Tracking Facial Features Using Gabor Wavelet Networks

ROGÉRIO S. FERIS AND ROBERTO M. CESAR JUNIOR

Departamento de Ciência da Computação
Instituto de Matemática e Estatística - USP
Rua do Matão, 1010, 05508-900 São Paulo - SP - Brasil
{rferis,cesar}@ime.usp.br

Abstract. This work presents a new method for automatic facial feature tracking in video sequences. In this method, a discrete face template is represented as a linear combination of continuous 2D odd-Gabor wavelet functions.The weights and 2D parameters (position, scale and orientation) of each wavelet are determined optimally so that the maximum of image information is preserved for a given number of wavelets. We have used this representation to achieve effective facial feature tracking that is robust to homogeneous illumination changes and affine deformations of the face image. Moreover, the tracking approach considers the overall geometry of the face, being robust to facial feature deformations such as eye blinking and smile. The number of wavelets in the representation may be chosen with respect to the available computational resources, even allowing real-time processing.

1 Introduction

The automatic tracking of faces and facial features in video sequences is a fundamental and challenging problem in computer vision. This research topic has many applications in human-computer interaction, model-based coding, gaze detection and teleconferencing. Furthermore, automatic face recognition from image sequences may require the tracking process in order to segment the face in each frame. Faster recognition may be achieved by using only facial features.

In this paper, we propose a method for tracking facial features that is based on the work of Kruger and Sommer [1], which uses a Gabor wavelet network (GWN) for face representation, allowing tracking robust to homogeneous illumination changes and affine deformations of the face image. This representation is generated by approximating the face template as a linear combination of continuous 2D odd-Gabor wavelet functions. Thus, we have a continuous wavelet representation of a discrete face template. The weights and 2D parameters (position, scale and orientation) of each wavelet are determined optimally so that the maximum of image information is preserved for a given number of wavelets in the representation.

Our tracking method is manually initialized by the user, who indicates, in the first frame, the face region and the position of the pupils, center of nose and center of mouth. It is worth saying that it could be done automatically by means of a skin-color blob information and some technique for automatic detection of facial features. We are still working on this problem [2], which will be addressed in a future paper.

After this initialization process, the wavelet representation for the face template is generated and the facial features are tracked along the video sequence by using it. The tracking approach considers the overall geometry of the face and, therefore, it is robust to facial feature deformations such as eye blinking and smile.

The number of wavelets that will approximate the face template may be chosen by the user. The representation becomes more specific as the number of wavelets increases. On the other hand, as the number of wavelets is decreased, the representation becomes more general, being more suitable to be applied to different individuals. It is interesting to note that a GWN is a RBF network, which provides generalization of the training data when a small number of basis functions are used.

The tracking algorithm may be even executed in real-time. For this, the user must choose the number of wavelets in the representation according to the available computational resources.

The remainder of this paper is organized as follows. Section 2 reviews some techniques related to our work. Section 3 introduces the wavelet networks as a powerful tool for function approximation. In section 4, the face representation obtained by a GWN is described and its advantages are discussed. Section 5 is concerned with the repositioning of a GWN, which allows face tracking. In section 6, facial feature tracking is presented as the major contribution of this paper. The experimental results are discussed in section 7. Finally, section 8 concludes this paper with some remarks on further research directions.

2 Related Work

Many approaches have been proposed to track faces and facial features in video sequences. Recently, color-based systems have been widely used to accomplish this task. The work of Jie Yang and Alex Waibel [3] presents a statistical skin-color model, which is invariant to people from differ-

0-7695-0878-2/00 $10.00 © 2000 IEEE

ent races. This work was extended to track faces in real-time [4].

Stiefelhagen and Yang [5] have used a color-based approach to track specific facial features (pupils, nostrils and lipcorners) in video sequences. The determined location of facial features in each frame was used to estimate face 3D position. The use of color to track faces and facial features has advantages such as face pose invariance and real-time processing. On the other hand, this approach is, in general, not robust to illumination changes.

