LETTERS

Hierarchy and adaptivity in segmenting visual scenes

Eitan Sharon[1], Meirav Galun[1], Dahlia Sharon[2], Ronen Basri[1] & Achi Brandt[1]

Finding salient, coherent regions in images is the basis for many visual tasks, and is especially important for object recognition. Human observers perform this task with ease, relying on a system in which hierarchical processing seems to have a critical role[1]. Despite many attempts, computerized algorithms[2–5] have so far not demonstrated robust segmentation capabilities under general viewing conditions. Here we describe a new, highly efficient approach that determines all salient regions of an image and builds them into a hierarchical structure. Our algorithm, segmentation by weighted aggregation, is derived from algebraic multigrid solvers for physical systems[6], and consists of fine-to-coarse pixel aggregation. Aggregates of various sizes, which may or may not overlap, are revealed as salient, without predetermining their number or scale. Results using this algorithm are markedly more accurate and significantly faster (linear in data size) than previous approaches.

We present an algorithm (flow chart in Fig. 1a) that adaptively assembles pixels into small aggregates according to resemblance in luminance. The small aggregates are then assembled in a similar manner, according to resemblance in their properties, into still larger, more complex aggregates. As the aggregates are formed at each level, their statistical properties are accumulated, and their salience is evaluated according to differences compared with their neighbours[7]. Figure 1b shows a salient segment (III) composed of a hierarchy of aggregates at lower levels (I, II). This increase in both the size and complexity of successive, higher-level aggregates is reminiscent of the primate visual system, in which neurons in successively higher visual areas respond to successively more complex stimulus features, within increasingly larger fields[8,9].

Computationally, it is useful to think of segmentation within the framework of cuts in graph theory[10]. Each pixel in an image (Fig. 2a) corresponds to a node in a graph (Fig. 2b), coupled to each of its four neighbours according to their similarity in luminance level. The goal is to 'cut' this graph into pieces. A salient segment in the image is one for which the similarity across its border is small, whereas the similarity within the segment is large (for a mathematical description, see Methods). We can thus seek a segment that minimizes the ratio of these two expressions. Despite its conceptual usefulness, minimizing this 'normalized cut' measure is computationally prohibitive, with cost that increases exponentially with image size[10]. Approximations to the optimal cut can be obtained using spectral methods, with the most efficient approximation to date having a computational cost proportional to $n^{\frac{3}{2}}$ (ref. 11) (where n is the number of pixels), but this supra-linear cost seems to be yet too demanding for the brain, which deals with very large images.

Moreover, finding salient segments in real images is by nature a problem of multiple scales. Pixel-scale measurements (for example, intensity and colour) alone are insufficient to characterize segments, and larger-scale measurements (for example, average intensity and texture) at multiple scales must be incorporated as well. However, there is an inherent 'chicken-and-egg' difficulty associated with applying coarser measurements to images: when coarse-scale measurements are taken near the boundaries of segments they mix their statistics, and smooth the transition between them (see Supplementary Fig. 1b). Consequently, it is difficult to locate the boundaries between the segments. Ideally, we should apply coarse measurements only within segments, but segment boundaries

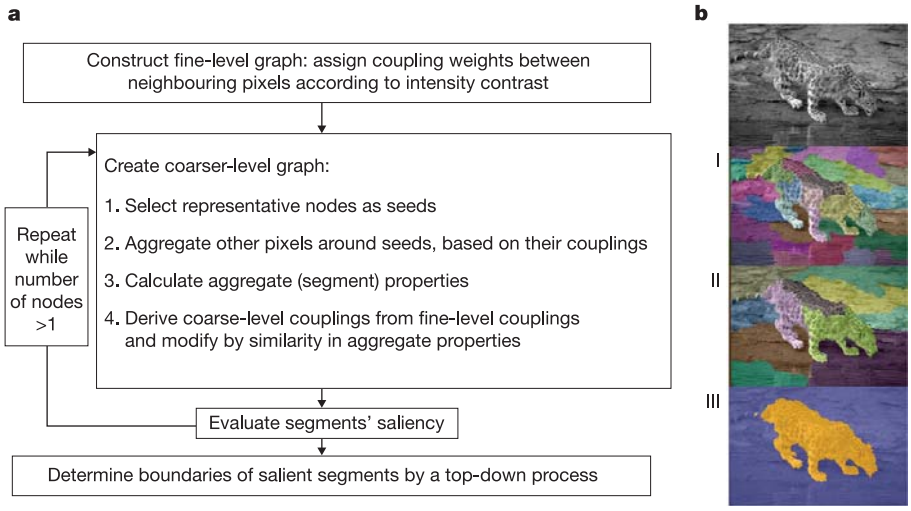

Figure 1 | SWA. a, Flow chart. **b**, A hierarchy composing a salient segment and its background. The leopard segment (III) is shown with two out of the ten levels of aggregates composing it (I, II). Original image is shown at the top.

