Understanding Visual Categorization from the Use of Information

Lizann Bonnar (LIZANN@PSY.GLA.AC.UK)
Philippe G. Schyns (PHILIPPE@PSY.GLA.AC.UK)
Frédéric Gosselin (GOSSELIF@PSY.GLA.AC.UK)
Department of Psychology, University of Glasgow, 58 Hillhead Street,
Glasgow, Scotland, G12 8QB

Abstract

We propose an approach that allows a rigorous understanding of the visual categorization and recognition process without asking direct questions about unobservable memory representations. Our approach builds on the selective use of information and a new method (Gosselin & Schyns, 2000, *Bubbles*) to depict and measure what this information is. We examine three face recognition tasks (identity, gender, expressive or not) and establish the information responsible for recognition performance. We compare the human use of information to ideal observers confronted to similar tasks. We finally derive a gradient of probability for the allocation of attention to the different regions of the face.

Introduction

In recent years, most face, object and scene recognition researchers have gathered around a common agenda: to understand the structure of representations in memory. A number of fundamental issues have been articulated, and researchers typically ask questions such as: "Are face, object and scene representations viewpoint-dependent? " (Hill, Schyns & Akamatsu, 1997; Perrett, Oram & Ashbridge, 1998; Troje & Bülthoff, 1996; Tarr & Pinker, 1989; Bülthoff & Edelman, 1992; Simons & Wang, 1998, among many others); "Are these representations holistic (e.g, view-based, Poggio & Edelman, 1990; Tarr & Pinker, 1991; Ullman, 1998), or made of smaller components? (e.g,, geons, Biederman, 1987; Biederman & Cooper, 1991)"; "Are internal representations complete (e.g., Cutzu & Edelman, 1996), or sparse? (Archambault, O'Donnell & Schyns, 1999; Rensink, O'Regan & Clark, 1997), "two- or three-dimensional?" (Liu, Knill & Kersten, 1995), "colored or not?" (Biederman & Ju, 1988; Oliva & Schyns, 2000; Tanaka & Presnell, 1999), "Are they hierarchically organized in memory?" (Jolicoeur, Gluck & Kosslyn, 1984; Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976), "Is there a fixed entry point into the hierarchy?" (Gosselin & Schyns, in press; Tanaka & Taylor, 1991) "Does expertise modify memory representations?" (Biederman & Shiffrar, 1987; Tanaka & Gauthier, 1998; Schyns & Rodet, 1997) and the entry point to recognition?" (Tanaka & Taylor, 1991); "What is the format of memory representations, and does it change uniformly across the levels of a hierarchy?" (Biederman & Gerhardstein, 1995; Jolicoeur, 1990; Tarr & Bülthoff, 1995).

To address these complex issues, recognition researchers should be equipped with methodologies of a commensurate power; methodologies that can assign the credit of behavioral performance (e.g., viewpoint-dependence, configural effects, color, speed of categorization, point of entry, expertise and so forth) to specific properties of the representations of visual events in memory. However, the relationship between behavior and representations is tenuous, making representational issues the most difficult to approach experimentally.

In this paper, we propose an alternative approach that allows a rigorous understanding of the recognition process, without directly asking questions about unobservable memory representations. Our analysis builds on the *selective use of diagnostic information*, an important but neglected stage of the recognition process. To recognize an object, people selectively use information from its projection on the retina. This information is not available to conscious experience, but the visual system knows what it is, and how to extract it from the visual array. Our approach interrogates the visual system to determine and to depict the information the system uses to recognize stimuli.

The aim of this paper is twofold. At an empirical level, we will use Gosselin and Schyns (2000) *Bubbles* technique to visualize the information used in three face categorization tasks (identity, gender and expressive or not). Faces are a good stimulus for our demonstrations: their compactness enables a tight control of presentation which limits the spatial extent of useful cues; the familiarity of their categorizations simplifies the experimental procedure which does not require prior learning of multiple categories--most people are "natural" face experts (Bruce, 1994). However, the principles developed with faces also apply to the more general cases of objects and scenes.

