From: R.P. Würtz and M. Lappe (eds.), Dynamic Perception, pp. 121–126. infix Verlag/IOS press, 2002.

Gabor-based Feature Point Tracking with Automatically Learned Constraints*

Jan Wieghardt[1], Rolf P. Würtz[2], and Christoph von der Malsburg[2,3]

[1] SIEMENS AG, CT SE 1, Otto-Hahn-Ring 6, D-81730 München
jan.wieghardt@mchp.siemens.de
[2] Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum
rolf.wuertz@neuroinformatik.ruhr-uni-bochum.de
[3] LCBV, University of Southern California, Los Angeles, USA

Abstract. All point tracking mechanisms sometimes fail due to ambiguities in the visual data, a problem which can be alleviated by introducing model knowledge in the form of constraints on groups of feature points. Starting from a point tracking mechanism based on Gabor phases we introduce model constraints, on the one hand by posterior regularization (externally) and on the other hand by incorporating them directly into the tracking mechanism (internally). In the special case of facial feature tracking we show how the necessary model knowledge expressed in the constraints can be learned without explicit user interaction. To this end typical transformations of point groups are learned from noisy but automatically determined correspondences via principal component analysis.

1 Introduction

Tracking feature points reliably through a sequence of images is a much desired skill for all applications where trajectories need to be measured and evaluated. In this context *Gabor wavelets* have turned out to be well suited to determine the disparity between two points from consecutive images [3, 4]. The phase of the complex response to a Gabor filter varies nearly linearly for small translations in the image plane [1], which allows disparity estimation with subpixel accuracy. Another important feature are the multi-scale properties providing a very flexible point description and the ability to robustify disparity estimation over a wide range of scales.

Despite these advantages the tracking of individual feature points using Gabor wavelets still suffers from local image ambiguities like the infamous *aperture problem* that cannot be resolved without taking a larger context into account. Such a context can often be provided by a set of constraints on a whole *group* of points to be tracked. We propose a method which allows to incorporate the constraints directly during disparity estimation. Full details about method and results can be found in [5].

* Funding by European Commission in the Research and Training Network MUHCI (HPRN-CT-2000-00111) and the German Federal Minister for Science and Education under the project LOKI (01 IN 504 E 9) is gratefully acknowledged.

2 Disparity estimation

In the tracking algorithm in [3] the *disparity* of a point from one frame to the next is estimated in terms of phase differences of single *Gabor jets* with amplitudes a and phases ϕ. Extracting two jets at positions $\boldsymbol{x}$ and $\boldsymbol{x}'$, their relative disparity $\boldsymbol{d}$ can be calculated by maximizing their similarity

$$s = \frac{\sum_{\boldsymbol{k}} a_{\boldsymbol{k}}(\boldsymbol{x}) a_{\boldsymbol{k}}(\boldsymbol{x}')(1 - 0.5(\phi_{\boldsymbol{k}}(\boldsymbol{x}) - \phi_{\boldsymbol{k}}(\boldsymbol{x}') - \boldsymbol{k}^T \boldsymbol{d})^2)}{|\boldsymbol{J}(\boldsymbol{x})||\boldsymbol{J}(\boldsymbol{x}')|} \,. \tag{1}$$

The disparity is first estimated using only the lowest center frequency $\boldsymbol{k}$. Afterwards, in each iteration one additional level is added, and the corresponding phase differences are corrected modulo 2π. Thus, the lower frequencies can resolve the natural ambiguity modulo the wavelength for the higher frequencies. In case the estimated intermediate disparity exceeds twice the actual width of the Gabor function on the next higher frequency level, the process is terminated.

3 Tracking individual feature points

A tracking algorithm can be based on this disparity estimation, by executing the following steps for each frame (the parameter α can be adjusted to the expected variability of the visual features during tracking).

1. Extract jets $\boldsymbol{J}_i(\boldsymbol{x}_1(t_i)), \ldots, \boldsymbol{J}_i(\boldsymbol{x}_m(t_i))$ at current positions in frame I_i.
2. Update model jets $\boldsymbol{J}_i^{\text{model}}(\boldsymbol{x}_n(t_i)) = (1-\alpha)\,\boldsymbol{J}_{i-1}^{\text{model}}(\boldsymbol{x}_n(t_{i-1})) + \alpha \boldsymbol{J}_i(\boldsymbol{x}_n(t_i))$.
3. Calculate disparity to the jets extracted from the next image I_{i+1} at the same image-coordinates $\boldsymbol{d}_n = \boldsymbol{d}_n\left(\boldsymbol{J}_i^{\text{model}}(\boldsymbol{x}_n(t_i)), \boldsymbol{J}_{i+1}(\boldsymbol{x}_n(t_i))\right)$.
4. Calculate new positions in image I_{i+1}: $\boldsymbol{x}_n(t_{i+1}) = \boldsymbol{x}_n(t_i) + \boldsymbol{d}_n$.

