Efficient Real-Time Face Tracking in Wavelet Subspace

Rogério S. Feris, Roberto M. Cesar Jr
Department of Computer Science
University of São Paulo
Rua do Matão, 1010, São Paulo-SP, Brazil
{*rferis* | *cesar*}*@ime.usp.br*

Volker Krüger
Center for Automation Research
University of Maryland
College Park, MD 20740
vok@cfar.umd.edu

Abstract

In this article we present a new method for visual face tracking that is carried out in wavelet subspace. Firstly, a wavelet representation for the face template is created, which spans a low-dimensional subspace of the image space. The video sequence frames where the face is tracked are then orthogonally projected into this low-dimensional subspace. This can be done efficiently through a small number of applications of the wavelet filters. All further computations are performed in wavelet subspace, which is isomorphic to the image subspace spanned by the sets of wavelets in the representation. Robustness w.r.t. facial expression and affine deformations, as well as the efficiency of our method, are demonstrated in various experiments.

1. Introduction

This paper addresses the issue of affine real-time face tracking. Real-time (RT, 25/30 Hz) is a major requisite for many human-computer-interface (HCI) and surveillance applications, as well as for tele-conferencing and teleteaching tasks. For gesture, gaze and pose estimation applications, tracking has to be not only fast, but also precise. A variety of tracking approaches already exists [5, 2, 1]. In this work, we will use Gabor Wavelet Networks (GWN) [6, 4] in order to represent the face to be tracked. It is worth saying that we have already discussed a GWN-based real-time tracking method in [5]. The method discussed in the present paper differs from that approach in the sense that it is performed in a low-dimensional wavelet subspace, posing a considerable enhancement in terms of efficiency.

Gabor Wavelet Networks (GWN) combine the advantages of RBF networks with the advantages of Gabor wavelets. Objects are represented through a linear combination of Gabor wavelets where the parameters of each of the Gabor functions (such as orientation, position and scale) are optimized to reflect the particular local image structure. The use of Gabor Wavelet Networks has several advantages, namely:

1. By their very nature, Gabor wavelet networks are invariant to some degree to affine deformations and homogeneous illumination changes,

2. Gabor filters are good feature detectors [8, 9] and the optimized parameters of each Gabor wavelet reflect the underlying image structure,

3. The Gabor wavelet weights are directly related to the Gabor filter responses and thus also reflect the underlying local image structure,

4. The precision of the representation can be varied to *any* desired degree ranging from a coarse to an almost photo-realistic representation by simply varying the number of used wavelets. Depending on the available computer power and the necessary tracking precision, the number of wavelets can be dynamically varied.

Properties 2-4 allow us to define a subspace of the image space. Its basis is given by the selectively chosen Gabor wavelets. Property 4 allows an easy projection of any image into the subspace, while property 3 assures the great sparseness of the wavelet representation. We will discuss each single point in section 2. In section 3, we will introduce our subspace tracking approach, discuss the details and conclude the paper with the experiments in section 4 and concluding remarks.

2. Introduction to Gabor Wavelet Networks

The basic idea of the wavelet networks was first stated in [11], and the use of Gabor functions is inspired by the fact that they are recognized to be good feature detectors [8, 9].

To define a GWN, we start out, generally speaking, by taking a family of N odd Gabor wavelet functions $\Psi =$

