Multi-scale keypoints in V1 and face detection

João Rodrigues[1] and J.M.Hans du Buf[2]

[1] University of Algarve – Escola Superior Tecnologia, Faro, Portugal
[2] University of Algarve – Vision Laboratory – FCT, Faro, Portugal

Abstract. End-stopped cells in cortical area V1, which combine outputs of complex cells tuned to different orientations, serve to detect line and edge crossings (junctions) and points with a large curvature. In this paper we study the importance of the multi-scale keypoint representation, i.e. retinotopic keypoint maps which are tuned to different spatial frequencies (scale or Level-of-Detail). We show that this representation provides important information for Focus-of-Attention (FoA) and object detection. In particular, we show that hierarchically-structured saliency maps for FoA can be obtained, and that combinations over scales in conjunction with spatial symmetries can lead to face detection through grouping operators that deal with keypoints at the eyes, nose and mouth, especially when non-classical receptive field inhibition is employed. Although a face detector can be based on feedforward and feedback loops within area V1, such an operator must be embedded into dorsal and ventral data streams to and from higher areas for obtaining translation-, rotation- and scale-invariant face (object) detection.

1 Introduction

Our visual system is still a huge puzzle with a lot of missing pieces. Even in the first processing layers in area V1 of the visual cortex there remain many open gaps, despite the amount of knowledge already compiled, e.g. [3, 5, 25]. Recently, models of cortical cells, i.e. simple, complex and end-stopped, have been developed, e.g. [7]. In addition, several inhibition models [2, 17], keypoint detection [7, 12, 22] and line/edge detection schemes [2, 12, 14, 15], including disparity models [6, 11], have become available. On the basis of these models and possible processing schemes, it is now possible to create a cortical architecture for figure-background segregation [16] and visual attention or Focus-of-Attention (FoA), bottom-up and/or top-down [4, 8, 13], and even for object categorisation and recognition.

In this paper we will focus exclusively on keypoints, for which Heitger et al. [7] developed a single-scale basis model of single and double end-stopped cells. Würtz and Lourens [22] and Rodrigues and du Buf [12] presented a "multi-scale" approach: detection stabilisation is obtained by averaging keypoint positions over a few neighbouring micro-scales. In [13] we introduced a truly multi-scale analysis: if there are simple and complex cells tuned to different spatial frequencies, spanning an interval of multiple octaves, it can be expected that there are also

end-stopped cells at all frequencies. We analysed the multi-scale keypoint representation, from very fine to very coarse scales, in order to study its importance and possibilities for developing a cortical architecture, with an emphasis on FoA. In addition, we included a new aspect, i.e. the application of non-classical receptive field (NCRF) inhibition to keypoint detection, in order to distinguish object structure from surface textures.

A difficult and still challenging application, even in machine vision, is face detection. Despite the impressive number of methods devised for faces and facial landmarks, which can be based on Gabor filters [18] or Gaussian derivative filters [26], colour [27], attention [19], morphology [9], behaviouristic AI [10], edges and keypoints [20], spiking neurons [1] and saliency maps [23], complicating factors that still remain are pose (frontal vs. profile), beards, moustaches and glasses, facial expression and image conditions (lighting, resolution). Despite these complications, in this paper we will study the multi-scale keypoint representation in the context of a possible cortical architecture. We add that (a) we will not employ the multi-scale line/edge representation that also exists in area V1, in order to emphasise the importance of the information provided by keypoints, and (b) we will not solve complications referred to above, because we will argue, in the Discussion, that low-level processing in area V1 needs to embedded in to a much wider context, including short-time memory, and this context is expected to solve many problems.

In Section 2 we present the models for end-stopped cells and non-classical receptive field inhibition, followed by keypoint detection with NCRF inhibition in Section 3, and the multi-scale keypoint representation with saliency maps in Section 4. In Section 5 we present facial landmark detection, and conclude with a discussion (Section 6).

