Multi-scale Cortical Keypoint Representation for Attention and Object Detection

João Rodrigues[1] and Hans du Buf[2]

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

Abstract. Keypoints (junctions) provide important information for focus-of-attention (FoA) and object categorization/recognition. In this paper we analyze the multi-scale keypoint representation, obtained by applying a linear and quasi-continuous scaling to an optimized model of cortical end-stopped cells, in order to study its importance and possibilities for developing a visual, cortical architecture. We show that keypoints, especially those which are stable over larger scale intervals, can provide a hierarchically structured saliency map for FoA and object recognition. In addition, the application of non-classical receptive field inhibition to keypoint detection allows to distinguish contour keypoints from texture (surface) keypoints.

1 Introduction

Models of cells in the visual cortex, i.e. simple, complex and end-stopped, have been developed, e.g. [4]. In addition, several inhibition models [3, 11], keypoint detection [4, 13, 15] and line/edge detection schemes [3, 13, 14], including disparity models [2, 9, 12], have become available. On the basis of these models and processing schemes, it is now possible to create a cortical architecture for figure-background separation [5, 6] and visual attention or focus-of-attention (FoA), bottom-up or top-down [1, 10], and even for object categorization and recognition.

In this paper we will focus on keypoints, for which Heitger et al. [4] developed a single-scale basis model of single and double end-stopped cells. Würtz and Lourens [15] presented a multi-scale approach: spatial stabilization is obtained by averaging keypoint positions over a few neighboring micro-scales. We [13] also applied multi-scale stabilization, but focused on integrating line/edge, keypoint and disparity detection, including the classification of keypoint structure (e.g. T, L, K junctions). Although the approaches in [13, 15] were multi-scale, the aim was stabilization at one (fine) scale. Here we will go into a truly multi-scale analysis: we will analyze 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 will include a new aspect, i.e. the application of non-classical receptive field (NCRF) inhibition [3] to keypoint detection, in order to distinguish between object structure and surface textures.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 255–262, 2005.

2 End-Stopped Models and NCRF Inhibition

Gabor quadrature filters provide a model of cortical simple cells [8]. 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. [3]):

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

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

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.

The responses of even and odd simple cells, which correspond to the real and imaginary parts of a Gabor filter, are obtained by the convolution of 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 and 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 [4, 15], 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 [3]. 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 response 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 as a convolution of the maximum response map $\tilde{C}_s$ with the weighting function w_σ, and the isotropic operator is $b^I_{s,\sigma} = [\tilde{C}_s - \alpha t^I_{s,\sigma}]^+$.

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 in relation to the complex cells (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 then the final keypoint map $K(x,y) = \max\{K^S(x,y), K^D(x,y)\}$.

Figure 1 presents, from left to right, input images and keypoints detected (single scale), before and after I-NCRF inhibition. The top image shows part of a building in Estoril ("Castle"). The middle images show two leaves, and the bottom one is a traffic sign (also showing, to the right, vertex classification with micro-scale stability, see [13]). Most important keypoints have been detected, and after inhibition contour-related ones remain. Almost all texture keypoints have been suppressed, although some remain (Castle image) because of strong local contrast and the difficulty of selecting a good threshold value without eliminating important contour keypoints (see Discussion).

4 Multi-scale Keypoint Representation

Here we focus on the multi-scale representation. Although NCRF inhibition can be applied at each scale, we will not do this for two reasons: (a) we want to

