Saliency Identified by Absence of Background Structure

Fred W. M. Stentiford
Electronics & Electrical Engineering Dept, University College London, Gower St, London, UK
f.stentiford@ucl.ac.uk

ABSTRACT

Visual attention is commonly modelled by attempting to characterise objects using features that make them special or in some way distinctive in a scene. These approaches have the disadvantage that it is never certain what features will be relevant in an object that has not been seen before. This paper provides a brief outline of the approaches to modeling human visual attention together with some of the problems that they face. A graphical representation for image similarity is described that relies on the size of maximally associative structures (cliques) that are found to be reflected in pairs of images. While comparing an image with itself, the similarity mechanism is shown to model pop-out effects when constraints are placed on the physical separation of pixels that correspond to nodes in the maximal cliques. Background regions are found to contain structure in common that is not present in the salient regions which are thereby identified by its absence. The approach is illustrated with figures that exemplify asymmetry in pop-out, the conjunction of features, orientation disturbances and the application to natural images.

Keywords: Visual attention, saliency, pattern recognition, similarity, pop-out, cliques

1. INTRODUCTION

Visual attention is commonly modelled by attempting to characterise objects using features that identify them as special or in some way distinctive in a scene. There is a considerable body of research that has pursued this methodology with a large measure of success in certain areas.

Osberger *et al*[1] identifies perceptually important regions by first segmenting images into homogeneous regions and then scoring each area using five intuitively selected measures. The approach is heavily dependent upon the success of the segmentation and in spite of this it is not clear that the method is able to identify important features in faces such as the eyes.

Luo *et al*[2] also devise a set of intuitive saliency features and weights and use them to segment images to depict regions of interest. Some higher level priors are used such as skin colour and selected images are used to normalise feature measurements. Itti *et al*[3] have defined a system which models visual search in primates. Features based upon linear filters and centre surround structures encoding intensity, orientation and colour, are used to construct a saliency map that reflects areas of high attention. Supervised learning is suggested as a strategy to bias the relative weights of the features in order to tune the system towards specific target detection tasks. Itti's work has provided a basis for performance comparisons reported in many papers on visual attention.

Stentiford[4] compared small local groups of pixels with others elsewhere in the image to detect unusual structure and hence a measure of saliency. Kadir *et al*[5] measure the entropy of the local distribution of image intensity. High entropy indicates high local complexity and hence high saliency. The study by Le Meur *et al*[6] lays emphasis on the considerable bias of observers towards looking at the central parts of images where perhaps the photographer usually places the subject. Le Meur *et al* also take account of visual masking in their model as it is known that the differential sensitivity of the human visual system is dependent on the absolute values of parameters such as spatial frequency. Harel *et al*[7] proposed a graphical model in which nodes corresponded to image locations and the edges represented feature based measures of dissimilarity.

Gao *et al*[8] use the feature decomposition of Itti *et al*[3] and saliency is determined from the discrimination obtained from the mutual information between centre and surround. Gopalakrishnan *et al*[9] apply features based on colour and orientation to characterise salient regions whereas Valenti *et al*[10] employ features based on the edges of colour regions and their curvature. Fang *et al*[11] divide the image into patches and identify saliency where a patch differs from those elsewhere in the image attaching a greater weight to patches that are closest.

Liu *et al*[12] use an intuitively selected set of features to train a classifier to identify salient objects. Zhang *et al*[13] also gathers statistics from a set of natural scenes to train sets of features used to estimate saliency. Saliency is indicated if features in a region are comparatively rare in the background. In a similar fashion Bruce *et al*[14] use 3600 natural images to prepare a set of basis functions and identify saliency using the likelihood of content within a region on the basis of the surround. It might be argued that methods employing the statistics of external images could be to some extent reflecting top-down information into the image being analysed.

Oliva *et al*[15] construct contextual features that guide attention towards specific targets such as people. Detecting irregularities as salient necessitates top-down knowledge of what is regular. Boiman *et al*[16] search for patch ensembles common to a database and the candidate image. Regions that cannot be composed from ensembles in the database are considered irregular. Patch configurations are compared according to their descriptors and their relative positions to an origin point. Hou *et al*[17] rely on frequency domain processing in which the difference between the original log spectrum and a prior averaged spectrum is transformed back into the spatial domain as the saliency map.

