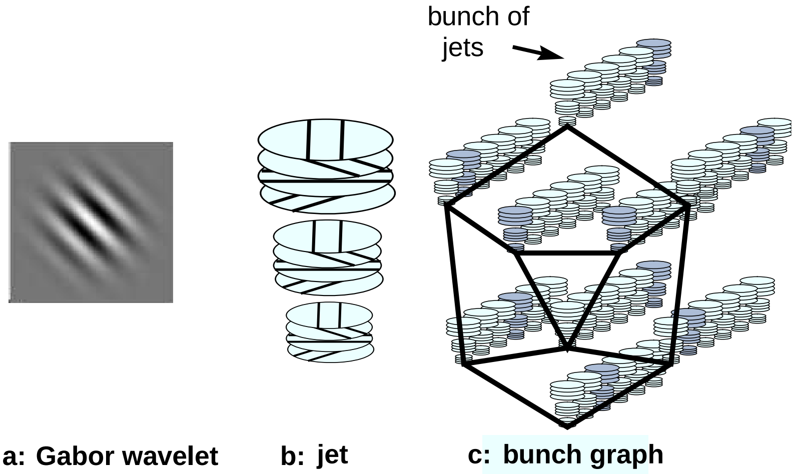

Figure 2: Representation of heads at a certain pose and size: **a**: the real part of a Gabor wavelet. **b**: a jet calculated as a set of Gabor wavelets (the discs symbolize the different frequencies and directions of $\vec{k}$). **c**: a bunch graph with the "expert–jets" found in a specific match symbolized as dark grey.

We create the bunch graph for a given pose by placing an appropriate graph (examples are shown in Fig. 1) over about 80 images of heads in that pose. All jets for a given landmark are attached as a bunch to that node. For each landmark, node positions are averaged, and the distance vectors between these average positions are stored as edge labels. One such bunch graph represents heads at a certain pose and size (see Fig. 1b,c and Fig. 2c). The collection of bunch graphs for five different poses p (frontal, left and right profile and half profile) and three different sizes s ($s \in \{$large, middle, small$\}$) form our raw knowledge base, to be optimized by the statistical methods described in section 3.

Elastic Graph Matching (EGM)

With the representation of faces described above we define a *total similarity* between a grid on a certain position on an image I and a bunch graph representing a certain pose with a certain size. This total similarity expresses the systems' confidence whether there is a face with a certain pose and size at a certain position on an image I. Here we give a informal description of this total similarity; a precise mathematical description can be found in [WFK, KPM]. Firstly, we define a similarity between two jets or *jet similarity.* We use two different jet similarities, the magnitude similarity [LVB] and the phase similarity [Wis]. The phase similarity is more sensitive to spatial displacements and gives as additional information to the similarity value an estimate of the displacement which allows more precise matching of landmarks. Utilizing the similarities of jets we define a *local similarity* between a jet extracted at a certain position in the Image J^I and a bunch of jets. The local similarity is defined simply as the maximum over all similarities between J^I and the jets in the bunchW we call the jet in the bunch graph for which this maximum is achieved an "expert–jet" (see Fig. 2c). Assuming a bunch represents, e.g., a left eye, the local similarity represents the systems' confidence whether J^I

represents a left eye. The total similarity between graphs is calculated as the sum of two terms, the average of the local similarities and a (negatively taken) measure for relative graph distortion, for details see [BLM, LVB].

A bunch graph is adapted to an image by EGM [LVB, WFK]. The total similarity is optimized by shifting, scaling and deforming the graph. The optimal similarity value for a graph gives the quality of its fit to the image. To estimate pose, the whole collection of bunch graphs is matched. The graph with the highest similarity determines the pose, but simultaneously it gives size and position of the face within the image, while the positions of its nodes identify the landmarks.

The complete graph matching process used in this paper proceeds in four steps. The matching procedure is performed for the bunch graphs of all poses and sizes in the collection and the one with the best similarity in the final step is selected, its identity determining pose and size. The four matching steps are: i) Rough location of the head in the image. ii) Adaptation of scale and improvement of location. iii) Independent scaling in x- and y-direction and further improvement of location. iv) Independent adaptation of node positions. In the first matching step we use the magnitude similarity and in steps ii)–iv) the phase similarity, utilizing the displacement estimation for fitting the grid to the correct location.

