Second-Order (non-Fourier) Attention-Based Face Detection

Albert L. Rothenstein[1], Andrei Zaharescu[2], and John K. Tsotsos[1]

[1] Dept. of Computer Science & Engineering and Centre for Vision Research
York University, Toronto, Canada
{albertlr,tsotsos}@cs.yorku.ca
[2] INRIA Rhone-Alpes, Montbonnot, France
andrei.zaharescu@inrialpes.fr

Abstract. We present an attention-based face detection and localization system. The system is biologically motivated, combining face detection based on second-order circular patterns with the localization capabilities of the Selective Tuning (ST) model of visual attention [1]. One of the characteristics of this system is that the face detectors are relatively insensitive to the scale and location of the face, and thus additional processing needs to be performed to localize the face for recognition. We extend ST's ability to recover spatial information to this object recognition system, and show how this can be used to precisely localize faces in images. The system presented in this paper exhibits temporal characteristics that are qualitatively similar to those of the primate visual system in that detection and categorization is performed early in the processing cycle, while detailed information needed for recognition is only available after additional processing, consistent with experimental data and with certain theories of visual object recognition[2].

1 Introduction

One of the major limitations in current object recognition schemes is the inherent difficulty of extracting reliable and repeatable features from highly textured real-world images. One of the sources of this limitation is the fact that methods rely mainly on linear filtering through various kernels (e.g. for edge detection). The major area in which non-linear feature extraction techniques have been used is perceptual grouping (e.g. [3]), inspired by Gestalt psychology [4, 5]. A continuous source of inspiration for researchers has been the study of the primate visual system, with results used mainly to augment edge detection algorithms [6, 7]. While these results are very promising, they generally limit themselves to simple edge-based perceptual grouping and center-surround competition.

In the current paper we propose a novel approach: non-linear processing targeted at object detection/recognition within a biologically plausible framework. In particular, we address the task of frontal face detection. This paper is organized as follows: In Sect. 2 we briefly describe previous work done in face detection and provide a brief overview of the Selective Tuning model of visual

attention. Section 3 describes our contribution – the algorithm proposed in order to perform face detection and the coupling with visual attention. Section 4 describes the implementation of the system and presents some of the results obtained. The results are discussed in Sect. 5.

2 Background

Detection is generally the first step in a face recognition system, an area that has received significant attention recently, especially for biometrics and security applications (see [8, 9] for recent reviews). The best results seem to come from appearance-based and learning approaches. The work of Turk and Pentland [10] on PCA-based eigenfaces has been very influential not only in face detection and recognition, but also in the more general context of object recognition. Subsequent work [11, 12] improves on the eigenfaces approach, mainly by using learning classifiers and clustering. The most successful recent face detection system, that of Viola and Jones [13, 14], uses AdaBoost learning to build a very rapid "cascade" classifier based on weak classifiers (Harr-like basis functions). The original work on frontal faces has been extended to detect tilted and non-frontal faces by extending the set of basic features and by the introduction of a pose estimator [15]. Variations of the framework that use different basis sets have been presented, e.g. Gabor wavelets [16], and local orientations of gradient and Laplacian based filters [17].

The primate visual system consists of a multi-layer hierarchy with pyramidal abstraction [18], a structure that makes computations tractable, but characterized by a loss of spatial information. As information progresses up the hierarchical structure, neurons represent more and more abstract information, but with less and less spatial accuracy. Due to the nature of the pyramidal structure, a neuron activated at the highest level of the pyramid corresponds to a large subregion in the first layer of the pyramid. In an extreme situation, the top layer can consist of a single neuron that only fires if a face exists in the input image. A mechanism is needed to be able to go down the pyramidal structure and locate the detected item at high spatial resolution, a mechanism provided by the Selective Tuning (ST) model of visual attention [1, 19] – see [20] for a comprehensive review of computational models of visual attention. This is performed in practice via a Winner-Take-All mechanism that will select the most activated region at the highest level. Results are propagated down the pyramidal structure through winner-take-all competitions within the winning receptive fields, until the first layer is reached. Regions that do not contribute to the high level decision are inhibited, thus eliminating distractors and improving the signal-to-noise ratio. This process effectively segments out the detected structure in the input layer, as demonstrated on video sequences (simulated and real) in [19]. A second feed-forward pass through the pyramidal structure will allow only the signals that participate in the detection task to propagate upwards. A second feed-forward pass through the network will provide a much cleaner detection result, since fea-

tures from the input layer that do not participate to the detection task and that would normally propagate up the pyramid are now blocked from the top layer.

