Locating and Tracking Facial Landmarks Using Gabor Wavelet Networks

Rogério S. Feris and Roberto M. Cesar Junior

Department of Computer Science, University of São Paulo,
Rua do Matão, 1010, 05508-900 São Paulo-SP, Brazil
{rferis,cesar}@ime.usp.br

Abstract. A new approach for locating and tracking facial landmarks in video sequences is introduced in this paper. Our method is based on Gabor wavelet networks, an effective technique that represents a discrete face template as a linear combination of 2D Gabor wavelet functions. This wavelet representation allows positioning of facial landmarks (e.g. eyes, nose and mouth), even in the presence of glasses, beard and different facial expressions. The feature tracking is robust to homogeneous illumination changes and affine deformations of the face image. Moreover, the tracking approach considers the overall geometry of the face, thus being robust to deformations such as eye blinking and smile, which is usually a critical situation to most local-based traditional methods.

Keywords: Computer Vision, Facial Feature Tracking, Gabor Wavelets.

1 Introduction

Computational face recognition is an important research problem in computer vision, presenting many applications such as in human-computer interaction, security systems and surveillance. There are two main approaches to face recognition by computers, namely, static and dynamic. While the former is related to recognizing people in still images, the latter addresses the problem of detecting, tracking, extracting information and recognizing moving people in digital video sequences. As far as the problem of tracking a face in a video sequence is concerned, there are three distinct approaches: (1) tracking the whole face region; (2) tracking the head outline (e.g. using active contour models); (3) tracking a set of feature points or facial landmarks, generally defined by the eyes, the nose and the mouth. The last approach presents, among its attractives, the potentiality for allowing faster recognition in real-time applications and the fact of presenting a psychophysical inspiration because of the well-known importance that these landmarks present to human perception. This paper introduces an approach for locating and tracking facial feature points, which has proven to be robust and even suitable for real-time applications.

The technique is based on a recent approach for face representation called Gabor wavelet network (GWN) [1]. The GWN represents the face as a linear combination of 2D Gabor wavelet functions, whose parameters (position,

scale and orientation) and weights are determined optimally so that the maximum of image information is preserved for a given number of wavelets.

Once that the GWN has been optimized, it can be repositioned with its wavelet parameters undergoing an affine transformation in order to match a target face image. The repositioning procedure is a key concept of this paper, being applied with two different purposes, namely, (1) for locating the facial feature points of the detected face-like blob in an initial frame and (2) for performing facial feature tracking.

Basically, our approach can be divided in three subsequent steps:

- *Face detection*, which is performed automatically by skin-color blob detection [2]. Once that a face-like blob is located, a simple correlation procedure is used to decide whether the blob actually corresponds or not to a true face.
- *Facial feature points positioning*, which is done automatically by matching a GWN,optimized considering a mean face, to the initial face-like blob;
- *Tracking of face and facial feature points* by the GWN. The tracking algorithm may be even executed in real-time. In this case, a suitable number of wavelets in the representation must be chosen with respect to to the available computational resources.

We show that our method is able to locate facial features points even in the presence of glasses, beard, and different facial expressions. The tracking approach is robust to homogeneous illumination changes and affine deformations of the face image. Moreover, since we consider the overall geometry of the face, it is robust to facial feature deformations such as eye blinking and smile.

The remainder of this paper is organized as follows. Section 2 reviews some techniques related to our work. Section 3 introduces the Gabor wavelet networks for face representation and advantages of this technique are discussed. Section 4 is concerned with the repositioning of a GWN, which allows face tracking. In section 5, the positioning of facial landmarks is presented. The facial feature tracking is described in section 6. Experimental results and some discussions are presented 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 locate and track faces and facial features in video sequences. Recently, color-based systems have been widely used to accomplish this task. For instance, we can cite the work of Jie Yang and Alex Waibel [3], which presents a real-time face tracker based on a statistical skin-color model [4]. Another example is the work of Stiefelhagen and Yang [5], that describes a color-based method for detection and tracking of specific facial features (pupils, nostrils and lipcorners). 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] used an edge-based approach to locate and track facial features in image sequences. This method is based on the fact that a higher edge concentration is observed in the center of facial features (eyes, nose and mouth), while slightly outside of such features a less edge concentration is observed. The method is simple but it fails in several situations, such as in the presence of cluttered backgrounds, glasses and hair covering the forehead.