The system as described up to this point already determines position, size and pose of faces with fairly high reliability. We achieve a performance of 87.9% on a set of 413 pictures and the processing of one picture requires 112 seconds on a SPARC 20. In the following we optimize the speed and performance of the system described above by statistical methods. Instead of using all nodes and jets of our bunch graphs in all four matching steps, we drive towards a system which reduces matching time and improves overall performance by ignoring nodes and jets in our face representation in the first three matching steps. The selection of these is the subject of the next sections. It is important for the performance of the pose estimation system that the full grid is positioned after the final matching step. Since the fourth step is relatively fast, we are able to perform it on the full set of nodes and jets without significant increase of computational time.

The output of the pose estimation thus described can be used as input to the face recognition system described in [WFK], in which the pose of a face was assumed to be known.

3 Formalization of the Principles of Learning

We now proceed to formalize the principles discussed in the introduction. For each pose and for each matching step we would like to have sparse but efficient bunch graphs. The number of possible choices of bunch graph entries and subgrids is very large. It seems impossible to perform a learning in this large space, especially because the evaluation of the suitability of a specific selection of parameters takes a long time. By making use of principle P0 about locality and independence of nodes, we reduce this enormous combinatoric space to a sum of small spaces, which by applying the properly formalized *a priori* principles P1 and P2 can further reduce the dimension of the search space to five.

In the following subsections, we use P1 and P2 to define the derived principles for the task of finding suitable nodes and bunch graph entries. The derived principles can be devided into derived principles for choosing suitable jets in the bunch graph (DM1: Section 3.1) and derived principles for choosing a suitable subgrid (DG1, DG2, DG3 and DG4: Section 3.2). The corresponding formalized principles are called FM1, FG1, FG2, FG3, and FG4. Here we only introduce the derived principles. The precise formulation of FM1, FG1, FG2, FG3, and FG4 can be found in [KPM]. Here we describe their application in a nutshell. The formalized principles allow us to reduce the search space of all possible subgrids and bunches to a five dimensional space parameterized by $\alpha_1, \ldots, \alpha_5$. We introduce a function $Q(\alpha_1, \ldots, \alpha_5)$ measuring the overall performance of the system depending on the $\alpha_1, \ldots, \alpha_5$. This function is optimized by a standard optimization method to deterimine an optimal representation. The subgrids found as optimal by our learning algorithm are shown in figure 3.

Differening from the algorithm described in section 2 we introduce for the final decision of the pose estimation algorithm as total similarity a weighted average of the local similarities which takes the importance of the different landmarks for the pose estimation problem into account (see Fig. 4b). These weights are learned by an algorithm introduced in [Krü] which also makes use of derived principles based on P1.

3.1 Principles for the Selection of Jets

In our bunch graph approach, we store jets, extracted from approximately 80 images of pose p and size s, at the k-th landmark to represent the shape of a landmark. In [WFK] we based the selection of suitable sample images on intuitive criteria such as balancing the database in terms of gender or race, hoping to cover the space of eye jets, nose jets, etc. appropriately while avoiding redundancy. The same number of jets was used for all nodes. Here we define a more systematical selection of jets.

Assuming a landmark is already represented by a number of jets, as a simple application of principle P2 we can formulate

DM1 A new jet should be added only if it isn't similar to an existing one in the representation.

We implement this principle with the help of a very simple clustering algorithm. Before including a new jet we check whether there is one already in the same bunch that is close to this entry according to our similarity function. If that is the case we do not include the jet in in our bunch graph. Figure 4a shows the number of entries in the bunch graph after clustering for each pose for the largest size. It is obvious that certain nodes need fewer entries than others to cover their landmark. This fact is especially important regarding DG2.

3.2 Principles for Selecting Nodes

In this subsection we define the derived principles to extract the nodes which are important for head finding and pose identification. The bunch of jets which we use to represent a landmark can be seen as a composite feature which during the match is compared to the jet centered at a given pixel to decide whether the given landmark is present or not. The similarity values of these local comparisons are averaged over the nodes of the graph as a basis for the decision whether there is a head in a specific pose and size is found at the actual position in the image or not.