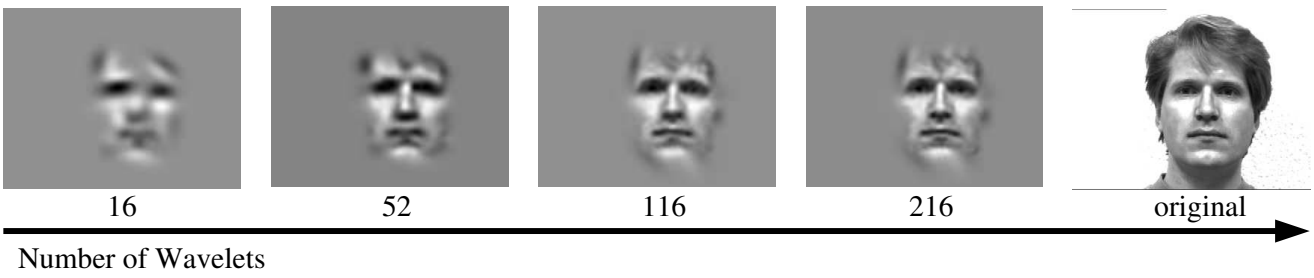

Figure 1. Face images reconstructed with different number of wavelets

$\{\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_N}\}$ of the form

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

with $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y)^T$. Here, c_x, c_y denote the translation of the Gabor wavelet, s_x, s_y denote the dilation and θ denotes the orientation. The choice of N is arbitrary and related to the degree of desired representation precision of the network. In order to find the GWN for a function $f \in \mathbb{L}^2(\mathbb{R}^2)$ (f dc-free, w.l.o.g.) the energy functional

$$E \quad = \quad \min_{\mathbf{n}_i, w_i \text{ for all } i} \|f - \sum_i w_i \psi_{\mathbf{n}_i}\|_2^2 \qquad (2)$$

is minimized with respect to the weights w_i and the wavelet parameter vector $\mathbf{n}_i$. The two vectors

$$\Psi = (\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_N})^T \quad \text{and} \quad \mathbf{w} = (w_1, \ldots, w_N)^T$$

define then the *Gabor wavelet network* $(\Psi, \mathbf{w})$ for function f.

In other words, considering a discrete image I[1], a Gabor wavelet network is defined through a N-dimensional vector of weights w_i and a N-dimensional vector of Gabor wavelets $\psi_{\mathbf{n}_i}$, where the weights w_i and the parameter vectors $\mathbf{n}_i$ are chosen such that the weighted sum of Gabor wavelets $\psi_{\mathbf{n}_i}$ approximates the discrete gray value image I optimally.

From the optimal wavelets Ψ and weights $\mathbf{w}$ of the Gabor wavelet network, the function f can be (closely) reconstructed by a linear combination of the weighted wavelets:

$$\hat{f} = \sum_{i=1}^{N} w_i \psi_{\mathbf{n}_i} = \Psi^T \mathbf{w} \ .$$

[1]We use the notation $f, g, \ldots$ to generally refer to continuous functions, while we use $I, J, \ldots$ when we explicitly refer to discrete gray value images.

Clearly, the quality of the image representation and reconstruction depends on the number N of wavelets and can be varied to reach almost any desired precision. An example of reconstruction can be seen in fig. 1. In this example, a family of 216 wavelets has been distributed over the inner face region of the very right image I by minimizing equation (2). Different reconstructions $\hat{I}$ obtained by the application of equation (3) for different values of N are shown.

2.1. Direct Calculation of Weights

The weights w_i of a GWN are directly related to the local filter responses of the Gabor filters $\psi_{\mathbf{n}_i}$. Gabor wavelet functions are not orthogonal, thus implying that, for a given family Ψ of Gabor wavelets, it is not possible to calculate a weight w_i by a simple projection of the Gabor wavelet $\psi_{\mathbf{n}_i}$ onto the image. In fact, a family of dual wavelets $\tilde{\Psi} = \{\tilde{\psi}_{\mathbf{n}_1} \ldots \tilde{\psi}_{\mathbf{n}_N}\}$ has to be considered. The wavelet $\tilde{\psi}_{\mathbf{n}_j}$ is the dual wavelet of the wavelet $\psi_{\mathbf{n}_i}$ iff $\langle \psi_{\mathbf{n}_i}, \tilde{\psi}_{\mathbf{n}_j} \rangle = \delta_{i,j}$. With $\tilde{\Psi} = (\tilde{\psi}_{\mathbf{n}_1}, \ldots, \tilde{\psi}_{\mathbf{n}_N})^T$, we can write $\left[\langle \Psi, \tilde{\Psi} \rangle\right] = \mathbb{I}$. In other words, given $g \in \mathbb{L}^2(\mathbb{R}^2)$ and a GWN $\Psi = \{\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_N}\}$, the optimal weights for g that minimize the energy in eq. (2) are given by $w_i = \langle g, \tilde{\psi}_{\mathbf{n}_i} \rangle$. It can be shown that $\tilde{\psi}_{\mathbf{n}_i} = \sum_j \left(\Psi^{-1}\right)_{i,j} \psi_{\mathbf{n}_j}$, where $\Psi_{i,j} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$.