Another different approach is presented in the work of Thomas Maurer and Christoph Malsburg [7], which describes a system for tracking facial features with the elastic graph matching technique. In this method, Gabor filters are applied in some facial feature locations (selected by hand), forming a feature vector, or a jet, for each face position. The face is then modeled as a graph, in which the nodes correspond to the jets and the edges encode face geometrical information. Facial feature tracking is performed by a graph matching procedure in each frame. The main disadvantage of this approach is the high computational cost required.

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 Face Representation Using GWN

Wavelet networks [10], or wavenets, were proposed as an alternative to feed-forward neural networks for function approximation. In this section we will show, with basis on the recent work of Kruger and Sommer [1], that this mathematical tool may be used to approximate a discrete face template, providing an effective face representation.

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

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

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 a wavelet representation for a face image f, the weights and parameters of each wavelet are determined optimally, by means of a fitting technique, which minimizes the energy function

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

with respect to the weights $w_i \in R$ and wavelet parameters $\mathbf{n}_i \in R^5$. In the equation above, $\bar{f}$ is the DC-value of f. The Levenberg-Marquard gradient descent method [11] 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 wavelet representation for f may be considered as the reconstruction of the original image and it is given by:

$$\hat{f} = \sum_{i=1}^{M} w_i \psi_{\mathbf{n}_i} + \bar{f} \tag{3}$$

The precision of the face representation is determined by the number M of used wavelets. For instance, if we use a short number of wavelets, we obtain a coarse GWN, which may work well in different individuals and may be suitable for real-time applications. As we increase the number of wavelets, the representation becomes more specific, encoding more precise object information. Another aspect of the GWN representation is that it is invariant to some degree to affine deformations of the face image, as we will see in the next section. Furthermore, since the odd-Gabor Wavelets are DC-free, they are invariant to some degree to homogeneous illumination changes.

The Gabor functions are biologically motivated [12] and provide the best possible tradeoff between spatial and frequency resolution (Heisenberg principle). Besides, they act as good feature detectors, enconding texture and geometrical information in the representation.

Figure 1(a) shows a face template and its discretized representation is illustrated in figure 1(b), 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. Figure 1(c) shows the position of the 16 largest wavelets, after optimization.

4 Face matching by the GWN

In the previous section, we have shown how a continuous wavelet representation for a face template is obtained based on a Gabor Wavelet Network. Now, we will see how this representation can match a new face image so that its wavelets are registered on the same facial features as in the original

Fig. 1. (**a**) The face template. (**b**) The wavelet representation obtained by the GWN. (**c**) Position of the 16 largest wavelets.

image. This process, which is called GWN repositioning, is done by applying a suitable affine deformation on the entire wavelet representation. It will be used both for positioning and for tracking the facial landmarks.

For instance, consider the face template shown in figure 1(a) and let G be its optimized GWN. Now, consider this face image in a different pose as shown in figure 2(a). In the repositioning process, the set of wavelets of G are positioned correctly on the same 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. Figure 2(b) shows the repositioned discrete face template representation (GWT), with 52 odd-Gabor wavelet functions. Figure 2(c) illustrates the position of the 16 largest wavelets of G in the image.

Fig. 2. (**a**) Face template in a different pose. (**b**) Repositioned wavelet representation. (**c**) Position of the 16 largest wavelets.

The repositioning of a GWN in a new image, i.e., the determination of the correct affine parameters, is established by using a superwavelet [13]. Let $\mathbf{\Psi} = (\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_M}), \mathbf{w} = (w_1, \ldots, w_M)$ be an optimized GWN. A Gabor superwavelet $\Psi_{\mathbf{n}}$ (GSW) may be defined as a linear combination of the wavelets $\psi_{\mathbf{n}_i}$ such that

$$\Psi_{\mathbf{n}}(\mathbf{x}) = \sum_i w_i \psi_{\mathbf{n}_i}(\mathbf{SR}(\mathbf{x} - \mathbf{c})) \tag{4}$$

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:

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

Thus, a Gabor superwavelet $\Psi_{\mathbf{n}}$ is again a wavelet that has the typical parameters, i.e. dilation s_x, s_y, translation c_x, c_y and rotation θ. So, for a given

new image g, we may arbitrarily deform the superwavelet by optimizing its parameter vector $\mathbf{n}$ so that the wavelet representation matches the face in image g. This is done by minimizing the energy function below, using the Levenberg-Marquard algorithm:

$$E = \min_{\mathbf{n}} ||g - \Psi_{\mathbf{n}}||_2^2 \tag{5}$$

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:

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

5 Locating Facial Landmarks

In this section, we address the problem of locating facial landmarks by using the GWN repositioning procedure described in the previous section. Our experiments were carried out on the Yale face database (http://cvc.yale.edu/projects/yalefaces/yalefaces.html), which consists of 15 different subjects with 11 images per person, showing different gestures. Because the database contains only grey-level images, we segmented the face region by hand. Experiments with color images and automatic face detection will be presented in the next section.

