Learning an Object Model for Feature Matching in Clutter

Toni Tamminen and Jouko Lampinen

Laboratory of Computational Engineering
Helsinki University of Technology
P.O. Box 9203
02015 HUT, Finland
{toni.tamminen, jouko.lampinen}@hut.fi

Abstract. We consider the problem of learning an object model for feature matching. The matching system is Bayesian in nature with separate likelihood and prior parts. The likelihood is based on Gabor filter responses, which are modelled as probability distributions in the filter response vector space. The prior model for the object shape is learnt in two stages: in the first stage we assume only the mean shape known, with independent variations for each feature point, and match 'easy' images. We then estimate the characteristics of the shape variations for a realistic prior on the shapes. We demonstrate how incorporating the shape variation prior into the matching model enhances matching performance in the presence of clutter.

1 Introduction

In feature based object recognition a central task is to find the matching locations for the stored features of an object in a novel image. Common approaches to the problem include template matching with pixel difference or correlation similarity measures [1], edge and corner point matching methods especially popular in stereo vision research [2] and line segment matching methods [3]. However, for more complex tasks such as face recognition a richer set of features is needed.

We use a Gabor filter-based approach to feature matching in this paper with human facial images as our data. Similar features have previously been used in, e.g., [4], with error minimization based matching. In our system we apply Bayesian inference, with explicit likelihood and prior terms, and sampling methods to find the posterior distribution of the matching locations given the image.

In this paper we describe how the prior model for the locations of the features can be learned. In the first stage of the matching process we use a simple prior with independently Gaussian deviations from the mean to match a number of faces in absence of clutter. These matches are used to build a covariance model of the face shape. We then proceed to demonstrate how a good model of face shape greatly improves matching performance in the presence of clutter, when there are multiple possible matches for each individual feature.

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 193–199, 2003.

2 The Matching Model

We approach the matching problem from the Bayesian point of view [5], i.e. we consider all observable and unobservable quantities random variables and analyze their dependencies statistically. Our observed variable is the perceived image $\mathbf{I}$ and our unobserved variables are the feature locations $\mathbf{X}$ and possible hyperparameters ξ, which are the parameters of the prior on $\mathbf{X}$. Furthermore, all variables are conditional on a number of model assumptions (e.g. noise models and such), denoted by M. Our aim is to infer the posterior distribution of the unobserved variables given the perceived image and the model assumptions

$$p(\mathbf{X}, \xi | \mathbf{I}, M). \tag{1}$$

The perceived image $\mathbf{I}$ is first transformed to the feature space, $\mathbf{I} \mapsto \mathbf{T}$, so that each pixel (u, v) has associated features $\mathbf{t}(u, v)$. By replacing $\mathbf{I}$ in Eq. 1 with the transformed image, applying Bayes's formula [5] and integrating over the hyperparameters (since we are not interested in their actual values), we arrive at the posterior distribution

$$p(\mathbf{X}|\mathbf{T}, M) \propto \int p(\mathbf{T}|\mathbf{X}, \xi, M) p(\mathbf{X}|\xi, M) p(\xi|M) d\xi. \tag{2}$$

In Eq. 2 the terms of the model can be seen clearly: the likelihood $p(\mathbf{T}|\mathbf{X}, \xi, M)$ is the probability of observing $\mathbf{T}$ given the feature locations and hyperparameters, $p(\mathbf{X}|\xi, M)$ is the prior distribution of the feature locations given the hyperparameters, and $p(\xi|M)$ is the hyperprior. The likelihood can be simplified by having it be independent of the hyperparameters, in which case it becomes $p(\mathbf{T}|\mathbf{x}, M)$.

