ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 85-90

Aspects of psychological computation in scene and face recognition

Philippe G. Schyns

Dept.of Psychology
University of Glasgow
Glasgow, UK
Email: philippe@psy.gla.ac.uk

Abstract. Different classifications of an identical visual stimulus may require different perceptual properties from the visual input. How do processes of object and scene categorization use the information associated with different perceptual spatial scales? One scenario suggests that recognition should use coarse blobs before fine scale edges because scale usage is perceptually determined. However, perceptual determination neglects one important aspect of any recognition task: The information demands of the considered classification of the input. We review evidence suggesting that scale usage could be flexibly determined by the diagnosticity of scale-specific cues for different categorizations of scenes and faces.

1. Introduction

The face picture and the scene of Figure 1 illustrate spatial scales, the perceptual materials that the human visual system might use for recognizing complex stimuli. High Spatial Frequencies (HSF) represent a woman with a neutral expression in the top picture, and New York city in the bottom picture. However, each of these pictures also represent another stimulus. If you squint, blink, or defocus while looking at Figure 1, a smiling man should appear in the top picture, and a highway scene should substitute for the city in the bottom picture (if this demonstration does not work, step back from the picture until your percept changes). Low Spatial Frequencies (LSF) represent the smiling man and the highway. In this paper, we will be concerned with the ways in which the human visual system uses variations of luminance at difference scales. We believe that this should inform the modelling of networks of face, object and scene recognition.

Evidence that perception filters the visual input at multiple scales resulted from many psychophysical studies on contrast detection and frequency-specific adaptation (see de Valois & de Valois, 1990, for an excellent review of spatial vision). Their conclusion was that the visual system comprises groups of independent, quasi-linear band-pass filters, each of which is narrowly tuned to particular frequency bands. Although recent psychophysical research showed that SF channels were interactive and nonlinear, it still remains that spatial filtering is prior to many early visual tasks such as motion, stereopsis, edge detection, depth perception and saccade programming.

The *raison d'être* of these spatial scales was probably most clearly argued in computational vision: Multi-scale processing is necessary to organize and simplify the description of visual events (e.g., Burt & Adelson, 1983; Marr & Hildreth, 1980; Watt, 1991; Witkin, 1986). For example, fine scale boundary edges (which tend to correspond to the precise contours of objects) are notoriously noisy, and they represent confusing details that would be absent in edges measured at a cruder spatial

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 85-90

resolution. However, details are often required for precise object classifications--for example for distinguishing two objects that look alike.