The above equations allow us to define the operator

$$\mathcal{T}_\Psi : \mathbb{L}^2(\mathbb{R}^2) \longmapsto < (\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_N}) > \qquad (3)$$

as follows: given a set Ψ of optimal wavelets of a GWN, the operator $\mathcal{T}_\Psi$ represents an orthogonal projection of a function g onto the closed linear span of Ψ (see eq. (3) and fig. 2), i.e.

$$\hat{g} = \mathcal{T}_\Psi(g) \ = \ g\tilde{\Psi}\Psi \ = \ \sum_{i=1}^{N} w_i \psi_{\mathbf{n}_i} \ , \text{with } \mathbf{w} = g\tilde{\Psi} \ . \quad (4)$$

3. Face Tracking in Wavelet Subspace

The wavelet representation described in the previous section may be effectively used for affine face tracking. Basi-

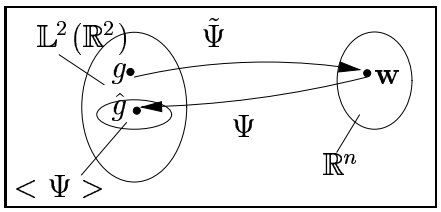

Figure 2. A function $g \in \mathbb{L}^2(\mathbb{R}^2)$ is mapped by the linear mapping $\tilde{\Psi}$ into the vector $\mathbf{w} \in \mathbb{R}^N$. The mapping of $\mathbf{w}$ into $\mathbb{L}^2(\mathbb{R}^2)$ is achieved with the linear mapping Ψ. Both mappings constitute an orthogonal projection of a function $g \in \mathbb{L}^2(\mathbb{R}^2)$ into the subspace $< \Psi > \subset \mathbb{L}^2(\mathbb{R}^2)$.

cally, this task is achieved by affinely deforming a GWN so that it matches the face image in each frame of a video sequence. The affine deformation of a GWN is carried out by considering the entire wavelet network as a single wavelet, which is also called *superwavelet* [5]. Let $(\Psi, \mathbf{w})$ be a Gabor wavelet network with $\Psi = (\psi_{\mathbf{n}_1}, \ldots, \psi_{\mathbf{n}_N})^T$ and $\mathbf{w} = (w_1, \ldots, w_N)^T$. A superwavelet $\Psi_{\mathbf{n}}$ is defined as a linear combination of the wavelets $\psi_{\mathbf{n}_i}$ such that

$$\Psi_{\mathbf{n}}(\mathbf{x}) \quad = \quad \sum_i w_i \psi_{\mathbf{n}_i}(\mathbf{SR}(\mathbf{x} - \mathbf{c})) \; , \tag{5}$$

where the vector $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y, s_{xy})$[2] defines the dilation matrix $\mathbf{S}$, the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{c}$, respectively. The affine face tracking is then achieved by deforming the superwavelet $\Psi_{\mathbf{n}}$ in each frame J, so that its parameters $\mathbf{n}$ are optimized with respect to the energy functional E (see eq. (2)):

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

Clearly, this method performs a typical pixel-wise pattern matching in image space, where the template corresponds to the wavelet representation, which is affinely distorted to match the face in each frame. It is interesting to note that the wavelet weights w_i (see eq. (5)) are constant under the deformation of the template. Therefore, the affine deformation is captured only by the deformation of the wavelets, while the weight vector remains invariant.