Liyanage Silva et. al. [6] proposed a method, which they called edge pixel counting, to detect and track facial features in image sequences. This method is based on the fact that the edge concentration is high near the facial features (eyes, nose and mouth) and low around them. The method is simple but it fails in several situations, such as in the presence of cluttered backgrounds, glasses and hair covering the forehead.

The work of Thomas Maurer and Christoph Malsburg [7] presents a system that tracks facial features with Gabor wavelet filter responses. Initially, feature positions are initialized by hand in the first frame of the sequence. Gabor filters are then used to extract feature vectors, or jets, from that positions. Finally, each feature point is individually tracked by phase-based displacement estimation. The main disadvantage of this approach is the high computational cost required, which leads to tracking with less than 1 fps.

Our approach uses a wavelet representation for the face image that is even sparser than the Gabor jet representation. Also it differs from the one introduced by Mallat or Daubechies [8, 9]. In fact, it is based on a wavelet network concept, which will be explained in the next section.

3 Function Approximation

The wavelet representation for the face template is obtained by a function approximation method. Our problem of function approximation consists in estimating an unknown continuous function $f : R^n \rightarrow R$ from scattered samples $\{(x_i, y_i)\}$ that are employed as training patterns, where $x_i \in R^n$ and $y_i \in R$.

We may consider the face image as an unknown continuous function $f : R^2 \rightarrow R$, assuming that we have a grey-level image. In this case, each pixel of the face template corresponds to a scattered sample (x_i, y_i) where x_i is the pixel position and y_i is the pixel intensity. Thus, our objective is to determine a continuous function $\hat{f} : R^2 \rightarrow R$ that approximates f, i.e., a continuous representation for the face template.

The method used to obtain the face representation is a wavelet network [10], which is an alternative to feedforward neural networks for approximating continuous functions. In the following subsections, we first discuss neural networks and after we introduce the wavelet networks for function approximation.

3.1 Neural Networks

Feedforward neural networks have been intensely studied as efficient tools for arbitrary function approximation. The work of Cybenko [11] shows a rigorous demonstration that multilayer perceptrons with only one hidden layer of processing elements is sufficient to approximate any continuous function with support in a hypercube. Figure 1 shows this network structure for approximation.

The neural network universal appproximation property follows: if σ is a non-linear continuous, limited and monotonically increasing function, then finite sums of the form:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{M} w_i \sigma(\mathbf{a}_i^T \mathbf{x} + b_i) \qquad (1)$$

are dense in the space of continuous functions defined on $[0,1]^n$, where $w_i, b_i \in R, a_i \in R^n$. In other words, given any continuous function f defined on $[0,1]^n$ and any $\epsilon > 0$, there is a sum $\hat{f}(x)$ of the form above, for which $|\hat{f}(x) - f(x)| < \epsilon$ for all $x \in [0,1]^n$.

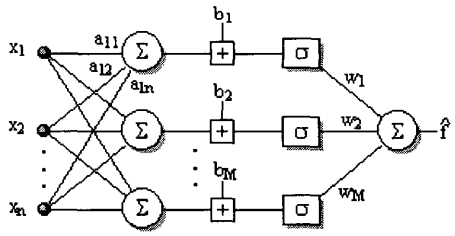

Figure 1 - Neural network for function approximation.

3.2 Wavelet Networks

Wavelet networks, or wavenets, were proposed as an alternative to feedforward neural networks for function approximation. This concept was inspired by both the wavelet decomposition and neural networks.

It is well know that wavelet decomposition allow us to decompose any function $f(x) \in L^2(R^n)$ using a family of functions obtained by dilating and translating a single mother wavelet function $\psi : R^n \rightarrow R$. Thus, $f(x)$ may be expressed as a linear combination of wavelet functions, where the wavelet coefficients (weights) are estimated by the decomposition process. In contrast, in the wavelet network, not only weights, but also the parameters of wavelet functions (translation, dilation and we may also consider

orientation) are jointly fitted from data. In this case, the number of wavelet functions may be choosen by the user and its parameters are optimized by a learning process. The more used wavelets, the more precise is the approximation.

