Visual Saliency and Categorisation of Abstract Images

M. Laine-Hernandez[1], T. Kinnunen[1], J.-K. Kamarainen[3], L. Lensu[2], H. Kälviäinen[2] and P. Oittinen[1]
[1]*Department of Media Technology, Aalto University, Finland (http://media.tkk.fi/)*
[2]*Machine Vision and Pattern Recognition Laboratory (http://www2.it.lut.fi/mvpr)*
Lappeenranta University of Technology (LUT) – [3]LUT Kouvola Unit, Finland

Abstract

Visual object categorisation problem has attracted significant attention during the last ten years, and the two main hypotheses adopted by virtually all methods are i) detection of visual saliency and ii) bag-of-visual-words based categorisation. It is, however, difficult to verify the hypotheses with humans since many recordings, such as gaze fixation locations, represent processing after the recognition and the object classification task is too easy for humans producing no information about uncertainties in the cognitive process.

To the authors' best knowledge, this work is the first attempt to study the main hypotheses and state-of-the-art algorithms for visual object categorisation with abstract images. These images inhibit rapid recognition and cause the observers' opinions differ substantially in assigning the images into "similar categories". Our work reveals interesting findings: the state-of-the-art methods' performances drop to almost pure chance while human observers remain surprisingly consistent.

1 Introduction

The most visual object categorisation (VOC) methods achieving state of the art performance are based on local feature detectors and object class models inferred from feature descriptions (e.g., [10, 1, 11]). These methods are based on the hypothesis that object detection and recognition are bottom-up processes driven by automatically detectable local image regions which are distinguishable due to their visual saliency. To capture the relevant characteristics of the human visual system (HVS) for computational methods, saliency detection has been under active investigation [5, 6]. Despite the research on well-founded visual models, there exists experimental evidence against them. For example, image centroids [6] and recognition of objects prior to saliency detection [3] predict gaze fixations better than any available detector, supporting top-down processing.

Computational methods and psychophysical experiments should be developed together. In this work, we advance toward this ideal situation with an approach easier to adopt: we introduce a dataset which consists of *i)* 100 abstract images (see Fig. 1), *ii)* visual attention (saliency) ground truth constructed from eye fixation patterns of 24 observers, and *iii)* visual category ground truth constructed from free-form class assignments of 20 participants. Abstract content inhibits rapid recognition and the observers' opinions differ substantially in assigning the images into "similar categories".

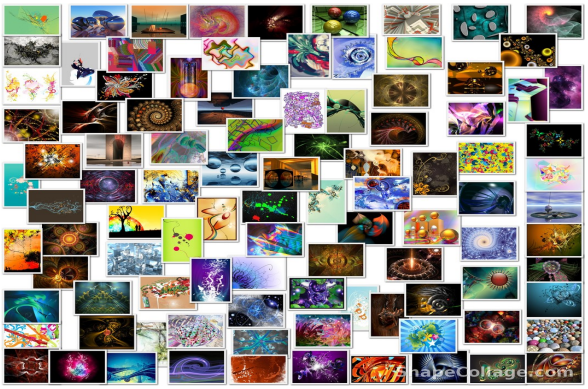

Figure 1: The 100 images of our abstract data set.

Ground truth recordings using human observers is time consuming, and thus, the number of images is limited. However, the proposed dataset will be made public and is easy to extend. Our experiments with the state-of-the-art saliency detection [5, 6] and unsupervised visual object categorisation [9] reveal interesting findings related to the limitations of the computational methods. Our main contributions are the construction of the new abstract image dataset and the special experimental setups for establishing the ground truth, and investigation of the state-of-the-art saliency detection and image categorisation methods which provided important new findings: human observers are surprisingly consistent even with abstract images while the performances of the computational methods collapse.

2 An abstract image dataset

Abstract images for any psychophysical experiments should be plausible for observers. Therefore, instead of using generated or random images, we downloaded images from various sources of artistic images[1]. The images were selected from categories labelled as *abstract* or *surreal*. The total of 250 colour images were of sufficient visual quality and non-recognisable abstract content and 100 images (Fig. 1) were selected randomly for the psychophysical experiments. 24 university students without a background in art and with normal vision were selected for our saliency experiment, and 20 students for the categorisation experiment.