3 Face Detection

The face detection system relies on circular pattern detectors based on second-order processing, corresponding to the behaviour of complex, end-stopped cortical neurons [21, 22]. Dobbins demonstrated that end-stopped cells can be used to encode boundary curvature [23], while Koendrink provided a theoretical basis [24]. Second-order filtering has been previously used in computer vision in motion analysis [25], texture boundary extraction [26] and non-Cartesian feature detection [27, 28]. The circular pattern detection is accomplished with the neural network presented in Fig. 1. Each pathway detects end-stopped segments, and these are combined spatially to detect circular patterns [29]. The rectification step makes the system insensitive to the direction of contrast in the input image. The equations describing the filters are presented in the Appendix.

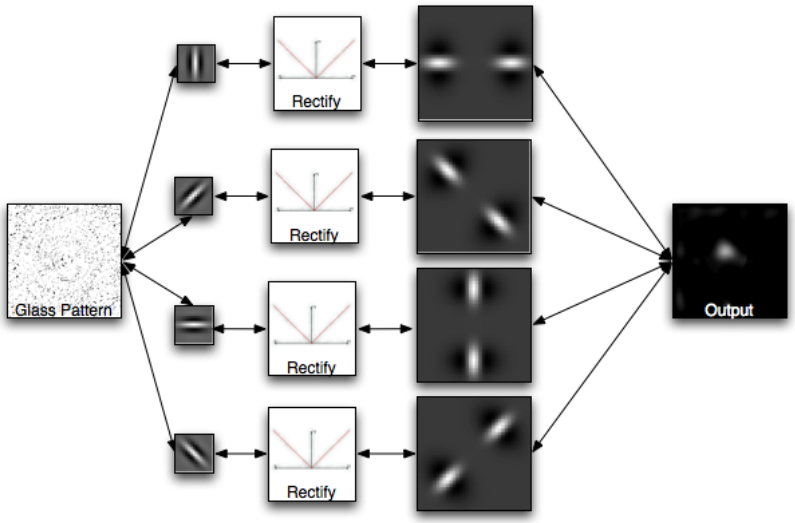


Fig. 1. Diagram of the circular pattern detectors. Only four pathways are represented for clarity. See text for details. Adapted from [29].

The idea behind the detection system presented here is that a face (in frontal view) is characterized by constellations of quasi-circular features at different scales (an idea originally proposed by Wilson [29]. Note that while the solution is biologically-inspired, we are not proposing that the primate visual system detects faces in this manner).

As it can be observed in Figure 2, we model a face by grouping circles at 3 spatial resolutions: small for eyes and nose-tip; medium for the eye-sockets and the mouth region and large for the overall face contour. A second-order circle detector is broadly tuned for a particular circle radius. By combining these circle detectors, we obtain a reliable face detector able to easily overcome changes in illumination, color and facial expressions.

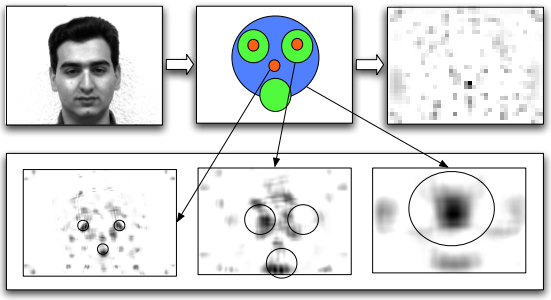


Fig. 2. Diagram illustrating the layout of the face detector based on the circular pattern detector at different three different scales.