We thus claim that the tracking in image space described above may also be achieved in the low-dimensional space $\mathbb{R}^N$, which is isomorphic to the image subspace $< \Psi >$, as illustrated by fig. 2. As it can be seen there, both spaces are related through the matrices Ψ and $\tilde{\Psi}$, respectively.

As the first step, consider a GWN $(\Psi, \mathbf{v})$ that is optimized for a certain face image. As previously mentioned, the optimal weight vector $\mathbf{v}$ is obtained by an orthogonal projection of the facial image into the closed linear span of Ψ. Hence, we say that the face template was mapped into the weights $\mathbf{v} \in \mathbb{R}^N$, which we will call *reference weights*.

We mentioned before that the wavelet template gets affinely deformed for tracking in image space. Analogously, the tracking in wavelet subspace is performed by affinely deforming the subspace $< \Psi >$, until the weight vector $\mathbf{w} \in \mathbb{R}^N$, obtained by the orthogonal mapping of the current frame into this subspace, is closest to the reference weight vector $\mathbf{v}$. In fact, this procedure performs roughly the same pattern matching as before, but this is done efficiently in the low-dimensional space $\mathbb{R}^N$.

The mapping of images into $\mathbb{R}^N$ is carried out with low computational cost through a small number of local filtrations with the wavelets. Recall from section 2.1 that $w_i = \langle I, \tilde{\psi}_{\mathbf{n}_i} \rangle$, where $\tilde{\psi}_{\mathbf{n}_i} = \sum_j \left(\Psi^{-1}\right)_{i,j} \psi_{\mathbf{n}_j}$. This is equal to the following equation:

$$w_i = \sum_j \left(\Psi^{-1}\right)_{i,j} \langle I, \psi_{\mathbf{n}_j} \rangle. \tag{7}$$

Thus, the optimal weights w_i are derived from a linear combination of wavelet filtrations, where the coefficients are given by the inverse of matrix $\Psi_{i,j} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$. It can be shown that the matrix $\Psi_{i,j}$ is, except for a scalar factor, invariant with respect to affine transformations of the wavelet network. It can therefore be computed off-line and beforehand. Hence, the weights w_i are computed efficiently with eq. (7) through a local application $\langle I, \psi_{\mathbf{n}_i} \rangle$ of each of the N Gabor wavelets $\psi_{\mathbf{n}_i}$, followed by a $N \times N$ matrix multiplication.

Let $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y, s_{xy})$ be an affine parameter vector which configures a parameterization for the subspace $< \Psi >$. As we described before, tracking in wavelet subspace is achieved by gradually changing these parameters until the projection of the current frame J onto $\mathbf{w} \in \mathbb{R}^N$ is closest to the reference weights $\mathbf{v}$. In other words, we must optimize the parameters $\mathbf{n}$ with respect to the energy functional:

$$E \quad = \quad \min_{\mathbf{n}} \|\mathbf{v} - \mathbf{w}\|_\Psi \text{ with} \tag{8}$$

$$w_i \quad = \quad \sum_j \frac{1}{s_x \cdot s_y} \left(\Psi^{-1}\right)_{i,j} \langle J, \psi_{\mathbf{n}_j}\left(\mathbf{SR}\left(\mathbf{x} - \mathbf{c}\right)\right)\rangle \tag{9}$$

where $\mathbf{S}$ is the dilation matrix, $\mathbf{R}$ is the rotation matrix and $\mathbf{c}$ is the translation vector, all defined through the vector $(c_x, c_y, \theta, s_x, s_y, s_{xy})$. During tracking, this optimization is done for each successive frame. As there is not much difference between adjacent frames, the optimization is fast. To minimize the energy functional (8), the Levenberg-Marquardt algorithm was used.

