Hybrid images

Aude Oliva* MIT-BCS
Antonio Torralba† MIT-CSAIL
Philippe. G. Schyns‡ University of Glasgow

Figure 1: A hybrid image is a picture that combines the low-spatial frequencies of one picture with the high spatial frequencies of another picture producing an image with an interpretation that changes with viewing distance. In this figure, the people may appear sad, up close, but step back a few meters and look at the expressions again.

Abstract

We present *hybrid images*, a technique that produces static images with two interpretations, which change as a function of viewing distance. Hybrid images are based on the multiscale processing of images by the human visual system and are motivated by masking studies in visual perception. These images can be used to create compelling displays in which the image appears to change as the viewing distance changes. We show that by taking into account perceptual grouping mechanisms it is possible to build compelling hybrid images with stable percepts at each distance. We show examples in which hybrid images are used to create textures that become visible only when seen up-close, to generate facial expressions whose interpretation changes with viewing distance, and to visualize changes over time within a single picture.

Keywords: Hybrid images, human perception, scale space

1 Introduction

Here we exploit the multiscale perceptual mechanisms of human vision to create visual illusions (*hybrid images*) where two different interpretations of a picture can be perceived by changing the viewing distance or the presentation time. We use and extend the method originally proposed by Schyns and Oliva [1994; 1997; 1999]. Fig. 1 shows an example of a hybrid image assembled from two images in which the faces displayed different emotions. High spatial frequencies correspond to faces with "sad" expressions. Low spatial frequencies correspond to the same faces with "happy" and "surprise" emotions (i.e., the emotions are, from left to right: happy, surprise, happy and happy). To switch from one interpretation to the other one can step away a few meters from the picture.

Artists have effectively employed low spatial frequency manipulation to elicit a percept that changes when relying on peripheral vision (e.g., [Livingstone 2000; Dali 1996]). Inspired by this work, Setlur and Gooch [2004] propose a technique that creates facial images with conflicting emotional states at different spatial frequencies. The images produce subtle expression variations with gaze changes. In this paper, we demonstrate the effectiveness of *hybrid images* in creating images with two very different possible interpretations.

Hybrid images are generated by superimposing two images at two different spatial scales: the low-spatial scale is obtained by filtering one image with a low-pass filter; the high spatial scale is obtained by filtering a second image with a high-pass filter. The final image is composed by adding these two filtered images. Note that *hybrid images* are a different technique than *picture mosaics* [Silvers 1997]. *Picture mosaics* have two interpretations: a local one (which is given by the content of each of the pictures that compose the mosaic) and a global one (which is best seen at some predefined distance). Hybrid images, however, contain two coherent global image interpretations, one of which is of the low spatial frequencies, the other of high spatial frequencies.

We illustrate this technique with several proof-of-concept examples. We show how this technique can be applied to create face pictures that change expression with viewing distance, to display two configurations of a scene in a single picture, and to present textures that disappear when viewed at a distance.

2 The design of hybrid images

A hybrid image (H) is obtained by combining two images (I_1 and I_2), one filtered with a low-pass filter (G_1) and the second one fil-

*e-mail: oliva@mit.edu
†e-mail:torralba@csail.mit.edu
‡e-mail: p.schyns@psy.gla.ac.uk