4 Tracking constrained groups of points

Constraints for the disparities $\boldsymbol{d}_n$ of the points n can only come from a *parameterized model* of the possible variations. They take the general form

$$\boldsymbol{d}_n - \boldsymbol{f}_n(\boldsymbol{\epsilon}) = 0. \tag{2}$$

In this situation $\boldsymbol{f}_n$ is a model of the possible group motion and $\boldsymbol{\epsilon}$ are the model-parameters. E.g., if only image plane rotations are possible, $\boldsymbol{\epsilon}$ would contain the center and angle of the rotation, and $\boldsymbol{f}_n(\boldsymbol{\epsilon})$ the resulting displacement of point n. In practice, the equality is relaxed to a minimization of the norm of the left hand side of (2).

These constraints can be incorporated by first estimating the disparities assuming all nodes to be mutually independent and then calculating the constrained disparity configuration that is closest, in a least square sense, to the estimated disparities. The disparities are subsequently changed to those given by the constrained configuration. This method, which we call *external constraints*,

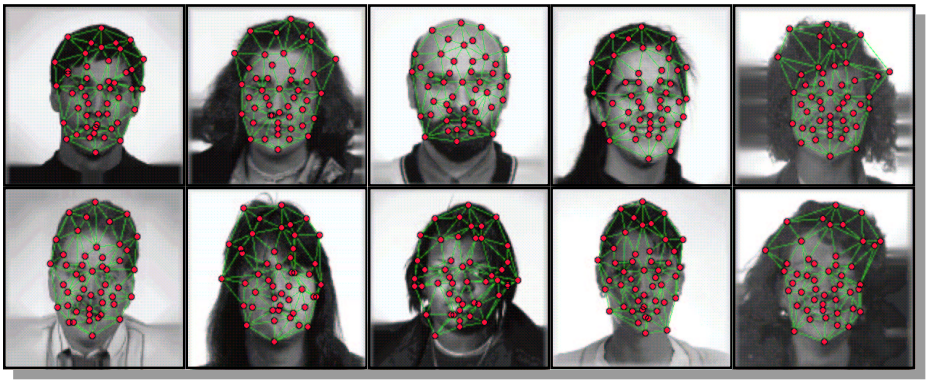


Fig. 1. Examples of automatically labeled faces: Displayed are 10 arbitrarily chosen examples of a set of approximately 1000 images. The retrieved correspondences are displayed by superimposing the *bunch graph*.

has serious drawbacks, as the separation of model knowledge and motion estimation forces decisions while estimating the initial disparities, even if the available image information is inadequate. This can cause small errors to accumulate and the total tracking result to deteriorate.

A better way is integrating the the model knowledge directly into motion estimation. Substituting constraints in the form of equation (2) into the phase-based disparity estimation of equation (1), the constrained disparities can be found by maximizing

$$s(\boldsymbol{\epsilon}) = \sum_n \frac{\sum_{\boldsymbol{k}} a_{\boldsymbol{k}}(\boldsymbol{x}_n) a_{\boldsymbol{k}}(\boldsymbol{x}'_n) \left(1 - 0.5 \left(\phi_{\boldsymbol{k}}(\boldsymbol{x}_n) - \phi_{\boldsymbol{k}}(\boldsymbol{x}'_n) - \boldsymbol{k}^T \boldsymbol{f}_n(\boldsymbol{\epsilon})\right)^2\right)}{|\boldsymbol{J}(\boldsymbol{x}_n)||\boldsymbol{J}(\boldsymbol{x}'_n)|}. \quad (3)$$

Applying a first order Taylor expansion and maximization in terms of $\Delta\boldsymbol{\epsilon}$ yields a linear equation system for $\Delta\boldsymbol{\epsilon}$, which can be solved during the coarse-to fine tracking described above. We term this use of model constraints *internal*.

5 Learning constraints from example data

Having established the need for constraints and a good way to apply them during tracking the question remains of how the correct constraints for an object class can be found. An analytical description will only be feasible in the simplest of cases, and it is desirable to learn the constraints from example images. We demonstrate a solution to this problem on face tracking. We match a bunch graph [6] onto a large set of more or less frontal faces. The resulting *correspondence fields* are converted into vectors and subjected to *Principal Component Analysis (PCA)* in a way similar to [2].

