

Eliciting Perceptual Ground Truth for Image Segmentation.

Victoria Hodge, Garry Hollier, John Eakins and Jim Austin.

Advanced Computer Architectures Group, Department of Computer Science,
University of York, York, UK
{vicky, hollier, eakins, austin}@cs.york.ac.uk

Abstract.

In this paper, we investigate human visual perception and establish a body of ground truth data elicited from human visual studies. We aim to build on the formative work of Ren, Eakins and Briggs who produced an initial ground truth database. Human participants were asked to draw and rank their perceptions of the parts of a series of figurative images. These rankings were then used to score the perceptions, identify the preferred human breakdowns and thus allow us to induce perceptual rules for human decomposition of figurative images. The results suggest that the human breakdowns follow well-known perceptual principles in particular the Gestalt laws.

Introduction

We hypothesise that perception and thus segmentation varies from person to person and also varies with the domain of application (context). This subjectivity is almost inevitably due to culture, education, expectation, domain of application, mood, age etc. but there must be a core set of commonalities across human judgements that we aim to distil out. There is currently no comprehensive theory of human or computational image and shape segmentation. One theory is that humans decompose images along Gestalt principles. There has been widespread investigation including human experimentation of individual Gestalt principles [W23],[K63],[K79],[G72].

Our work forms part of the PROF I (Perceptually-Relevant Retrieval of Figurative Images) project¹. In PROF I, we aim to develop new techniques for the retrieval of figurative images (i.e. abstract trademarks and logos) from large databases. The techniques will be based on the extraction of perceptually relevant shape features and the matching of these features in the target image against features in the stored images, thereby overcoming many of the limitations of existing methods. In this paper we focus on the perceptual segmentation of raw images and grouping shape elements.

Existing systems, for example trademark search systems, attempt to match a target against stored images such as those shown in Figs. 1-3 in one of two ways: (a) com-

¹ PROF I web page: <http://www.cs.uu.nl/profi/>

paring features generated from the images as a whole, or (b) matching features from individual parts of the images [E01].



Fig. 1



Fig. 2



Fig. 3

The principal difficulty in matching by parts is the selection of parts that accurately reflect the image's appearance to a human observer. In Fig. 1 this is reasonably clear (2 triangles and a circle). But in Fig. 2, should the central bars be matched as six individual components, or as two groups of three? And in Fig. 3, should matching be based on a circle and a triangle - neither of which are actually present in the image itself? These are the questions which this current research aims to answer.

For present purposes, therefore, we are primarily interested in clarifying two aspects of human segmentation behaviour: the formation of intermediate-level groupings of image parts; and, the generation of perceived elements not explicitly present in the original image. Our hypothesis is that these will allow us to identify the most salient image elements for matching more accurately than has hitherto been possible.

The seminal decomposition paper for this aspect of the PROFI project is Ren et al. [REB00]. The paper evaluates human participants when segmenting trademark images into their perceived constituent parts. The participants initially breakdown trademark images into a set of components in as many ways as they see fit. These breakdowns are then fed into the second part of the experiment where participants rank these breakdowns by their perceived likelihood. The paper's main discoveries are that humans partition trademark images into disjoint regions most commonly, then into overlapping or nested regions and partition into separate line segments or groups least commonly. The breakdowns generated are similar to the Gestalt principles [W23],[K63],[K79],[G72] of human perceptual organisation. The authors posit that perceptual line grouping, closed-region identification, texture processing, identifying familiar shapes (such as triangles, squares etc.) and uncovering 'hidden' image features (such as figure-ground reversal) are areas requiring further investigation. We aim to augment and complement these results in the current paper and use the results in our development of a computerized image retrieval system.