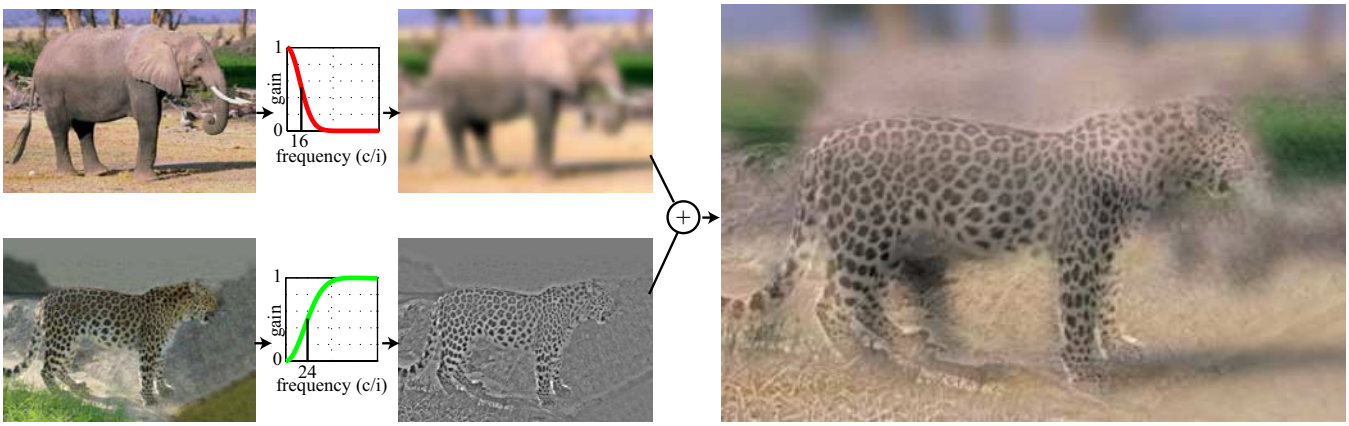

Figure 2: hybrid images are generated by superimposing two images at two different spatial scales: the low-spatial scale is obtained by filtering one image with a low-pass filter, and the high spatial scale is obtained by filtering a second image with a high-pass filter. The final hybrid image is composed by adding these two filtered images.

tered with a high-pass filter $(1-G_2)$: $H = I_1 \cdot G_1 + I_2 \cdot (1-G_2)$, the operations are defined in the Fourier domain. Hybrid images are defined by two parameters: the frequency cut of the low resolution image (the one to be seen at a far distance), and the frequency cut of the high resolution image (the one to be seen up close). An additional parameter can be added by introducing a different gain for each frequency channel. For the hybrids shown in this paper we have set the gain to 1 for both spatial channels. We use gaussian filters (G_1 and G_2) for the low-pass and the high-pass filters. We define the cut-off frequency of each filter as the frequency for with the amplitude gain of the filter is $1/2$.

Figure 2 illustrates the process used to create one hybrid image. The distance at which each component of a hybrid image is best seen and the distance at which the hybrid percept alternates can be fully determined as a function of the image size and the cut-off frequencies of the filters (expressed in cycles/image[1]). When viewing the images in this paper, switch between interpretations by stepping a few meters away from the picture. Note that the larger you display the images, the farther you will have to go in order to see the alternative image interpretation.

2.1 The perception of hybrid images

In the following section we describe the motivation behind hybrid images, as they relate to studies in human perception. We will provide the framework for understanding the mechanisms involved in perception of double image percepts.

Visual psychophysics research has shown that human observers are able to comprehend the meaning of a novel image within a short glance (100 msec [Potter 1975]). This phenomenal performance of rapid image understanding can be experienced while watching fast scene edits in an action movie or in a music video. Research in human perception has suggested that image understanding efficiency is based on a multi-scale, global to local analysis of the visual input [Burt and Adelson 1983; Majaj et al. 2002]: an initial analysis of the global structure and the spatial relationships between components guides the analysis of local details [Schyns and Oliva 1994; Watt 1987]. The global precedence hypothesis of image analysis ("seeing the forest before the trees", [Navon 1977]) implies a coarse-to-fine frequency analysis of an image, where the low spatial frequency components, which are contrasted and carried by the fast magnocellular pathway, dominate early visual processing [Hughes et al. 1996; Lindeberg 1993; Parker et al. 1992; Schyns and Oliva 1994; Sugase et al. 1999].