2 End-stopped cells and NCRF inhibition

Gabor quadrature filters provide a model of cortical simple cells [24]. In the spatial domain (x, y) they consist of a real cosine and an imaginary sine, both with a Gaussian envelope. A receptive field (RF) is denoted by (see e.g. [2]):

$$g_{\lambda,\sigma,\theta,\varphi}(x,y) = \exp\left(-\frac{\tilde{x}^2+\gamma\tilde{y}^2}{2\sigma^2}\right)\cdot\cos(2\pi\frac{\tilde{x}}{\lambda}+\varphi),$$

$$\tilde{x} = x\cos\theta + y\sin\theta \;\;;\;\; \tilde{y} = y\cos\theta - x\sin\theta,$$

where the aspect ratio $\gamma = 0.5$ and σ determines the size of the RF. The spatial frequency is $1/\lambda$, λ being the wavelength. For the bandwidth σ/λ we use 0.56, which yields a half-response width of one octave. The angle θ determines the orientation (we use 8 orientations), and φ the symmetry (0 or $\pi/2$). We apply a linear scaling between $f_{\min}$ and $f_{\max}$ with, at the moment, hundreds of contiguous scales.

Responses of even and odd simple cells, which correspond to real and imaginary parts of a Gabor filter, are obtained by convolving the input image with the

RF, and are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, s being the scale, i the orientation ($\theta_i = i\pi/(N_\theta - 1)$) and N_θ the number of orientations. In order to simplify the notation, and because the same processing is done at all scales, we drop the subscript s. The responses of complex cells are modelled by the modulus

$$C_i(x,y) = [\{R^E_i(x,y)\}^2 + \{R^O_i(x,y)\}^2]^{1/2}.$$

There are two types of end-stopped cells [7, 22], i.e. single (S) and double (D). If $[\cdot]^+$ denotes the suppression of negative values, and $\mathcal{C}_i = \cos\theta_i$ and $\mathcal{S}_i = \sin\theta_i$, then

$$S_i(x,y) = [C_i(x + d\mathcal{S}_i, y - d\mathcal{C}_i) - C_i(x - d\mathcal{S}_i, y + d\mathcal{C}_i)]^+ ;$$

$$D_i(x,y) = \left[C_i(x,y) - \frac{1}{2}C_i(x + 2d\mathcal{S}_i, y - 2d\mathcal{C}_i) - \frac{1}{2}C_i(x - 2d\mathcal{S}_i, y + 2d\mathcal{C}_i)\right]^+ .$$

The distance d is scaled linearly with the filter scale s, i.e. $d = 0.6s$. All end-stopped responses along straight lines and edges need to be suppressed, for which we use tangential (T) and radial (R) inhibition:

$$I^T(x,y) = \sum_{i=0}^{2N_\theta - 1} [-C_{i \bmod N_\theta}(x,y) + C_{i \bmod N_\theta}(x + d\mathcal{C}_i, y + d\mathcal{S}_i)]^+ ;$$

$$I^R(x,y) = \sum_{i=0}^{2N_\theta - 1} \left[C_{i \bmod N_\theta}(x,y) - 4 \cdot C_{(i+N_\theta/2) \bmod N_\theta}(x + \frac{d}{2}\mathcal{C}_i, y + \frac{d}{2}\mathcal{S}_i)\right]^+ ,$$

where $(i + N_\theta/2) \bmod N_\theta \perp i \bmod N_\theta$.

The model of non-classical receptive field (NCRF) inhibition is explained in more detail in [2]. We will use two types: (a) anisotropic, in which only responses obtained for the same preferred RF orientation contribute to the suppression, and (b) isotropic, in which all responses over all orientations equally contribute to the suppression.