The outline of research into visual attention given here is by no mean exhaustive as there is currently a great deal of activity in the field. For example, many of the above approaches and others are surveyed by Borji *et al*[18] and Frintrop et al[19]. Many researchers [e.g.[20,21]] now feel that although much pre-attentive visual behaviour can be modelled fairly accurately, human visual attention is also driven by other factors.

2. ATTENTION MECHANISM

Detecting saliency is easiest when the properties of the interesting region are known and appropriate top-down features can be relied upon to identify the region. In the absence of such knowledge bottom-up processes may be appropriate, but in any case certain preselected features are used. In cases where nothing is known in advance about the attentive object, one can never be sure that the chosen features will be suitable for the task.

However, it is not unreasonable to assume that where some saliency is present, the background in the scene possesses significant self similarity, for if not, no region would be more salient than any other. This means that if sufficient of the background can be identified and isolated, whatever is left must be salient. This approach therefore makes no assumption about the object and only requires a mechanism for measuring how well (background) regions reflect into themselves.

A measure of similarity that does not employ *a priori* features has been used to recognize movie posters[22] and to model the Poggendorff and Kanizsa illusions[23]. By detecting commonality between pairs of images the problems of representative training data and feature extraction are avoided. Furthermore the restrictions imposed by the geometry of some feature spaces allow a greater ability to model real pattern relationships. e.g. the triangle inequality prevents A being similar to B, and B to C, but not C to A. However, this does mean that similarity is defined only between pairs of patterns and the details of the commonality are different between all pairs. In this paper instead of comparing different patterns and measuring their similarity, patterns are compared with *themselves*. In this case maximal regions within the same image are compared and if matched and sufficiently large are considered part of the background leaving any remaining regions as salient candidates.

The similarity measure used in this paper is determined by extracting maximal structure that is common to pairs of regions in an image. This structure is not necessarily the largest possible but is optimal in terms of the search algorithm.

Pairs of pixels $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in region 1 and pairs of pixels in region 2 $(\boldsymbol{x}_m, \boldsymbol{x}_n)$ match if brightness, local gradient orientation, and relative orientation lie within certain thresholds where

$$\left|u^{x_i} - u^{x_m}\right| < \delta, \ \left|u^{x_j} - u^{x_n}\right| < \delta \text{ and } u^{x_k} \text{ is the brightness at pixel } \boldsymbol{x}_k, \quad (1)$$

$$\left|\theta_i - \theta_m\right| < \varepsilon_1, \ \left|\theta_j - \theta_m\right| < \varepsilon_1 \text{ and } \theta_k \text{ is the local gradient orientation of pixel } \boldsymbol{x}_k, \quad (2)$$

$$\text{and } \frac{(\boldsymbol{x}_j - \boldsymbol{x}_i)\bullet(\boldsymbol{x}_n - \boldsymbol{x}_m)}{\left|\boldsymbol{x}_j - \boldsymbol{x}_i\right| * \left|\boldsymbol{x}_n - \boldsymbol{x}_m\right|} \geq \lambda \quad (3)$$

The inner product in equation (3) constrains the difference in slopes between the pairs of points in each region to be less than a certain angle $\varepsilon_2 \geq \left|\varphi_{ij} - \varphi_{nm}\right|$ where $\lambda = \cos \varepsilon_2$.

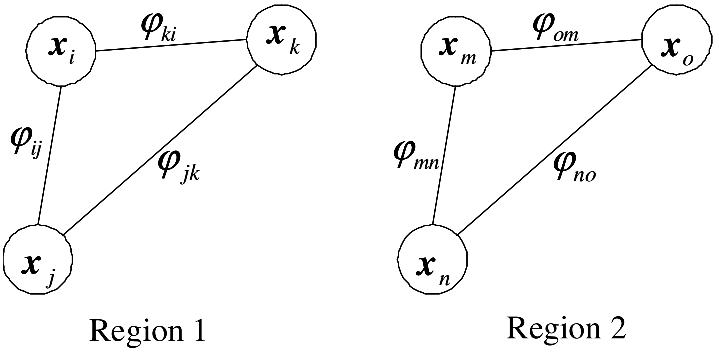


Figure 1. Matching cliques of size 3.

The matching of the pairs of pixels $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ has greater reliability if the pair $\boldsymbol{x}_k$ and $\boldsymbol{x}_i$ also match as this shows that the properties of all three points match according to (1) and (2) and are in the same relative angular position according to (3) in both regions. The three pixels are represented by nodes in a fully connected graph or clique (Fig. 1) with edges representing their angular relationship. Greater reliability is obtained through the associative power of maximal cliques that may be traded off against the precision of the thresholds $\delta, \varepsilon_1, \varepsilon_2$ and therefore obtain easier matching. The measure of similarity of regions is defined as the number of matching pairs of points $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in one of the regions.