Using hybrid stimuli, Schyns and Oliva [1994] tested the role that spatial frequency bands play for the interpretation of natural images. When the task required identifying a scene image quickly, human observers interpreted the low spatial frequency band (at a frequency cutoff of 8 cycles/image) before the high spatial frequency band (from 24 cycles/image): when showed hybrid images for 30 ms only, observers identified the low spatial scale (e.g., they would answer "cheetah" when presented with the image from Fig. 3) whereas for 150 ms duration, they identified the high spatial scale first (e.g., tiger in Fig. 3). Interestingly, participants were unaware that the visual stimuli had two interpretations. Additional experiments suggested that the spatial frequency band preferentially selected for interpreting an image depends on the task the viewer must solve. Using hybrid faces similar to the one in Fig. 5.b, Schyns and Oliva [1999] showed that when participants were asked to determine the emotion of an hybrid face image displayed for only 50 ms (happy, angry or neutral), they selected the low spatial frequency face (angry in Fig. 5.b), but when they had to determine the gender of the same image, they used the low spatial frequency components of the hybrid as often as the high. Again, participants did not report noticing presence of two emotions or two genders in these images. These results demonstrated that the selection of frequency bands for fast image recognition is a flexible mechanism: The image analysis might still unfold according to a low to high spatial scale processing, but human observers are able to quickly select the frequency band, low or high, that conveyed the most information to solve a given task and interpret the image. Importantly, when selecting a spatial frequency, observers were not conscious of the information in the other spatial scale.

In the study of human perception, hybrid images allow characterizing the role of different frequency channels for image recognition, and evaluate the time course of spatial frequency processing. Hybrid images provide a new paradigm in which images interpretation can be modulated by playing with viewing distance or presenta-

[1] We use the units *cycle/image* for spatial frequencies as they are independent of the image resolution. The output of a gaussian filter with a cutoff frequency of 16 *cycles/image* will be the same independently of the resolution of the original image. The units *cycle/degree of visual angle* are used to describe the resolution observed when the image has a fixed size and is seen from a fixed distance.

Figure 3: Perceptual grouping between edges and blobs. The three images are perceived as a tiger when seen up-close and as a cheetah from far away. The differences among the three images is the degrees of alignment between the edges and blobs. Image a) contains two images superimposed without alignment. In image b), the eyes are aligned. And in image c), the head pose and the locations of eyes and mouth are aligned. Under proper alignment, the residual frequency band does not manage to build a percept. When seen up-close, it is difficult to see the cheetah's face, which is perfectly masked by the tiger's face. From far away, the tiger's edges are assimilated to the cheetah's face.

Figure 4: Color at high spatial frequencies is used to enhance the bicycle up-close. From a distance, one sees a motorcycle. The shape of the motorcycle is interpreted as shadows up-close.

tion time. For a given distance of viewing, or a given temporal frequency a particular band of spatial frequency dominates visual processing. Visual analysis of the hybrid image still unfolds from global to local perception, but within the selected frequency band, for a given viewing distance, the observer will perceive the global structure of the hybrid first (that the image in Fig. 3 represents a head), and take an additional hundred milliseconds to organize the local information into a coherent percept (organization of blobs if the image is viewed at a far distance, or organization of edges for close viewing).

2.2 Rules of perceptual grouping and hybrid images

In theory, one can combine any two images to create a hybrid picture. In practice, aesthetically pleasing hybrid images require following some rules that we describe in this section. In successful Hybrid images, when one percept dominates, consciously switching to the alternative interpretation becomes almost impossible. Only when the viewing distance changes can we switch to the alternative interpretation. In a hybrid image it is important that the alternative image is perceived as noise (lacking internal organization) or that it blends with the dominant subband.

Rules of perceptual grouping modulate the effectiveness of hybrid images. Low spatial frequencies (blobs) lack a precise definition of object shapes and region boundaries, which require the visual system to group the blobs together to form a meaningful interpretation of the coarse scale. When observers are presented with ambiguous forms they interpret the elements in the simplest way. Observers prefer an arrangement having fewer rather than more elements, having a symmetrical rather than an asymmetrical composition and generally respecting other Gestalt rules of perception.

