Real-Time Scale Invariant Object and Face Tracking using Gabor Wavelet Templates

Alexander Mojaev, Andreas Zell

University of Tuebingen, Computer Science Dept.,
Computer Architecture,
Sand 1, D - 72076 Tuebingen, Germany
{mojaev,zell}@informatik.uni-tuebingen.de

Abstract. This paper describes a real-time technique for scale invariant object or face tracking with standard PC hardware. The tracking method is based on a low redundancy (compressed) object image representation. For image decomposition a fast non-iterative transform based on the odd-symmetric gabor functions is used, which guarantees on the one hand a low redundancy of the resulting representation and on the other hand an automatic detection of the significant features in the image. Tracking control is realized by a scale factor discriminator control technique and convolution based 2D cross-correlation. Experiments show that this method enables real-time object or face tracking in a wide scale changing range and provides strong robustness against camera shaking, which is important for visual robot control.

1 Introduction

Various applications in robotics such as human following, object grasping and manipulating, gesture recognition and visual navigation need a robust visual object tracking control. A big problem in visual data processing is that a robot must store a large amount of visual information to be able to operate in the environment, recognize and track objects and persons [9]. The goal of this work is to develop an effective way to use a low-redundancy image representation for robust tracking of objects or structures appearing in images. Such a representation must be able to appropriately detect and store the significant object features, which is important for robust object recognition and tracking. We can use the compressed object descriptor to quickly manipulate the object template and adapt it to the changes in the input video stream using a discriminator technique well-known from analogue signal processing. The discriminator based tracking control allows to follow very fast movements in the object position and scale.

2 Related Work

Many tracking techniques use colour, contour or geometric templates for object or face tracking [2]. Using a previously given model of the object reduces flexibility of the tracking system and makes the learning process more complicated. Other tracking methods use well-defined object features for the estimation of object displacements. Unlike

the model based feature extracting algorithms such as the Hough transform, the use of 2D-Gabor filters allows to detect and isolate disturbations in the spatial-frequency domain with no apriori knowledge about feature structure.

Daugman used 2D Gabor functions to approximate the impulse response of the simple cells in the striate cortex [1]. Lee describes in [6] the use of an ensemble of 2D Gabor wavelets (frame), whose structure is also biologically inspired, for image coding, and derived the conditions, under which the set will provide a complete representation of any image. Those frames allow a good quality of reconstructed images with quantized wavelet coefficients but are hardly suitable for extraction of specific object features. Manjunath, Shekhar and Chellappa show in [7], that object features can be extracted and tracked using Gabor filters. Zhang et al [12] have compared geometry based and gabor wavelet based approaches for facial expression recognition using a multi-layer perceptron. They showed, that the recognition rate with the Gabor wavelet coefficients is much better than the analysis of geometric positions of fiducial points on a face.

The well-known “Elastic bunch graph matching” [11] uses fixed points for calculating the wavelet coefficients of the Gabor filter. The points are precisely positioned on the original image to represent the topological object information which is a time expensive technique.

However, due to the non-orthogonality of the elementary Gabor functions the computation of the coefficients for the transform is complicated and time-consuming. Various iterative optimization techniques [3, 4] have been proposed as solution. Krüger describes in [4] a network based approach (“Gabor Wave-let Network”) whose node represents a wavelet coefficient with variable translation, dilation and orientation. The network is optimized with the Levenberg-Marquardt algorithm. During face tracking this optimisation is done for each frame [5]. Unfortunately, almost all iterative optimization methods have common disadvantages: the running time increases strongly with the number of wavelets (nodes) used for representation; in addition there exists a convergence problem.

In this work a fast non-iterative method for initial Gabor wavelet representation of an image [8], which is free from the above disadvantages are used. The Gabor coefficients are selected directly from the 2D filter magnitude responses by searching for local extrema. The local maxima and minima in a filter response denote a large correlation between the wavelet and input image. Places with large numbers of such extrema with different frequencies and orientations denote significant features of the object. We demonstrate that the extracted representation can be used as object template for scale invariant real-time object tracking using a discriminator loop technique and a convolution based cross-correlation.

