ENHANCED PHASE–BASED DISPLACEMENT ESTIMATION
An Application to Facial Feature Extraction and Tracking

Mohamed Dahmane and Jean Meunier
Diro, Université de Montréal, CP 6128, Succursale Centre-Ville
2920 Chemin de la tour, Montréal, Québec, Canada, H3C 3J7

Keywords: Facial feature extraction, Facial analysis, Gabor wavelets, Tracking.

Abstract: In this work, we develop a multi-scale approach for automatic facial feature detection and tracking. The method is based on a coarse to fine paradigm to characterize a set of facial fiducial points using a bank of Gabor filters that have interesting properties such as directionality, scalability and hierarchy. When the first face image is captured, a trained grid is used on the coarsest level to estimate a rough position for each facial feature. Afterward, a refinement stage is performed from the coarsest to the finest (original) image level to get accurate positions. These are then tracked over the subsequent frames using a modification of a fast phase–based technique. This includes a redefinition of the confidence measure and introduces a conditional disparity estimation procedure. Experimental results show that facial features can be localized with high accuracy and that their tracking can be kept during long periods of free head motion.

1 INTRODUCTION

The computer vision community is interested in the development of techniques to figure out the main element of facial human communication in particular for HCI applications or, with additional complexity, meeting video analysis. In both cases, automatic facial analysis is highly sensitive to face tracking performance, a task which is rendered difficult due principally to environment changes and particularly to its great appearance variability under different head orientations, its non–rigidity adds yet another degree of difficulty. To overcome these problems, a great number of techniques have been developed which can be divided into four categories: knowledge–, feature–, template– and appearance–based (Yang, 2004).

Among these techniques, it is known that face analysis by feature point tracking demonstrates high concurrent validity with manual FACS (Facial Action Coding System) coding (Cohen et al., 1999), which is promising for facial analysis (Cottrell et al., 2003). Moreover, when facial attributes are correctly extracted, geometric feature–based methods typically share some common advantages, such as explicit face structure, practical implementation, collaborative feature-wide error elimination (Hu et al., 2004). In this context, several concepts were developed.

The classical matching technique extracts features from two frames and tries to establish a correspondence, whereas correlation-based techniques compare windowed areas in two frames, and the maximum cross correlation value provides the new relative position. However, recent techniques have been developed to determine the correct relative position (disparity[1]) without any searching process as it is required by the conventional ones. In this category, phase–based approaches have attracted attention because of their biological motivation and robustness (Theimer and Mallot, 1994; Fleet and Jepson, 1993).

In the literature, one can find several attempts at designing non–holistic methods based on Gabor wavelets (Shen and Bai, 2006). Due to their interesting and desirable properties including spatial locality, self similar hierarchical representation, optimal joint uncertainty in space and frequency as well as biological plausibility (Flaton and Toborg, 1989). However,

[1] we use interchangeably the words "disparity" and "displacement"

Dahmane M., Meunier J. and Meunier J. (2008).
ENHANCED PHASE–BASED DISPLACEMENT ESTIMATION - An Application to Facial Feature Extraction and Tracking.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 427-433
DOI: 10.5220/0001081804270433
Copyright © SciTePress

most of them are based on the magnitude part of the filter response (Lades et al., 1993; Tian et al., 2002; Liu and Wechsler, 2003; Valstar and Pantic, 2006). In fact, under special consideration, particularly because of shift–variant property, the Gabor phase can be a very discriminative information source (Zhang et al., 2007).

In this paper, we use this property of Gabor phase for facial feature tracking. In section 2, we describe the Gabor-kernel family we are using. In section 3, we introduce the adopted strategy for facial features extraction. The tracking algorithm is given in section 4, including technical details and a discussion on its derivation. Finally, we apply the approach to a facial expression database, in section 5.