Symmetry and repetitiveness of a pattern in the low spatial frequencies are bad: they form a strong percept that it is difficult to eliminate perceptually. If the image in the high spatial frequencies lacks the same strong grouping cues, the image interpretation corresponding to the low spatial frequencies will always be available, even when viewing from a short distance. By introducing accidental alignments it is possible to reduce the influence of one spatial channel over the other. For instance, in Fig. 2 the top of the elephant (low spatial frequencies) is aligned with the horizon line (both low and high spatial frequencies). Therefore, when seeing the image up close, the top edge of the elephant can be explained by some of the fine edges. This reduces the saliency of the elephant. Fig. 3 shows several examples of hybrid images with different degrees of agreement between the low and high spatial frequencies.

Color provides a very strong grouping cue that can be used to create more compelling illusions. For instance, in Fig. 4 color is used only in the high spatial frequencies to enhance the bicycle and to reinforce the interpretation of the motorcycle as shadows when the image is viewed up close.

The importance of correctly choosing the cut-off frequencies for the filters is illustrated in Fig. 5. In Fig. 5.a, both filters have a strong overlap, and consequently, there is not a clean transition between the two faces. For the hybrid image on Fig. 5.b, the two filters have little overlap. The result is a cleaner image that produces an

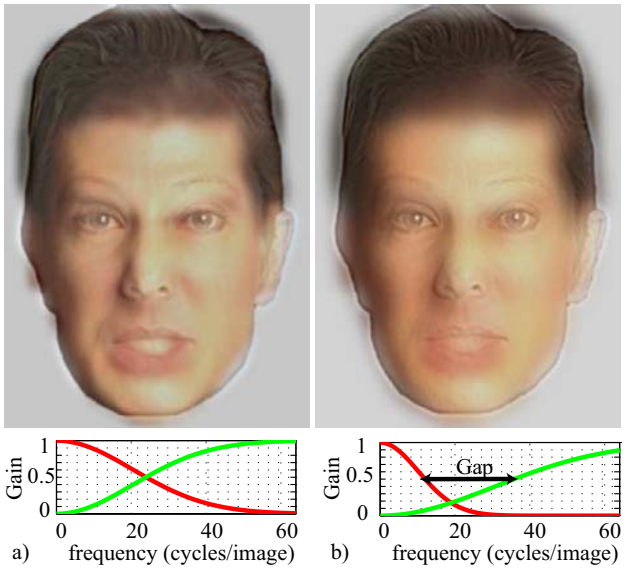

Figure 5: An angry man or a thoughtful woman? Both hybrid images are produced by combining the faces of an angry man (low spatial frequencies) and a stern woman (high spatial frequencies). You can switch the percept by watching the picture from a few meters. a) Bad hybrid image. The image looks ambiguous from up close due to the filter overlap. b) Good Hybrid image.

unambiguous interpretation (it looks like a woman from up close and as a man from far away). This is especially important when the images are not perfectly aligned.

One interesting observation is that when the images are properly constructed, the observer seems to perceive the masked image a noise. Hybrid images break one important statistical property of real-world natural images (Fig. 6), i.e., the correlations between outputs of pass-band filters at consecutive spatial scales. Fig. 6.a shows the cross-correlation matrix obtained between the different levels of a Laplacian pyramid for a natural image. The edges found at one scale are correlated with the edges found in the scales below and above. The same thing is obtained when two images are superimposed (additive transparency). In this case there is not a simple filter to separate both images (and the percept of the two images is mixed independently of the distance at which we observe the image). Fig. 6.c shows the correlation matrix obtained when an image is blurred (with a cutoff frequency of $16c/i$) and then corrupted with additive white noise. The correlation matrix reveals which scales are dominated by the noise, as they do not have the cross-scale correlations we'd expect from a natural image. In the case of a hybrid image, the correlation matrix (Fig. 6.d) reveals the existence of two groups.