In current computational approaches, shapes may be segmented using either the shape's boundary or the shape's interior (fill area) but rarely both compared to the holistic viewpoint used by humans. Some examples of shape segmentation approaches, which are founded on geometrical properties, include Hoffman & Richards [HR84] who subdivide shapes based on the notion that concavities arise when two convex parts are joined and hence, divide the surface into parts at loci of negative minima. Siddiqi & Kimia [SK95] proposed a similar approach using limbs and necks: negative curvature minima and local minima of inscribed circles. Singh et al. [SSH99] use minimum distance and skeletal axes to determine segmentation lines between boundaries where at least one boundary is a concave vertex. Tanase & Veltkamp [TV02] also propose a segmentation approach using skeletons. The shape is initially segmented using the skeletal bifurcation points and the boundaries of these segments are

then simplified and protrusions removed. Leung & Chen [LC02] aim to unify skeletons and edge detection approaches thus going some way to a boundary-based/fill-area combination technique. The system either performs edge detection or thinning. The authors note that for a solid region where the shape conveys much visual information, edge detection is preferable to thinning as it extracts the contour of the region. However, for a region containing curves, thinning is preferable as it extracts the skeleton and “produces a better representation”.

The central premise for the investigations in this paper is to identify how humans decompose images, the degree of commonality across a range of human subjects and to provide a set of ground truth images. These ground truth images may be further analysed to elicit statistics and preference scores regarding the decomposition preferences of humans: i.e., which decomposition is generally preferred for each image, a ranked order of decompositions for each image, how many potential decompositions there should be for each image. We aim to investigate symmetry, texture, singularities and also to some extent the effect of figure/ground phenomena. We note that it is extremely difficult to isolate Gestalt principles within the trademark images. For example, altering an image along symmetrical lines will inevitably alter other Gestalt properties such as familiarity, continuity or perhaps grouping. We attempted to provide as wide a variety of symmetry, texture or singularity alterations as possible. We aim to use the results from our experimental analyses to drive the formation of an integrated computational system that mimics human segmentation. We need to ensure that our resultant computerised technique will not produce too many decompositions for a particular image as multiplicity implies that a Gestalt factor can only be active if it does not produce too many decompositions [DMM04].

In the remainder of this paper we detail the development and implementation of the experimental methodology and provide some analysis.

Experimental Methodology

The experimental methodology was developed in conjunction with the Psychology Department at the University of York, UK who advised on methodology, ethical considerations, and best practice and provided general advice and guidance.

We performed an initial pilot study to select suitable trademark or other figurative images and to revise and improve the experimental methodology.

For our experiment, a set of images was presented to University of York staff, students and their relatives and friends. Each image used is 4.5 cm high including any white space. All images are monochrome TIFFs. 28 subjects completed the experiment unsupervised in their own time and 25 subjects attended a 1 hour supervised session giving **53** subjects in total. Each of the **53** subjects received a printed booklet containing: a front sheet and 16 pages with 2 images per page in 2 columns giving **32** images in total in each booklet. The subjects also received a copy of the experiment instructions. The subjects were requested to draw (using pen or pencil) their perceived decompositions of each image in turn on to the booklet and to rank each decomposition (1st, 2nd, 3rd etc.) according to the order in which they perceived that decomposition. All completed booklets were anonymized and labelled with a subject ID

number. All subjects who completed the experiment were entered into a prize draw where the prizes were a £200, £50 and 5 x £10 shopping vouchers. The statistics of the subjects are: age range: 14 – 70; gender: *mixed*; nationality: *mixed international*.

There were 3 sets of **32** images. Each set contains some images present in the other sets to act as controls and thus to verify that the subjects in each group are statistically similar. The trademarks were in pairs (14 pairs in each set, $p_1 \dots p_{14}$) along with 4 other images (i1 .. i4). The unpaired images are supplementary control images (i1, i2) and buffer images (i3, i4) in case the subjects do not complete the exercise. The paired images were ordered $p_1^1, p_2^1, p_3^1, \dots, p_{14}^1, i1, i2, p_1^2, p_2^2, p_3^2, \dots, p_{14}^2, i3, i4$. The subjects received the first image of a pair and then later, a second paired image: the same image but altered according to symmetry, texture or singularity principles. These 3 sets of images were further divided into forward and backward sets giving 6 sets in total (A-Forward, A-Reverse, B-Forward, B-Reverse, C-Forward and C-Reverse). The forward and reverse sets have the order of the images reversed to prevent order bias where the order of image presentation affects the perception:

- Forward - $p_1^1, p_2^1, p_3^1, \dots, p_{14}^1, i1, i2, p_1^2, p_2^2, p_3^2, \dots, p_{14}^2, i3, i4$ and then
- Reverse - $p_{14}^2, p_{13}^2, p_{12}^2, \dots, p_1^2, i1, i2, p_{14}^1, p_{13}^1, p_{12}^1, \dots, p_1^1, i3, i4$.