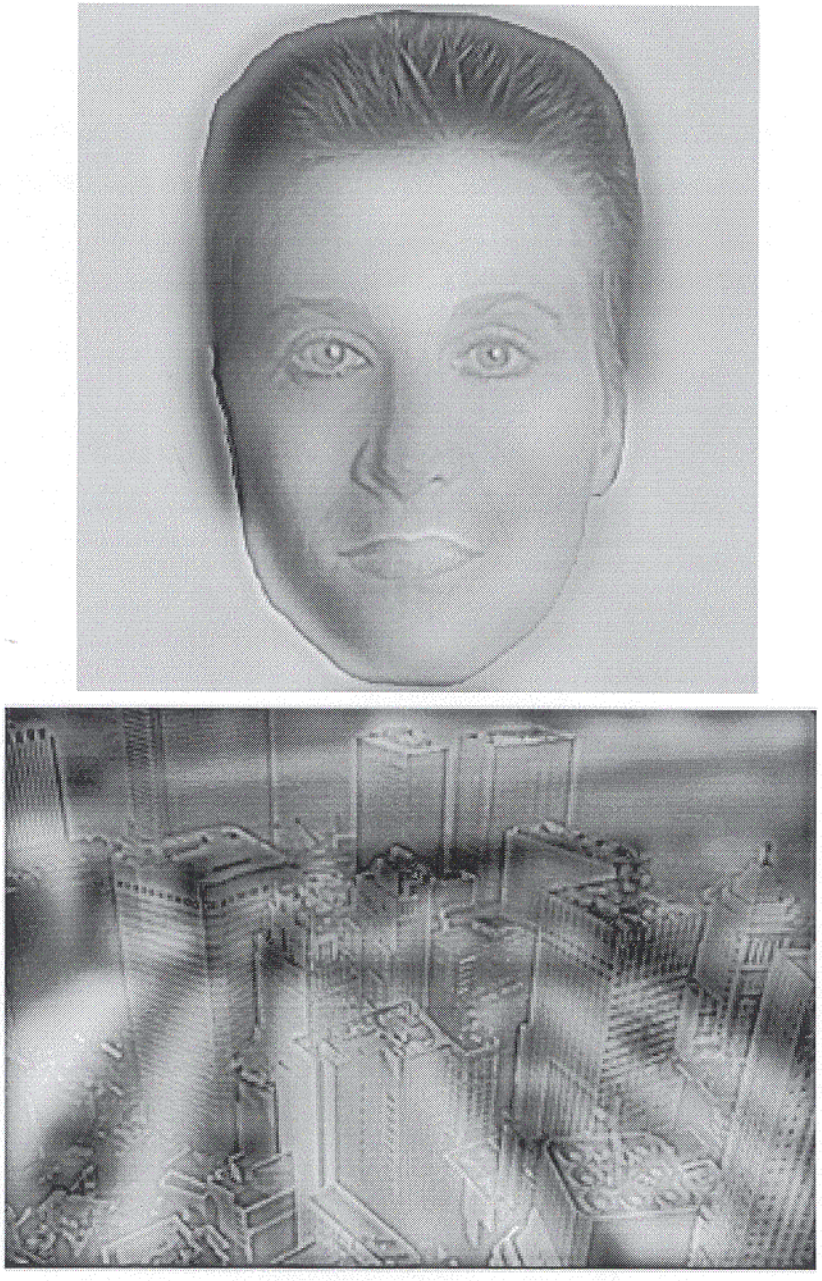

Fig. 1. Example of a hybrid face and a hybrid scene. Fine information reveals a nonexpressive female face and the scene is a view of New York. However, coarse scale information reveal s another interpretation of the stimuli. The face is a smiling male, and the scene is a highway.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 85-90

Hence, boundary edges that would correspond across resolutions could form a skeleton describing a coarse object structure which would later be fleshed out by fine scale edges. *Coarse-to-fine processing* suggests that it is computationally more efficient to first extract a coarse image description before extracting detailed, but noisier, information.

2 Coarse-to-fine usage of spatial scales for recognition.

Few studies exist that have studied scale-specific recognition of real-world stimuli (though see Costen, Parker & Craw, 1994; Parker, Lishman & Hughes, 1992; Schyns & Oliva, 1994). Their results were generally compatible with a *coarse-to-fine* usage of scale information. It is generally argued that recognition should use coarse blobs before fine edges because this is the order in which the scales are perceived.

However, such a fixed scenario neglects one important aspect of any recognition task: The information demands of the considered recognition. For example, the face of Figure 1 can be recognized as a face, as a woman, as a young face or as a non-expressive face. Similarly the picture of the scene, can be classified as an outdoor scene, as a city, or as New York. Flexible recognition allows people to place an identical visual input into one category or another. However, the cues that enable these distinct recognitions could reside at different spatial scales. If this was the case, one might argue that scale usage could be partially determined by the task at hand. Although this would assume that top-down influence can affect lower-level perception, recent studies showed that such influence indeed existed (e.g., Schyns, Goldstone & Thibaut, in press). In this paper, we argue that different sorts of recognitions of identical pictures can change scale usage, and we will discuss the implications of flexible scale usage on models of recognition from spatial scales.

From an experimental viewpoint, typical stimuli do not separate spatial scales, and so one would not know which scale was utilized for which categorization. However, the hybrid stimuli presented on Figure 1 multiplex meaningful information in scale space and thereby permit the study of scale-dependent recognition. Schyns and Oliva (1994) showed that such hybrid scenes tended to be recognized in a coarse-to-fine sequence. These early results were compatible with other studies of scale-based recognition. However, neither these, nor our experiments did, in fact, tested different recognitions of these pictures. Consequently, they could not distinguish between a mandatory, perceptually-determined coarse-to-fine recognition scheme, and a flexible scale usage.