When we optimize a coarse GWN (with a short number of wavelets) on a face template, we obtain a wavelet representation that may be repositioned in different individuals. According to our experiments, the repositioning procedure never fails when the target face image is obtained from the same person with different facial expressions in which the GWN was optimized. However, we can not guarantee that a coarse GWN optimized considering a specific individual can always be repositioned in any person. The repositioning process depends heavily on the similarity between the person in which the GWN was optimized and the test person as well as on the number of used wavelets.

The solution that we propose for this problem is to optimize the GWN considering a mean face. We have normalized and averaged a set of 15 faces of the Yale database, corresponding each one to a different individual.

Using the GWN optimized considering this mean face, the repositioning worked well on all individual images of the Yale database. We are now investigating how to address the repositioning problem in any person considering very large face databases.

The proposed method for locating facial landmarks is now described. Initially, facial feature points are located in the mean face. We have considered the pupils, center of nose and center of mouth. Then, the GWN related to the mean face is repositioned in the target face image. In order to determine

the facial landmarks in the test face image, we apply a suitable affine transformation to the initial facial feature points of the mean face. The correct parameters of this transformation are obtained from the superwavelet parameter vector $(s_x, s_y, s_{xy}, c_x, c_y, \theta)$, which is determined by equation (5) in the repositioning process.

The obtained results show the robustness of the method. Figure 3 illustrates the positioning of facial landmarks in three individuals of the Yale database. It is worth saying that the method may work even in the presence of glasses and beard. Furthermore, figure 4 illustrates the results considering face images of the same person under different expressions and illumination changes. Section 7 presents more discussions about this technique.

Fig. 3. Facial landmarks positioning in different individuals.

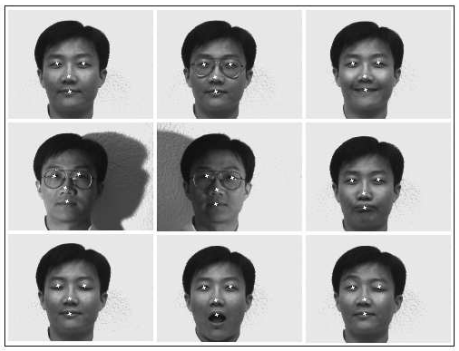

Fig. 4. Facial feature positioning under different expressions and illumination changes.

6 Tracking Facial Features

The GWN repositioning described in section 4 may be applied to an image sequence, allowing affine face tracking. Thus, for each frame J_t at time step t, the Gabor superwavelet $\Psi_{\mathbf{n}_t}$ is optimized according to the energy function:

$$E = \min_{\mathbf{n}_t} ||J_t - \Psi_{\mathbf{n}_t}||_2^2 \tag{6}$$

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 quickly. Initial values for $\mathbf{n}_0$ in the first frame are derived from a color blob information.

Our tracking method assumes that the facial landmarks have been correctly determined in the first frame of the sequence, as described in the previous section. We then apply, in each frame, the suitable affine transformation to the located feature points, performing facial landmark tracking. The parameters of the affine transformation are obtained by means of the superwavelet parameter vector in each frame of the sequence.

It is important to emphasize that the procedure to track facial landmarks considers the overall face geometry, thus being robust to deformations such as eye blinking and smile, which is usually a critical situation to most local-based traditional approaches. It is also important to note that the introduced approach can be straighforwardly generalized in order to track additional feature points or even regions, such as arbitrary polygons around the eyes, nose and mouth.

7 Experiments and Discussion

As discussed in section 1, our approach can be divided in three subsequent steps: face detection, facial landmarks positioning and tracking of face and facial landmarks. The first step is performed by a skin-color approach as well as by a simple correlation procedure to verify the presence of a face in the located skin-blob [2]. Once the face was detected, its scale information is obtained and the color face region is converted to a grey-level image. Facial landmarks are then located by repositioning a GWN into the face region. The position and scale of the face-like blob are used as initial parameters in the repositioning procedure. Finally, face and facial landmarks are tracked along the video sequence as described in the previous section.