[1]Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel. [2]Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA.

© 2006 Nature Publishing Group

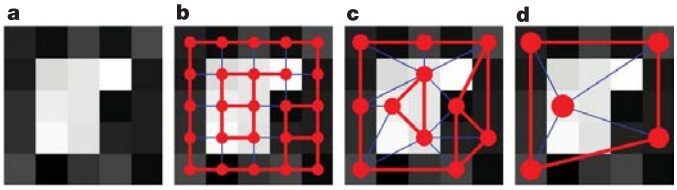

Figure 2 | The multiscale normalized cut graph approach. a, A simple image. **b**, Pixels of the image are nodes, represented by filled circles; strong coupling is represented by thick red lines, and weak coupling by thin blue lines. **c**, Adaptive coarsening. Each pixel in **b** is strongly coupled to one of the chosen seeds shown here (thus, pixels strongly coupled to a given seed form an aggregate). Couplings between the seeds are shown. **d**, An additional coarsening level. In this case, this is the level at which the salient segment is detected.

are unknown before the measurements are taken—hence, the chicken-and-egg difficulty. As detailed below, our hierarchical algorithm applies an adaptive coarsening process that solves this problem. (See Supplementary Fig. 1 for an example of the preservation of boundaries by aggregates in our algorithm, compared to regular coarsening.) In addition, the complexity of our algorithm is linear with the number of pixels in the image, on a regular serial computer, and only logarithmic with the image size on a parallel computer. Therefore both the results and computational speed for obtaining them are qualitatively improved under our segmentation by weighted aggregation (SWA) algorithm.

Inspired by algebraic multigrid solvers for physical systems[12] we apply a multiscale approach for recursively reducing the normalized-cut minimization problem in a non-iterative manner. We start by choosing about half of the pixels as representatives, which we call seeds: these are chosen so that every pixel in the original image is strongly coupled (that is, similar) to at least one seed adjacent to it[6] (Fig. 2c). We then define the minimization problem only for the seeds, and subsequently approximate the solution for the whole image using an interpolation matrix that is set according to the coupling between neighbouring pixels in the fine scale (see Methods).

In addition to significantly reducing the number of nodes in the graph, this coarsening creates small aggregates of pixels adapted to the image at hand, the intensities of which are similar. Every pixel belongs to either one or several aggregates, each centred at one seed, by an interpolation weight proportional to the coupling between that pixel and the seed. We continue recursively (Fig. 2d), aggregating collections of nodes into much fewer nodes in the coarser-level graphs, thus creating a pyramid of graphs with larger aggregates of pixels in its coarser levels. Salient segments emerge in the appropriate level of the pyramid as nodes that are coupled weakly to their neighbours (for example, Figs 1b and 2d). Consequently, the minimization problem is simplified into one of looking for salient nodes at all levels of the pyramid.

Moreover, we use this approach to go far beyond cut minimization. After each coarsening step, coarse measurements are taken over the newly formed aggregates and are subsequently used to affect the aggregation process. (One such measure, shown in Supplementary Fig. 1, is the average intensity of the aggregate.) These multiscale measures for each aggregate include statistics characterizing the textures that appear in the image: the variance of the mean intensities and the second-order shape moments of its sub-aggregates at each finer level[13]. Consequently, each aggregate is represented by a multiscale set of characteristic measurements, efficiently summarizing its detailed pixel information. These measures are calculated recursively, and hence very efficiently: every measure at any level is determined directly from measures computed at the previous, finer levels. This treatment of texture replaces the need to apply many filters to the image[4,14,15], which suffers from the chicken-and-egg difficulty and from higher complexity.