3 Evidence for a flexible usage of spatial scales

Of course, the idea that spatial scales can be flexibly use would be of little interest if it was shown that scales are always perceived from coarse to fine, or if the fine scale was not available early enough. In their first experiment, Oliva and Schyns (1996) tested this issue of the availability of all spatial scales. One hybrid picture was presented and it was tested whether it could successfully facilitate (prime) the recognition of two scenes--the LSF and the HSF scenes that simultaneously compose hybrids. Hybrids were presented for 30 ms, immediately followed by a full spectrum noise mask, which was followed by a normal scene picture. The normal scene picture could either be related to the HSF or the LSF of the hybrid, or be unrelated. Subjects were instructed to categorize the normal scene picture as soon as they possibly could, and reaction times were recorded. Results revealed that a single hybrid picture could

facilitate the recognition of two stimuli in such very brief, masked conditions of presentation. Consequently, Oliva and Schyns (1996) concluded that the LSF component of an hybrid did not interfere with the perceptual registration of the HSF component, and that both were available at the onset of visual processing.

In their second experiment, Oliva and Schyns (1995) aimed for an "existence proof" that flexible, diagnosticity-driven scale usage was possible in visual cognition. Their strategy was to assign diagnosticity to the information associated with one spatial scale, and to observe subsequent scale selection strategies for recognition. Their experiment was a two-phase design, without discontinuity between the phases. In the sensitization phase, two groups of subjects were exposed to hybrids which were only meaningful at one scale; the other scale was "structured noise". The LSF group was initially exposed to hybrids meaningful in LSF and meaningless in HSF. The HSF group initially saw the opposite hybrids--i.e., meaningless in LSF and meaningful in HSF. Subjects saw several of these stimuli, one at a time, and their task was to categorize them. As only one scale of the hybrids was diagnostic, we expected categorization to become tuned to information at this scale.

In the testing phase, hybrids were presented that were meaningful at both scales. Each hybrid was presented in a brief, three-frame animation (45 ms per frame, for a total presentation time of 135 ms). In the animations, the cut-off points of LSF and HSF were gradually changed. In Frame 1, LSF encoded SF below 2 cycles/deg. of visual angle, and HSF represented SF above 6 cycles/deg of visual angle. These thresholds were respectively changed to 3 and 5 in Frame 2, and 4 and 4 in Frame 3. Together, the three frames presented a *coarse-to-fine* and a *fine-to-coarse* animations in scale space. Studies that directly fed visual cognition with animations all reported a coarse-to-fine bias (Parker et al., 1992; Schyns & Oliva, 1994). Furthermore, testing showed that this technique produced a motion in scale space which "locked" attention to the scale that was first selected. Henceforth, when we refer to hybrids, we mean these animations.

Categorizations of the test hybrids revealed that subjects maintained their categorizations at the scale congruent with their sensitization phase. That is, identical hybrids were orthogonally categorized. In summary, it appears that the constraint of locating scale-specific diagnostic information can drive scale selection for scene recognition. When categorization processes use the information content of a particular scale, the unattended scale is nonetheless perceptually registered and it facilitates categorization across scale, and across trials. Consequently, these results refute the idea that a low-level perceptual bias determines scale usage for recognition. Instead, there is evidence that a mandatory perception of multiple spatial scales promotes flexible scene encodings, perceptions and categorizations.

4 Evidence that task-specific information exists at different scales.

It should be noted that the experiments of Oliva and Schyns (1996a) only showed that spatial scales can be flexibly used in recognition tasks. However, these experiments did not test the *raison d'être* of flexible perception--namely, that cues at different scales can effectively distinguish between different object categories. Evidence for a flexible usage of scales does not necessarily imply that they represent different object categorizations.

Schyns and Oliva (1996) explicitly tested that scale-specific cues were associated with different categorizations of identical faces. The idea was that judging the gender vs. the expression of identical face hybrids (see Figure 1) might change the scale at

which these are preferentially apprehended. Each hybrid used in the experiment combined a man and a woman, and each one of the component faces could either be expressive (happy, angry) or not (neutral). Two different categorization tasks was applied to the same set of hybrid pictures: male vs. female and expressive vs. non-expressive. Each stimulus was presented for 50 ms on the screen, and subjects were instructed either to categorize the stimulus into male vs. female, or into expressive vs. nonexpressive. Categorization responses were used to trace the scale preference for gender and expression categorizations.

No bias was found for a particular scale for gender categorizations. That is, the percentage of LSF and HSF categorizations was precisely 50%. However, judgments of the same stimuli according to whether or not they were expressive induced a HSF bias (HSF = 65%). This contrast showed that different categorizations of identical stimuli can change the scale information that is used.