2 LOCAL FEATURE MODEL BASED ON GABOR WAVELETS

2.1 Gabor Wavelets

A Gabor jet $J(\mathbf{x})$ describes via a set of filtering operation (eq. 1), the spatial frequency structure around the pixel $\mathbf{x}$, as a set of complex coefficients.

$$J_j(\mathbf{x}) = \int_{N^2} I(\mathbf{x}')\Psi_j(\mathbf{x}-\mathbf{x}')\,d\mathbf{x}' \tag{1}$$

A Gabor wavelet is a complex plane wave modulated by a Gaussian envelope:

$$\Psi_j(\mathbf{x}) = \eta_j e^{-\frac{\|\mathbf{k}_j\|^2\|\mathbf{x}\|^2}{2\sigma^2}}\left[e^{\imath\mathbf{k}_j\cdot\mathbf{x}} - e^{-\frac{\sigma^2}{2}}\right] \tag{2}$$

where $\sigma = 2\pi$, and $\mathbf{k}_j = (k_{jx}, k_{jy}) = (k_\nu\cos(\phi_\mu), k_\nu\sin(\phi_\mu))$ defines the wave vector, with

$$k_\nu = 2^{-\frac{\nu+2}{2}}\,\pi \quad \text{and} \quad \phi_\mu = \mu\frac{\pi}{8}$$

Notice that the last term of equation 2 compensates for the non-null average value of the cosine component. We choose the term η_j so that the energy of the wavelet Ψ_j is unity (eq. 3).

$$\int_{N^2} \left|\Psi_j(\mathbf{x})\,d\mathbf{x}\right|^2 = 1 \tag{3}$$

A *jet* $J(\mathbf{x}) = \{a_j\, e^{\imath\phi_j} \;/\; j = \mu + 8\nu\}$, is commonly defined as a set of 40 complex coefficients constructed from different Gabor filters spanning different orientations ($\mu \in [0,7]$) under different scales ($\nu \in [0,4]$).

3 AUTOMATIC VISUAL ATTRIBUTE DETECTION

3.1 Rough Face Localization

When the first face image is captured, a pyramidal image representation is created, where the coarsest level is used to find near optimal starting points for the subsequent individual facial feature localization stage. Each trained grid (Fig. 1) from a set of prestored face grids is displaced as a rigid object over the image. The grid position that maximizes the weighted magnitude–based similarity function (eq. 4 and 5) provides the best fitting node positions.

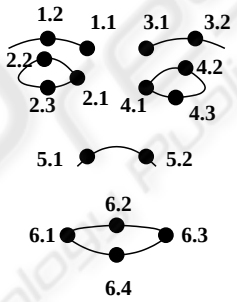

Figure 1: Facial nodes with their respective code.

$$Sim(\mathrm{I},\mathrm{G}) = \prod_{l}^{L} S(J_l, J'_l) \tag{4}$$

$S(J,J')$ refers to the similarity between the jets of the corresponding nodes (eq. 5), L stands for the total number of nodes.

$$S(J,J') = \sum_j c_j \frac{a_j a'_j}{\sqrt{\sum a_j{}^2 \sum a'_j{}^2}} \text{ with } c_j = \left(1 - \frac{\left|a_j - a'_j\right|}{a_j + a'_j}\right)^2 \tag{5}$$

The role of the weighting factor c_j is to model the amplitude–distortion δ as illustrated in figure 2.

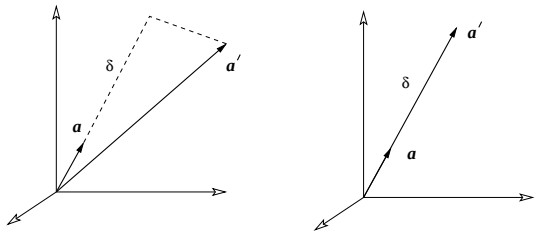

Figure 2: Two different 3–dimensional jets. In the right subfigure, a not–weighted amplitude–based similarity $S(J,J')$ would have given an incorrect perfect match value 1..

3.2 Local Facial Feature Position Refinement

The rough facial grid-node positions are then independently refined by estimating the displacement using a hierarchical selective search. The calculated displacements are propagated to subsequent hierarchy level, and a refinement operation is again performed. The optimal displacements are, finally, given at the finest image level.

The selective local search can be described as a local 3×3 neighborhood search, which allows distorting the grid until the maximum similarity value is reached. The search is then refined by propagating, to the next finer level, the three positions giving the highest similarity values. For each propagated potential position $P(x,y)$ the three adjacent neighboring positions $P(x+1,y), P(x,y+1)$ and $P(x+1,y+1)$ are also explored. The selective search continues downward until the finest level of the pyramid image is reached, where the optimal position is maximum (eq. 5).