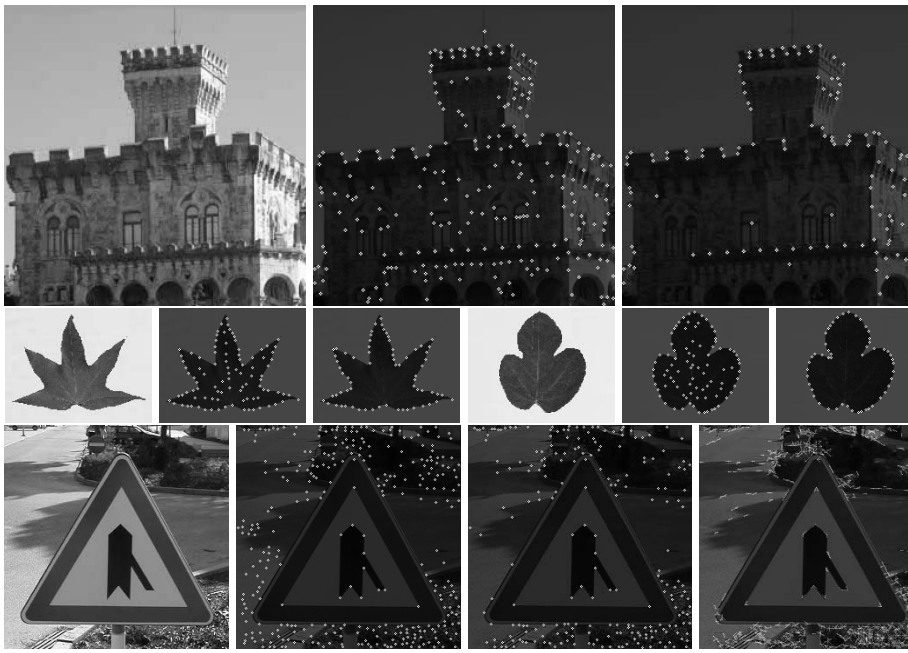

Fig. 1. Keypoints without and with NCRF inhibition; see text.

study keypoint behavior in scale space for applications like FoA, and (b) in many cases a coarser scale, i.e. increased RF size, will automatically eliminate detail (texture) keypoints. 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 $K_s^D(x,y) = D_s(x,y) - gI_s(x,y)$.

For analyzing keypoint stability we create an almost continuous, linear, scale space. In the case of Fig. 2, which shows the (projected) trajectories of detected keypoints over scale in the case of a square and a star, we applied 288 scales with $4 \leq \lambda \leq 40$. Figure 2 illustrates the general behavior: at small scales contour keypoints are detected, at coarser scales their trajectories converge, and at very coarse scales there is only one keypoint left near the center of the object. However, it also can be seen (star object) that there are scale intervals where keypoints are unstable, even scales at which keypoints disappear and other scales at which they appear. (Dis)appearing keypoints are due to the size of the RFs in relation to the structure of the objects, in analogy with Gaussian scale space [7]. Unstable keypoints can be eliminated by (a) requiring stability over a few neighboring micro-scales [13], i.e. keep keypoints that do not change position over 5 scales, the center one and two above and two below (Fig. 2e), or (b) requiring stability over at least N_s neighboring scales (Fig. 2f and 2g with $N_s = 10$ and 40, respectively).

The leftmost five columns in Fig. 3 illustrate that similar results are obtained after blurring, adding noise, rotation and scaling of an object (a leaf), whereas

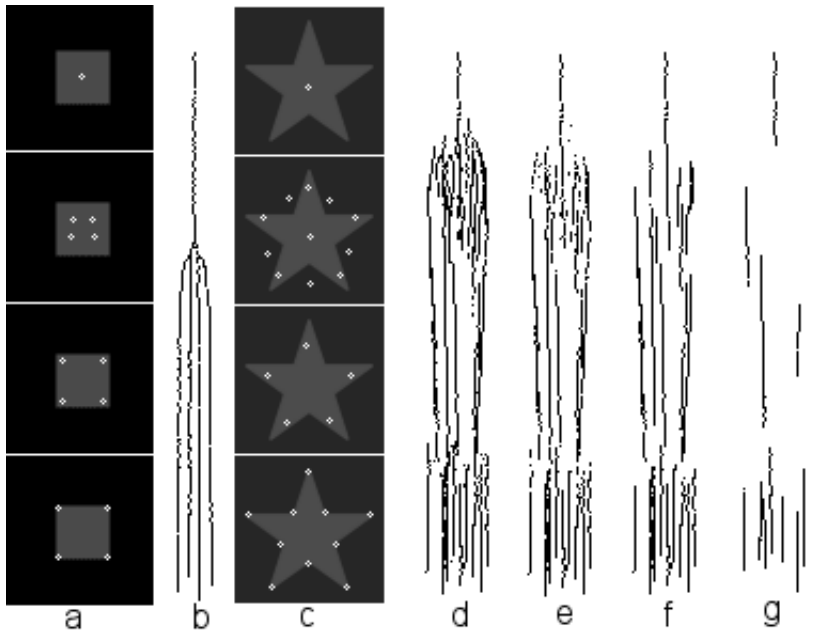

Fig. 2. Keypoint scale space, with finest scale at the bottom. From left to right: (a) square; (b) projected 3D keypoint trajectories of square; (c) and (d) star and projected trajectories; (e) micro-scale stability; (f) and (g) stability over at least 10 and 40 scales respectively.

the last two columns show results for other leave shapes. In all cases, important contour keypoints remain at medium scales and texture keypoints disappear, without applying NCRF inhibition.

With respect to object categorization/recognition, a coarse-to-fine scale strategy appears to be feasible. Figure 4 shows an image with four objects, i.e. two leaves, a star and a van from a traffic sign (see [13]). At very coarse scales the keypoints indicate centers of objects. In the case of the elongated van, an even coarser scale is required. Going from coarse to fine scales, keypoints will indicate more and more detail, until the finest scale at which essential landmarks on contours remain. In reality, the keypoint information shown will be completed by line/edge and disparity (3D) information.