Thus, if we want to approximate a continuous function $f : R^n \rightarrow R$ from scattered data, we may use a wavelet network, computing the approximation function $\hat{f}$ acoording to the equation below:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{M} w_i \psi_{\mathbf{n}_i}(\mathbf{x}) + \bar{f} \qquad (2)$$

where $w_i \in R, \psi_{\mathbf{n}_i}$ is a wavelet function, $\mathbf{n}_i$ is the parameter vector of each wavelet and $\bar{f}$ is introduced in order to approximate functions with nonzero average. Weights and parameter vectors of wavelets are determined optimally by a learning process. Figure 2 illustrates the typical structure of a wavenet for a function approximation problem.

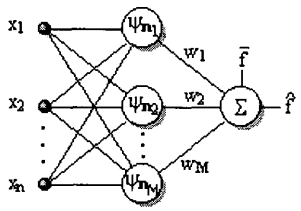

Figure 2 - Wavelet network for function approximation.

4 Face Representation Using Gabor Wavelet Networks

The face representation is obtained by using a wavelet network in which the mother wavelet is a Gabor function. The use of Gabor filters in image analysis is biologically motivated as they model the response of the receptive fields of the orientation-selective simple cells in the human visual cortex [12]. Furthermore, they provide the best possible tradeoff between spatial and frequency resolution (Heisenberg principle). Figure 3 shows an illustration of a 2D odd-Gabor function.

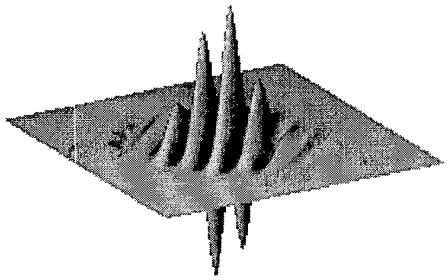

Figure 3 - 2D odd-Gabor wavelet function.

To define a Gabor wavelet network, we start by taking a family of M 2D odd-Gabor wavelet functions $\Psi = \{\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_M}\}$ of the form

$$\begin{aligned}
&\psi_{\mathbf{n}}(x,y) = \\
&\exp(-\frac{1}{2}[s_x((x-c_x)cos\theta - (y-c_y)sin\theta)]^2 \\
&+[s_y((x-c_x)sin\theta + (y-c_y)cos\theta]^2]) \\
&\times sin(s_x((x-c_x)cos\theta - (y-c_y)sin\theta)) \qquad (3)
\end{aligned}$$

with the parameter vector $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y)$, where c_x, c_y denote the translation (position) of the Gabor wavelet, s_x, s_y denote the dilation (scale) and θ denotes the orientation.

In order to obtain the wavelet representation for a face image f, the weights and parameters of each wavelet are determined optimally, by means of a learning process, which minimizes the energy function

$$E = \min_{\mathbf{n}_i, w_i \forall i} ||f - (\sum_i w_i \psi_{\mathbf{n}_i} + dc(f))||_2^2 \qquad (4)$$

with respect to the weights $w_i \in R$ and wavelet parameters $\mathbf{n}_i \in R^5$. In the equation above, $dc(f)$ is the DC-value of f. The Levenberg-Marquard gradient descent learning method [13] was employed to determine the optimal wavelet network for the face template. The method might get stuck in local minima and a careful selection of the initial parameters is important.

Then, we can say that the two optimized vectors $\mathbf{\Psi} = (\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_M})^T$ and $\mathbf{w} = (w_1, \ldots, w_M)^T$ define an optimized Gabor Wavelet Network $(\mathbf{\Psi}, \mathbf{w})$ for a specific face image f. The continuous representation for f may be considered as the reconstruction of the original image and it is given by:

$$\hat{f} = \sum_{i=1}^{M} w_i \psi_{\mathbf{n}_i} + dc(f) \qquad (5)$$

Of course the quality of the reconstruction depends on the number M of used wavelets. Figure 4 shows a face template (left) and its discretized representation (middle), which we call the Gabor wavelet template (GWT). This representation was obtained by using a GWN of just $M = 52$ odd-Gabor wavelets, initialized in the inner face region. The right illustration shows the position of the 16 largest wavelets, after optimization.