Fig. 7 shows the output of a Laplacian pyramid applied to the hybrid image from Fig. 5.b. Low frequency channels and high frequency channels see different images. Note that each subband is also an hybrid image itself. If you move away from the page, you will see that, one by one, the subbands take the identity of the low-scales. At reading distance, the four images on the top row are interpreted as an angry man; the bottom, a stern woman. As you step back from the images, you will see that the angry man's face begins to appear in more subbands. The finer the scale of each subband, the farther you have to go in order to see the switch of images.

In summary, two primary mechanisms can be exploited to create compelling hybrid images. The first is maximizing the correlation

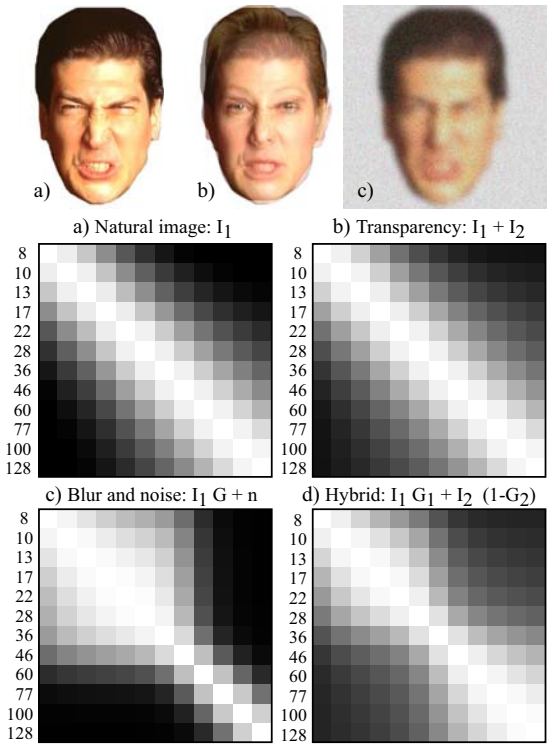

Figure 6: Correlations across levels of a Laplacian pyramid for images following several manipulations. a) Natural image, b) two images added, c) blurry image with additive white noise, and d) hybrid image ($f_1 = 16$ *cycles/image*, $f_2 = 48$ *cycles/image*).

between edges in the two scales so that they blend. The second resides in the fact that the remaining edges that do not correlate with other edges across scales can be perceived as noise. This is the case in Fig. 5.b, for which there is a very compelling blending of edges across scales, but, when viewing the image up close, there seems to be some low-spatial frequency noise.

2.3 Capacity of scale space

Up to now, hybrid images have been obtained by mixing two images, but could it be possible to combine more than two images and still have a coherent percept that transitions as we change viewing distance? In a study about text masking, Majaj et al. [2002] created a stimulus superimposing 4 letters, each containing energy at different spatial scales. As the observer moves away from the stimuli, they report the image switching from one letter to another. The results are interesting, but the lack of good grouping cues between the multiple scales creates an image that looks distorted. Also, multiple letters are visible at any given time. Superposition of multiple images remains an open issue.

3 Applications

In this section we discuss some applications (see video complementing the paper for additional examples).

Private font: We can use the hybrid images to display text that is not visible for people standing at some distance from the