3 The Likelihood

In our model the likelihood is the probability of observing the feature image $\mathbf{T}$ given the feature locations $\mathbf{X}$ and the model assumptions M. We make the simplifying assumption that the likelihoods of the transformed pixels $\mathbf{t}$ are independent of each other and dependent only on the individual feature locations $\mathbf{x}_i$ and the background knowledge M_i pertaining to the ith feature, which allows us to write

$$p(\mathbf{T}|\mathbf{X}, M) = \prod_i p(\mathbf{t}_i|\mathbf{X}, M) = \prod_i p(\mathbf{t}_i|\mathbf{x}_i, M_i). \tag{3}$$

Thus we need to determine $p(\mathbf{t}_i|\mathbf{x}_i, M_i)$, the likelihood of observing a transformed pixel $\mathbf{t}_i$ given that feature i is located at $\mathbf{x}_i$, and compute this for all pixels $\mathbf{t}_i$.

As the transformation $\mathbf{I} \mapsto \mathbf{T}$ we employ a Gabor filter bank [6] with 3 frequencies and 6 orientations. The filter responses are stacked as vectors, or jets, and to define $p(\mathbf{t}_i|\mathbf{x}_i, M_i)$ we need to compare the perceived jets and the jets we would expect the feature to have. We do this by assuming that the distribution of the amplitude and phase jets in the vector response space are both Gaussian, and determine the distribution parameters by fitting a number of faces by hand

and measuring the jets at the feature locations. (These distribution parameters are contained in M_i.) It is then easy to compute the likelihood for each pixel of the perceived image by using the probability density function of the Gaussian distribution.

For contrast-independence, we transform the Cartesian jets into $n-1$ hyperspherical angle coordinates (i.e. jet length becomes 1) [7]. However, complete contrast-independence is not desirable as it causes the system to be sensitive to faint patterns and noise in uniform areas. To avoid this we add a term measuring the energy of the Gabor jet. This term also has a Gaussian form, with variance multiplied by an *ad hoc* constant to reduce its effect.

By combining the amplitude, phase, and energy components, we get the total likelihood, as follows:

$$\begin{aligned} p(\mathbf{t}_i|\mathbf{x}_i, M_i) \propto \quad & \exp(-\tfrac{1}{2}(G_{amp}-\mu_{amp})^T \Sigma_{amp}^{-1}(G_{amp}-\mu_{amp})) \cdot \\ & \exp(-\tfrac{1}{2}(G_{phase}-\mu_{phase})^T \Sigma_{phase}^{-1}(G_{phase}-\mu_{phase})) \cdot \\ & \exp(-\tfrac{1}{2\sigma_{energy}^2}(G_{energy}-\mu_{energy})^2), \end{aligned} \tag{4}$$

where the G_{amp}, G_{phase} and G_{energy} are the Gabor jet properties corresponding to $\mathbf{t}_i$ and $\mu_{amp}, \mu_{phase}, \mu_{energy}, \Sigma_{amp}, \Sigma_{phase}$ and σ_{energy} the corresponding distribution means and (co)variances. A sample likelihood field is illustrated in Figure 1. It can be seen that the likelihood is insufficient by itself for unambigous matching of the features, as there are many spurious responses in regions that have locally similar spatial frequencies.

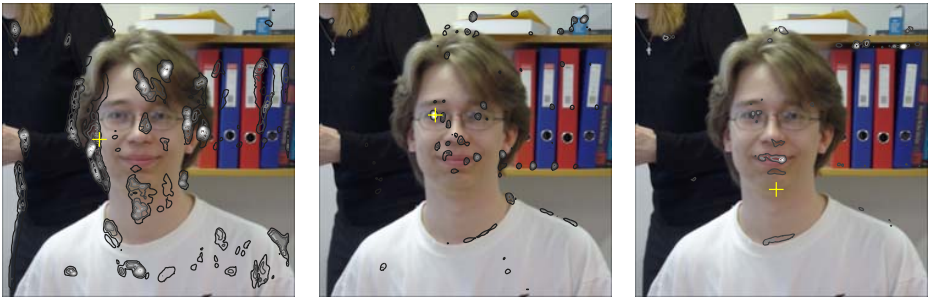

Fig. 1. The likelihood fields in sample images. Manually selected location of the features are marked as + in the images. There are multiple false matches in all the images, and the maximum likelihood for the chin point occurs in an incorrect location, as due to lighting the chin shadow is almost non-existent.