The obtained continuous face representation has several advantages: it provides generalization depending on the number of used wavelets; it is invariant to some degree to affine deformations of the face image, as we will see in the next section; and since the odd-Gabor Wavelets are DC-free, they are invariant to some degree to homogeneous illumination changes.

5 Repositioning a Gabor Wavelet Network

In the previous section we have shown how a continuous wavelet representation for a face template is obtained based

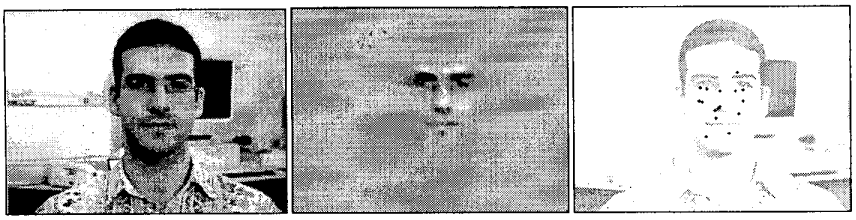

Figure 4 - Face template representation.

on a Gabor wavelet network. Now, we will see how this representation can be affinely repositioned in a new face image so that its wavelets are registered on the same facial features as in the original image. This process is called GWN repositioning.

For instance, consider the face template shown in figure 5 (top left) and let G be its optimized GWN. Now, consider this face image in a different pose as shown in figure 5 (bottom left). In the repositioning process, the set of wavelets of G are positioned correctly on the same facial facial features in the distorted image. It is important to emphasize that the GWN repositioning may determine the parameters (translation, scale, rotation and shearing) of any affine deformation applied to the original image. The right illustrations of figure 5 show the position of the 16 largest wavelets of G in each image, whereas the middle illustrations show the original and repositioned discrete face template representation (GWT), which was obtained with 52 odd-Gabor wavelet functions.

The repositioning of a GWN in a new image, i.e., the determination of the correct affine parameters, is established by using a superwavelet [14]. Let $\mathbf{\Psi} = (\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_M}), \mathbf{w} = (w_1, \ldots, w_M)$ be a GWN. A Gabor superwavelet $\Psi_{\mathbf{n}}$ (GSW) may be defined as a linear combination of the wavelets $\psi_{\mathbf{n}_i}$ such that

$$\Psi_{\mathbf{n}}(\mathbf{x}) = \sum_i w_i \psi_{\mathbf{n}_i}(\mathbf{SR}(\mathbf{x} - \mathbf{c})) \tag{6}$$

where the parameters of vector $\mathbf{n}$ of the GSW Ψ define the dilation matrix $\mathbf{S}$, the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{c}$ with:

$$\begin{aligned} \mathbf{S} &= \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}, \\ \mathbf{R} &= \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}, \\ \mathbf{c} &= (c_x, c_y)^T. \end{aligned}$$

Thus, a Gabor superwavelet $\Psi_{\mathbf{n}}$ is again a wavelet that has the typical wavelet parameters dilation s_x, s_y, translation c_x, c_y and rotation θ. So, the GSW $\Psi_{\mathbf{n}}$ may be handled in the same way as we handled each single Gabor wavelet in the previous section. For a given new image g, we may arbitrarily deform the superwavelet by optimizing its parameter vector $\mathbf{n}$ according to the energy function below:

$$E = \min_{\mathbf{n}} ||g - \Psi_{\mathbf{n}}||_2^2 \tag{7}$$

It is important to note that the parameters of a wavelet include only translation, dilation and rotation. Even so, we may include shearing and thus allow any affine deformation of GSW $\Psi_{\mathbf{n}}$. For this, we add the parameter s_{xy} to vector $\mathbf{n}$ and rewrite the scaling matrix:

$$\mathbf{S} = \begin{pmatrix} s_x & s_{xy} \\ 0 & s_y \end{pmatrix}$$

In order to minimize the energy function and to determine the optimal parameter vector $\mathbf{n}$, we may use the same Levenberg-Marquard algorithm as in the previous section. In general, the initialization that must be supplied to the gradient descent method may be within the range of approximately $\pm 10px$ in position, $\pm 20\%$ in scale and $\pm 10°$ in orientation.