If all subjects receive p_1^1 before p_1^2 then this may influence their perception of p_1^2 .

The first stage of analysing the images was to collate the breakdowns drawn by the subjects and to note the rank (1st, 2nd, 3rd etc.). For each image, each breakdown had a list of the ID of the subjects who perceived that breakdown and the rank they awarded it. For each image, if two subjects had drawn identical or extremely similar breakdowns then the breakdowns were marked as the same and the subjects' IDs and the rank they awarded the breakdown added to the list for that specific breakdown. Otherwise, the breakdowns were marked as two separate breakdowns and the subjects' IDs and ranks added to the respective breakdowns' lists. The output from this analysis is a listing of all breakdowns for each image in turn along with a listing of all subjects who drew that breakdown and the rank that each subject gave it.

Preference Scoring Mechanism

Ren et al. [REB00] used a slightly different experimental methodology compared to us. Ren et al. used subjects to elicit the breakdowns in stage 1 and then used a second set of subjects to rank the breakdowns in stage 2. We collapsed this into a single stage due to time constraints with all **53** subjects drawing and ranking their own breakdowns. This also required a slightly different scoring mechanism

For 74 of the 84 images, the subjects each drew 1, 2 or 3 breakdowns². The remaining 10 images produced 4 or 5 breakdowns³. Therefore, for all images we awarded scores of 3, 2, 1, 0.5 & 0.25 for ranks 1 to 5 respectively.

² For 2 images the maximum number of breakdowns drawn by one subject was 1, for 32 images the maximum number was 2 breakdowns and for 40 images the maximum number was 3.

³ For 7 images the maximum number of breakdowns drawn by one subject was 4. 3 images produced 5 breakdowns. 2 subjects drew most of these 4 or 5 breakdowns per image with another 3 subjects drawing 4 breakdowns per image once each.

For each breakdown the scores are totalled and divided by the total of the scores across all breakdowns for that image. This gives the preference score for each breakdown of each image.

Overview

Results from the analysis of the perceptions derived from the various sets of subjects indicate that the number of breakdowns perceived varies quite widely from image to image. If the number of human breakdowns is large then the search space required for any computerised shape decomposition system will be large to allow an identical decomposition to be created by the computerised system. The search space will also be large for a computerised system matching components from one image against components in other stored images due to the large potential search space.

One factor that we expect to affect the number of breakdowns is the number of degrees of freedom available within an image. 9 images produced at least 17 breakdowns seen by at least one subject and each of these images had a large number of potential components and a large number of possible arrangements of components. The search space for a computerized decomposition system or image component matching system processing these images would be large.

However, the number of breakdowns seen by 2 or more subjects is much more closely grouped than the number of breakdowns perceived by 1 or more subjects. The mode number of breakdowns perceived by 2 or more people is 3 and only one image had more than 8 breakdowns perceived across all **53** subjects. This indicates that there are individual breakdowns seen by only one person but that there exists a core set of breakdowns that is more tightly grouped which will be seen by 2 or more people. These core breakdowns are the breakdowns we aim to focus on and ensure that any computerized system can reproduce them.

Ren et al. [REB00] had between 1 and 4 breakdowns for each image in their analyses. We found our unrestricted breakdown policy coupled with consolidating Ren et al.'s two-stage experimental process into a single stage allowed more scope for subject variation.

Analysis

In the following, we analyse the core set of breakdowns for each image seen by 2 or more subjects. Qualitative analysis of individual results yields a number of insights that we expect to prove useful in subsequent phases of the PROFIT project.