2.1 Visual saliency ground truth

The stimuli (images) were presented using a 22" LCD display located at the distance of 62cm. The viewing distance was controlled with a forehead rest. The display resolution was 1680×1050px. The stimuli subtended horizontally and vertically between 13.0 to 23.5 degrees. The participants' eye positions were recorded with the SMI iViewX 250 infrared binocular remote eye-tracker with a sampling frequency of 250Hz. The data were acquired binocularly, but the eye-tracker was calibrated always to the left eye. In post-test processing, the more accurate eye was identified for each stimulus trial based on the amount of deviation from a pre-stimulus fixation cross, and the data from that eye was used in the analysis. To study the effect of colour, random images were presented in grey-level (luminance).

The eye-tracker was calibrated using nine-point calibration. After the calibration, as well as after every 20 images, a four-point calibration validation was carried out. If the gaze deviated from any of the four validation points more than 2 degrees of the visual angle, a re-calibration was evoked. Each stimulus image was preceded by a central fixation cross that appeared for two seconds. Again, if the gaze deviated from the fixation cross more than 2 degrees, the re-calibration was automatically evoked. The participants were also encouraged to take a break after every ten images in order to rest their eyes, stretch, etc.

The stimulus presentation was programmed using the Presentation™ software (Neurobehavioral Systems, www.neurobs.com). Post-test saccade and fixation detection was carried out using BeGaze 2.4. Thresholds to detect saccades were set to the minimum velocity of 40deg/s and the minimum duration of 22ms. The minimum fixation duration was set to 50ms. With 1.5 degrees of the viewing angle as the accuracy threshold, 96% of the trials produced accurate fixations.

[1] digitalart.org, caedes.net and sxc.hu

Each participant saw all the 100 (50 colour and 50 grey-level) test images in a random order for five seconds each. The task was to *"study the images carefully"*. This instruction is used in other similar eye movement studies (e.g., [2, 7]) being fairly general, but preventing the subjects from adopting individual viewing strategies (such as trying to guess the purpose of the experiment) more than the instruction to "freely view".

To form a single ground truth per image, the fixation locations of the participants were summed up for each image. These fixation locations were then filtered using a Gaussian filter to produce a more continuous saliency map. The variance of the filter was set so that the width at half of its maximum height equals one degree of the visual angle in the setup used in the experiment. The final ground truth saliency map was normalised by dividing its values by the total sum (see Fig. 2).

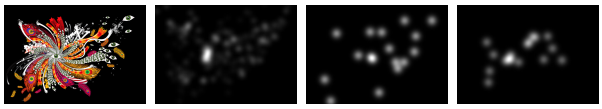

Figure 2: Image under test (left), average (middle) and two individual saliency maps.

2.2 Visual categorisation ground truth

For the image categorisation task, the test images were printed at the resolution of 180dpi and placed onto grey cardboards for easier handling. The length of the longer edge of the printed images varied between 11.2 and 12.7cm. The image orientation was marked with a small circle at the bottom of each cardboard. Again, the task was defined to try to avoid any bias due to the observers guessing the purpose of the experiment. The images were handed to the participant in a random-ordered pile together with the following written instruction:
"*Sort the images into piles according to their similarity so that similar to each other are in the same pile. There are no 'correct' answers, but it is about what you experience as similar. You decide the number of piles and how many images there are in each pile. An individual image can also form a pile. There is no time limit for the experiment.*"

For this task, the opinions varied significantly: the number of stacks varied from 4 to 39 with the average of 21. Respectively, the number of images in stacks varied between 1-37 (mean 5). Using the counts of image pairs in the same stacks, a similarity matrix was made. The matrix was normalised by the number of participants.

The similarity matrix forms the basis of ground truth. The remaining question is how to evaluate the subjects or computational methods producing a different number of categories (stacks). This was solved by using the normalised pair-wise similarities to construct the optimal hierarchy tree of a requested number of leaf nodes.

In our experiments, the construction was made by the agglomerative clustering with the average link distance criterion [4] (see Fig. 3 for an example).