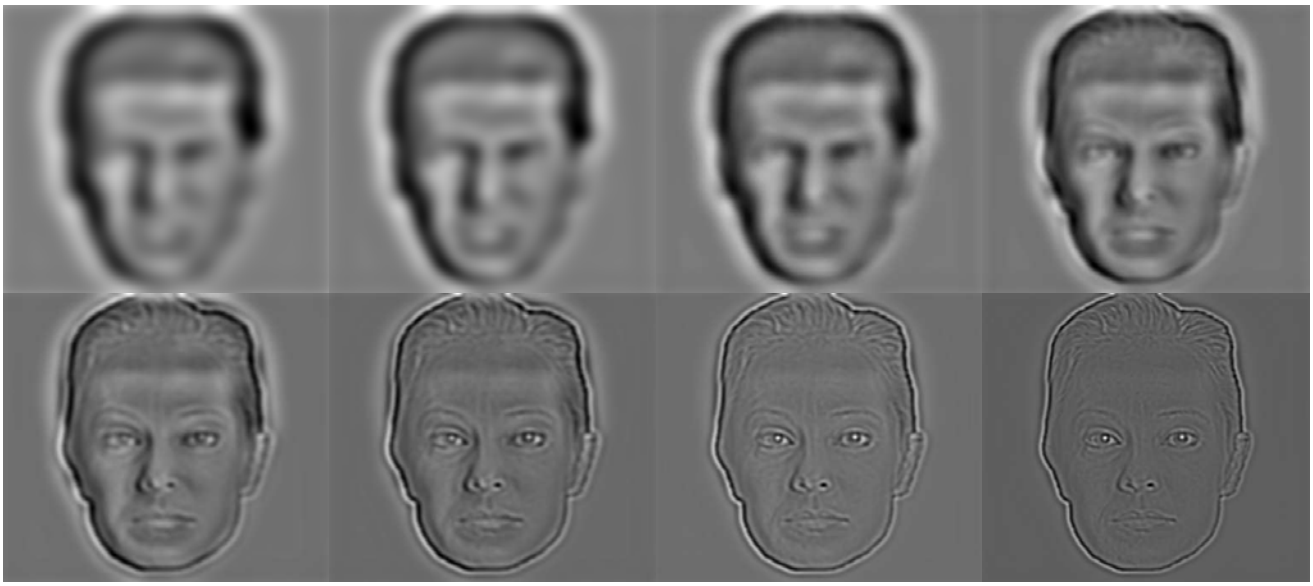

Figure 7: Output of a Laplacian pyramid revealing the components of the hybrid image of Fig. 5.b.

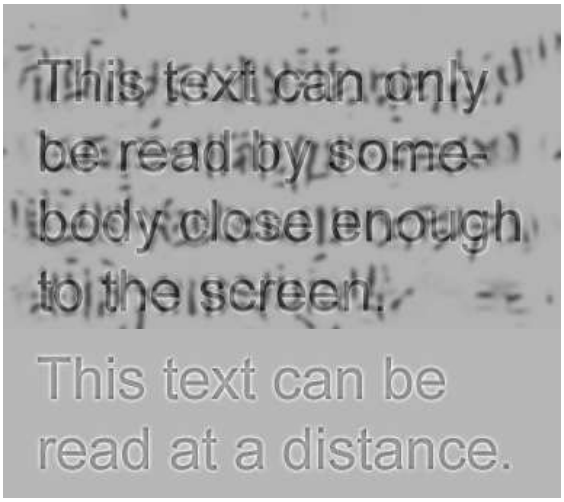

Figure 8: The hybrid font becomes invisible at few meters. The bottom text remains easy to read at relatively long distances.

screen. Commercial products for user privacy generally rely on head mounted displays or on polarized screens for which visibility decreases with viewing angle. Hybrid fonts comprises two components: the high spatial frequencies (which will contain the text) and the low spatial frequencies (which will contain the masking image).

For the high pass filter we use a gaussian filter with a width (σ) adjusted so that $\sigma < n_p$, where n_p is the thickness of a letter's stroke measured in pixels. The low-frequency channel (masking signal) contains a text-like texture [Portilla and Simoncelli 2000]. Solomon and Pelli [1994] have shown that letters are more effectively masked by a noise in the frequency band of 3 cycles per letter. Therefore we adjust the cut-off frequency of the low-pass filter to be $3 * n$ with n being the number of letters in a text line. The goal is to reduce the interference of the noise with the text when we viewing up close, while having an effective masking noise when looking from further away. In the example shown in Fig. 8 the text is only readable from a distance below one meter. From a distance of about two meters, the text is unreadable. Masking of the low spatial frequencies is very important in producing this effect (Fig. 8). The text in the bottom has only been high-pass filtered, and there is no masking at low spatial frequencies, therefore it remains easy to read at relatively long distances.