In equation (8) we used the notation $\|\mathbf{v} - \mathbf{w}\|_\Psi$ to refer to a distance between vectors $\mathbf{v}$ and $\mathbf{w}$ of the subspace

[2] To include the shear in the parameter set of the wavelets, see [5].

$< \Psi >$. However, it is not yet clear how this distance measure should be calculated. One could use the Euclidean distance between two vectors, as done in [10]. However, that distance measure misses any interpretation in this context and all wavelets are treated as equal, even though they might be of different scales.

Therefore we propose a different distance measure, which is based on the Euclidean distance between the two corresponding points in the wavelet subspace $< \Psi >$. We thus define the difference $\|\mathbf{v} - \mathbf{w}\|_\Psi$ as the Euclidean distance between the two respective images $\hat{f}$ and $\hat{g}$:

$$\begin{aligned} \|\mathbf{v} - \mathbf{w}\|_\Psi &= \|\mathcal{T}_\Psi(f) - \mathcal{T}_\Psi(g)\|_2 \\ &= \|\sum_{i=1}^{N} v_i \psi_{\mathbf{n}_i} - \sum_{j=1}^{N} w_j \psi_{\mathbf{n}_j}\|_2 \end{aligned} \quad (10)$$

Various transformations lead to

$$\begin{aligned} \|\mathbf{v} - \mathbf{w}\|_\Psi &= \left[\sum_{i,j} (v_i - w_i)(v_j - w_j)\langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j}\rangle\right]^{\frac{1}{2}} \\ &= (\mathbf{v} - \mathbf{w})^t \; \Psi_{i,j} \; (\mathbf{v} - \mathbf{w}). \end{aligned} \quad (11)$$

The matrix of pairwise scalar products $\Psi_{i,j} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j}\rangle$ is the same matrix as the one in section 2.1 and eq. (7). Note that if the wavelets $\{\psi_{\mathbf{n}_i}\}$ were orthogonal, then Ψ would be the unity matrix and eq. (11) would describe the same distance measure as proposed in [10].

Compared to tracking in image space, the method presented here poses a considerable enhancement in terms of efficiency, as it provides a great data reduction, considering that tracking is performed in a low dimensional wavelet subspace. Moreover, it spares out the computationally demanding template reconstruction and pixel-wise sum-of-squared-difference computation required in the image-based tracking.

4. Experiments

The proposed approach for affine face tracking was successfully tested on various image sequences. All test sequences showed a person in motion, more or less frontal to the camera, so that the facial features are always visible.

Our experiments demonstrate the ability of our method to track a face as it undergoes changes in pose and expression. Although we have represented the face as a rigid object undergoing limited motion, we realized that different face expressions and small depth variations exhibited by facial features are enough to be well-approximated by the affine wavelet model. Furthermore, since Gabor wavelets are DC free, our approach also showed robustness with respect to homogeneous illumination variations.

Concerning tracking precision, we have evaluated the proposed method with respect to the number of wavelets used in the representation. As the first step, we recorded a video sequence that shows a person under different poses and facial expressions [3]. On the face of that person, a GWN with 116 Gabor wavelets was optimized and used to estimate the "ground truth" affine parameters in each frame. Figure 3 illustrates some frames of our subspace tracking results, using this wavelet representation. Note that the tracking method is robust to facial expressions variations as well as affine deformations of the face image.

Now, we want to analyze the decrease of precision when using a smaller number of wavelets in the representation. For that, from the large GWN above, GWNs that contained only the largest 51, 22 and 9 wavelets, sorted according to decreasing normalized weights, were used for tracking. The graphs in figure 4 depict the estimation of the face parameters x-position, y-position and angle θ as well as the ground-truth in each frame.

In the graphs we can see that the tracking precision increases with an increasing number of applied wavelets. On the other hand, the tracking speed is improved when the number of wavelets is small. Clearly, the number of applied wavelets is task dependent and can be dynamically changed.