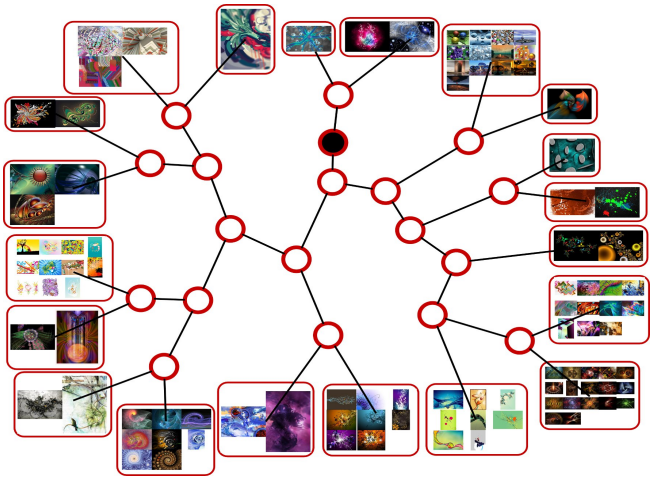

Figure 3: Categorisation ground truth for 20 categories. The black node denotes the root node (Sec. 2.2).

3 Visual saliency of abstract images

The state-of-the-art performance of visual saliency detection for real scene images with fixation ground truth similar to our case was reported by Judd et al. [6]. They pointed out the importance of the image centre to predict the saliency pattern of an image. Judd et al. proposed a saliency model which is based on learnt cues and the distance-from-centre rule. We selected the following methods for our experiments (see Fig. 4): i) Judd et al. [6] (state-of-the-art), ii) Itti and Koch [5] (baseline), and iii) Distance-from-centre rule. Our performance evaluation was based on the receiver operating characteristic (ROC) curve similar to [6]. Since our abstract images contain strong colour cues for which the saliency methods are insensitive the inter-subject evaluation is reported for the both colour and grey-level images using leave-one-out. The performance curves for all methods and inter-subject experiments are in Fig. 5.

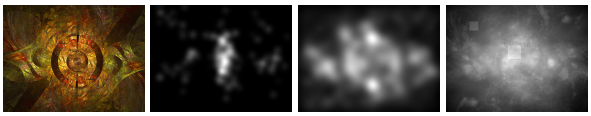

Figure 4: From the left: the input image, ground truth saliency, Itti&Koch [5] and Judd et al. [6].

The first important finding is that colour does not play an important role in saliency detection as the inter-subject performances for colour images (the red curve) and grey-level images (magenta) are almost identical. It is noteworthy that these curves were much better for the real scene images in [6] (> 0.7 for 10% of the pixels detected as salient). However, while the difference between the distance-from-centre and the Judd et al. method was clear in [6], for the abstract images the curves are congruent (green and blue) until a very high number of pixels are detected as salient. The state-of-the-art method does not seem to add anything relevant to the distance-from-centre rule.

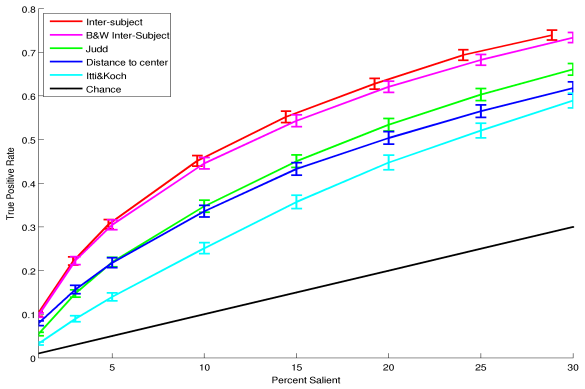

Figure 5: ROC curves for saliency detectors. The vertical bars denote the standard deviations.

4 Visual categories of abstract images

The visual object categorisation (VOC) problem has been one of the most actively investigated topics in computer vision in the recent years. For the state-of-the-art methods using discriminative learning our case has the following difficulties: the lack of sufficient number of training examples and the remarkable uncertainty in the assignments used to construct the ground truth (cf. the practically 100% performance for real objects). Therefore, our data and task resembles more the problem of *unsupervised* visual object categorisation (UVOC) where the classes, the assignments per image and/or hierarchies, should be found automatically. Tuytelaars et al. [9] provided state-of-the-art evaluation for 20 object categories from the popular Caltech-256 dataset. The best results were achieved with a variant of the bag of visual words (BoW) which we also selected. The method uses the Hessian-Laplace interest point detector and the SIFT descriptor to extract and encode local image features. A visual vocabulary (codebook) was built by clustering the local features into 1000 clusters with the K-means algorithm. For each image, a feature was constructed by computing a histogram of the found codes and normalising with the L2 norm. To form the categories, the histogram features were clustered by the K-means with a fixed number of clusters.