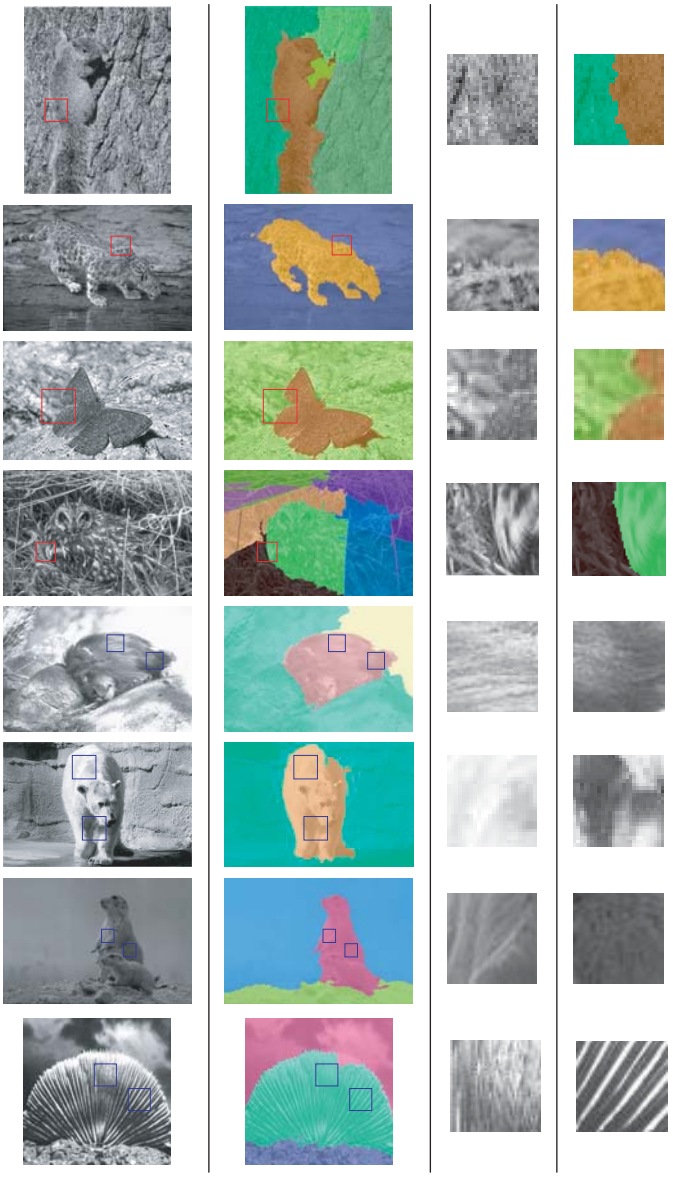

Figure 3 | Segmentation results for eight challenging images of animals on cluttered backgrounds. Transparent colours are overlaid over the original image (left) to mark segments (second left). In the top four images, we zoom into regions (marked with red squares) that are strikingly ambiguous to segment, and show our successful global-segmentation results. In the lower four images we zoom into regions (marked with blue squares) that have very different textural properties, but were nevertheless correctly grouped together into the same segment.

All of these measurements are incorporated into the segmentation process by affecting the coupling (similarity) between neighbouring aggregates (see Methods). Thus, analogous to the coupling between pixels according to similarity in luminance level, the coupling between aggregates reflects statistical similarities that could not be detected at finer levels. Note that this means that we are using contrast between region properties as a qualitative improvement of the process, in addition to contrast across boundaries in the simple normalized-cut approach.

The coarsening process detailed above keeps—out of the vast amount of data in the image—only the information necessary to segment it. In particular, nodes at coarse levels do not carry with them the precise boundary location of the corresponding segment. A top-down process can be used upon the detection of any salient segment in order to locate its exact boundary. We roll down the

© 2006 Nature Publishing Group

Figure 4 | Similarity search by parts. Top: three example queries (cyan, pink, orange frames) with the original images on the left. The segmentation of each image is shown on the right in each frame, with the target query segment (user-selected) outlined in yellow: lower flank of left frame, right lens and left lens for the left, middle and right examples, respectively. Bottom: 27 images from the database (out of 110; see Supplementary Fig. 3), with coloured boxes indicating the most similar sunglasses found for each query.

pyramid to a desired fine level, by multiplying the interpolation matrices relating each two consecutive levels. We take advantage of this top-down procedure to improve further the segmentation by an additional energy-minimization sweep at each level as it is encountered (see Methods), together with prodding the weighted interpolations towards boolean associations, thus sharply delineating the segment boundary.