Hybrid textures: We can create textures that disappear with viewing distance. An example of this idea is shown in Fig. 9. This figure shows an example of a woman's face that turns into a cat when looking close. Note that this effect can not be obtained by superimposing the woman's face and the cat's face using transparency. Using transparency (additive superposition) creates a face that will not change with distance.

Changing faces: Hybrid images are especially powerful to create images of faces that change expressions, identity, or pose as we vary the viewing distance. Fig. 1 shows a compelling example of changes of facial expression. The edges at multiple scales blend producing images that look natural at all distances. In the case of face images, correct alignment between facial features is important in order to create pictures that seem unaltered. In case of misalignment, the best is to apply a distortion (affine warping) to the face that will be in the low spatial frequencies.

Time changes: Fig. 9 shows an example of using an hybrid image to show two states of a house by combining two picture taken at two different instants.

4 Conclusion

We have described the technique, hybrid images, which permits creating images with two interpretations that change as a function of viewing distance. Despite the simplicity of the technique, the images produce very compelling surprise effects on naive observers. They also provide an interesting new visualization tool to morph two complementary images into one. Creating compelling hybrid images is an open and challenging problem, as it relies on perceptual grouping mechanisms that interact across different spatial scales.

Figure 9: right) Cat woman: the texture corresponding to the cat's face disappears when the image is viewed from a few meters. Left) The house under construction. When you view the image at a short distance, the house is seen under construction, but if you step away from the picture you will see its final state.

References

BURT, D. C., AND ADELSON, E. H. 1983. The laplacian pyramid as a compact image code. *IEEE Transaction on Communications 31*, 532–540.

DALI, S. 1996. *The Salvador Dali Museum Collection*. Bulfinch Press.

HUGHES, H. C., NOZAWA, G., AND KITTERLE, F. 1996. Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience 8*, 197–230.

LINDEBERG, T. 1993. Detecting salient blob-like images structures and their spatial scales with a scale-space primal sketch: a method for focus of attention. *Int. J. Comp. Vis 11*, 283–318.

LIVINGSTONE, M. S. 2000. Is it warm? is it real? or just low spatial frequency? *Science 290*, 5495, 1299.

MAJAJ, N., PELLI, D., KURSHAN, P., AND PALOMARES, M. 2002. The role of spatial frequency channels in letter identification. *Vision Research 42*, 1165–1184.

NAVON, D. 1977. Forest before trees: the precedence of global features in visual perception. *Cognitive psychology 9*, 353–383.

OLIVA, A., AND SCHYNS, P. 1997. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive psychology 34*, 1, 72–107.

PARKER, D., LISHMAN, J., AND HUGHES, J. 1992. Temporal integration of spatially filtered visual images. *Perception 21*, 147–160.

PARKER, D., LISHMAN, J., AND HUGHES, J. 1996. Role of coarse and fine information in face and object processing. *J. Exp. Psychol. Hum. Percept. Perform. 22*, 1448–1466.

PORTILLA, J., AND SIMONCELLI, E. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comp. Vis 40*, 49–71.

POTTER, M. 1975. Meaning in visual scenes. *Science 187*, 965–966.

SCHYNS, P., AND OLIVA, A. 1994. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science 5*, 195–200.

SCHYNS, P., AND OLIVA, A. 1999. Dr. angry and mr. smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition 69*, 243–265.

SETLUR, V., AND GOOCH, B. 2004. Is that a smile?: gaze dependent facial expressions. In *NPAR '04: Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering*, ACM Press, New York, NY, USA, 79–151.

SILVERS, R. 1997. *Photomosaics*. Henry Holt and Company, Inc.

SOLOMON, J., AND PELLI, A. 1994. The visual filter mediating letter identification recognition. *Nature 369*, 395–397.

SUGASE, Y., YAMANE, S., UENO, S., AND KAWANO, K. 1999. Global and fine information coded by single neurons in the temporal visual cortex. *Nature 400*, 869–873.

WATT, R. 1987. Scanning from coarse to fine spatial scales in the human visual system after onset of a stimulus. *J. Opt.Soc.Am: A, 4*, 2006–2021.