From analyzing the **53** subjects' drawings, we noticed that the subjects may be focused purely on eliciting the component breakdowns of each image probably due to the experiment focusing on image decomposition. We feel they may concentrate on the individual components and do not always see the "larger picture". For example, where 6 triangles are arranged in a hexagonal shape many subjects drew 6 triangles but not the overall hexagonal shape. We feel that this should be taken into consideration when using the component breakdowns.

The main empirical findings from the human decompositions produced from our experiments are:

Singularity – changing the orientation of image components changes the perception. This is particularly true for textures where altering the angle of the texture can change the figure/ground perception (see the discussion below regarding figure/ground for an example). Also, familiar image components such as human figures or aircraft (see Table 1) are less often perceived when not in their natural orientation.

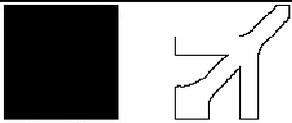
			
	Score		Score
	0.567		0.242

Table 1 showing the image (left column) and modified image (right column) in the top row and the top decomposition for each image (as seen by 2 or more subjects) along with the associated scores in the lower row.

Familiarity – when elements of an image are gradually removed/reorganized so as to destroy familiarity of the image then the human breakdowns change to be based on individual components rather than the entire image and tend to proximity-based grouping as shown in the example in Table 2.

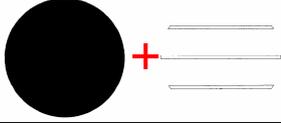
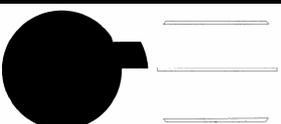
			
	Score		Score
	0.451		0.481
	0.407		0.346
	0.055		0.123

Table 2 showing the familiar image (left column) and the less-familiar modified image (right column) with the top 3 decompositions and their associated scores below.

Symmetry – when symmetry is removed from an image, the human decompositions tend to individual components or image halves. This is particularly true for illusory contours and images where axial symmetry is removed (as shown in the example in Table 3) although there are exceptions where the removal of symmetry has little effect on the decompositions particularly for images that trace the outlines of shapes as seen in Table 4.

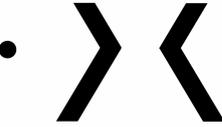
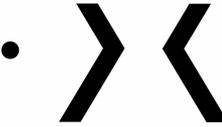
			
	Score		Score
	0.333		0.564
	0.311		

Table 3 showing the symmetrical image (left column) and asymmetrically modified image (right column) with the top decompositions for each image below and their associated scores.

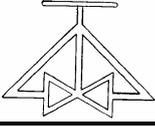
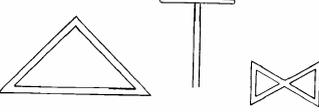
			
	Score		Score
	0.237		0.205

Table 4 showing the symmetrical image (left column) and asymmetrically modified image (right column), with the top decomposition and associated scores in the lower row.

Continuity – reducing the continuity alters the human perceptions with a tendency to proximity grouping and decomposition into individual components. This is particularly true for illusory contours such as Necker cubes where only relatively minor perturbations of the image remove the perception of the cube (see Table 5).

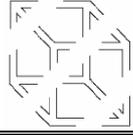
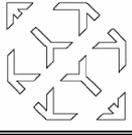
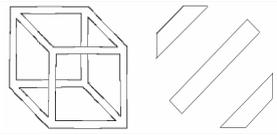
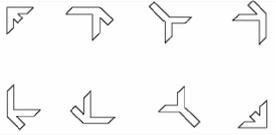
			
	Score		Score
	0.309		0.466

Table 5 showing the Necker Cube image (left column) and modified discontinuous image (right column) with the top decomposition and associated scores in the lower row.

When continuity is reduced in conjunction with symmetry removal then the decomposition differs from when continuity alone is removed. An asymmetric image promotes the perception of good continuity whereas a symmetric variant of the image promotes proximity grouping as seen in Table 6.

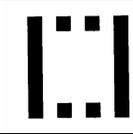
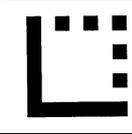
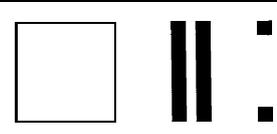
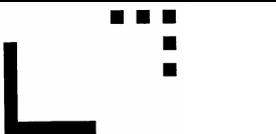
			
	Score		Score
	0.433		0.656

Table 6 showing the symmetric image (left column) and modified asymmetric image (right column) with the top decomposition for each image below and their associated scores.