The algorithm is extremely efficient, as only its initial, very simple aggregation is done at the level of individual pixels. Complexity per aggregate increases linearly with the level, whereas their number drops geometrically. Moreover, the algorithm can use massive parallel processing, especially at the finer (the most expensive) levels. Our current implementation takes 2 s to complete the bottom-up aggregation of a 450×450 image using a Pentium 4, 1.6-GHz processor. Further optimization is still possible for the runtime to take significantly less than a second, and even down to a tiny fraction of a second on a parallel computer.

We tested our method on a set of challenging natural images containing animals camouflaged against their backgrounds. Whereas humans may use object knowledge aided by memory to segment such images, our approach has been to find out how far segmentation can reach using input-driven processing only. Figure 3 shows a typical set of results. In each case we present the automatically detected level in the pyramid containing the most salient segment. In all examples the animal was segmented in one piece by our method, out-performing other leading algorithms (see Supplementary Fig. 2). Our method succeeds in these challenging images, in which a cluttered background is often locally difficult to distinguish from the animal segment. In the top four examples we zoom in on an area where local differences in the original image (second panel from right) are hard to detect, yet the multiscale considerations in our algorithm yield a successful segmentation (far right panel). Note, for example, the fine difference detected along the squirrel's back. Furthermore, despite differences between areas within a segment, our method captures the essential similarities well enough to join them together into a coherent segment. This is exemplified in the bottom four examples, which show two widely differing areas that are segmented together (two right panels). Two regions with very different luminance levels are segmented together in the lion example, because of texture similarity. Even such seemingly different areas as those belonging to the shell (bottom example) are joined together because their orderly oriented texture makes them more similar to each other than to neighbouring areas. In Supplementary Fig. 2 we compare SWA over the same set of images to several available state-of-the-art segmentation algorithms, and demonstrate its superior performance.

To demonstrate the utility of the hierarchical segment representation of the image provided by the SWA for higher-level visual tasks, we next applied it to a search task. The task was to find within a database of objects—sunglasses in this case—those that most resemble a target item in some feature (for example, shape of lenses), by matching aggregated properties of salient segments in the SWA hierarchy. The system pre-segments all images, and a query consists of selecting a segment within the target image. The segment with the most similar aggregated properties (shape, colour) within all database images is instantaneously found by searching through all summarized properties, and the sunglasses it belongs to are presented as the most similar to the query image. Figure 4 presents the results of three queries using this system, and Supplementary Fig. 3 presents the full database. The success of this system highlights the value of a robust hierarchical segmentation for higher-level tasks, importantly allowing a comparison of the same semantic object parts (for example, lenses) within different images, thus comparing 'apples to apples' and 'oranges to oranges'.

Our SWA algorithm constructs a representation of the image as a hierarchy of adaptive segments and finds the most salient segments without predetermining their number or size. Many common objects can naturally be described as a hierarchical collection of segments—thus, obtaining a hierarchy of salient segments is a useful novel framework within which to tackle object recognition. Our framework naturally allows the incorporation of top-down effects, expressing prior knowledge about properties of visual objects, as well as effects of context and attention, through the modification of couplings in the adaptive structure. For example, a probabilistic preference for smooth edges can be used in a top-down manner to join together aggregates with co-aligned boundaries even when lighting conditions make them appear different[16]. The simple top-down process already implemented in our algorithm suggests that although bottom-up mechanisms quickly segment the image into meaningful regions, feedback is needed for the accurate delineation of segment boundaries, as suggested also for the human visual system[17]. The increase in the size of aggregates and the complexity of their characteristics when moving up the hierarchy resemble well known properties of the primate visual system, as does the importance of interactions between the different levels[1]. Precise localization of segment borders may well be an important role for the massive feedback projections in the visual cortex.