So far have presented the face detection system focusing on the overall layout of the system. We have assumed that all the feature planes are of the same dimensions, and we have not worried about the computational cost of each feature and layer. To overcome the performance limitations of this approach, we implemented the system in a pyramidal fashion, coupled with The Selective Tuning (ST) model, which is able to recover the correct location of the detected stimuli. In Figure 3 we show the final layout of current the system. The sizes of the feature planes are also depicted, in order to illustrate overall pyramidal structure. All the connecting arrows between the feature planes are bi-directional, denoting the presence of top-down connections.

4 Implementation and Results

All simulations were implemented in the TarzaNN neural network simulator [30]. The simulator, instructions and all additional files needed to reproduce the results presented here are available online at http://www.TarzaNN.org. The simulations were performed on a Macintosh PowerMac G5. Note that the simulator will also work on other platforms, including Windows, Linux and Solaris. Testing has been performed on images from the Yale face database [31, 32], on composite images derived from the database, and on a group photo. The size and spatial distribution of the circular detectors was tuned manually based on

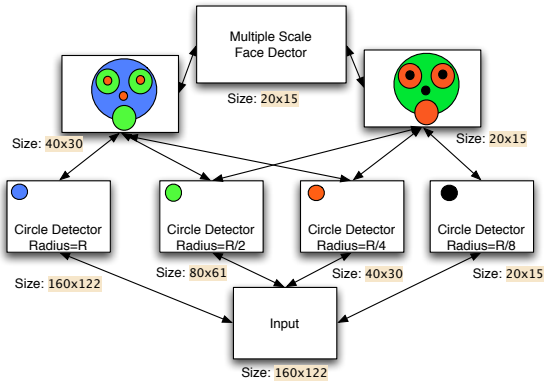


Fig. 3. Diagram illustrating the layout of the face detector based on the circular pattern detector at different scales.

three examples from the Yale database, and we expect results to improve with the inclusion of a learning algorithm. All the results presented are "out of the box" using the system as designed, the only exception being the thresholding of the outputs in some figures, which was tuned manually where indicated. The manual thresholding was performed to enhance the graphical representation of the results, and has no functional role in the system.

Figure 4 illustrates the responses of the system to two composite images, including faces at two scales and other objects. In both cases, responses to faces are significantly stronger than those to other objects, including other circular features such as the wheels and front of the car. The thresholded results show that the system is able to detect and localize the faces.

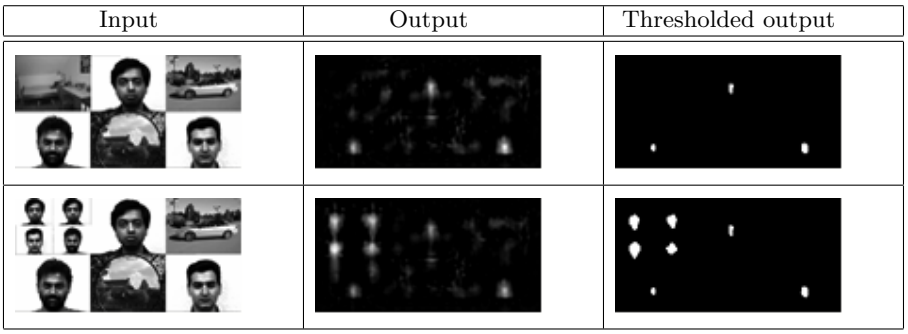


Fig. 4. Responses of the system to two composite images, including faces at two scales and other objects.

Figure 5 is a group photo, with superimposed thresholded system output. Thresholding parameter was adjusted in favour of false-positives, so that all faces are shown as detected. Most false-positives are in the neck and chest areas. In the same Figure we present a couple of representative false-positives. Figure 5(b) illustrates a shirt, where the collars, shirt patterns, and occlusions form patterns that the system classifies as face-like. In Figure 5(a), symmetrical chin shadows and shirt neck line form a pattern that the system responds to. This pattern is caused mainly by the strong vertical lighting from above (the picture was taken in an atrium with glass ceiling).