The anisotropic NCRF (A-NCRF) model is computed by an inhibition term $t^A_{s,\sigma,i}$ for each orientation i, as a convolution of the complex cell responses C_i with the weighting function w_σ, with $w_\sigma(x,y) = [DoG_\sigma(x,y)]^+ / \|[DoG_\sigma]^+\|_1$, $\|\cdot\|_1$ being the L_1 norm, and

$$DoG_\sigma(x,y) = \frac{1}{2\pi(4\sigma)^2}\exp(-\frac{x^2+y^2}{2(4\sigma)^2}) - \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}).$$

The operator $b^A_{s,\sigma,i}$ corresponds to the inhibition of $C_{s,i}$, i.e. $b^A_{s,\sigma,i} = [C_{s,i} - \alpha t^A_{s,\sigma,i}]^+$, with α controlling the strength of the inhibition.

The isotropic NCRF (I-NCRF) model is obtained by computing the inhibition term $t^I_{s,\sigma}$ which does not dependent on orientation i. For this we construct the maximum response map of the complex cells $\tilde{C}_s = \max\{C_{s,i}\}$, with $i = 0, ...N_\theta - 1$. The isotropic inhibition term $t^I_{s,\sigma}$ is computed by the convolution of the maximum response map $\tilde{C}_s$ with the weighting function w_σ, and the isotropic operator is $b^I_{s,\sigma} = [\tilde{C}_s - \alpha t^I_{s,\sigma}]^+$.

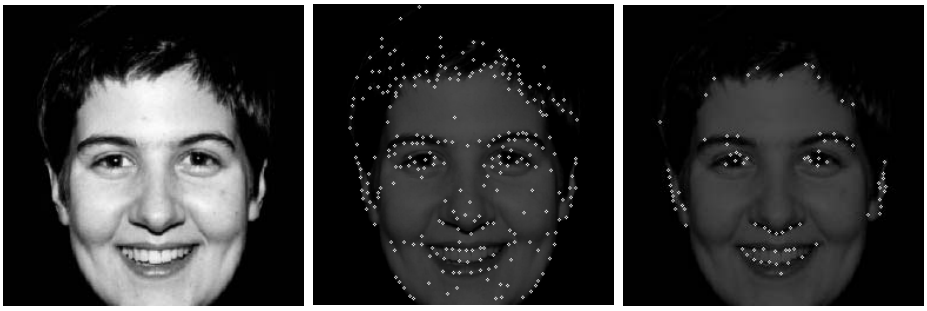

Fig. 1. Centre and right: keypoints without and with NCRF inhibition (face196).

3 Keypoint detection with NCRF inhibition

NCRF inhibition permits to suppress keypoints which are due to texture, i.e. textured parts of an object surface. We experimented with the two types of NCRF inhibition introduced above, but here we only present the best results which were obtained by I-NCRF at the finest scale.

All responses of the end-stopped cells $S(x,y) = \sum_{i=0}^{N_\theta - 1} S_i(x,y)$ and $D(x,y) = \sum_{i=0}^{N_\theta - 1} D_i(x,y)$ are inhibited by $b^I_{s,\sigma}$, i.e. we use $\alpha = 1$, and obtain the responses $\tilde{S}$ and $\tilde{D}$ of S and D that are above a small threshold of $b^I_{s,\sigma}$. Then we apply $I = I^T + I^R$ for obtaining the keypoint maps $K^S(x,y) = \tilde{S}(x,y) - gI(x,y)$ and $K^D(x,y) = \tilde{D}(x,y) - gI(x,y)$, with $g \approx 1.0$, and the final keypoint map $K(x,y) = \max\{K^S(x,y),\, K^D(x,y)\}$.

Figure 1 shows, from left to right, an input image and keypoints detected (single, finest scale), before and after I-NCRF inhibition. After inhibition, only contour-related keypoints remain. Almost all texture keypoints have been suppressed, although some may still remain because of strong local contrast (see [13]).