PCA yields the mean deformation and the deformations with the largest variation in the dataset. The first 6 are visualized in figure 2. As it turns out the

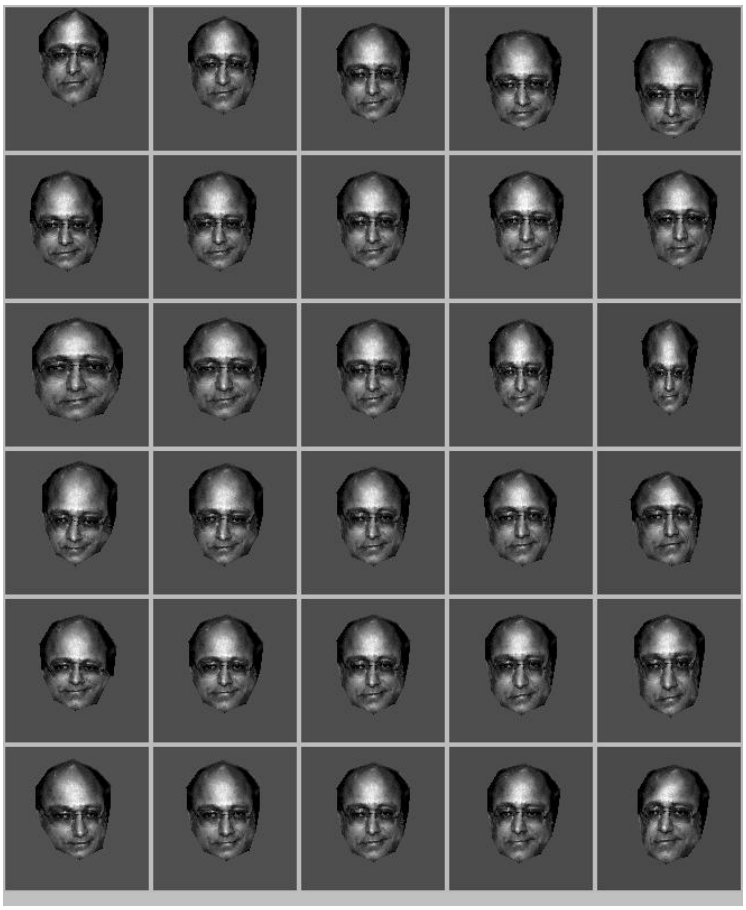
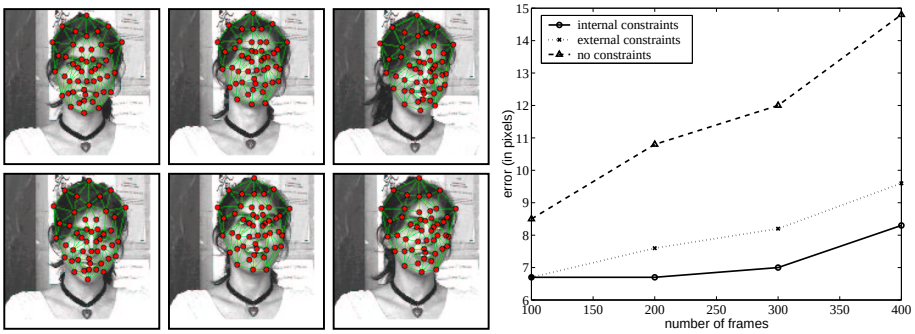


Fig. 2. Textured principal components of correspondence fields: The principal components $\boldsymbol{P}_1$ through $\boldsymbol{P}_6$ (top to bottom) of the feature point locations are illustrated here in terms of the mapping they perform on the standard gray value image shown in the central column. Each row shows the deformation from the mean along one principal component by -4,-2,0,2 and 4 standard deviations, respectively.

principal components are readily interpretable. They code transformations that are easily identified and named by visual inspection. The first one is a mixture of vertical translation and tilt, the second is horizontal translation, the remaining four contain scaling and rotation in depth. This is remarkable for several reasons. First, the results are based on a noisy database of automatically resolved correspondences. Although the database contained a lot of different individuals and was restricted to approximately frontal pose, the inter-individual variations (such as, e.g., jaw size or eye distance) are not dominant. The main variations seem to stem from geometrical variations. The only inter-individual variation visible in the first six components is expressed in the independence of scaling in x- and



(a) Example sequence

(b) Tracking error over time

Fig. 3. Tracking of faces: (a) shows the point positions on selected frames of a sequence, (b) the tracking error over time for internal, external, and no constraints, respectively. The constraints were derived from the first six principal components