At a more theoretical level, we will reveal the information used in recognition tasks without asking questions (or even making assumptions) about memory representations. This is nonetheless a powerful approach because the information used encompasses all the visual features that mediate the recognition task at

hand. These features therefore reflect the information required from memory to recognize the stimulus; their extraction from the visual array specifies the job of low-level vision. Shortly put, the features involved in a recognition task bridge between memory and the visual array. Now, show me the features!

Experiment

This experiment was cast as a standard face categorization and recognition experiment. In a between-subjects design, a different subject group resolved one of three possible categorizations (identity, gender, expressive or not) on the same set of ten faces (5 males, 5 females), each displaying two possible expressions (neutral vs. happy). Prior to the experiment, all subjects learned the identity of the ten faces, in order to normalize exposure to the stimuli.

To determine the specific use of face information in each task, we applied Gosselin and Schyns' (2000) *Bubbles* technique. *Bubbles* samples an input space to present as stimuli sparse versions of the faces. Subjects categorize the sparse stimuli and *Bubbles* keeps track of the samples of information that lead to correct and incorrect categorization responses. From this information, we can derive the usage of each region of the input space for the categorization task at hand (see Figure 1). In a nutshell, *Bubbles* performs an exhaustive search in a specified image generation space (here, the image plane x spatial scales), using human recognition responses to determine the diagnostic information.

Methods

Participants.

Participants were forty-five paid University of Glasgow students, with normal, or corrected to normal vision. Each participant was randomly assigned to one of three possible experimental groups (IDENTITY; male vs. female, GENDER; expressive or not, EXNEX) with the constraint that the number of participants be equal in each group.

Stimuli.

All experiments reported in this paper ran on a Macintosh G4 using a program written with the Psychophysics Toolbox for Matlab (Brainard, 1997; Pelli, 1997). Stimuli were computed from the greyscale faces of Schyns and Oliva (1999) (5 males, 5 females each of whom displayed two different expressions, neutral and happy, with normalized hairstyle, global orientation and lighting).

To search for diagnostic information, we used Gosselin and Schyns' (2000) *Bubbles* technique applied to an image generation space composed of three dimensions (the standard X and Y axes of the image plane, plus a third Z axis representing 6 bands of spatial frequencies of one octave each). Figure 1 illustrates the stimulus generation process.

To compute each stimulus, we first decomposed an original face into 6 bands of spatial frequencies of one octave each—at 2.81, 5.62, 11.25, 22.5, 45 and 90 cycles per face, from coarse to fine, respectively (computations were made with the Matlab Pyramid Toolbox, Simoncelli, 1997). The coarsest band served as a constant background, as a prior study revealed that it does not contain face identification information.

The face represented at each band was then partly revealed by a mid-grey mask punctured by a number of randomly located Gaussian windows (henceforth called "bubbles"). The size of the Gaussian differed for each frequency band, to normalize to 3 the number of cycles per face that any bubble could reveal (standard deviations of bubbles were 2.15, 1.08, .54, .27, and .13 cycles/deg of visual angle, from coarse to fine scales). Since the size of the bubbles decreases from coarse to fine scales, we multiplied the number of bubbles at each scale to normalize the average area of the face revealed at each scale.

To generate an experimental stimulus, we simply added the information revealed at each scale. The total subspace revealed by the bubbles (and therefore the number of bubbles per scale) was adjusted to maintain categorization of the sparse faces at a 75% correct criterion.