The resolution of the image frames showed in figure 3 is 160x120 pixels and the size of the inner face region in which the GWN was optimized is 50x65 pixels. Using only 9 wavelets, the computing time for each Levenberg-Marquardt cycle was 15ms on a 1GHz Linux-Athlon. Higher performance is achieved for smaller face regions, or fewer parameters (e.g. just translation and rotation). It is worth saying that the number of cycles for each frame depends on the distance of the initial parameters from the local minimum, which is directly related to object speed. Increasing the number of wavelets in the representation leads to a more precise but slower tracking. For instance, using 51 wavelets, a computing time of 85ms per cycle was required in each frame.

Our method is closely related to the one presented in [5]. In comparison to that method we observed in our experiments a speedup of a factor two for each single Levenberg-Marquardt cycle, but with a slight increase in the number of needed cycles. However, we do believe that the use of other algorithms, such as the Condensation method [3] or the Sequential Importance Sampling (SIS) method [7], will even lead to better results in terms of efficiency.

5. Conclusions

In this paper we have presented a tracking method, that is based on Gabor Wavelet Networks. The GWNs are in-

[3] see http://www.ime.usp.br/~rferis

Figure 3. Sample frames of our wavelet subspace tracking. Note that the tracking method is robust to facial expressions variations as well as affine deformations of the face image.

variant to affine deformations so that they leave the reference vector of a template face constant. Furthermore, the direct relationship between the wavelet coefficients and the wavelet filter responses, which is widely used for multiresolution analysis and for motion estimation, allows to map an image into the low-dimensional space $\mathbb{R}^N$, where all subsequent computations can be carried out. Furthermore, the careful selection of the wavelets for defining a GWN ensures that almost all visual information is preserved (see fig. 1 and [4]).

The method has the further advantage, that its precision and computation time can be adapted to the needs of a given task. When fast and only approximate tracking is needed, a small number of filtrations is usually sufficient to realize tracking. When high precision tracking is needed, the number of wavelets can be gradually increased. This implies on the one hand more filtrations, but ensures on the other hand a higher precision, as we have shown in the experimental section.

So far, the optimization of the wavelet networks has been done with a Levenberg-Marquardt algorithm. We think that the application of Condensation or Sequential Importance Sampling (SIS) instead would lead to a further speed-up. This will be evaluated in future work.

6. Acknowledgements

Rogerio Feris is grateful to FAPESP (99/01487-1). Roberto M. Cesar Jr. is grateful to FAPESP (99/12765-2), CNPq (300722/98-2) and TSI-ENST Paris.

References

[1] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, volume 2, pages 142–149, Hilton Head Island, SC, June 13-15, 2000.

[2] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

[3] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[4] V Krüger, Sven Bruns, and G. Sommer. Efficient head pose estimation with gabor wavelet networks. In *Proc. British Machine Vision Conference*, Bristol, UK, Sept. 12-14, 2000.

[5] V Krüger and G. Sommer. Affine real-time face tracking using gabor wavelet networks. In *Proc. Int. Conf. on Pattern Recognition*, Barcelona, Spain, Sept. 3-8, 2000.

[6] V. Krüger and G. Sommer. Gabor wavelet networks for object representation. Technical Report 2002, Institute of Computer Science, University of Kiel, 2000.

[7] Baoxin Li. Human and object tracking and verification in video. Technical Report CS-TR-4140, Center for Automation Research, University of Maryland, May 2000.

[8] B.S. Manjunath and R. Chellappa. A unified approach to boundary perception: edges, textures, and illusory contours. *IEEE Trans. Neural Networks*, 4(1):96–107, 1993.

[9] R. Mehrotra, K.R. Namuduri, and R. Ranganathan. Gabor filter-based edge detection. *Pattern Recognition*, 52(12):1479–1494, 1992.