It is interesting to note that a wavelet representation with a small number of wavelets may work well in different individuals. For instance, the optimized GWN for the face template showed in figure 4 may be repositioned in other individuals, since only 52 wavelets are used, providing generalization. Figure 6 illustrates this property, showing the 16 largest wavelets repositioned in other person.

6 Tracking of Facial Features

The GWN repositioning described in the previous section may be applied to an image sequence, allowing affine face tracking. We consider the face as a planar object that is viewed under ortographic projection.

Thus, for each frame J_t at time step t, the Gabor superwavelet $\Psi_{\mathbf{n}_t}$ is optimized according to the energy function:

$$E = \min_{\mathbf{n}_t} ||J_t - \Psi_{\mathbf{n}_t}||_2^2 \tag{8}$$

The parameter vector $\mathbf{n}_{t-1}$ is used as initial value for optimization in the frame J_t. As image changes are small from frame to frame, the optimization process converges

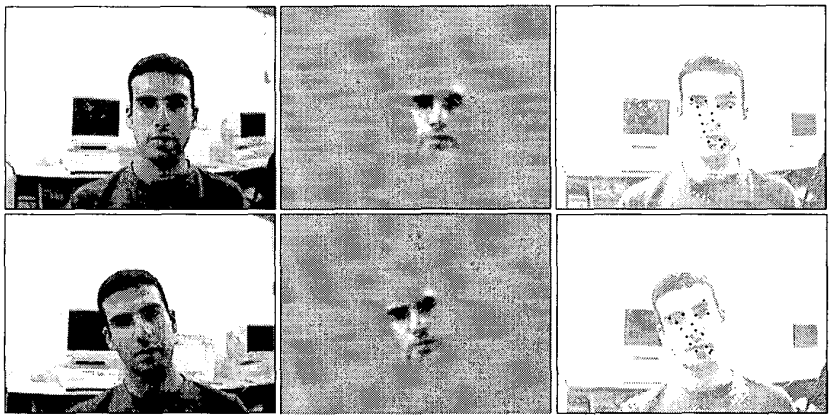

Figure 5 - GWN repositioning.

quickly. Initial values for $\mathbf{n}_0$ in the first frame are choosen by hand, but it could be derived from a color blob information.

In order to track facial features, the user manually selects, in the first frame, the pupils, center of nose and center of mouth. The face region must also be indicated, since this information is necessary to the Gabor wavelet network initialization.

After this initialization process, the wavelet representation for the face template is obtained by using a GWN. This representation is then affinely repositioned in the next frames, as described above. Facial feature tracking is then performed by applying, in each frame, the correct affine transformation to the selected points. The parameters of the affine transformation are obtained by means of the superwavelet parameter vector $(s_x, s_y, s_{xy}, c_x, c_y, \theta)$ in each frame.

Thus, this facial feature tracking approach considers the overall geometry of the face and, therefore, it is robust to facial feature deformations such as eye blinking and smile. In other words, the method does not require the condition of a high inter frame correlation around the feature areas as it is required in template matching.

Instead of tracking feature points, we may choose to track facial areas such as eyes, nose and mouth. For this, a rectangle is drawn around each feature point selected by the user in the first frame. The size of each rectangle is determined according to the specific facial feature and face region size. In the following frames, the rectangle is affinely deformated according to the superwavelet parameters. This tracking process may be useful for recognition from video sequences, since facial features are segmented in each frame.

It is worth saying that the GWN technique has recently been used to perform face tracking, face recognition and pose estimation. Our facial feature tracker is a contribution over face tracking, in the sense that it may be a good alternative to most facial feature tracking systems, which are either not robust or computationally expensive.

7 Experimental Results

Facial feature tracking was tested in different video sequences and the obtained results confirmed the robustness of the method. Figure 7 illustrates 3 frames of a test image sequence during tracking. Other examples can be seen in http://www.ime.usp.br/~rferis. It is important to note that each wavelet of the GSW has to be evaluated during the repositioning process. Then, using less wavelets results in a respective speedup.