This procedure permits to decrease the inherent complexity required to calculate the convolution under an exhaustive search, first by reducing the search area (e.g. a 12×12 neighborhood on the finest level will correspond only to a 3×3 on the coarsest one) (Fig. 3), and second by using smaller–size jets in coarser levels.

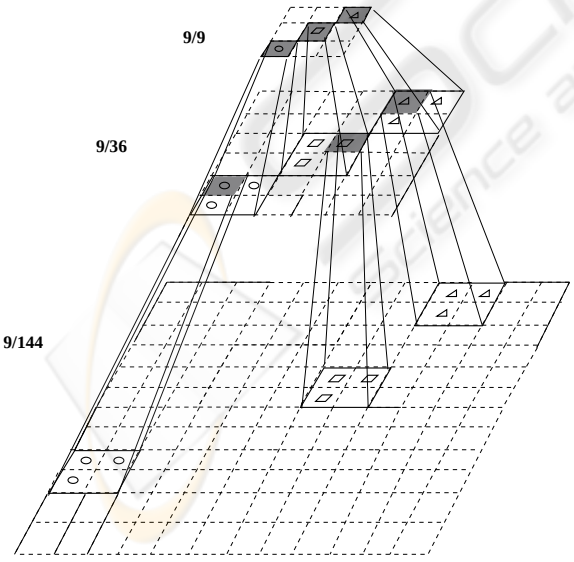

Figure 3: Hierarchical–selective search. The values in left side denote the number of explored positions vs. the total number that would be explored in the case of an exhaustive search.

4 FACIAL ATTRIBUTES TRACKING

Facial features tracking is performed by estimating a displacement $\mathbf{d}$ via a disparity estimation technique (Theimer and Mallot, 1994), that exploits the strong variation of the phases of the complex filter response (Maurer and von der Malsburg, 1996).

Later adopted by (Zhu and Ji, 2006), this framework investigated in (Maurer and von der Malsburg, 1996; Wiskott et al., 1997) is based on the maximization of a phase–based similarity function which is nothing else than a modified way to minimize the squared error, within each frequency scale ν given two jets J and J' (eq. 6), as it has been proposed in (Theimer and Mallot, 1994).

$$e_\nu^2 = \sum_\mu c_{\nu,\mu}(\Delta\phi_{\nu,\mu} - \mathbf{k}_{\nu,\mu} \cdot \mathbf{d}_\nu)^2 \tag{6}$$

However, we assume that the merit of that framework is the use of a *saliency* term (eq. 7) as weighting factor $c_{\nu,\mu}$, privileging displacement estimation from filters with higher amplitude response. Also, for such response it seems that phase is more stable (McKenna et al., 1997).

$$c_j = a_j a'_j \tag{7}$$

In (Theimer and Mallot, 1994), the weighting factor c_j represents a confidence value (eq. 8), that assesses the relevance of a single disparity estimate, and tends to reduce the influence of erroneous filter responses.

$$c_j = 1 - \frac{\left|a_j - a'_j\right|}{a_j + a'_j} \tag{8}$$

Both saliency term and normalized confidence ignore the phase of the filter response. In the present work, we try to penalize the response of the erroneous filters by using a new confidence measure that combines both amplitude and phase (eq. 9).

$$c_j = {a_j}^2 \left(1 - \frac{\left|a_j - a'_j\right|}{a_j + a'_j}\right)^2 \frac{\pi - \left|\left\lfloor \Delta\phi_j \right\rfloor_{2\pi}\right|}{\pi} \tag{9}$$

The first term in this formulation represents the saliency term that is incorporated as a squared value of only the amplitude of the *reference jet J* which – contrary to the *probe jet J'*– necessarily ensures high confidence. We mean here by the reference jet the jet calculated from the previous frame or even a prestored one. The second bracket squared-term holds the normalized magnitude confidence. While, the last term, where $\left\lfloor \Delta\phi_j \right\rfloor_{2\pi}$ denotes the principal part of the phase difference within the interval $[-\pi, \pi)$, allows giving more weight to filters where the phase difference has a favorable convergence while, at the same time, limiting the influence of outlier filters.