Figure/ground – if the components of an image are tilted or inverted then the figure/ground perception changes as exemplified in Table 7. If the components are textured with stripes then the figure/ground perception changes from the untextured image and if the texture is strengthened with a darker texture then the figure/ground perception changes even more. This is shown in the example in Table 8. A uniform background enhances the perception of figure/ground reversal whereas familiarity of image components reduces the figure/ground reversal.

	Score		Score
	0.355		0.527
	0.289		0.418
	0.276		

Table 7 showing the image (left column) and modified image (right column) and all decompositions for each image with their associated scores in the lower rows.

	Score		Score
	0.575		0.466
	Score		Score
	0.424		0.25

Table 8 showing the image (top left) and 3 modified textured images coupled with the top decomposition for each image and its associated score.

Conclusion & Future Work

Our results concur with previous investigations such as [REB00] in that image decomposition appears to follow a set of perceptual principles analogous to the Gestalt laws. The experiments and analyses show that these Gestalt laws interact and possibly conflict as noted by [DMM04]. The experiments also indicate that there are a core set of decompositions for each image perceived by 2 or more people along with a set of decompositions seen only by individuals.

We have identified some possibilities for additional work that would generate useful data. The experimental analyses detailed in this paper are very human-oriented. Humans generate all the breakdowns with no recourse as to whether they are feasible for a computer system to generate. Therefore, after we have used the data from these analyses to develop and refine our computational system, we could use the resultant system to generate a set of breakdowns for further images. We can then present these sets of breakdowns, for each image in turn, to human subjects who can rank them *1 to n* where *n* is the number of images in the set. This will allow us to fine-tune the computational system further using tangible computer-generated breakdowns.

Acknowledgement

This work was supported by E.U. FP6 IST **Project Reference: 511572 - PROFIL**.

References

- [DMM04] Desolneux, A., Moisan, L., and Morel, J.-M. A theory of digital image analysis. 2004. Book in preparation
- [E01] Eakins, J.P. Trademark image retrieval. In M. Lew (Ed.), *Principles of Visual Information Retrieval* (Ch 13). Springer-Verlag, Berlin, (2001).
- [G72] Goldmeier, E. Similarity in Visually Perceived Forms [1936], in *Psychological Issues VIII*(1), ed. Herbert J. Schlesinger, International Universities Press, 1972.
- [HR84] Hoffman, D.D. and Richards, W.A. Parts of recognition. *Cognition*, 18:65-96, 1984
- [K79] Kanizsa, G. *Organization in Vision: Essays in Gestalt Perception*, Praeger, NY, 1979.
- [K63] Koffka, K.. *Principles of Gestalt Psychology*. Harcourt Brace. New York, 1963.
- [LC02] Leung, W.H. and Chen, T. "Trademark retrieval using contour-skeleton stroke classification", *IEEE Intl. Conf. on Multimedia and Expo. (ICME 2002)*, Lausanne, August 2002.
- [REB00] Ren, M., Eakins, J. P. and Briggs, P. Human perception of trademark images: implications for retrieval system design. *Journal of Electronic Imaging*, 9 (4):564-575, 2000.
- [SK95] Siddiqi, K. and Kimia, B. B. Parts of visual form: Computational aspects. *Pattern Analysis And Machine Intelligence*, 17(3):239-251, March 1995
- [SSH99] Singh, M., Seyranian, G. & Hoffman D.D. Parsing silhouettes: The short-cut rule. *Perception & Psychophysics*, 61:636-660, 1999.
- [TV02] Tanase, M and Veltkamp, R.C. Polygon Decomposition Based on the Straight Line Skeleton. *Theoretical Foundations of Computer Vision 2002*: 247-267.
- [W23] Wertheimer, M. *Laws of Organization in Perceptual Forms* (1923). In, Ellis (ed) *A Source Book of Gestalt Psychology*, Routledge & Kegan Paul, London 1938.