In another experiment, subjects were also asked to perform these two categorization tasks, but only after learning the identity of the faces--i.e., subjects learning an arbitrary mapping between names and faces. There is evidence that identification is mandatorily processed in gender judgments. Hence, if identity judgments required specific scale cues, we might expect this bias to be carried over to the judgments of gender--which was not biased in the other experiment. Results showed that this was indeed the case. Biases for the different categorizations were in this experiment HSF for gender (HSF = 67%) and HSF for expression (HSF = 71%).

In sum, results of this experiment with hybrid faces revealed that task-specific information could reside at different scales. Thus, the flexible scale usage that was reported for scenes is here grounded on the presence of visual cues at different scales for different categorizations. The visual system can flexibly adjust to pick information at the scale that is most informative (i.e., diagnostic) of the categorization task.

5 Consequences for network modelling

In this paper, we framed the usage of spatial scales in terms of diagnostic recognition (Schyns, 1996), a framework which explains recognition performance as an interaction between the information demands of different categorizations and the availability of perceptual information. We showed that information demands could drive the selection of different spatial scales for recognition. This view is in marked contrast with the recognition literature which has often assumed that scale selection was perceptually determined by an early bias for the coarse scale in low-level perception. However, the evidence reviewed in this paper suggests that flexible, diagnosticity-driven, rather than fixed, perceptually determined, scale selection is a better explanation of recognition performance.

Recognition performance is the modelling target of artificial recognition systems. It is a goal of many neural network models to come as close as possible to human recognition performance in face, object and scene recognition tasks. In other words, their constructed representations should enable the same flexibility as the human visual system. However, there is an inherent problem with constructing representations from high-dimensional space data because these are virtually empty. If the input distribution varies along many degrees of freedom, a learning problem in high-dimensional space may require an unrealistically large training set to discover robust representations, even if an asymptotic solution exists in principle. Thus, any network that will recognize real-world images will need to be properly constrained.

We believe that most existing network models have neglected an important source of constraints: the multiplicity of recognition tasks.

The reviewed evidence suggests two main sources of constraints: perceptual spatial scales, and recognition tasks. Many modellers are aware they they should sufficiently engineer their input space so as to reduce its dimensionality, and "fill up" the space with sufficiently many data to derive robust estimators. The is the problem of measurement of the input. However, fewer modellers are aware that the multiplicity of recognition tasks might also offer another, powerful source of constraints on the discovery of relevant, low-dimensional representations.

Generally speaking, in dimensionality reduction techniques, the feature extraction stage is independent of the higher-level recognition that takes place, and thus there is no garantee that the extracted features will be at all useful for the considered recognition. The psychological evidence on flexible scale usage suggests that the different categorizations of objects should constrain the search for a relevant lower-dimensional space to reproject the data. Thus, the serial process of (1) project high-dimensional data on a new lower-dimensional space, then (2) determine categorization with new dimensions, will have to be modified so that the second process informs the first (Schyns et al., in press).

References

Burt, P., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31, 532-540.

Costen, N. P, Parker, D. M., & Craw, I. (1994). Spatial content and spatial quantization effects in face recognition. *Perception*, 23, 129-146.

De Valois, R. L., & De Valois, K. K. (1990). *Spatial Vision*. Oxford University Press: New York.

Marr, D., & Hildreth, E. C. (1980). Theory of edge detection. *Proceedings of the Royal Society of London, Series B*, 207, 187-217.

Oliva, A., & Schyns, P. G. (1996). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Submitted for publication.

Parker, D.M., Lishman, J.R., & Hughes, J. (1992). Temporal integration of spatially filtered visual images. *Perception*, 21, 147-160.

Schyns, P.G. (1996). Diagnostic recognition: Task constraints, object information and their interactions. Submitted for publication.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (in press). The development of features in object concepts. *Brain and Behavioral Sciences*.

Schyns, P.G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychological Science*, 5, 195-200.

Schyns, P. G., & Oliva, A. (1996). Coarse faces and fine smiles. Submitted for publication.

Watt, R. J. (1991). *Understanding vision*. Academic Press, London.

Witkin, A. (1986). Scale-space filtering. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1019-1022. Los Altos, CA: Morgan Kauffman.