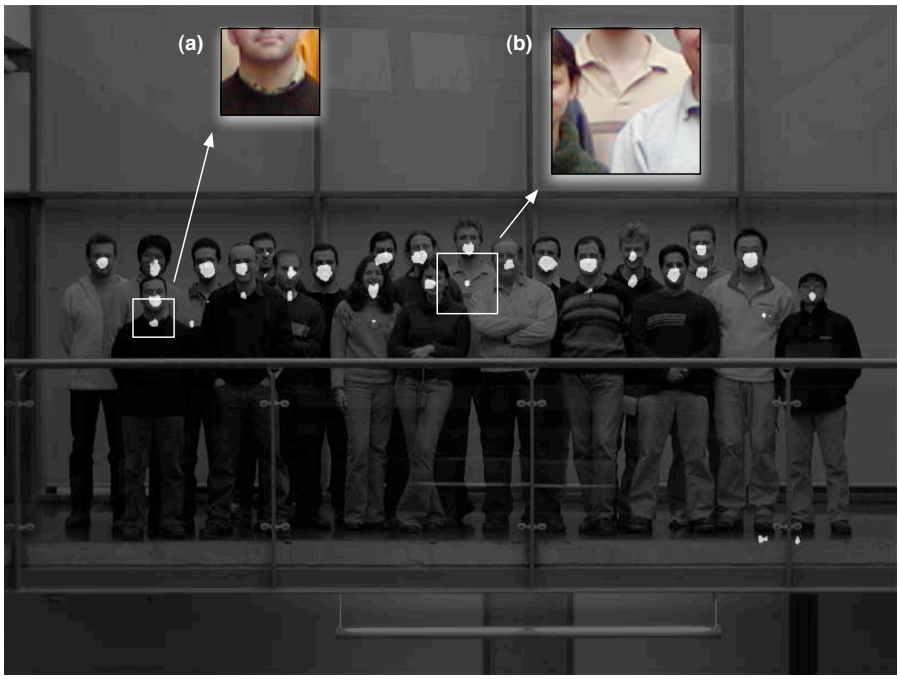
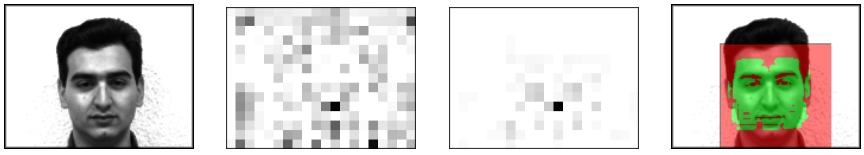


Fig. 5. Group photo with superimposed thresholded system output. Thresholding parameter was adjusted in favour of false-positives, so that all faces are shown as detected. Most false-positives are in the neck area and fall into one of the two categories illustrated above. These errors demonstrate both the flexibilty of the feature extraction process and the frailty of the template matching process. (a) "Face" created by chin shadows and shirt neck line (dominant lighting from above) and (b) "Face"-like shirt.

The inclusion of the attentional mechanism is demonstrated in Fig. 6 and 7. Fig. 6(b) shows the output without attention, note the very noisy output. Fig. 6(c) illustrates the effect of the ST attentional filtering on the face detector

output, with a much clearer peak of activation corresponding to the detected face. Fig. 6(d) presents the localization of the face in the original input image, together with the inhibited region.



(a) (b) (c) (d)

Fig. 6. Effects of attention on the face detection task. (a) Input image. (b) Output of the face detector layer without attention.(c) Output of face detector layer with attention. (d) Location of the face in the input layer using the attentional beam (the outer square represents the inhibited region).

Figure 7 shows the behavior of the system with overlapping faces. In the first fixation, (Fig. 7(b)) the first face is detected and localized. Following this, ST inhibits the connections corresponding to the winning units, and a second pass through the network detects the and localizes the second face (Fig. 7(c)). Note that only the visible part of the second face is selected, since only those pixels contributed to the second detection.

The focus of the current implementation was on demonstrating the principles and feasibility of the method, and little effort has been invested in the performance aspects. In general the computational load of the method is significantly higher than that of other current face detection methods due to the multi-scale convolutions with fairly large kernels, and the ST adds a significant memory load due to the need to have all the intermediate results of the convolution available for the feedback pass.