The displacements can then be estimated with sufficient accuracy by minimizing (eq. 6) which leads to a set of linear equations for $\mathbf{d}$, that can be directly resolved from (eq. 10).

$$\mathbf{d}(J,J') = \begin{pmatrix} \sum_j c_j {k_{jx}}^2 & -\sum_j c_j k_{jx} k_{jy} \\ -\sum_j c_j k_{jx} k_{jy} & \sum_j c_j {k_{jy}}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_j c_j k_{jx} \lfloor \Delta\phi_j \rfloor_{2\pi} \\ \sum_j c_j k_{jy} \lfloor \Delta\phi_j \rfloor_{2\pi} \end{pmatrix} \quad (10)$$

4.1 Iterative Disparity Computation

In (Theimer and Mallot, 1994), to obtain the disparity within one scale, the feature displacement estimates for each orientation were combined into one displacement per scale ($\mathbf{d}_\nu$) using the least squared error criterion (eq. 6). The optimal disparity is then calculated by a combination of these estimates as an average value over all scales with appropriate weights (eq. 8). Whereas in various approaches, a least squared solution is obtained in one pass, over the overall considered frequencies (Wiskott et al., 1997), some of them propose at first to use the lower frequencies subset (e.g. $\nu \in [2,4]$), and then to resolve for higher frequencies subset (e.g. $\nu \in [0,2]$).

These resolutions may carry an additive risk of unfavorable results; that is knowing that at each scale, there exists a displacement value above which its estimation would not be reliable, due to the lack of a large overlap of the Gabor kernels. Obviously, this value depends on the radius (σ/k_ν) of the Gaussian envelope.

As the power spectrum of the Gabor signal (eq. 2) is concentrated in the interval $[-\sigma/(2k_\nu), \sigma/(2k_\nu)]$, we can compute the maximum disparity $\mathbf{d}_\nu^{max}$ that can be estimated within one scale (eq. 11).

$$d_\nu^{max} = \frac{\sigma}{2k_\nu} = \frac{\pi}{k_\nu} \quad (11)$$

If for example the true displacement is $d = 7\,pixels$, then according to the Gabor–kernel family we used (section 2.1), only the lowest frequency band filter gives a reliable estimation of the disparity.
So, the trick consists in estimating the disparity iteratively, from the lowest frequency to a highest *critical* frequency, depending on a stopping criterion involving the maximum allowed disparity value that can be effectively estimated. Some values are shown in table 1 as a function of scale.

Given $J(\mathbf{x}) = \{a_j e^{\imath\phi_j}\}$ the *reference jet* and $J'(\mathbf{x}+\mathbf{d}) = \{a'_j e^{\imath\phi'_j}\}$ the *probe jet* i.e. the jet calculated at the probe position $(\mathbf{x}+\mathbf{d})$, using

Table 1: Critical displacement for each frequency.

ν	0	1	2	3	4
$d_\nu^{\max}(pixel)$	2	≈ 3	4	≈ 6	8

the j^{th} wavelet, an iterative disparity estimation algorithm (Fig. 4) gives the optimal displacement $\mathbf{d}_{opt}$, that makes the two jets the most similar possible.

Algorithm 1. ITERATIVEDISPARITYESTIMATION $(\mathbf{x})$

1 Initially set ν with the lowest frequency index;
2 Calculate $J'_\nu(\mathbf{x})$ for the components that refer to ν at different orientations;
3 Estimate the disparity $\delta\mathbf{d}$ using equation (10) by considering all the processed frequencies at different orientations;
4 Compensate for the phase $\phi'_j = \lfloor \phi'_j - \mathbf{k}_j \cdot \delta\mathbf{d} \rfloor_{2\pi}$;
5 Cumulate the disparity $\mathbf{d} = \mathbf{d} + \delta\mathbf{d}$;
6 Perform the convergence test, if δd is greater than a threshold goto (3);
7 If the stopping criterion is not met, i.e. the overall displacement d is less than the critical displacement value $\mathbf{d}_\nu^{\max}$, see Table (1), then put $\nu = \nu + 1$ (the next higher frequency) and goto (2).