[10] L. Wiskott, J. M. Fellous, N. Krüger, and C. v. d. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.

[11] Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. Neural Networks*, 3(6):889–898, Nov. 1992.

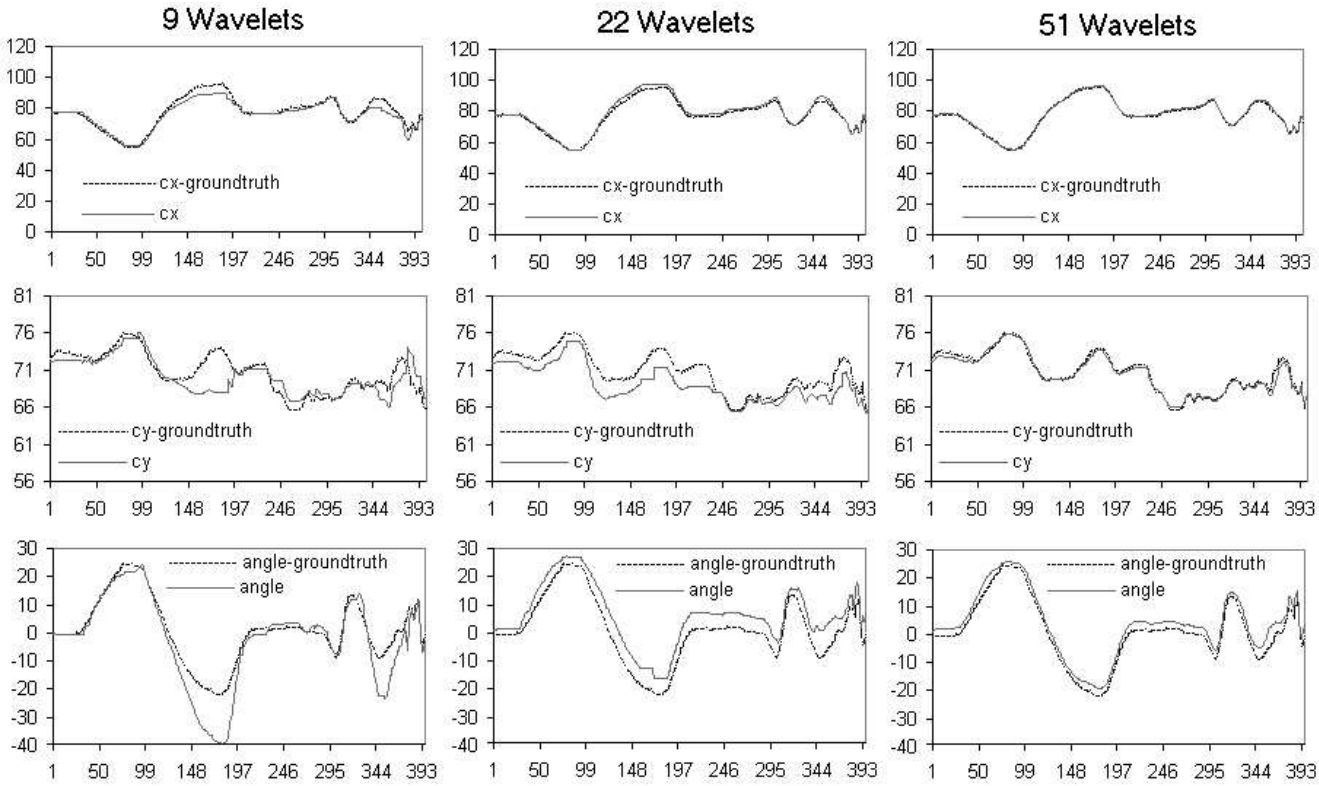

Figure 4. Estimation of the face parameters x-position, y-position and angle θ in each frame, using GWNS with 9, 22 and 51 wavelets. The ground-truth is depicted to illustrate the decrease of precision when considering few wavelets.