Figure 5 shows that a coarse-to-fine strategy is also feasible in the case of real scenes, i.e. the tower of the Castle image. At coarse scales keypoints indicate the shape of the tower; at finer scales appear structures like the battlements, whereas the corners of the battlements appear at the finest scales. Here we did not apply NCRF inhibition to all scales in order to show that the multi-scale approach selectively "sieves" according to structure detail and contrast.

Another element of an object detection scheme is focus-of-attention by means of a saliency map, i.e. the possibility to inspect, serially or in parallel, the most important parts of objects or scenes. If we assume that retinotopic projection is

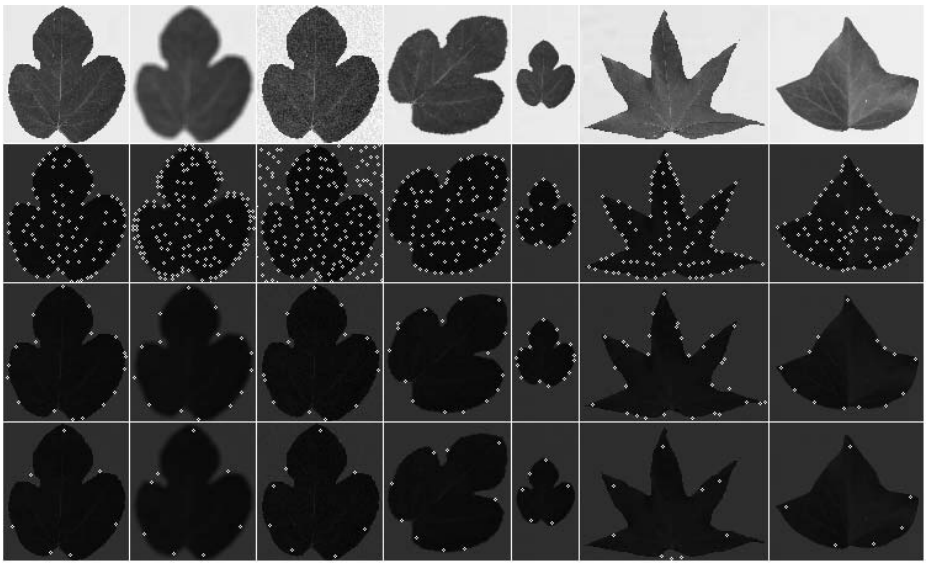

Fig. 3. From left to right: ideal image, blurred, with added noise, rotated and scaled leaf, plus two other leaves. From fine (2nd line) to medium scale (bottom line).

Fig. 4. Object detection from coarse (right) to fine (left) scales ($4 \leq \lambda \leq 50$).

maintained throughout the visual cortex, the activity of keypoint cells at position (x, y) can be easily summed over scale s. At the positions where keypoints are stable over many scales, this summation map, which could replace or contribute to a saliency map [10], will show distinct peaks at centers of objects, important structures and contour landmarks. The height of the peaks can provide information about the relative importance. This is shown in Fig. 6. In addition, this summation map, with some simple processing of the projected trajectories of unstable keypoints, like lowpass filtering and non-maximum suppression, might solve the segmentation problem: the object center is linked to important structures, and these are linked to contour landmarks. Such a 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 objects in the visual field [1]. In addition, the summation map with links between the peaks might be available at higher cortical levels, where serial processing occurs for e.g. visual search.

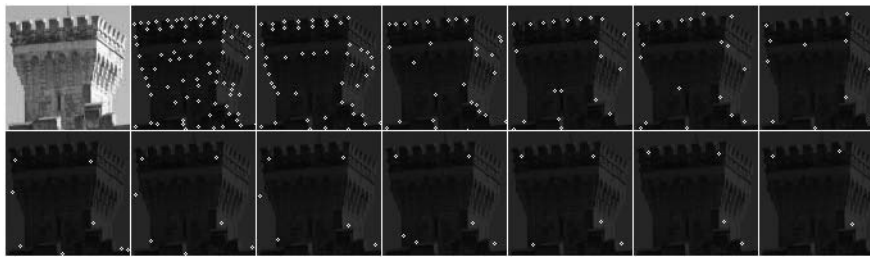

Fig. 5. Keypoint scale-space without NCRF inhibition. From left to right and top to bottom increasing scale ($4 \leq \lambda \leq 50$).