3 Wavelet basis

We use only the odd Gabor wavelets to analyse an input image rather than the complex form of the Gabor functions, although it is necessary to achieve the maximal resolution in the time-frequency space. The use of both real and imaginary components improves the representation quality only insignificantly.

3.1 Mother wavelet

As mother wavelet, the odd-symmetric form of the Gabor function [4] has been used, given by

$$\Psi(x,y,\omega,\theta,\Omega,\gamma) = \frac{1}{\sqrt{\pi\gamma}k} e^{-\frac{\omega^2}{2\gamma^2 k^2}\left(\gamma^2(x\cos\theta - y\sin\theta)^2 + (x\sin\theta + y\cos\theta)^2\right)} \sin\left(-\omega\left(x\cos\theta + y\sin\theta\right)\right), \tag{1}$$

where

$$k = \sqrt{2\ln 2}\left(\frac{2^{\Omega}+1}{2^{\Omega}-1}\right), \tag{2}$$

where Ω is the wavelet bandwidth in octaves, γ is the standard deviations ratio or Gaussian ellipse dilation $\frac{\sigma_y}{\sigma_x}$, ω is the frequency of the modulated sine, θ is the wavelet rotation angle. The octave bandwidth Ω and standard deviations ratio γ for all filter kernels are constant.

3.2 Filter definition

We use the octave scaling of filter frequency to achieve optimal signal representation with a small number of channels. A filter bank with $N_\omega = 8$ frequencies and $N_\theta = 4$ orientations is defined by:

$$\omega_i = 2\pi m_\omega^i \sqrt{m_\omega},\ \ i \in [0 : N_\omega - 1]\,. \tag{3}$$

$$\theta_i = \frac{i\pi}{N_\omega},\ \ i \in [0 : N_\theta - 1]\,, \tag{4}$$

where

$$m_\omega = \left(\frac{N}{2}\right)^{\frac{1}{N_\omega}} \tag{5}$$

is the frequency multiplicator, $N = 64$ is the horizontal/vertical image size. The channels with the two lowest frequencies have only two orientations. Because only the frequency ω and orientation θ are varied, we denote the kernel functions as follows:

$$\Psi_{\omega,\theta} \equiv \Psi(x,y,\omega,\theta,\Omega,\gamma)\,. \tag{6}$$

The wavelet images of the whole filter bank are shown in Fig. 1.

4 Decomposition

4.1 Image preprocessing

The original image I_0 is normalized to minimize the contrast/brightness influence and to eliminate the DC component as follows:

$$I(x,y) = \frac{I_0(x,y) - \overline{I}_0}{\max(I_0) - \min(I_0)}\,, \tag{7}$$

where $\overline{I}_0$ is the mean of the image intensity.

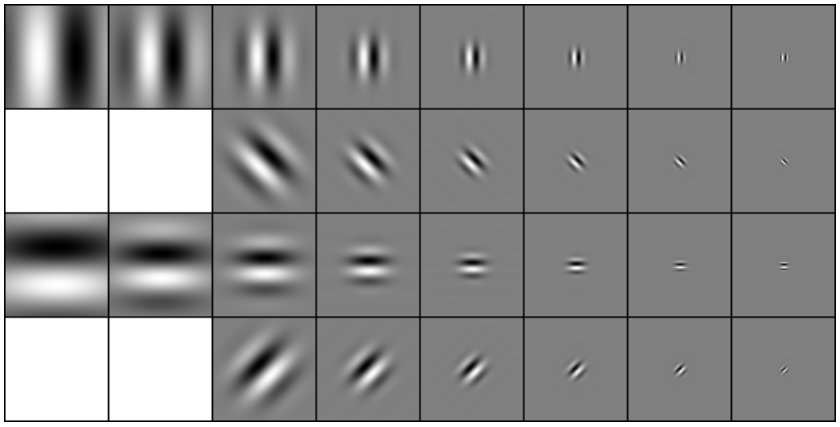

Fig. 1. The filter bank wavelets

4.2 Filtering

The filter responses are obtained by the convolution of the input image I with the filter kernels $\Psi_{\omega,\theta}$:

$$I_{\Psi_{\omega,\theta}} = I \otimes \Psi_{\omega,\theta} = \mathrm{FFT}^{-1}\left(\mathrm{FFT}\left(I\right) \times \mathrm{FFT}\left(\Psi_{\omega,\theta}\right)\right) , \tag{8}$$

where $\otimes$ denotes the convolution operator, $\times$ denotes the complex multiplication. The 2D convolution is realized with the 2D Fast Fourier Transform (FFT). The FTs of filter kernels can be computed and stored before.

4.3 Coefficient selection

Due to the non-orthogonality of the Gabor functions the resulting wavelet filter responses contain a large information redundancy in comparison to the original image, so not all coefficients are needed to reconstruct the image.

Unlike the widely used slow iterative methods to define the wavelet coefficients we have implemented an algorithm for directly extracting the gabor coefficients from filtered images. The idea was to reduce the data redundancy by searching for only significant local extrema in the filter responses for building the object representation. It is based on the observation that the magnitude of the filter response varies approximately with the wavelet frequency ω (while using only the odd gabor functions). The coefficients extraction technique is described in [8] in detail. The example of the extracted points in the filter responses is shown in the Fig. 2

4.4 Object templates

For each extracted point i the following kernel parameters θ_i, ω_i (or kernel number), the wavelet center coordinates x_i, y_i and the projection coefficient (filter response) $K_i = I_{\Psi_{\omega,\theta}}(x_i, y_i)$ are stored. The set of the extracted wavelet parameters describes the object image in the wavelet space.

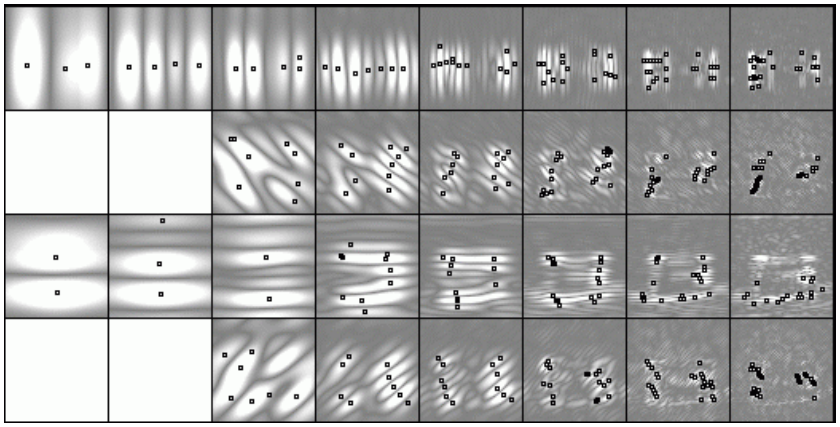

Fig. 2. The response images of the filter bank and the extracted local maxima

4.5 Object image reconstruction

The object image can be reconstructed by linear summation of the extracted Gabor wavelets using the projection coefficients as follows:

$$\hat{I} = \sum_{i=0}^{N_G-1} K_i \Psi(x + x_i, y + y_i, \omega_i, \theta_i, \Omega, \gamma) \,, \tag{9}$$

where N_G is the number of the extracted wavelets.