4 Multiscale keypoint representation

Although NCRF inhibition can be applied at all scales, this will not be done for two reasons: (a) we want to illustrate keypoint behaviour in scale space for the application of FoA, and (b) at coarser scales, i.e. increased RF sizes, most detail (texture) keypoints will be eliminated automatically. In the multi-scale case, keypoints are detected the same way as done above, but now by using $K^S_s(x,y) = S_s(x,y) - gI_s(x,y)$ and $K^D_s(x,y) = D_s(x,y) - gI_s(x,y)$.

An important aspect of a face detection scheme is Focus-of-Attention by means of a saliency map, i.e. the possibility to draw attention to and to inspect, serially or in parallel, the most important parts of faces, objects or scenes. In terms of visual search, this includes overt attention and pop-out. If we assume that retinotopic projection is maintained throughout the visual cortex, the activities of all keypoint cells at the same position (x,y) can be easily summed over scale s, which leads to a very compact, single-layer map. At the positions where

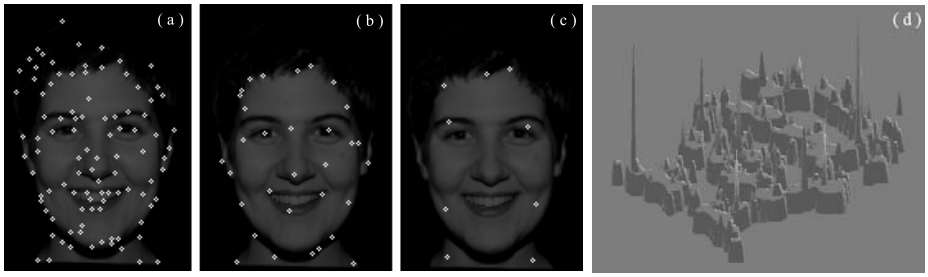

Fig. 2. Keypoints at fine (a), medium (b) and coarse (c) scales, with saliency map (d).

keypoints are stable over many scales, this summation map, which could replace or contribute to a saliency map [4], will show distinct peaks at centres of objects, important sub-structures and contour landmarks. The height of the peaks (summation cell activity) can provide information about the relative importance. In addition, this summation map, with some simple processing of the projected trajectories of unstable keypoints, like a dynamic lowpass filtering related to the scale and non-maximum suppression, might solve the segmentation problem: the object centre is linked to important sub-structures, and these are linked to contour landmarks. This is shown in Fig. 2(d) by means of a 3D perspective projection. Such a mapping or data stream is data-driven and bottom-up, and could be combined with top-down processing from inferior temporal cortex (IT) in order to actively probe the presence of certain objects in the visual field [8]. In addition, the summation map with links between the peaks might be available at higher brain areas where serial processing occurs for e.g. visual search.

In order to illustrate keypoint behaviour in the case of human faces we created an almost continuous, linear, scale space. Figure 2 ("face196"), shows three different scales from scale space: (a) fine scale with $\lambda = 4$, (b) medium scale with $\lambda = 20$, and (c) coarse scale with $\lambda = 40$. At even coarser scales there will remain only a single keypoint more or less in the centre of the face (not shown). Most if not all faces show a distinct keypoint the middle of the line that connects the two eyes, like in Fig. 2(b). Figure 2(d) shows the saliency map of the entire scale space ($\lambda = [4, 40]$) with 288 different scales. Important peaks are found at the eyes, nose and mouth, but also at the hairline and even the chin and neck. For a detailed analysis of keypoint behaviour and stability we refer to [13].

5 Detection of facial landmarks

In Fig. 2(d) we can see the regions where important features are located, but it is quite difficult to see which peaks correspond to important facial landmarks. On the other hand, looking at Fig. 2(b) it is easy to see that some keypoints correspond to landmarks that we pretend to find (in this study limited to eyes, nose and mouth), but (a) there are many more keypoints and (b) at other scales (e.g. Fig. 2(c)) they are located at other structures. Presumably, the visual system