5 Conclusions

One of the major limitations in current object recognition schemes is the inherent difficulty of extracting reliable and repeatable features from highly textured real-world images. Here we propose the use of second order processing as a solution to this problem, and demonstrate the validity of the approach by applying it to the problem of face detection, with encouraging results. The currently presented detection system is able to correctly detect faces from the Yale Face database [31, 32], under numerous variations (changes in illumination, colour, etc). The main limitation specific to this system is imposed by the ad-hoc nature of the template, generated manually and based on visual inspection. Stronger templates, based on learning and statistical analysis of face images would most likely improve the performance of the system. Implementing the system in a pyramidal, hierarchical fashion has posed the additional problem of recovering the exact location of the

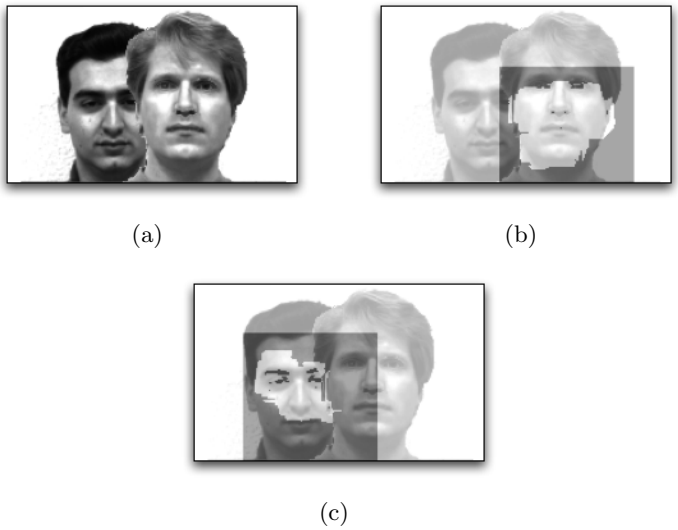


Fig. 7. Input with overlapping faces. (a) Input image. (b) First attentional fixation selects the first face (the outer square represents the inhibited region). (c) Second attentional fixation. Note that the second fixation only selects the visible part of the occluded face.

face in the input images, task accomplished by using the attentional feedback mechanism of Selective Tuning [1]. This is the first demonstration of Selective Tuning in a complex object recognition network.

While the current system does not solve the generic object recognition problem, it provides certain intuitions for how higher level cognition tasks can be performed within a biologically plausible framework. It is important to observe that the temporal structure of the proposed solution (i.e. detection followed by selection and recognition) is consistent with recent psychophysical results [2] that show a temporal lag between the detection and identification of faces. Also, see [33] for a review and discussion of results that highlight the importance of feedback connections and of early visual areas in conscious perception.

6 Appendix

Eq. 1 represents an edge detector composed of a central elongated excitatory lobe flanked by two inhibitory areas. Eq. 2 represents the second stage filters, composed of a similar arrangement of activation lobes, but each filter has two detectors symmetrically shifted by the radius of the circle for which the detector is tuned. Filters for each pathway are rotated by an appropriate amount, and the second stage filters are orthogonal to those in the first stage. See [29] for details on the choice of parameters and on the weighting of the pathways.

$$F(x,y) = \left[Ae^{-\frac{x^2}{\sigma_1^2}} - Be^{-\frac{x^2}{\sigma_2^2}} - Ce^{-\frac{x^2}{\sigma_3^2}} \right] e^{-\frac{y^2}{\sigma_y^2}} \quad (1)$$

$$W(x,y) = \left(Ae^{-\frac{x^2}{\sigma_E^2}} - Be^{-\frac{x^2}{\sigma_I^2}} \right) \left(e^{-\frac{(y-\Delta)^2}{\sigma_y^2}} + e^{-\frac{(y+\Delta)^2}{\sigma_y^2}} \right) \quad (2)$$

Acknowledgements

The authors would like to thank Kosta Derpanis, Sven Dickinson, and Radu Horaud for helpful comments and suggestions.