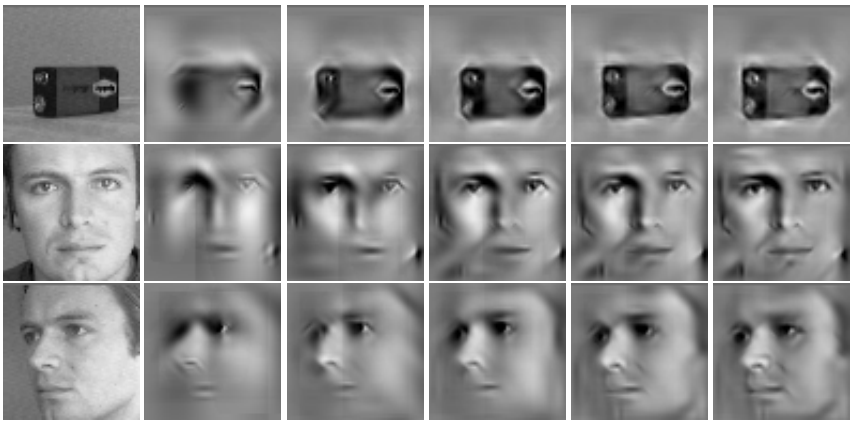

Fig. 3. The original (left) and the reconstructed images with the following number of Gabor wavelets (from left to right): 28 (only one point per kernel), 85, 150, 250, 450.

To test the feature extraction and approximation abilities of the decomposition method we vary the relative threshold in range [0.05:1] to manipulate the number of the wavelet

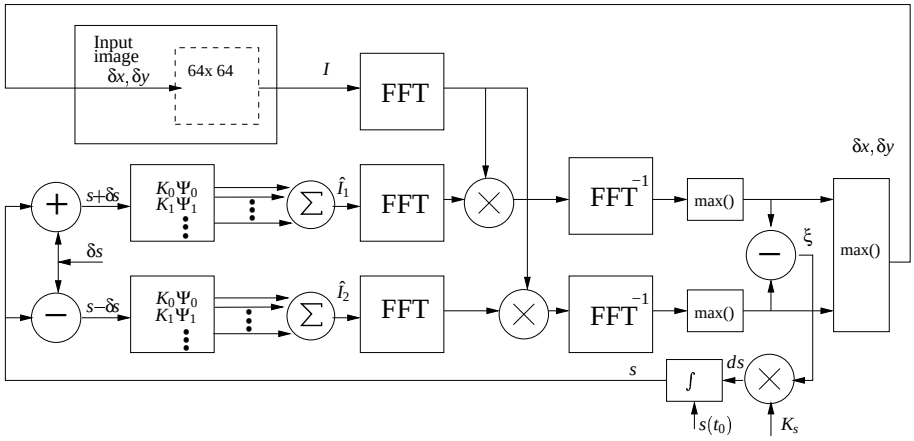

Fig. 4. Tracking by the scale factor discriminator loop

coefficients in the representation. Other parameters are: the octave bandwidth $\Omega = 1.1$, Gaussian ellipse dilation $\gamma = 1.1$. Fig. 3 presents the reconstruction results for three object images, "battery", "face" and "face in profile" for different numbers of wavelets. At low approximation rates only high contrast features in the image have been detected, with 80-150 wavelets already most objects can be easily recognized.

5 Tracking Control

The resulting Gabor representation of an object image is stored and used as object model. Positions of particular wavelets can only be changed by varying the translation, rotation, scale or other model parameters.

A scale invariant tracking technique using the Gabor object representation is shown in Fig. 4. We implement a fast Fourier convolution to calculate a cross-correlation between reproduced object images $\hat{I}_1, \hat{I}_2$ and the input image I. The object images $\hat{I}_1, \hat{I}_2$ are reconstructed by linear summation of the extracted Gabor wavelets using the projection coefficients and different scale coefficients $s \pm \delta s$.

After the initial image decomposition to obtain the current object template (needed only once during initialization), the tracking control with initial scale factor $s(t_0) = 1.0$ can be started. As shown in Fig. 4 we use two discriminator channels and reconstruct template images with different scale factors $s \pm \delta s$. New coordinates of wavelet nodes of the template are calculated as follows:

$$\begin{aligned} \tilde{x}_i &= \delta x + x_c + s(x_i - x_c) \\ \tilde{y}_i &= \delta y + y_c + s(y_i - y_c), \quad i \in [0 : N_G - 1] \,, \end{aligned} \tag{10}$$

where x_c, y_c are template center coordinates and $\delta x, \delta y$ are translation deviations. So we obtain two template images $\hat{I}_1$ and $\hat{I}_2$ with symmetric scale factors $s \pm \delta s$. Then we extract global maxima of the convolution results of both channels and use their values as