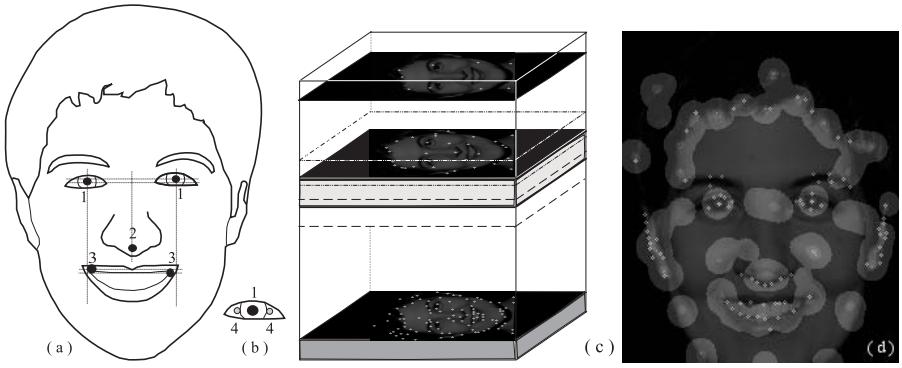

Fig. 3. Left to right: (a) facial landmarks, (b) eye landmarks, (c) impression of keypoint scale space, and (d) saliency map with single-scale keypoints and NCRF inhibition.

uses a "global" saliency map in combination with "partial" ones obtained by summing keypoints over smaller scale intervals, or even keypoints at individual scales, in order to optimise detection. This process can be "steered" by higher brain areas, which may contain prototype object maps with expected patterns (with approximate distances of eyes and nose and mouth), which is part of the fast "where path." The actual "steering" may consist of excitation and inhibition of pre-wired connections in keypoint scale space, i.e. grouping cells that combine end-stopped cells in approximate areas and at certain scales, which is part of the slower "what path."

In our simulations we explored one possible scenario. We assume the existence of very few layers of grouping cells, with dendritic fields in partial saliency maps that map keypoints in specific scale intervals. The top layer with "face" cells groups axons of "eyes" (plural!), "nose" and "mouth" grouping cells. The "eyes" cells group axons of pairs of "eye" cells. Only the "eye," "nose" and "mouth" cells connect to the saliency maps, the "face" and "eyes" cells do not. The scenario consists of detecting possible positions of eyes, linking two eyes, then two eyes plus nose, and two eyes plus nose plus mouth. This is done dynamically by activating synaptic connections in the partial saliency maps.

In our simulations, in which we experimented with faces of different sizes (Fig. 5), we used 7 partial saliency maps, each covering 40 scales distributed over $\Delta\lambda = 5$, but the scale intervals were overlapping 20 scales. The finest scale was at $\lambda = 4$. The search process starts at the coarsest scale interval, because there are much less candidate eye positions than there are at the finest scale interval. A feedback loop will activate connections to finer scale intervals, until at least one eye candidate is detected.

First, "eye" cells respond to significant peaks (non-maximum suppression and thresholding) in the selected saliency map (in the case of "face196" $\lambda = [13, 18]$, see Fig. 4 (left)), as indicated by Fig. 3(b)-1, but only if there are also two stable symmetric keypoints at the 40 finest scales (Fig. 3(b)-4). In order to reduce false positives, the latter is done after NCRF inhibition (Fig. 3(d)). If not a single eye

Fig. 4. Left: the saliency map of face196 ($\lambda = [13, 18]$); Right: result of face196.

cell responds, the scale interval of the saliency map is not appropriate and the feedback loop will step through all saliency maps (Fig. 3(c)), until at least one eye cell responds.

Second, "eyes" cells respond if two "eye" cells are active on an approximately horizontal line (Fig. 3(a)-1), each "eyes" cell being a grouping cell with two dendritic fields. If no eye pair is found, a new saliency map is selected (feedback loop).

Third, when two eyes can be grouped, a "nose" cell is activated, its dendritic field covering an area below the "eyes" cell in the saliency map (Fig. 3(a)-2). If no peak is detected, a new saliency map is selected (feedback loop).