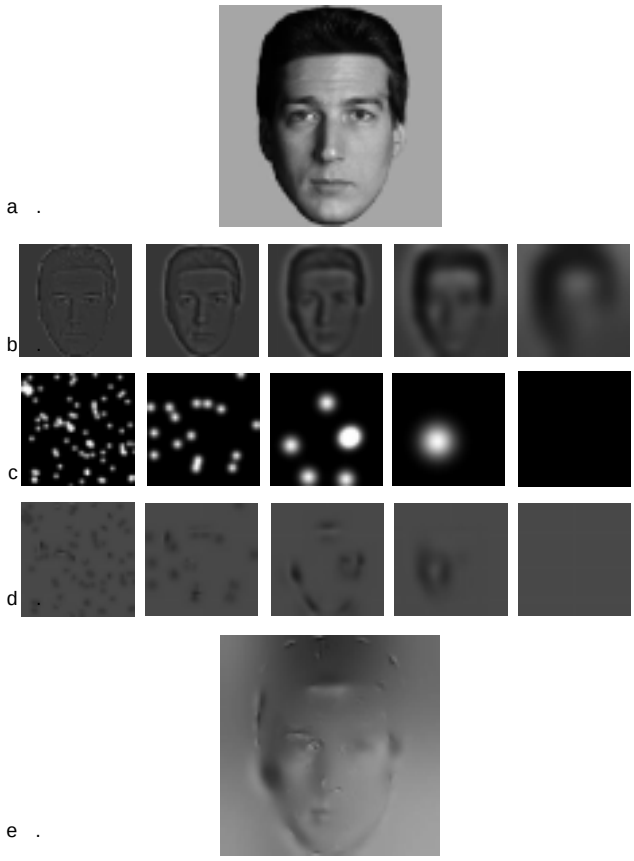

Figure 1 illustrates the application of Bubbles to the 3D space composed of a 2D face in Experiment 2. Pictures in (b) represent five different scales of (a); (c) illustrate the bubbles applied to each scale; (d) are the revealed information of (b) by the bubbles of (c). Note that on this trial there is no revealed information at the fifth scale. By integrating the pictures in (d) we obtain (e), a stimulus subjects actually saw.

Procedure

Prior to experimentation, all participants learned to criterion (perfect identification of all faces twice in a row) the gender, expression and the name attached to each face from printed pictures with corresponding name at the bottom. Each participant was then randomly assigned to one of the three different categorization tasks. In IDENTITY, participants had to determine the identity of each face stimulus. In the GENDER task, participants were instructed to decide whether the stimulus was male or female. In EXNEX, participants had to judge whether the face was expressive or not. Thus, each group performed different categorizations on the same stimulus set.

In a trial, one sparse face computed as just described appeared on the screen. To respond, participants pressed labelled computer-keyboard keys. No feedback was provided. The experiment comprised two sessions of 500 trials (25 presentations of the 20 faces), but we only used the data of the last 500 trials, when subjects were really familiar with the faces and experimental procedure. A chinrest was used to maintain subjects at a constant viewing distance (of 100 cm). Stimuli subtended 5.72 x 5.72 deg of visual angle on the screen.

Results

On average, a total of 33, 20 and 15 bubbles were needed for subjects to reach the 75% performance criterion in the identity, gender and expressive or not task, respectively. Remember that these bubbles resided at different scales of the same stimulus, and were randomly distributed within each scale. Thus, *Bubbles* performs a random search of the input space that is exhaustive after many trials.

Following Gosselin and Schyns' (2000) methodology, we used subjects responses to determine which stimulus information was, and was not diagnostic. The correct categorization of one sparse stimulus indicates that the information revealed in the bubbles was sufficient for its categorization. When this happened, we added the mask of bubbles to a CorrectPlane, for each scale—henceforth, CorrectPlane(scale), for scale = 1 to 5. We also added these masks to a TotalPlane(scale), for each scale. Across trials, TotalPlane(scale) represents the addition of all masks leading to a correct categorization *and* a miscategorization.

From CorrectPlane(scale) and TotalPlane(scale), we can compute for each subject the diagnositicity of each region of the input space with ProportionPlane(scale) = CorrectPlane(scale)/TotalPlane(scale). For each scale, the ProportionPlane(scale) is the ratio of the number of times a specific region of the input space has led to a successful categorization over the number of times this region has been presented. Across subjects, the averaged ProportionPlane(scale) weighs the importance of the regions of each scale for the categorization task at hand (Gosselin & Schyns, 2000). If all regions had equal diagnosticity, ProportionPlane(scale) would be uniformly grey. That is, the probability that any randomly chosen bubble of information led to a correct categorization of the input would be equal to the performance criterion—here, .75. By the same reasoning, whiter regions are significantly above the performance criterion, and therefore more diagnostic of these tasks.

To compute the significance of diagnostic regions, a confidence interval is built around the mean of the ProportionPlane(scale), for each proportion ($p < .01$). To depict the complex interaction between categorization tasks, spatial scales and use of information, we can visualize the *effective stimulus* of each task (see Figure 2). The effective stimulus is a concrete image of the information the visual system uses in each task. It is obtained by multiplying the face information in Figure 2 with the diagnostic masks.