References

1. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y.H., Davis, N., Nuflo, F.: Modeling visual-attention via selective tuning. Artif. Intell. **78**(1-2) (1995) 507–545
2. Grill-Spector, K., Kanwisher, N.: Visual recognition: as soon as you see it, you know what it is. Psychological Science **16**(2) (2005) 152–160
3. Lowe, D.G.: Perceptual organization and Visual Recognition. Kluwer (1985)
4. Kanizsa, G.: Organization in Vision: Essays on Gestalt Perception. Praeger (1979)
5. Koffka, K.: Principles of Gestalt Psychology. Kegan Paul, London (1936)
6. Zucker, S.W.: Computational and psychophysical experiments in grouping: Early orientation selection. In Beck, J., Hope, B., Rosenfeld, A., eds.: Human and Machine Vision. Academic Press (1983) 545–567
7. Grigorescu, C., Petkov, N., Westenberg, M.A.: Contour and boundary detection improved by surround suppression of texture edges. Image and Vision Computing **22**(8) (2004) 609–622
8. Hjelmåsa, E., Low, B.K.: Face detection: A survey. Computer Vision and Image Understanding **83**(3) (2001) 236–274
9. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(1) (2002) 34–58
10. Turk, M., Pentland, A.: Eigenfaces for recognition. Cognitive Neuroscience **13**(1) (1991) 71–96
11. Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7) (1997) 696–710
12. Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(1) (1998) 39–51
13. Viola, P., Jones, M.: Robust real-time object detection. In: ICCV 2001 Workshop on Statistical and Computation Theories of Vision. (2001)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conf. Computer Vision and Pattern Recognition. Volume 1. (2001) 511–518
15. Jones, M., Viola, P.: Fast multi-view face detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2003)

16. Zhang, L., Li, S.Z., Qu, Z.Y., Huang, X.: Boosting local feature based classifiers for face recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Volume 5., Washington, D.C., USA (2004) 87
17. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In Pajdla, T., Matas, J., eds.: European Conference on Computer Vision. Volume 1., Springer Verlag (2004) 69–82
18. Tsotsos, J.K.: A complexity level analysis of immediate vision. International Journal of Computer Vision **1**(4) (1987) 303–320
19. Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J.C., Pomplun, M., Simine, E., Zhou, K.: Attending to visual motion. Comput. Vis. Image Und. **100**(1-2) (2005) 3–40
20. Rothenstein, A.L., Tsotsos, J.K.: Attention links sensing to recognition. Image and Vision Computing (in press doi:10.1016/j.imavis.2005.08.011) (2006)
21. Hubel, D., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology (160) (1962) 106–154
22. Dreher, B.: Hypercomplex cells in the cat's striate cortex. Invest Ophthalmol. **5**(11) (1972) 355–356
23. Dobbins, A., Zucker, S.W., Cynader, M.S.: Endstopped neurons in the visual cortex as a substrate for calculating curvature. Nature **329** (1987) 438–441
24. Koenderink, J.J., Richards, W.A.: Two-dimensional curvature operators. J. Opt. Soc. Am. A **52** (1988) 1136–1141
25. Fleet, D., Black, M., Jepson, A.: Motion feature detection using steerable flow fields. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR). (1998) 274–281
26. von der Heydt, R., Peterhans, E., Baumgartner, G.: Illusory contours and cortical neuron responses. Science **224** (1984) 1260–1262
27. Gallant, J., Braun, J., Van Essen, D.C.: Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. Science **259** (1993) 100–103
28. Gallant, J.L., Connor, C.E., Rakshit, S., Lewis, J., Van Essen, D.C.: Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey. Journal of Neurophysiology **76** (1996) 2718–2737
29. Wilson, H.R.: Non-Fourier cortical processes in texture, form, and motion perception. In Ulinski, P.S., Jones, E.G., eds.: Cerebral Cortex. Volume 13. Kluwer Academic/ Plenum Publishers, New York (1999)
30. Rothenstein, A.L., Zaharescu, A., Tsotsos, J.K.: Tarzann : A general purpose neural network simulator for visual attention modeling. In Paletta, L., Tsotsos, J.K., Rome, E., Humphreys, G., eds.: Lecture Notes in Computer Science. Volume 3368. Springer Verlag (2005) 159–167
31. Bellhumer, P.N., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(7) (1997) 711–720
32. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6) (2001) 643–660
33. Lamme, V.A.F., Roelfsema, P.R.: The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences **23**(11) (2000) 571–579