Fourth, if both "eyes" and "nose" cells respond, a "mouth" cell with two dendritic fields at approximate positions of the two mouth corners (Fig. 3(a)-3) is activated. If keypoints are found, a "face" cell will be excitated. If not, a new saliency map is selected (feedback loop).

The process stops when one face has been detected, but in reality it might continue at finer scale intervals (there may be more faces with different sizes in the visual field). However, see the Discussion section. The result obtained in the case of "face196" is shown in Fig. 4, where $+$, $\square$ and $\times$ symbols indicate detected and used keypoints at eyes, nose and mouth corners (actual positions of face and eyes cells are less important). More results are shown in Fig. 5, which includes a correctly detected (!) fake face. Obviously, more features must be used, including the multi-scale line/edge representation.

6 Discussion

As Rensink [21] pointed out, the detailed and rich impression of our visual surround may not be caused by a rich representation in our "visual memory," because the stable, physical surround already "acts" like memory. In addition, focused attention is likely to deal with only one object at a time. His triadic architecture therefore separates focused attention to coherent objects (System II)

from nonattentional scene interpretation (Layout and Gist subsystems in System III), but both Systems are fed by low-level feature detectors, e.g. of edges, in System I.

In this paper we showed that keypoints detected by end-stopped operators, and in particular a few partial keypoint maps that cover overlapping scale intervals, may provide very important information for object detection. Exploring a very simple processing scheme, faces can be detected by grouping together axons of end-stopped cells at *approximate* retinotopic positions, and this leads to robust detection in the case of different facial expressions. However, the simple scheme explored only works if the eyes are open, if the view is frontal, and if the faces are approximately vertical. For pose-, rotation- and occlusion-invariant detection, the scheme must be fed by Rensink's short-term Layout and Gist subsystems, but also the long-term Scene Schema system that is supposed to build and store collections of object representations, for example non-frontal faces.

Owing to the impressive performance of current computers, it is now possible to test Rensink's [21] triadic architecture in terms of e.g. Deco and Rolls' [8] cortical architecture. The ventral WHAT data stream (V1, V2, V4, IT) is supposed to be involved in object recognition, independently of position and scaling. The dorsal WHERE stream (V1, V2, MT, PP) is responsible for maintaining a spatial map of an object's location and/or the spatial relationship of an object's parts as well as moving the spatial allocation of attention. Both data streams are bottom-up and top-down. Apart from input via V1, both streams receive top-down input from a *postulated* short-term memory for shape features or objects in prefrontal cortex area 46, i.e. the more ventral part PF46v generates an object-based attentional component, whereas the more dorsal part PF46d specifies the location. As for now, we do not know *how* PF46 works. It might be the neurophysiological equivalent of the cognitive Scene Schema system mentioned above, but apparently the WHAT and WHERE data streams are necessary for obtaining view-independent object detection through cells with receptive fields of 50 degrees or more [8]. However, instead of receiving input directly from simple cells, the data streams should receive input from feature extraction engines, including end-stopped cells.

Acknowledgments: The images used are from the Psychological Image Collection at Stirling University (http://pics.psych.stir.ac.uk/). Research is partly financed by PRODEP III Medida 5, Action 5.3, and by the FCT program POSI, framework QCA III.

References

1. A. Delorme and S.J. Thorpe. Face identification using one spike per neuron: resistance to image degradations. *Neur. Net.*, 14(6-7):795–804, 2001.
2. C. Grigorescu, N. Petkov and M.A. Westenberg. Contour detection based on nonclassical receptive field inhibition. *IEEE Tr. Im. Proc.*, 12(7):729–739, 2003.
3. C. Rasche. *The making of a neuromorphic visual system.* Springer, 2005.