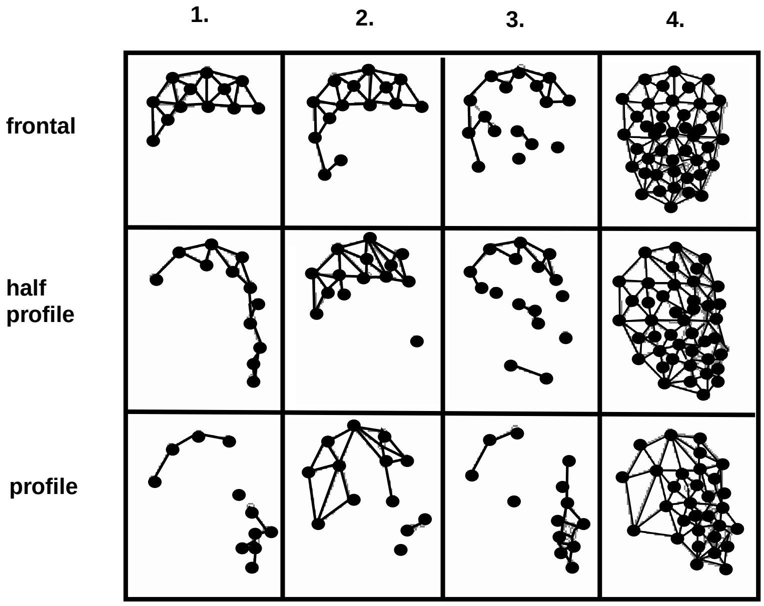

Figure 3: Grids used during the different matching steps. The columns correspond to the four matching steps, the rows to the three types of poses considered. We use the same bunch graph structure to represent a (half) profile left and a (half) profile right.

Furthermore at each node an estimate of the displacement to the correct location is computed. Bearing these remarks in mind we can formulate derived principles for the usefulness of a certain landmark of the grid to represent a face. A certain node is useful if

DG1 it provides high values at the correct position and low values at incorrect positions,

DG2 the number of jets needed to cover all variations of its local region is low,

DG3 the estimate of displacements of the node relative to the correct position is in most cases correct, and if

DG4 there is no other node nearby in the graph covering parts of the information of this landmark.

DG1 is a simple application of P1. DG2, DG3 and DG4 are applications of P2. DG2 is especially important for speed.

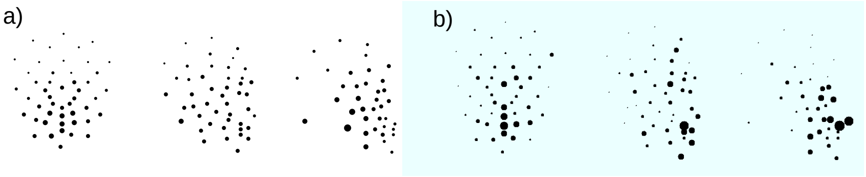

Figure 4: **a:** Number of entries in the clustered bunch graph corresponding to the largest size. There are significant differences in the number of jets needed to cover the different landmarks. For instance, for the frontal pose the mouth region has many more entries than the outline of the face or the eye region. Many entries are needed at the top of the head for the different poses. **b:** The learned weights for the pose estimation task for frontal (left), half profile (middle) and profile (right) views. The tip of the nose is very significant for the recognition of the frontal and half profile poses, the lips are very significant for the discrimination of frontals and half profiles. The eyes are not very significant for our pose estimation algorithm.

4 Results and Discussion

We tested our head finding and pose estimation algorithm on a set containing 413 pictures, of which 180 are frontals, 108 halfs (46 half left and 61 half right) and 126 profiles (26 left and 100 right). This set is completely independent of the data used for generation of the bunch graphs or our learning algorithm. When we do head finding and pose estimation with the system described in section 2 without any introduction of learning (i.e., with all nodes in each matching step and bunch graphs which are a not clustered) we achieve a performance of 87.9% and the processing of each picture requires 112 seconds. By introducing a sparse representation which is learned as described in section 3, we can improve performance to 92% correct pose identification and simultaneously achieve a speed-up by a factor of 6.2 (18.8 sec. per picture). With an even sparser representation we achieved a speed-up factor of 9.7 (11.5 sec. per picture), still maintaining a performance similar to the one obtained with the unreduced graph collection. Concerning the assessment of the result of 92% we like to remark that our algorithm has to simultaneously solve three problems: head finding, scale normalization and pose estimation.