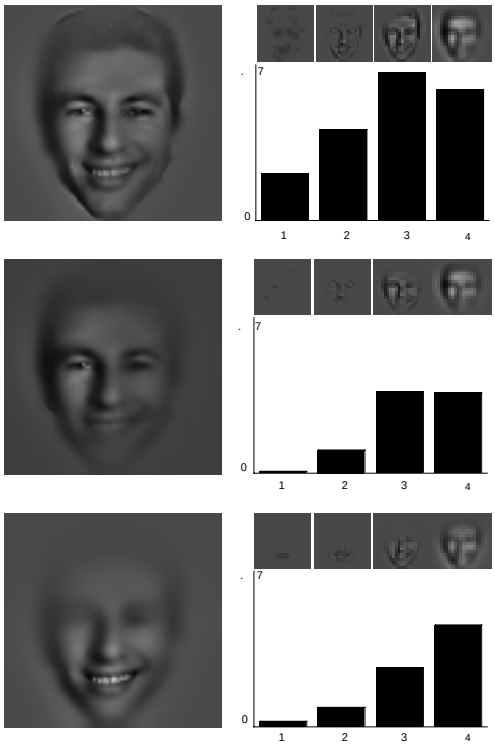

Figure 2. (a) The larger face depicts the effective face stimulus for the identity task. The smaller pictures illustrate the diagnostic information used to resolve the identity task at each independent scale from fine to coarse, respectively. The coarsest scale is not depicted as it contains no meaningful information. The bar chart provides a quantitative illustration of the proportion of the face area used to resolve the task at each scale. Figures (b) and (c) follow the same format as figure (a) illustrating the potent face for the gender task and expressive or not task respectively, the diagnostic information for each task at each scale and a quantitative account of the use of information is illustrated in the bar charts.

Discussion

Use of scale information between categorization tasks. Figure 2 presents a comparison of the relative use of scale information across tasks. From top to bottom, the large face pictures depict the information used in identity, gender and expressive or not. The figure reveals that the use of diagnostic information differs across categorization tasks, and scales. For example, whereas the mouth is well-defined at all scales in the identity and expressive tasks it is neglected at the finest scales in the gender task. In a related vein, the eyes are both represented at all scales in identity, but only one of them is well represented in gender, and both are neglected in expressive. The chin is well defined in identity, but not in expressive and gender. Compared to the mouth and the eyes, the nose is less well defined in all tasks.

To quantify the use of spatial scales across tasks, we computed the diagnostic areas revealed at each scale over the total area covered by the face in the image plane. The histograms in Figure 2 plot the use of diagnostic information across spatial scales--1 means finest, and 4 coarsest scale. The small face pictures corresponding to each scale illustrate what this face information is. The pictures reveal that the use of fine scale information (labelled 1 in the histograms, and depicted in the leftmost small picture) is most differentiated across the three tasks. In identity, it depicts the eyes, the mouth and the chin, whereas in gender it is only used for the left side eye, and the mouth in expressive. In contrast to the finest scale, the coarsest scale (i.e., the fourth scale) is much less differentiated, revealing only a holistic representation of the face features. This forms a skeleton that is progressively distinguished and fleshed out with increasing spatial resolution (see the progression of face information from coarse to fine in the small pictures of Figure 2, from right to left.) The asymmetry in extracting diagnostic information to resolve the gender task is consistent with studies showing that there is a right-hemisphere bias (the left-side of the image) in processing various facial attributes, including gender (Burt & Perrett, 1997).