4. D. Parkhurst, K. Law and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Res.*, 42(1):107–123, 2002.
5. D.H. Hubel. *Eye, brain and vision*. Scientific American Library, 1995.
6. D.J. Fleet, A.D. Jepson and M.R.M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.
7. F. Heitger et al. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Res.*, 32(5):963–981, 1992.
8. G. Deco and E.T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res.*, (44):621–642, 2004.
9. H. Han, T. Kawaguchi and R. Nagata. Eye detection based on grayscale morphology. *Proc. IEEE Conf. Comp., Comm., Cont. Pow. Eng.*, 1:498–502, 2002.
10. J. Huang and Wechsler H. Visual routines for eye location using learning and evolution. *IEEE Trans. Evol. Comp.*, 4(1):73–82, 2000.
11. J. Rodrigues and J.M.H. du Buf. Vision frontend with a new disparity model. *Early Cogn. Vision Workshop, Isle of Skye, Scotland*, 28 May - 1 June 2004.
12. J. Rodrigues and J.M.H. du Buf. Visual cortex frontend: integrating lines, edges, keypoints and disparity. *Proc. Int. Conf. Image Anal. Recogn.*, Springer LNCS 3211(1):664–671, 2004.
13. J. Rodrigues and J.M.H. du Buf. Multi-scale cortical keypoint representation for attention and object detection. *2nd Iberian Conf. on Patt. Recogn. and Image Anal.*, Springer LNCS 3523:255–262, 2005.
14. J.H. Elder and A.J. Sachs. Psychophysical receptive fields of edge detection mechanisms. *Vision Research*, 44:795813, 2004.
15. J.H. van Deemter and J.M.H. du Buf. Simultaneous detection of lines and edges using compound Gabor filters. *Int. J. Patt. Rec. Artif. Intell.*, 14(6):757–777, 1996.
16. J.M. Hupe et al. Feedback connections act on the early part of the responses in monkey visual cortex. *J. Neurophysiol.*, 85(1):134–144, 2001.
17. N. Petkov, T. Lourens and P. Kruizinga. Lateral inhibition in cortical filters. *Proc. Int. Conf. Digital Signal Processing and Int. Conf. Computer Applications Engineering Systems*, Nicosa, Cyprus:122–129, July 14-16 1993.
18. P. Kruizinga and N. Petkov. Person identification based on multiscale matching of cortical images. *Proc. Int. Conf. and Exhib. High-Perf. Comp. Net.*, Springer LNCS 919:420–427, 1994.
19. R. Herpers and G. Sommer. An attentive processing strategy for the analysis of facial features. *Face Recog.: From Theory to Applications, H. Wechsler et al. (eds), NATO ASI Series F, Springer-Verlag*, 163:457–468, 1998.
20. R. Herpers et al. Edge and keypoint detection in facial regions. *Int. Conf. Automatic Face Gest. Recogn.*, pages 212–217, 1996.
21. R. Rensink. The dynamic representation of scenes. *Vis. Cog.*, 7(1-3):17–42, 2000.
22. R.P. Würtz and T. Lourens. Corner detection in color images by multiscale combination of end-stopped cortical cells. *Im. and Vis. Comp.*, 18(6-7):531–541, 2000.
23. S. Ban, J. Skin and M. Lee. Face detection using biologically motivated saliency map model. *Proc. Int. Joint Conf. Neural Netw.*, (1):119–124, 2003.
24. T.S. Lee. Image representation using 2D Gabor wavelets. *IEEE Tr. PAMI*, 18(10):pp. 13, 1996.
25. V. Bruce, P.R. Green and M.A. Geargeson. *Visual Perception Physiology, Psychology and Ecology*. Psychology Press Ltd, 2000.
26. W. Huang and R. Mariani. Face detection and precise eyes location. *Int. Conf. on Patt. Recogn.*, 4:722–727, 2000.
27. Z. Liu, Z. You and Y. Wang. Face detection and facial feature extraction in color image. *Proc. 5th Int. Conf. Comp. Intell. Mult. Appl.*, pages 126– 130, 2003.