We are still verifying the performance of the system so that future work will cover quantitative experimental results (related to efficiency, invariance properties, etc.) as well as comparison with other systems.

Figure 6 - Repositioning in different individuals.

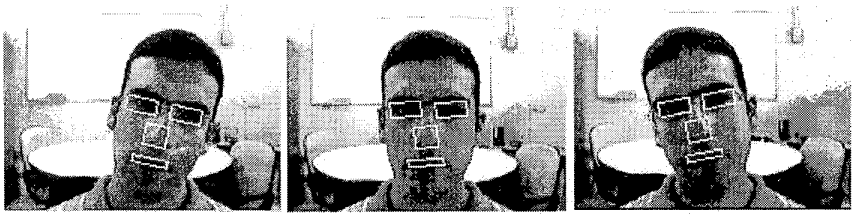

Figure 7 - Facial feature tracking.

8 Conclusions

This paper described a method for facial feature tracking using Gabor Wavelet Networks. The method is based on a continuous wavelet representation of a discrete face template, which is invariant to some degree to illumination changes and affine deformations of the face image. This representation may be specific or generic, depending on the number of used wavelets.

Facial feature tracking was achieved by repositioning the wavelet representation in each frame. Since the overall geometry of the face is considered, the method is robust to facial feature deformations such as eye blinking and smile.

As future work, we intend to use skin-color and Gabor wavelet networks to perform automatic detection of faces and facial features.

Acknowledgements

Roberto M. Cesar Junior is grateful to FAPESP for the financial support (98/07722-0), as well as to CNPq (300722/98-2). Rogerio Feris is grateful to FAPESP (99/01487-1).

We are grateful to Volker Kruger for providing source code related to the GWN technique and for discussions.

References

[1] V. Kruger and G. Sommer, "Affine real-time face tracking using a wavelet network". Presented at the *ICCV'99 Workshop Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece, September 1999.

[2] R. Feris, T. Campos and R. Cesar, "Detection and tracking of facial features in video sequences", *Lecture Notes in Artificial Intelligence*, vol. 1793, pp. 127-135, Springer-Verlag, April 2000.

[3] J. Yang, W. Lu and A. Waibel, "Skin-color modeling and adaptation", CMU CS Technical Report, CMU-CS-97-146, May 1997.

[4] J. Yang and A. Waibel, "A real-time face tracker", *Proc. of the Third IEEE Workshop on Applications of Computer Vision*, pp. 142-147, Sarasota, Florida, 1996.

[5] R. Stiefelhagen and J. Yang, "Gaze tracking for multimodal human computer interaction", University of Karlsruhe, 1996. Avaiable at http://werner.ira.uka.de/ISL.multimodal.publications.html

[6] L. Silva, K. Aizawa and M. Hatori, "Detection and tracking of facial features", *Proc. of SPIE Visual Communications and Image Processing*, Taiwan, May 1995.

[7] T. Maurer and C. Malsburg, "Tracking and learning graphs and pose on image sequences of faces", *Proc. Int. Conf. on Artificial Neural Networks*, Bochum, 1996.

[8] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11(7):674-693, 1989.

[9] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis", *IEEE Trans. Informat. Theory*, 36, September 1990.

[10] Q. Zhang and A. Benviste, "Wavelet networks", *IEEE Trans. on Neural Networks*, 3(6):889-898, November 1992.

[11] G. Cybenko, "Approximation by superposition of a sigmoidal function", *Mathematics of Control, Signals and Systems*, 2:303-314, 1989.

[12] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized two-dimensional visual cortical filters.", *Journal Opt. Soc. Am.*, 2(7):1160-1168, 1985.

[13] W. Press, B. Flannery, S. Teukolsky and W. Vetterling, *Numerical Recipes, The Art of Scientific Computing*, Cambridge University Press, UK, 1986.

[14] H. Szu, B. Telfer and S. Kadambe, "Neural network adaptive wavelets for signal representation and adaptation. *Optical Engineering*, 31(9):1907-1961, 1992.