Turning to the relative use of scales within each task, there is a clear advantage for the third scale in identity, corresponding to face information comprised between 11.25 and 22.5 cycles per face. This is consistent with the face recognition literature where the best scale for face recognition is between 8 and 32 cycles per face, depending on authors (see Morrison & Schyns, in press, for a review). Note, however, that our analysis is more refined because not only can we specify what the best scale is, but also where this information is located in the image plane. In contrast, the best scale for expressive or not (here, the discrimination between neutral and happy) is information comprised between 5.62 and 11.25 cycles per face (the fourth scale). This is in line with Jenkins et al. (1997) and Bayer, Scwartz & Pelli, (1998) who also found that the detection of the happy expression was most resilient to changes in viewing distances (i.e., with information corresponding to coarser scales). For gender, scales 3 and 4 were most used, and across task, there appears to be a bias for face information comprised between 5.62 and 22.5 cycles per face (the coarser scales) when information was available from the entire scale spectrum. At this stage, it is worth pointing out that the self-calibration property of Gosselin and Schyns' (2000) technique ensures that if subjects required only information from the finest scale to resolve the tasks, they would not reach the performance criterion of 75% and the average number of bubbles would increase at each scale, until they displayed enough information at the finest scale to reach criterion. In other words, the reported biases for the coarser scales do not arise from the technique, which is unbiased, but from the biases of observers who use information in categorization tasks.

Ideal Observers. In *Bubbles*, the observer determines the informative subset of a randomly, and sparsely sampled search space. To highlight this unique property, we here contrast human and ideal observers (Tjan, Braje, Legge & Kersten, 1987). The ideal observer will provide a benchmark of the information available in the stimulus set to resolve each task. We have biased the ideal to capture all the regions of the image that have highest local variance between the considered categories (identity, male vs. female, and neutral vs. expressive). This ideal considers the stimuli as images (not as faces composed of eyes, a nose and a mouth, as humans do). The ideal might not necessarily be sensitive to the regions that humans find most useful (the diagnostic regions), but to the information that is mostly available in the data set for the task at hand. We constructed a different ideal observer for the tasks of identity, gender, and expressive or not and submitted them to *Bubbles*, using the same parameters as those of our experiment with humans. Here, however, the number of bubbles remained constant (equal to the average required in each task), and we added to the face

stimuli a varying percentage of white noise to maintain categorization performance at 75% correct. In a Winner-Take-All algorithm, the ideal matched the information revealed in the bubbles with the same bubbles applied to the 32 memorized face pictures. The identity, gender or expressive or not categorization response of the ideal was the best matched picture. We then computed the ProportionPlane(scale) and DiagnosticPlane(scale), as explained earlier, to derive the effective face of each categorization task (see Figure 3). A comparison between the human and the ideal effective faces reveal only a partial correlation of use of information. This indicates that the highest variations of information in the image were not necessarily used by humans, who instead focused on the diagnostic face information. It further stresses that *Bubbles* is a human, partially efficient, not a formal, optimally efficient, feature extraction algorithm (Gosselin & Schyns, 2000).

Figure 3. The effective face stimulus of the Ideal Observer for each categorization task, identity, gender and expressive or not, respectively.

Deriving a two-dimensional map of attention. So far, we have examined the use of information across the different spatial scales of a face. We can now derive a precise measure of the diagnosticity of each image locations for the different face categorizations. Remember that the DiagnosticPlane(scale) represent a with value of 1 the presence of diagnostic information at all image locations. To measure the gradient of probability of finding diagnostic information at any image location, we simply multiply the normalized probability of using a scale with the DiagnosticPlane of this scale, and add together all the DiagnosticPlane(scale). When diagnostic information was present (vs. absent) at all scales for this image location, it has a probability of 1 (vs. 0). Figure 4 renders with a grey scale the gradient of probability (white = 1, black = 0) of finding diagnostic information at any location of the image in identity, gender, and expressive or not. If the attention is alloctated (or eye movements are guided) to the most relevant image locations in a task, the maps of Figure 4 have a predictive value. For example, Figure 2 reveals that the regions of the eyes and the mouth are diagnostic across the entire scale spectrum, and so these locations have highest probability in Figure 4. From the seminal work of Yarbus (1965), studies in eye movements have consistently demonstrated that the eyes and the mouth were mostly scanned in face identification tasks.

Figure 4. The 2D attentional maps for each categorization task, identity, gender and expressive or not, respectively.

Concluding Remarks

Our goal was to address the problem of recognition without directly asking questions about internal representations. Our analysis established how three face categorization tasks selectively used information from a three-dimensional input space (the two-dimensional image plane x spatial scales). From this selective use, we derived a gradient of probability of locating diagnostic information in the image plane. A rational categorizer should selectively allocate its attention to the regions of the image that maximize this probability thus minimizing the uncertainty of locating diagnostic information, see Figure 4.