4 The Priors

In Bayesian inference, the prior for the unknown parameters in any given task is the posterior of the parameters given all the previously observed information on

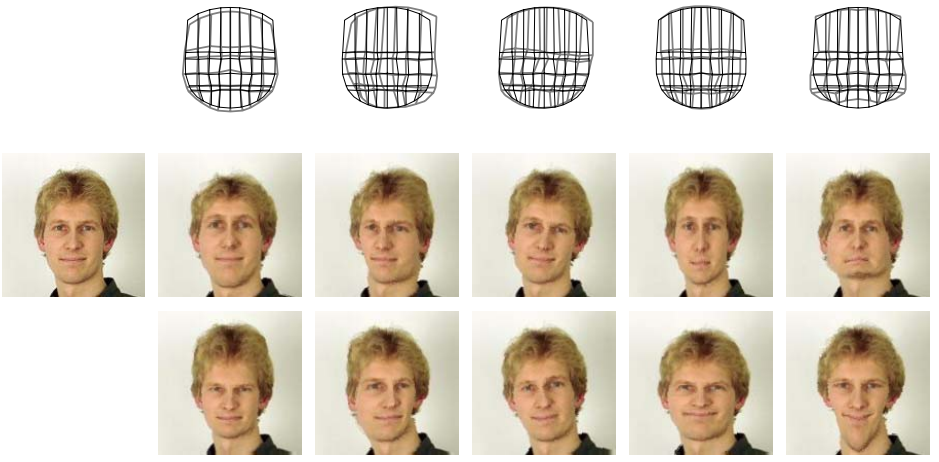

Fig. 2. The leading principal components. In the upper row, the dark grids show the mean shape and the lighter grids show the results of adding the leading eigenvectors to the mean. In the other rows, the face on the left has been morphed according to the principal components, both in the positive (middle row) and negative (lower row) directions. Components 2 and 3 appear to be related to rotations, while components 1, 4, and 5 are shape-related.

the parameters. So, in theory, we could start with a non-informative prior on the object shapes and update the prior after each new recognized shape to reflect the observed data. This would require a robust assessment of the goodness of the shape recognition, which is a rather difficult issue. For simplicity, we build the prior model in two stages. The first stage prior contains no information about the shape variations, but it is easy to define. This prior is used to match a number of faces in a controlled environment (clutter-free background, relatively standard lighting conditions), after which the matched faces are used to learn the covariance prior (or, to be more correct, the posterior $p(\mathbf{X}|D, M)$) which encodes the typical shape variations.

4.1 The Simple Prior

In the simple prior we assume that the features are ordered in a planar graph with independent Gaussian variations. A reference graph is computed from the manually annotated images used in the likelihood parameter estimation and during recognition, the prior mean on feature location $\mathbf{x}_i$ is obtained by scaling and translating the reference grid according to the other features. The variance parameter of the distribution determines the scale of alloved deviations from the mean. In the Bayesian approach we avoid fixing this value by setting an inverse-gamma prior on it and integrating the parameter out in the MCMC sampling of the parameters. For details on the sampling techniques see [8].