The UVOC performance evaluation is also more

complicated and two evaluation approaches have been proposed: the conditional entropy of class assignments by Tuytelaars et al. [9] and the hierarchy-based performance by Sivic et al. [8]. Since these two performance measures report results differently, we report them both in Fig. 6. Note that the shapes of the graphs are not of interest since the selection of the number of categories is still an open problem in UVOC and varied also between the human observers (blue dots in Fig. 6). There are two important findings: 1) every human observer performs much better (leave-one-out) than the computational method (green curve) and 2) the state-of-the-art performance is much closer to the pure chance (red curve) than to the average human performance (blue curve denoting the linear fit of the individual inter-subject performances), which was not the case with real scene images. These findings hold for the both performance measures.

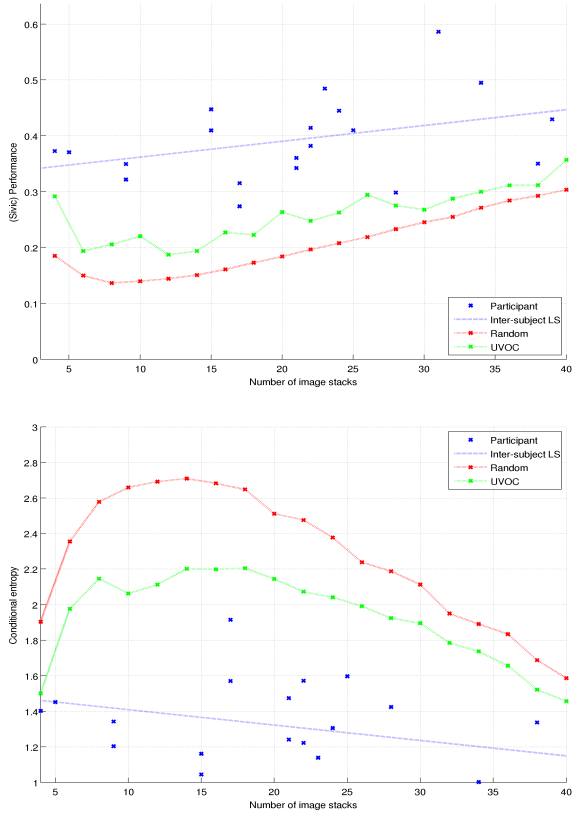

Figure 6: Categorisation performance. Up: Sivic et al. performance (larger value is better). Bottom: Tuytelaars et al. conditional entropy (smaller value is better).

5 Conclusion

We investigated uncertainties involved in visual saliency detection and object categorisation by introducing a novel dataset which inhibits the ground truth bias related to datasets of natural scenes, more precisely, easily recognisable objects which rapidly attract our attention and lead to fast decision making without uncertainty. Our dataset contains abstract images for which the ground truth of saliency and class assignments were collected by appropriate experimental setups. Our experiments revealed the interesting findings:

- Inter-subject saliency is clearly worse for the abstract images than for natural scenes (see [6]).
- Performance of the saliency detectors collapse, in particular, their performance is indistinguishable from a simple distance-to-centre rule, which was not the case with real images [6].
- The collapse cannot be explained by the omittance of colour information since observers viewing only grey-level versions of images performed equally.
- Inter-subject categorisation is clearly worse for the abstract images than for real objects (see [9]).
- Performance of a state-of-the-art method [9] was poor, actually close to pure chance.

References

[1] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial bag-of-features. In *CVPR*, 2010.

[2] W. Einhauser and P. König. Does luminance-contrast contribute to a saliency map for overt visual attention? *European J of Neuroscience*, 17(5):1089–1097, 2003.

[3] W. Einhauser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14)(18), 2008.

[4] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 3 edition, 2011.

[5] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[6] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[7] M. Nyström and K. Holmqvist. Semantic override of low-level features in image viewing - both initially and overall. *Journal of Eye-Movement Research*, 2(2), 2008.

[8] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *Proc. CVPR*, pages 1–8, 2008.

[9] T. Tuytelaars, C. Lampert, M. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *Int J Comput Vis*, 88(2), 2010.

[10] J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE PAMI*, 32(7), 2010.

[11] L. Zhu, Y. Chen, and A. Yuille. Unsupervised learning of probabilistic Grammar-Markov models for object categories. *IEEE PAMI*, 31(1), 2009.