Figure 4: Conditional iterative disparity estimation algorithm.

Iteratively, the conditional iterative disparity estimation (Fig. 4) will unroll on the novel position $\mathbf{x}^{new} \leftarrow \mathbf{x} + \mathbf{d}_{opt}$ until a convergence criterion is achieved i.e. $\mathbf{d}_{opt}$ tends to $\mathbf{0}$ or the maximum number of iterations $l_{max_{iter}}$ is reached. Herein, ν_{critic} could keep its previous value, instead of starting, for each new position, with the coarsest scale (i.e. $\nu_{critic} = N_f - 1$).

5 EXPERIMENTAL RESULTS

The Hammal–Caplier face database (Hammal et al., 2007) is used to test the proposed approach. In this database, each video contains about 120 frames for each of the 15 distinct subjects that are acting different facial expressions (neutral, surprise, disgust and joy) with some tolerance on rotation and tilting. We used 30 videos with spatial resolution of (320×240).

Table 2: Percentage of used frames to handle local facial deformations.

facial feature	1.1	1.2	2.1	2.2	2.3	3.1	3.2	4.1	4.2	4.3	5.1	5.2	6.1	6.2	6.3	6.4
(%) of used frames	2.5	1.8	3.9	4	3	2.3	3.4	4.2	3.7	2.7	1.5	2.4	3.8	8	2	9

A generic face grid (Fig. 1) is created using one frame from each subject (frontal view). In order to handle the facial deformation and prevent drifting, facial feature bunches are generated. Table 2 shows each landmark and the percentage of the total number of frames required to create its own representative facial bunch. As we can see the number increases with the degree of variability of the local deformation that can be observed for each facial feature. These percentages were set empirically.

To locate the face grid, a search is performed over the coarsest level of the 3 image-levels that we used. Then a hierarchical selective refinement is performed using a weighted magnitude–based similarity to get the optimal node positions. Figure 5 shows the results corresponding to the position refinement after rough node positioning.

Figure 5: Nodes position refinement (bottom) after rough positioning (top).

Figure 6 shows the magnitude profile corresponding to $(\mu, \nu) = (0,0)$ for node 2.1 (right inner–eye) from a video where the subject is performing a disgust expression. Figure 7 illustrates the phase profile of the same subject with and without phase compensation $(\phi'_j \leftarrow \left\lfloor \phi'_j - \mathbf{k}_j \cdot \mathbf{d}_I \right\rfloor_{2\pi})$ in Algorithm 1.
One can observe some large and sharp phase variations when non compensation is used, corresponding to tracking failure.
Figure 8 shows three shots of a video showing a subject performing a disgust expression, the top subfigure presents the last frame. In this figure, we can see that the tracking has failed with a single jet (instead of a bunch). It's easy to see that the drifting can not be measured from the magnitude profile only (middle row), because the magnitude changes smoothly with the position. This is not the case for the phase (bottom row) which is shift–variant, however by using a shift–compensation and facial bunches as described in Algorithm 1, we can correctly track the facial landmarks (Fig. 9). In comparison with figure 8, the bottom graph shows a horizontal and correct phase profile (without node drifting). The reader can appreciate the impact of such correction by looking in particular at node

Figure 6: Amplitude profile over time of Node 2.1 (right inner–eye).

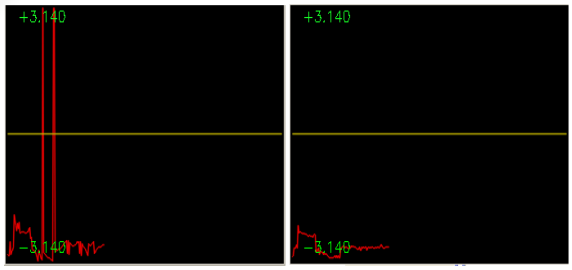

Figure 7: phase profile : not–corrected (left) vs. corrected (right) phase.

2.1 (right inner–eye) and 2.3 (right lower eyelid) in figures 8 and 9.

In table 3, we summarize the tracking results of 16 facial features of 10 different persons with different expressions. The mean error of node positions using the proposed approach is presented in pixels. From the last column, we can see how the use of facial bunches appreciably increases nodes positioning and consequently the tracking accuracy.