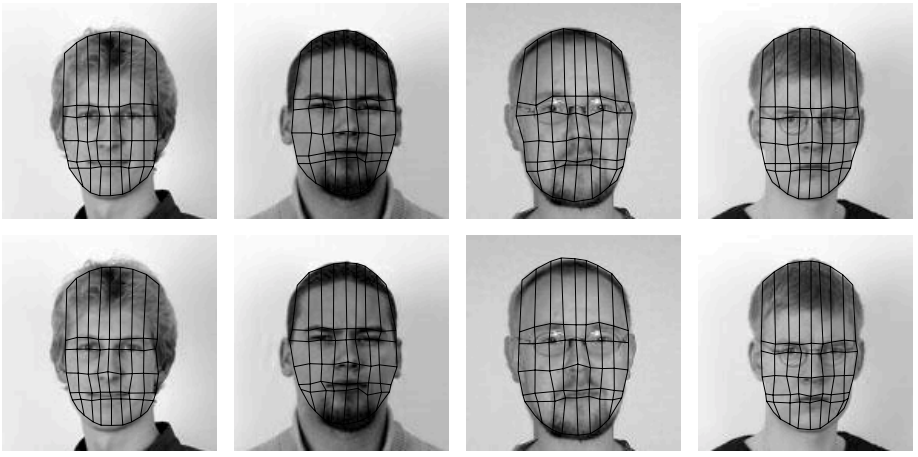

Fig. 3. Matching in the absence of clutter. The images in the upper row are matched using the simple prior, while the ones in the lower row have utilized the covariance prior. The priors perform approximately equally.

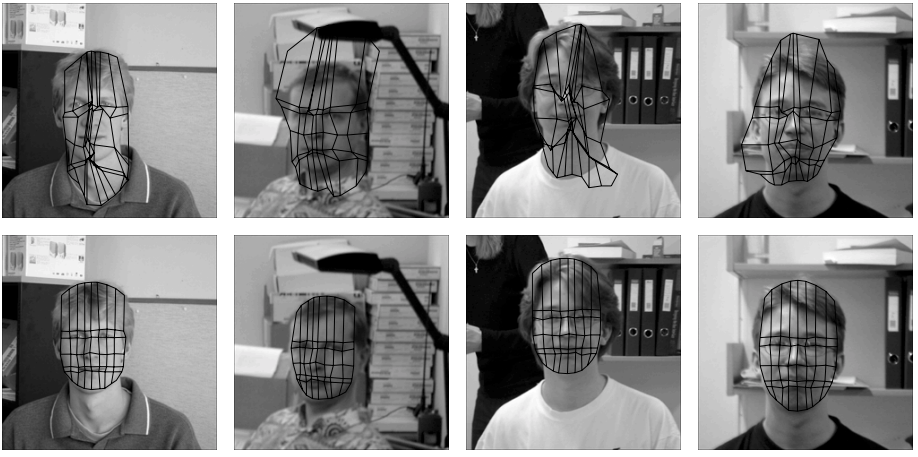

Fig. 4. The system tested with real images. The images in the upper row are matched using the simple prior, while the ones in the lower row have utilized the covariance prior. The superiority of the covariance prior is evident.

4.2 The Covariance Prior

Once we have matched a set of images (larger than the manually annotated one) using the simple prior, we can compute the posterior of the feature locations given the obtained matches $p(\mathbf{X}|D, M)$. We again assume this distribution to be Gaussian. However, since it is very high-dimensional (with 63 feature points, as

used here, the distribution is 126-dimensional), a very large amount of data is required to estimate the distribution parameters accurately. To deal with this, we find the most significant variations of the face shape by principal components analysis and project the estimated covariance matrix onto the principal components. The first five principal components are illustrated in Figure 2.

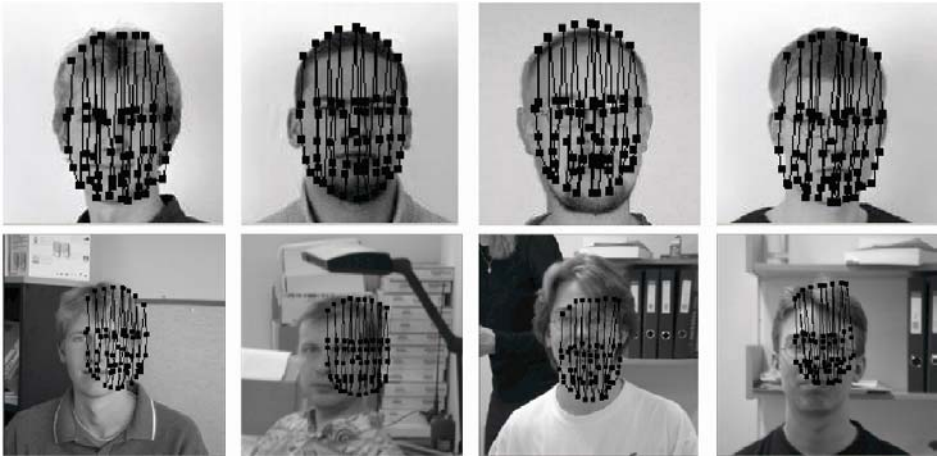

Fig. 5. Results with AAM. Without clutter (upper row) the matching result is slightly inferior to our system. In cluttered images (lower row) the matching fails completely.