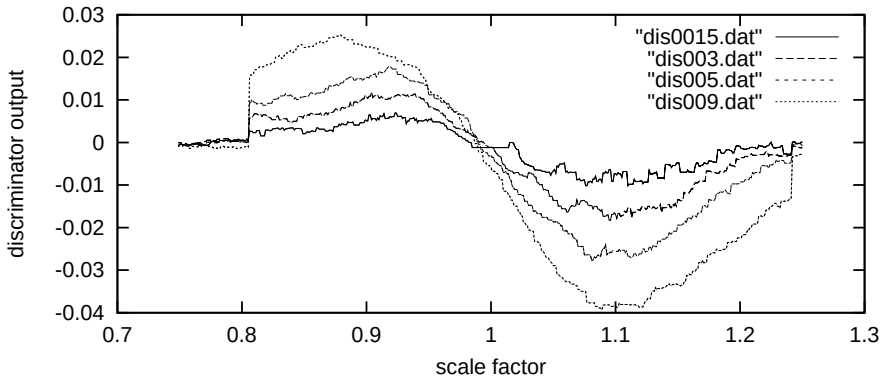

Fig. 5. Discriminator characteristica for different δs: $\delta s = 0.015$, $\delta s = 0.03$, $\delta s = 0.05$ and $\delta s = 0.09$

the discriminator inputs. The output of the discriminator is the scale factor error signal ξ. The gained error (with the gain factor K_s) is integrated and results in the new scale factor value $s(t_1)$, which is used for next frame analysis and closes the control loop. New coordinates of the search frame are obtained using the coordinates of the position of the global maximum $\delta x, \delta y$ in the convolution image with the maximal magnitude (comparing outputs of both channels). These values denote the relative translation vector of the search frame.

We have experimentally calculated the discriminator characteristic using a "face" template for different δs. The results are shown in Fig. 5. We can see the high linearity of the discriminator output ξ in a wide range of the scale factor with $\delta s > 0.3$. However, the larger the channel scale factor divergence δs is, the less exactly can we determine the translation vector $(\delta x, \delta y)$ because of the higher cross-correlation error. So we use a trade-off value $\delta s = 0.03 - 0.04$ to achieve a smooth robust face tracking.

6 Experimental results

We tested the proposed decomposition and tracking technique on various video sequences, obtained with one of the two cameras of our mobile robot "Robin", an RWI B21 robot. The tracking control runs in real-time (>25fps) with image size 168x128 (allthough the successive tracking process is independent of the image size) on a Linux PC with Pentium 1.3GHz using 250-750 wavelets for object description. Fig. 6 shows a tracking example (additional videos can be obtained from the author's homepage [10]). Experiments demonstrate high robustness of tracking control against vertikal/horizontal camera shaking and wide scale changing range [0.2-1.5].

Tab. 1 presents the computation time of the initial Gabor wavelet decomposition (executed once during initialization for building the object template) and tracking depending on the number of extracted wavelets N_G.

In contrast to other existing tracking methods based on the Gabor wavelet image representation such as "Gabor Wavelet Network" [4, 5] the computation time of the

Fig. 6. Tracking of a face image

N_G	init. Gabor decomp.	tracking
250	250-270 msec	8.7-12 msec
350	270-290 msec	9.0-12.5 msec
550	330-360 msec	9.3-13.5 msec
750	390-430 msec	9.5-14.7 msec

Table 1. Computation time of initial Gabor decomposition and scale invariant face tracking

initial image decomposition and of the tracking process increases only insignificantly with the number of wavelet nodes.

Fig. 7 shows the computation time of the face tracking dependent on the template scale factor.