METHODS

The image is regarded as a weighted graph $G = (V, E)$, V being its set of n nodes, each corresponding to a pixel, v_i ($i = 1,\ldots n$), and E the set of undirected weighted edges w_{ij}, connecting neighbouring nodes v_i, v_j. $w_{ij} = e^{-\alpha|I_i - I_j|}$, where I_i, I_j are the intensities of pixels i, j, respectively, and α is a globally set, positive real constant. We conveniently treat W as a symmetric matrix, with all $w_{ii} = 0$

© 2006 Nature Publishing Group

and $w_{ij} = 0$ if v_i and v_j are not neighbours. We further associate with the nodes a state vector $\mathbf{u} = (u_1, u_2, \ldots, u_n)$, and define the energy

$$\Gamma(\mathbf{u}) = \frac{\sum_{i>j} w_{ij}(u_i - u_j)^2}{\sum_{i>j} w_{ij} u_i u_j} = \frac{\mathbf{u}^{\mathrm{T}} L \mathbf{u}}{\frac{1}{2}\mathbf{u}^{\mathrm{T}} W \mathbf{u}} \quad (1)$$

where L, the so-called laplacian matrix of the graph, is set to satisfy the equality between the numerators, and W between the denominators. Any boolean assignment of $\mathbf{u}$ that yields a low-energy value $\Gamma(\mathbf{u})$ corresponds to a salient segment S in the image: the pixels $i \in \{1, 2, \ldots, n\}$ for which $u_i = 1$ are the pixels in S, otherwise $u_i = 0$.

If we permit real-value assignments to $\mathbf{u}$, the minimum for Γ is obtained by the solution of the generalized eigen problem $L\mathbf{u} = \lambda W\mathbf{u}$ with minimal positive eigenvalue λ. The algebraic multigrid procedure for solving this eigen problem consists of choosing a representative subset of $N \simeq \frac{n}{2}$ pixels in the image, which we call seeds and denote by renaming their corresponding state variables u_is to be $\mathbf{U} = (U_1, U_2, \ldots, U_N)$. For suitable choice of representatives[6], the minimizing eigenvector satisfies $\mathbf{u} \simeq P\mathbf{U}$, where the interpolation P is a sparse matrix $\{p_{ij}\}$; assuming for simplicity of notation that $U_k = u_k$, $(k = 1, 2, \ldots, N)$, then $p_{ik} = w_{ik}/\sum_j w_{ij}$ for $i > N$; whereas for $i \le N$, $p_{ik} = 0$ except for $p_{ii} = 1$, $(1 \le i, k \le N)$. Substituting the interpolation relation in equation (1), we seek to minimize

$$\Gamma(\mathbf{U}) \simeq \frac{\mathbf{U}^{\mathrm{T}}(P^{\mathrm{T}} L P)\mathbf{U}}{\frac{1}{2}\mathbf{U}^{\mathrm{T}}(P^{\mathrm{T}} W P)\mathbf{U}} = \frac{\sum_{k>1} \tilde{w}_{kl}(U_k - U_l)^2}{\sum_{k>1} \hat{w}_{kl} U_k U_l} \quad (2)$$

where $\{\tilde{w}_{kl}\}$ are set to satisfy the equality between the numerators, and $\{\hat{w}_{kl}\}$ between the denominators. $\Gamma(\mathbf{U})$ may effectively be approximated by replacing $\{\tilde{w}_{kl}\}$ with the simpler $\{\hat{w}_{kl}\}$. We call $\{\hat{w}_{kl}\}$ the coarse graph weights. By this coarsening we reduce the original minimization problem in $\mathbf{u}$ to a much smaller minimization problem in $\mathbf{U}$, the solution of which approximates the solution in $\mathbf{u}$ via the interpolation relation. Finally, this coarsening procedure can be repeated recursively, level after level[7].

We next modify the coarse graph weights to reflect also contrast in properties at the current coarse level, in addition to contrast at the finer level. This is done by collecting a vector of multilevel properties $\mathbf{f}_k = (f_{k1}, f_{k2}, \ldots f_{km})$ for every coarse node k and using these properties to reduce the coupling between neighbouring aggregates k and l by a factor proportional to $\exp(-\mathbf{f}_k^{\mathrm{T}} \Lambda \mathbf{f}_l)$, where the diagonal matrix Λ weighs the importance of every property. Entries in $\mathbf{f}_k$ represent statistics over the properties of the set of sub-aggregates of k. These are computed from the sub-aggregates using averaging weighted according to the interpolation matrix, and include its average intensity and the variances in the average intensities of its sub-aggregates at all finer scales, as well as its low-order shape moments (based on averaging $x_i^l y_i^k$, where (x_i, y_i) is the location of node v_i and k, $l = 0, 1, \ldots$).