5 Results

Figure 3 illustrates the performance of the two priors when the images are clutter-free. We produced samples from the posterior distribution of feature locations with Gibbs sampling (the simple prior) and Metropolis-Hastings sampling (the covariance prior) [9], and the shown grids are the medians of the samples. The system was also tested with a number of realistic images containing clutter. Some results are shown in Figure 4. In the clutter-free case the results are roughly equal. In cluttered scenes the performance of the covariance prior is much better. With the simple prior, some features tend to be matched to the false local maxima of the likelihood, because the simple prior does not enforce the graphs to be face-shaped.

For comparison, we used the same training data to build an Active Appearance Model (AAM) [10] and matched both the clutter-free and cluttered images with it. The matching was done using the AMMLab implementation by M. B. Stegmann [11]. The results are shown in Figure 5. In the clutter-free images AAM performs approximately equally to our system. However, when matching faces in cluttered scenes, the performance of the proposed system is clearly superior. It must be noted, though, that the quality of the training image set

was apparently insufficient for proper training of the AAM model - there were slight pose variations in the images, hampering especially the texture principal components, and the number of images was rather low. On the other hand, the results suggest that the proposed system has tolerable performance also on small imperfect training data.

6 Conclusion

We have presented a face matching framework in which the shape of the objects to be matched is learned in a bootstrap fashion. With a good measure of the goodness of fit, this kind of system could be used in unsupervised learning of image contents so that the system would match objects automatically and improve the prior according to them. However, in practice the efficiency of such a learning process would depend critically on the used measure. In addition to this, in the future we are going to work on an occlusion model and extend the system so that it is able to cope with larger distortions of pose.

References

1. Cox, G.S.: Template matching and measures of match in image processing. Department of Computer Science, University of Cape Town (1995)
2. Xu, G., Zhang, Z.: Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach. Computational Imaging and Vision, Vol. 6. Kluwer Academic Publishers (1996)
3. Zhang, Z.: Token tracking in a cluttered scene. Technical report RR-2072, Institut National de Recherce en Informatique et en Automatique (1994)
4. Wiskott, L., Fellous, J.-M., Krüger, N., von der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching. In Jain, L.C., Halici, U., Hayashi, I., and Lee, S.B. (eds.), Intelligent Biometric Techniques in Fingerprint and Face Recognition. CRC Press (1999)
5. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.R.: Bayesian Data Analysis. Texts in Statistical Science. Chapman & Hall (1995)
6. Daugman, J.G.: Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. IEEE Trans. on Acoustics, Speech, and Signal Processing **36** (1988) 1169–1179
7. Hassani, S.: Mathematical Physics. A Modern Introduction to its Foundations. Springer (1998)
8. Tamminen, T., Lampinen, J.: Bayesian Object Matching with Hierarchical Priors and Markov Chain Monte Carlo. In Bernardo, J.M., Dawid, A.P., Berger, J.O., West, M. (eds.), Bayesian Statistics 7. In press. Oxford University Press (2003)
9. Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds.): Markov Chain Monte Carlo in Practice. Chapman & Hall (1996)
10. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. IEEE Trans. on Pattern Analysis and Machine Intelligence **23** (2001) 681–685.
11. Stegmann, M.B.: Analysis and Segmentation of Face Images using Point Annotations and Linear Subspace Techniques. Technical Report, Informatics and Mathematical Modelling, Technical University of Denmark (2002)