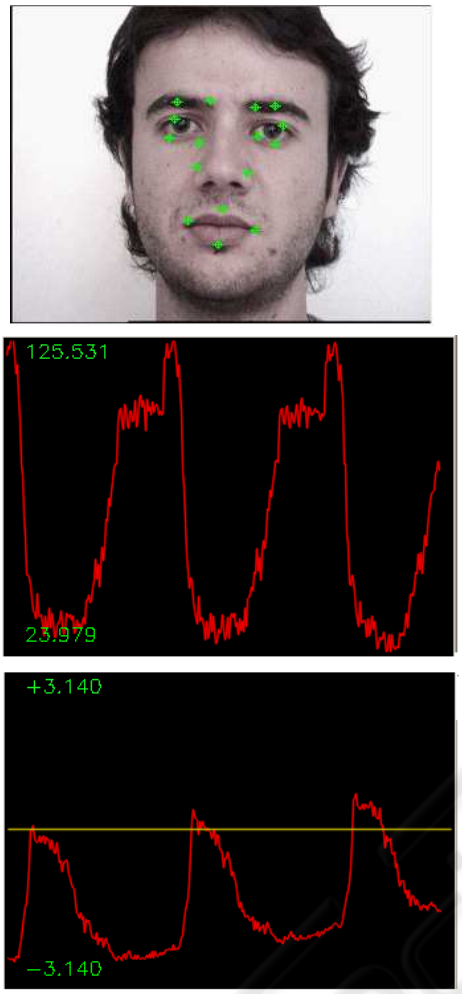

Figure 8: A drifting case : Magnitude vs. Phase profile.

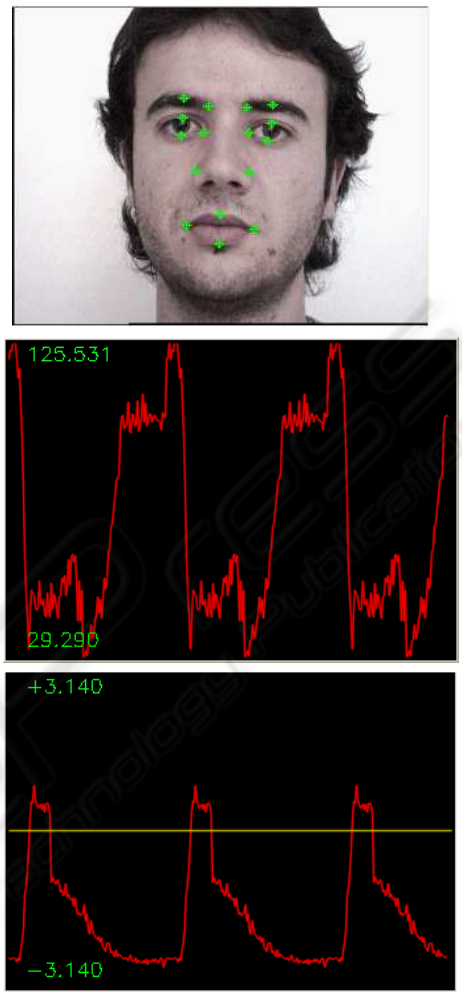

Figure 9: Drift avoidance.

Table 3: Mean position error (pixels).

Subject	Without bunches	With bunches
#1	4.28	1.78
#2	3.98	1.37
#3	5.07	2.03
#4	4.44	1.9
#5	4.17	1.7
#6	4.05	1.63
#7	4.69	1.5
#8	4.1	1.75
#9	5.85	2.49
#10	6.93	2.47

6 CONCLUSIONS

In this work, we present a modification of a phase–based displacement estimation technique using a new confidence measure and a conditional disparity estimation. The proposed tracking algorithm permits to eliminate accumulation of tracking errors to avoid drifting, so offering a good facial landmark localization, which is a crucial task in a feature–based facial expression recognition system. We notice that in these experiments, excepts for the first frame, no geometry constraints were used to enforce the facial shape configuration, especially for features that are difficult to track.