We have tested our method in different color video sequences, obtaining good results (http://www.ime.usp.br/~rferis). Figure 5 shows, in the left illustration, the detection of a face in the initial frame. The right illustration shows the facial landmarks positioning, which was achieved by repositioning the GWN optimized considering the mean face of the Yale database. Tracking of regions around the facial landmarks (eyes, nose and mouth) is illustrated in figure 6, which presents the frames 60, 98 and 221 of a specific video sequence. Note that the method is robust to eye blinking, homogeneous illumination changes and different facial expressions.We are still verifying the performance of the system so that future work will cover more experimental results as well as comparison with other systems.

Concerning facial landmarks positioning, our method has basically two limitations. The first one is related to GWN repositioning, for this procedure may fail when large databases are involved. The second limitation is derived

from the fact that the distance among facial features varies from person to person. This may lead to imprecise facial landmarks location in some cases. So, it would be interesting to have another procedure that could adjust the position of the selected facial feature points after the GWN repositioning. These topics belong to our ongoing research work.

The tracking of facial landmarks may be imprecise under strong 3D face pose variation. On the other hand, it showed to be robust and even suitable for real-time applications, when a small number of wavelets is considered in the representation. We intend to use this approach for face recognition from video sequences.

Fig. 5. Locating a face and its facial landmarks.

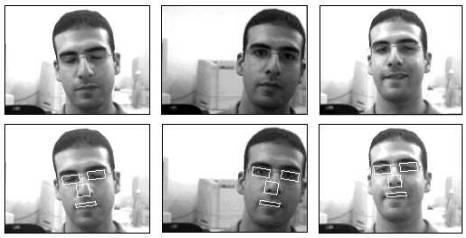

Fig. 6. Tracking eyes, nose and mouth.

8 Conclusions

This paper described a method for locating and tracking facial landmarks 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.

The obtained results confirmed the robustness of the method. Positioning of facial landmarks in the initial frame may be accomplished even in the presence of glasses, beard and different facial expressions. The tracking approach considers the overall geometry of the face so that it is robust to facial feature deformations such as eye blinking and smile. As future work, we intend to use the GWN to perform face detection.

Acknowledgements

Roberto M. Cesar Junior is grateful to FAPESP for the financial support (98/07722-0 and 99/12765-2), to "pro-reitoria de pesquisa" and to "pro-reitoria de pós-graduação" - USP, 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 the source code of GWN technique and for discussions. Some images in the paper were derived from the Yale face database.

References

1. Kruger V. and Sommer G. (1999) Affine real-time face tracking using a wavelet network. Proceedings of ICCV'99 Workshop Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece.
2. Feris R., Campos T. and Cesar R. (2000) Detection and tracking of facial features in video sequences. Lecture Notes in Artificial Intelligence, vol. 1793, pp. 127-135, Springer-Verlag.
3. Yang J. and Waibel A. (1996) A real-time face tracker. Proceedings of the Third IEEE Workshop on Applications of Computer Vision, pp. 142-147, Sarasota, Florida.
4. Yang J., Lu W. and Waibel A. (1997) Skin-color modeling and adaptation. CMU CS Technical Report, CMU-CS-97-146.
5. Stiefelhagen R. and Yang J. (1996) Gaze tracking for multimodal human computer interaction. University of Karlsruhe. Available at http://werner.ira.uka.de/ ISL.multimodal.publications.html
6. Silva L., Aizawa K. and Hatori M. (1995) Detection and tracking of facial features. Proceedings of SPIE Visual Communications and Image Processing, Taiwan.
7. Maurer T. and Malsburg C. (1996) Tracking and learning graphs and pose on image sequences of faces. Proceedings of Int. Conf. on Artificial Neural Networks, Bochum.
8. Mallat S. (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Analysis and Machine Intelligence, 11(7):674-693.
9. Daubechies I. (1990) The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Informat. Theory, 36(5):961-1004.
10. Zhang Q. and Benviste A. (1992) Wavelet networks. IEEE Trans. on Neural Networks, 3(6):889-898.
11. Press W., Flannery B., Teukolsky S. and Vetterling W. (1986) Numerical Recipes, The Art of Scientific Computing, Cambridge University Press, UK.
12. Daugman J. (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized two-dimensional visual cortical filters. Journal Opt. Soc. Am., 2(7):1160-1168.
13. Szu H., Telfer B. and Kadambe S. (1992) Neural network adaptive wavelets for signal representation and adaptation. Optical Engineering, 31(9):1907-1961.