Acknowledgements

This research was supported by ESRC grant R000223179.

References

Archambault, A., O'Donnell, C., & Schyns, P.G. (1999). Blind to object changes: When learning one object at different levels of categorization modifies its perception. *Psychological Science,* **10**, 249-255.

Bayer, H.M., Schwartz, O., & Pelli, D. (1998). Recognizing facial expressions efficiently. *IOVS,* **39**, S172.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, **94**, 115-147.

Biederman, I., Shiffrar, M.M. (1987). Sexing day-old chicks: a case study and expert systems analysis of a difficult perceptual leaning task. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **13**, 640-645.

Biederman, I., & Cooper, E.E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, **23,** 393-419.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, **20**, 38-64.

Biederman, I. & Gerhardstein, P.C. (1995). Viewpoint-dependent mechanisms in visual object recognition: a critical analysis. *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 1506-1514.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436.

Bruce, V. (1994). What the human face tells the human mind: Some challenges for the robot-human interface. *Advanced Robotics*, **8**, 341-355.

Bülthoff, H.H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view theory of object recognition. *Proceedings of the National Academy of Science USA*, **89**, 60-64.

Burt, D.M. & Perrett, D.I. (1997). Perceptual asymmetries in judgements of facial attractiveness, age, gender, speech and expression. *Neuropsychologia*, **35**, 685-693.

Cutzu, F., & Edelman, S. (1996). Faithful representations of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Science*, **93**, 12046-12050.

Gosselin, F. & Schyns, P.G. (2000). Bubbles: A new technique to reveal the use of visual information in recognition tasks. Submitted for publication.

Gosselin, F & Schyns, P.G. (in press). Why do we SLIP to the basic-level? Computational constraints and their implementation. *Psychological Review.*

Hill, H., Schyns, P.G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition.* **62**, 201-222.

Jenkins, J., Craven, B., Bruce, V., & Akamatsu, S. (1997). Methods for detecting social signals from the face. Technical Report of IECE, HIP96-39. The Institute of Electronics, Information and Communication Engineers, Japan.

Jolicoeur, P. (1990). Identification of disoriented objects: A dual-systems theory. *Mind and Language*, **5**, 387-410.

Jolicoeur, P., Gluck, M., & Kosslyn, S.M. (1984). Pictures and names: Making the connexion. *Cognitive Psychology*, **19**, 31-53.

Liu, Z., Knill, D.C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, **35**, 549-568.

Morrisson, D. & Schyns, P.G. (in press). Usage of spatial scales for the categorization of faces, object and scenes. *Psychological Bulletin and Review*.

Oliva, A. & Schyns, P.G. (2000). Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology*, **41**, 176-210.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442.

Perrett, D.I., Oram, M.W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformation. *Cognition*, **67**, 111-145.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, **8**, 368-373.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

Schyns, P.G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **23**, 681-696.

Simoncelli, E.P. (1999). Image and Multi-scale Pyramid Tools [Computer software]. NewYork: Author.

Simons, D., & Wang, R.F. (1998). Perceiving real-world viewpoint changes. *Psychological Science.* **9**, 315-320.

Tanaka, J., & Gauthier, I. (1997). Expertise in object and face recognition. In R.L. Goldstone, D.L. Medin, & P.G. Schyns (Eds.), *Perceptual Learning*. San Diego: Academic Press.

Tanaka, J.W., & Presnell, L.M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics*, **61**, 1140-1153.

Tanaka, J., & Taylor, M.E. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, **15**, 121-149.

Tarr, M.J., & Bülthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 1494-1505.

Tarr, M.J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, **21**, 233-282.

Tarr, M.J., & Pinker, S. (1991). Orientation-dependent mechanisms in shape recognition: Further issues. *Psychological Science*, **2**, 207-209.

Troje, N. & Bülthoff, H.H. (1996) Face recognition under varying pose: The role of texture and shape. *Vision Research*, **36**, 1761-1771.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, **67**, 21-44.

Yarbus, A.L. (1965). *Role of eye movements in the visual process*. Nauka: Moscow, USSR.