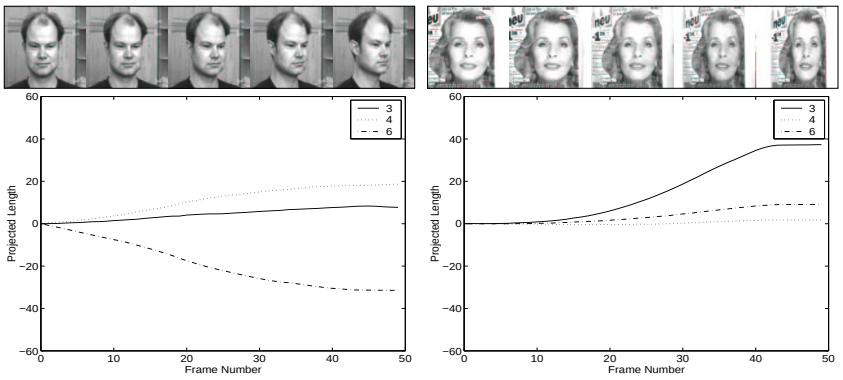
y-direction ($\boldsymbol{P}_3$ and $\boldsymbol{P}_5$), which might be attributed to different head shapes. Although no explicit knowledge about the three-dimensional transformations of rigid objects went into the constraint construction, their main properties were captured. Moreover, the degrees of freedom are nicely separated in an intuitive fashion.

An accurate model of the group motion of the selected feature points can thus be derived by assuming that the whole motion is restricted (or close) to the space spanned by the first principal components $\boldsymbol{P}_1$ through $\boldsymbol{P}_6$. Thus, the projection onto these components can serve directly as model parameters $\boldsymbol{\epsilon}$.

6 Results

Although the correspondences derived from bunch graph matching are far from perfect, the components with the highest eigenvalues seem to capture the major transformations that a face undergoes (see figure 2). They can directly serve as constraints and result in improved tracking performance. The results of three tracking procedures, namely unconstrained tracking, tracking with external constraints and the method proposed here using internal constraints are compared in figure 3(b) and clearly show the superiority of the latter.

Furthermore, the same constraints can be used to give rough pose information and distinguish 3D-motion of a true face from a rotated image of a face. It is remarkable how well the model captures the transformations of a moving face although no image sequences were provided when deriving the model. If the model parameters estimated by projecting the flow fields onto the first PCs are plotted over time for a sequence showing a moving head, as it was done in figure 4, it can be clearly seen that the derived motion model can be used for more than



(a) Head rotated by 30° (b) Journal cover rotated by 45°

Fig. 4. Principal components under rotation in depth: Shown are the projections of the correspondence fields on $\boldsymbol{P}_3$, $\boldsymbol{P}_4$, and $\boldsymbol{P}_6$, respectively, component as functions of the frame number for a real head (a) and a flat photograph of a head (c) monotonously rotating in depth. It can be clearly seen that $\boldsymbol{P}_4$ and $\boldsymbol{P}_6$ are closely related to a head's rotation in depth and its three-dimensional structure.

constraining the tracking. The transformation properties of faces, especially their behavior under rotation in depth, are so well captured that the model parameters themselves can be exploited to yield at least a qualitative pose estimation. The experiment with the journal cover shows that the resulting horizontal scaling can be clearly separated from the 3-D rotation of a real face.

References

1. D. J. Fleet and A. Jepson. Computation of component image velocity from local phase infor mation. *International Journal of Computer Vision*, 5(1):77–104, 1990.
2. A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding od face images using flexible models. *IEEE Trans. PAMI*, 19(7):743–756, July 1997.
3. T. Maurer and C. von der Malsburg. Tracking and learning graphs and pose on image sequences of faces. In I. Essa, editor, *Proc. 2nd AFGR*, pages 176–181. IEEE Computer Society Press, 1996.
4. S. J. McKenna, S. Gong, R. P. Würtz, J. Tanner, and D. Banin. Tracking facial feature points with Gabor wavelets and shape models. In J. Bigün, G. Chollet, and G. Borgefors, editors, *Proc. 1st AVBPA*, pages 35–42. Springer, 1997.
5. J. Wieghardt. *Learning the Topology of Views: From Images to Objects.* Shaker Verlag, Aachen, 2001.
6. L. Wiskott, J.-M. Fellous, N. Krüger, and C. v.d. Malsburg. Face recognition by elastic graph matching. *IEEE Trans. PAMI*, 19(7):775–779, 1997.