The algorithm as described in section 2 already works with high reliabilty. By applying our learning algorithm described in section 3 we can achieve further gains in performance and a significant speed–up. The improvement of performance can be explained by the introduction of weighting of the nodes for the classification of poses. Furthermore, the accuracy of matching is improved by the fact that our learning algorithm allows us to use only important nodes for the head finding problem. The improvement of speed is caused from the sparseness of the learned representation.

5 Outlook

We admit that at the present stage our object representation is still very dependent on manual construction of the initial pose graphs and on manual provision of ground truth in the form of correct landmark positioning in training images. Our eventual goal must be to reduce the importance of these knowledge sources and to achieve a system learning autonomously, which can then recognize and track arbitrary objects in complex scenes. Furthermore, we believe that for further performance improvements our features, which are now rigidly confined to the form of wavelets, have to be replaced by some more flexible type which can flexibly adapt to the "essence" of landmarks, representing, for instance, an eye brow as a horizontal line which is perhaps slightly bending downwards [KrP, ShS]. Such a representation could reduce memory requirements compared to our bunch-of-jets approach, may increase speed of matching, and could be more reliable, because the system could more narrowly focus on essential features.

We believe that with a more extensive application of the concepts introduced in this paper, a system could autonomously learn the necessary representations and mechanisms to deal with arbitray objects in complex scenes.

Acknowledgement: We wish to thank Christoph von der Malsburg, Laurenz Wiskott, Michael Lyons, Ladan Shams, Jan Vorbrüggen and Gabi Peters for fruitful disussions.

References

[BLM] J. Buhmann, M. Lades, C vd. Malsburg. Size and Distortion Invariant Object Recognition by Hierarchical Graph Matching. Proceedings of the IJCNN International Joint Conference on Neural Networks, 1990. p:411-416.

[Dau] J.D. Daugman. Complete discrete 2-d Gabor Transforms by Neural Networks for Image Analysis and Compression. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 36:1169–1179, 1988.

[LTC] A. Lanitis, C.J. Taylor, T.F.Cootes, T.Ahmed; Automatic Interpretation of Human Faces and Hand Gestures Using Flexible Models. Proceedings of the International Workshop on Automatic Face- and Gesture recognition. Zürich 1995.

[KDN] D. Koller, K. Daniilidis, H.H. Nagel; Model-Based Tracking in Monocular Image Sequences of Road Traffic Scenes; International Journal of Computer Vision 10:3 (1993) 257–281.

[Krü] N. Krüger. Learning Weights in Discrimination Functions using a priori Constraints, G. Sagerer, S.Posch, F. Kummert: Mustererkennung 1995. Springer Verlag, p:110–117.
(See also WWW: http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLIST/1995/html/pub95.html)

[KrP] N. Krüger, G. Peters. Elastic Graph Matching with Learned local Filters (work in progress).

[KPM] N. Krüger, M. Pötzsch, C. v.d. Malsburg. Determination of Face Position and Pose with a Learned Representation based on Labeled Graphs, IRINI 96–03.
(See also WWW: http://www.neuroinformatik.ruhr-uni-bochum.de/ini/ALL/PUBLICATIONS/IRINI/irinis96.html)

[LVB] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, W. Konen. Distortion Invariant Object Recognition in the Dynamik Link Architecture. IEEE Transactions on Computers 1992, 42(3):300-311.

[ShS] L. Shams and J. Spoelstra. Learning Gabor-based Features for Face Detection. Submitted to World Congress on Neural Networks (WCNN) 1996 (San Diego).

[WFK] L. Wiskott, J.-M. Fellous, N. Krüger, C. von der Malsburg. Face Recognition and Gender Determination. Proceedings of the International Workshop on Automatic Face- and Gesture recognition. Zürich 1995.

[Wis] L. Wiskott. Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis. Verlag Harry Deutsch (Frankfurt am Main) 1995.