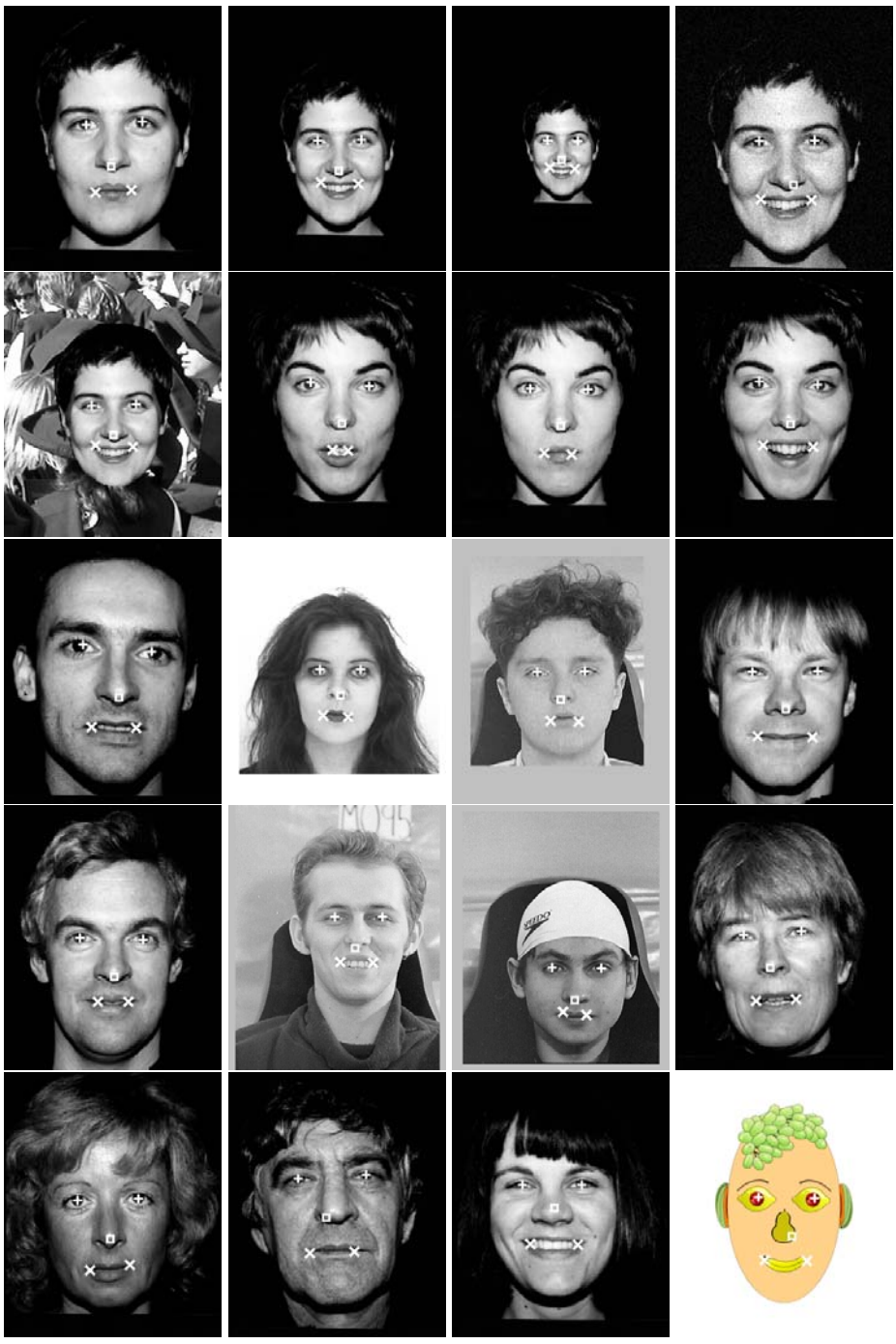

Fig. 5. Results obtained with different faces and expressions.