We detect the salient segments at all levels of the pyramid as those aggregates k for which $\Gamma(\mathbf{U})$ has low values, with the state vector $\mathbf{U}$ set to 1 at the kth entry and 0 elsewhere. For each such aggregate k we use a top-down process to delineate its boundary. We start at the level at which k was detected with its characteristic state vector $\mathbf{U}$. By repeating interpolations from this level down to any finer level, using successively the interpolation relations given by the matrices P, we obtain for each finer aggregate the relative weight by which it relates to the aggregate k. Our goal at this stage is to bring $\mathbf{u}$ closer to a boolean state vector. We do so by modifying, at each of the finer levels, the interpolated $\mathbf{u}$ before interpolating it to the next finer level. Values $u_i > 0.9$ are set to 1, and values $u_i < 0.1$ are set to 0. Finally, at the finest level each pixel is associated solely with the aggregate k at the coarsest level for which its weight turned out largest. Note that despite the incorporation of the top-down process, complexity remains linear because the finest scale of aggregates needed to detail the segment boundary is proportional to its size.

Received 25 March; accepted 13 June 2006.
Published online 28 June 2006.

1. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1,** 1–47 (1991).
2. Pietikainen, M., Rosenfeld, A. & Walter, I. Split-and-link algorithms for image segmentation. *Patt. Recog.* **15,** 287–298 (1982).
3. Comanicu, D. & Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Patt. Anal. Machine Intell.* **24,** 603–619 (2002).
4. Malik, J., Belongie, S., Leung, T. & Shi, J. Contour and texture analysis for image segmentation. *Int. J. Comp. Vision* **43,** 7–27 (2001).
5. Felzenszwalb, P. & Huttenlocher, D. Efficient graph-based image segmentation. *Int. J. Comput. Vision* **59,** 167–181 (2004).
6. Brandt, A. Algebraic multigrid theory: the symmetric case. *Appl. Math. Comput.* **19,** 23–56 (1986).
7. Sharon, E., Brandt, A. & Basri, R. Fast multiscale image segmentation. *Proc. IEEE Conf. Comput. Vision Patt. Recog.* **1,** 70–77 (2000).
8. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195,** 215–243 (1968).
9. Tanaka, K. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* **19,** 109–139 (1996).
10. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Machine Intell.* **22,** 888–905 (2000).
11. Golub, G. H. & Van Loan, C. F. *Matrix Computations* (Johns Hopkins Univ. Press, Baltimore, 1989).
12. Brandt, A., McCormick, S. & Ruge, J. In *Sparsity and its Applications* (ed. Evans, D. J.) 257–284 (Cambridge Univ. Press, Cambridge, 1984).
13. Galun, M., Sharon, E., Basri, R. & Brandt, A. Texture segmentation by multiscale aggregation of filter responses and shape elements. *Proc. Int. Conf. Comput. Vision* **1,** 469–476 (2003).
14. Julesz, B. Textons, the elements of texture perception, and their interactions. *Nature* **290,** 91–97 (1981).
15. Voorhees, H. & Poggio, T. Computing texture boundaries from images. *Nature* **333,** 364–367 (1988).
16. Sharon, E., Brandt, A. & Basri, R. Segmentation and boundary detection using multiscale intensity measurements. *Proc. IEEE Conf. Comput. Vision Patt. Recog.* **1,** 469–476 (2001).
17. Stanley, D. A. & Rubin, N. fMRI activation in response to illusory contours and salient regions in the human lateral occipital complex. *Neuron* **37,** 323–331 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Research was supported in part by the European Commission Project Aim Shape, the Binational Science foundation, and by the German–Israeli Foundation. D.S. was supported by a grant from the National Institutes of Health. The research was conducted at the Moross Laboratory for Vision and Motor Control at the Weizmann Institute of Science. We thank N. Rubin and D. Jacobs for many useful remarks, and S. Geman for commenting on an earlier version of the manuscript. We are grateful to E. Borenstein for his help with constructing the sunglasses search system. We also thank M. Varma and R. Deitch for help with the comparisons presented in the Supplementary Information and N. Brandt for help with the graphics.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.S. (eitan.sharon@weizmann.ac.il).

© 2006 Nature Publishing Group