More training sessions could be needed to obtain pre-stored grids and features bunches that are rep-

resentative of the variability of the human face appearance for initialisation and tracking respectively. In this context, through available face databases, advanced statistical models of data can be obtained using learning algorithms, such as EM (Jiao et al., 2003).

To reinforce the refinement step we are working on improving the local structure by providing an alternative appearance model which focuses more on high frequency domain without necessarily altering the relevant low frequency texture information, instead of modeling the grey level appearance (Zhang et al., 2003) or exploiting the global shape constraint (McKenna et al., 1997) which tends to smooth out important details.

As future work, we plan to use facial feature bunches to generate for each facial expression and for each facial attribute what could constitute "Expression Bunches" for facial expression analysis.

ACKNOWLEDGEMENTS

This research was supported by the National Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

Cohen, J., Zlochower, A., Lien, J., and Kanade, T. (1999). *Face Analysis by Feature Point Tracking Has Concurrent Validity with Manual FACS Coding*. Psychophysiology 36(1):35–43.

Cottrell, G., Dailey, M., and Padgett, C. (2003). *Is All Faces Processing Holistic? The view from UCSD*. M. Wenger, J Twnsend (Eds), Computational, Geometric and Process Perspectives on Facial Recognition, Contexts and Challenges: Contexts and Challenges, Erlbaum.

Flaton, K. and Toborg, S. (1989). An approach to image recognition using sparse filter graphs. In *International Joint Conference on Neural Networks, (1):313–320*.

Fleet, D. and Jepson, A. (1993). Stability of phase information. In *IEEE Trans. on PAMI, 15(12):1253–1268*.

Hammal, Z., Couvreur, L., Caplier, A., and Rombaut, M. (2007). Facial expression classification: An approach based on the fusion of facial deformation unsing the transferable belief model. In *Int. Jour. of Approximate Reasonning*.

Hu, Y., Chen, L., Zhou, Y., and Zhang, H. (2004). Estimating face pose by facial asymmetry and geometry. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Jiao, F., Li, S., Shum, H.-Y., and Schuurmans, D. (2003). Face alignment using statistical models and wavelet features. In *Computer Vision and Pattern Recognition (1) p. 321–327*.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. In *IEEE Transactions on Computers 3(42):300–311*.

Liu, C. and Wechsler, H. (2003). Independent component analysis of gabor features for face recognition. In *IEEE Trans. on Neural Networks, (14):4, 919–928*.

Maurer, T. and von der Malsburg, C. (1996). Tracking and learning graphs and pose on image sequences of faces. In *2nd International Conference on Automatic Face and Gesture Recognition, p. 76*.

McKenna, S., Gong, S., Würtz, R., Tanner, J., and Banin, D. (1997). Tracking facial feature points with gabor wavelets and shape models. In *Proceedings of the First International Conference on Audio– and Video–based Biometric Person Authentication, 1206(3):35–42*. Springer Verlag.

Shen, L. and Bai, L. (2006). A review on gabor wavelets for face recognition. In *Pattern Analysis and Applications, (9):2,273–292*.

Theimer, W. and Mallot, H. (1994). Phase–based binocular vergence control and depth reconstruction using active vision. In *CVGIP: Image Understanding, 60(3):343–358*.

Tian, Y., Kanade, T., and Cohn, J. (2002). Evaluation of gabor wavelet–based facial action unit recognition in image sequences of increasing complexity. In *In Proc. of the 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*.

Valstar, M. and Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. In *CVPRW, p. 149*.

Wiskott, L., Fellous, J., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. In *IEEE Transactions on Pattern Analysis and Machine Intelligence. 19(7):775–779*.

Yang, M. (2004). Recent advances in face detection. In *Tutorial of IEEE Conferece on Pattern Recognition*.

Zhang, B., Gao, W., Shan, S., and Wang, W. (2003). Constraint shape model using edge constraint and gabor wavelet based search. In *AVBPA03, 52–61*.

Zhang, B., Shan, S., Chen, X., and Gao, W. (2007). Histogram of gabor phase patterns (HGPP): A novel object representation approach for face recognition. In *IEEE Tran. on Image Processing (16):1, pp.57-68*.

Zhu, Z. and Ji, Q. (2006). Robust pose invariant facial feature detection and tracking in real-time. In *ICPR, 1092-1095*.