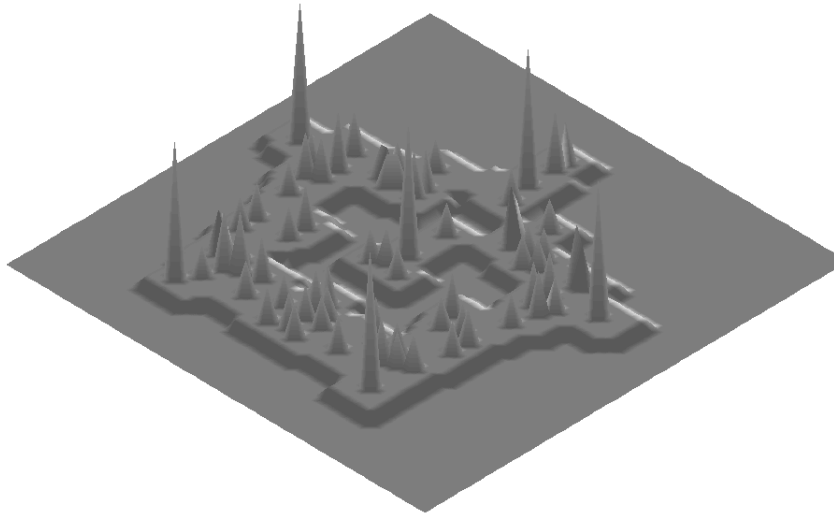

Fig. 6. 3D visualization of the keypoint summation map of the star.

5 Conclusions

The primary visual cortex contains low-level processing "engines" for retinotopic feature extraction. These include multi-scale lines and edges, bars and gratings, disparity and keypoints. Mainly being data-driven, these engines feed higher processing levels, for example for translation, rotation and scale invariant object representations, visual attention and search, until object recognition.

To the best of our knowledge, this is the first study to analyze the importance of multi-scale keypoint representation for e.g. focus-of-attention and object recognition. We showed that the trajectories of keypoints in scale space may be quite complex, but also that keypoints are stable at important structures. In general, at coarse scales keypoints can be expected at centers of objects, at finer scales at important structures, until they cover finest details. We also showed that retinotopic summation of "keypoint-cell activity" over scale provides very useful information for a saliency map (FoA), and even could solve the segmen-

tation problem by bounding objects and linking structures within objects. It seems that the multi-scale keypoint representation, obtained by a linear scaling of cortical end-stopped operators, might be the most important component in building a complete cortical architecture. However, much more information is available through line and edge cells, bar and grating cells, and disparity-tuned cells. In addition, data-driven and bottom-up signals must be used together with top-down or feedback signals coming from higher processing levels.

Finally, it should be mentioned that the hundreds of quasi-continuous scales used here, which is computationally very expensive, can be seen as an abstraction of cortical reality: in reality, there may be an octave or half-octave RF organization, with at each level adaptivity (plasticity) in order to stabilize detection results. Such a scheme, and its application to e.g. FoA, has not yet been explored.

References

1. G. Deco and E.T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res.*, (44):621–642, 2004.
2. D.J. Fleet, A.D. Jepson, and M.R.M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.
3. 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.
4. F. Heitger et al. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Res.*, 32(5):963–981, 1992.
5. J.M. Hupe et al. Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature*, 394(6695):784–787, 1998.
6. 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.
7. J.J. Koenderink. The structure of images. *Biol. Cybern.*, 50(5):363–370, 1984.
8. T.S. Lee. Image representation using 2D Gabor wavelets. *IEEE Tr. PAMI*, 18(10):pp. 13, 1996.
9. I. Ohzawa, G.C. DeAngelis, and R.D. Freeman. Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol.*, 18(77):2879–2909, 1997.
10. D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.
11. N. Petkov, T. Lourens, and P. Kruizinga. Lateral inhibition in cortical filters. *Proc. Int. Conf. Dig. Sig. Proc. and Inter. Conf. on Comp. Appl. to Eng. Sys.*, Nicosa, Cyprus:122–129, July 14-16 1993.
12. J. Rodrigues and J.M.H. du Buf. Vision frontend with a new disparity model. *Early Cognitive Vision Workshop, Isle of Skye, Scotland*, 28 May - 1 June 2004.
13. J. Rodrigues and J.M.H. du Buf. Visual cortex frontend: integrating lines, edges, keypoints and disparity. *Proc. Int. Conf. Image Anal. Recogn.(ICIAR)*, Springer LNCS 3211(1):664–671, 2004.
14. 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.
15. R.P. Würtz and T. Lourens. Corner detection in color images by multiscale combination of end-stopped cortical cells. *Image and vision computing*, 18(6-7):531–541, 2000.