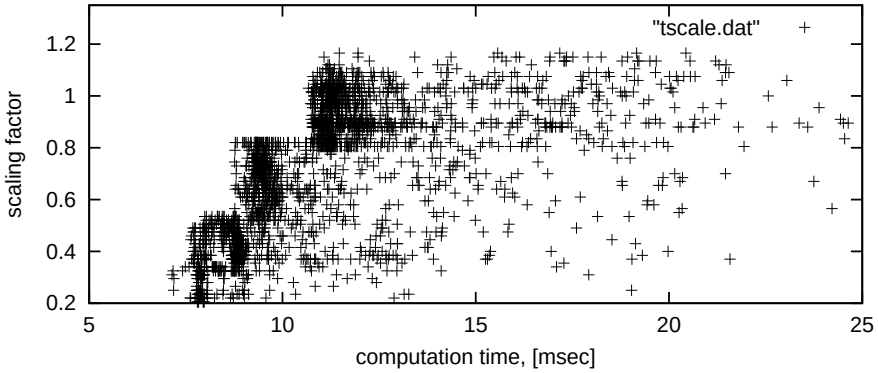

Fig. 7. Tracking computation time for $N_G = 382$

7 Conclusion

In this paper we present a real-time technique for scale invariant object or face tracking with standard PC hardware. The tracking method is based on a low redundancy object image representation (template). For initial image decomposition a fast non-iterative transform based on the odd-symmetric gabor functions is used, which guarantees a low redundancy of the resulting template and an automatic detection of the significant features in the image. Tracking control is realized by a scale factor discriminator loop and a convolution based 2D cross-correlation. The implemented tracking technique shows the robustness of the system against large scale variations and camera shaking, which is important in robot vision.

References

1. Daugman, J.G.: Two-dimensional spectral analysis of cortical receptive field profiles. Vision Res., vol.20, pp.847-856, 1980
2. Feyrer, S., Zell, A.: Detection, Tracking, and Pursuit of Humans with an Autonomous Mobile Robot. In Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'99), (1999) 864–869
3. Fischer, S., Cristobal, G.: Minimum entropy transform using gabor wavelets for image compression. In: Int. Conf on Image Analysis and Processing, Palermo, Italy, 2001
4. Krüger, V., Sommer, G.: Gabor wavelet networks for object representation. In 22. DAGM-Symposium Kiel, Germany (2000) 309–316
5. Krüger, V., Feris, R.: Wavelet Subspace Method for Real-time Face Tracking In Proc. Pattern Recognition, 23rd DAGM Symposium, Munich, Germany, 2001
6. Tai Sing Lee: Image Representation Using 2D Gabor Wavelets IEEE Transactions on pattern analysis and machine intelligence, Vol. 18, No. 10, October 1996
7. Manjunath, B.S., Shekhar, C., Chellappa, R.: A new approach to image feature detection with applications. Pattern Recognition (1996) 31:627–640
8. Mojaev, A., Zell, A.: Real-Time Object and Face Tracking with Gabor Wavelets Proceedings of the IEEE International Conference on Advanced Robotics (ICAR 2003), Vol. 2, 1178–1183
9. Mojaev, A.: Umgebungswahrnehmung, Selbstlokalisation und Navigation mit einem mobilen Roboter. Ph.D. thesis, University of Tbingen, WSI, Shaker Verlag ISBN 3-8265-7865-1, (2000)
10. Alexander Mojaev's Homepage: http://www-ra.informatik.uni-tuebingen.de/mitarb/mojaev/welcome_e.html
11. Wiskott, L., Fellous, J.-M., Krüger, N., von der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching In: Proc. 7th Intern. Conf. on Computer Analysis of Images and Patterns, CAIP'97, Kiel. G. Sommer and K. Daniilidis and J. Pauli (eds.), No. 1296, Springer-Verlag, Heidelberg (1997) 456–463
12. Zhang, Z., Lyons, M., Schuster, M., Akamatsu, S.: Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expressions Recognition Using Multi-Layer Perceptron. In: Proc. 3rd IEEE International Conference on Automatic Faceand Gesture Recognition, April 14-16 1998, Nara Japan, IEEE Computer Society, pp. 454-459