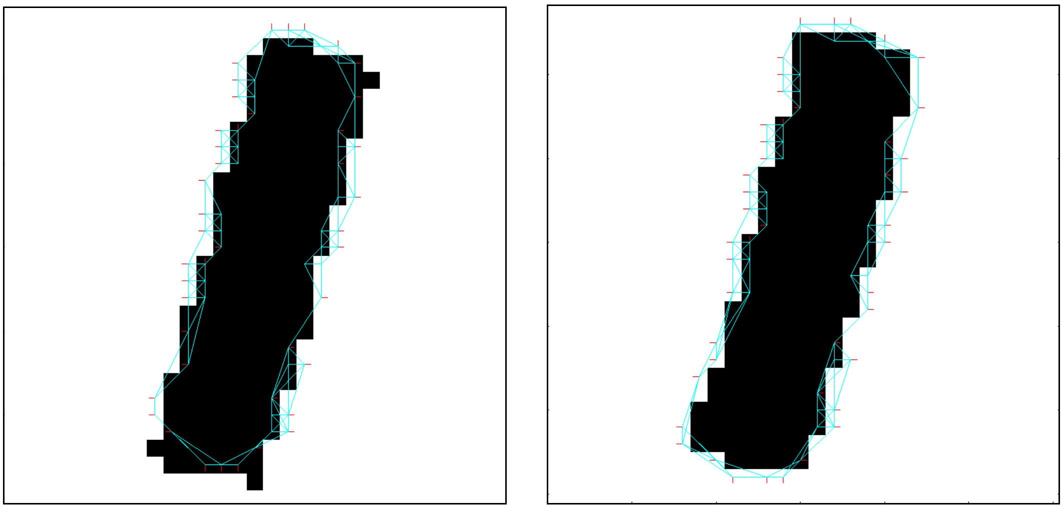
3. APPLICATION TO POP-OUT FIGURES

The similarity measure is used to analyse a number of black and white images exhibiting pop-out effects. Gradient orientations are quantized into just 4 values (0°, 90°, 180°, 270°). Intensities in (1) are not used with black and white images. Thresholds are set as $(\varepsilon_1, \varepsilon_2) = (0, 20°)$. In practice the check on the ε_2 threshold is only applied to the four closest nodes because as distance increases virtually all nodes satisfy the condition if the first four do. The images are compared with themselves and an additional restriction is placed on matching pixel separations

$$\left|\boldsymbol{x}_i - \boldsymbol{x}_j\right| < R \text{ and } \left|\boldsymbol{x}_m - \boldsymbol{x}_n\right| < R \quad (4)$$

where R is chosen to limit the size of the regions being compared.

Figure 2 shows two maximal matching cliques identified in regions 1 and 2 in Figure 3. The edges are represented by cyan lines that connect only the four closest matching pixel pairs (for clarity) and the red lines show the orientation of the local gradient. It can be seen that although the detailed structure of the two shapes is different, the effectively wide thresholds enable a pair of 64-node matching cliques to be found. Figure 4 shows a composite set of cliques of size > 50 nodes that embraces all the similarly shaped objects; no large clique fits the salient target shape that pops out.



a) b)

Figure 2. Matching cliques from a) region 1 and b) region 2 in Figure 3.

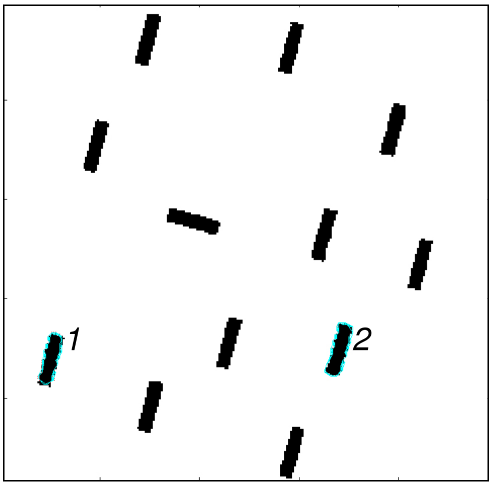


Figure 3. Matching cliques of size 3.

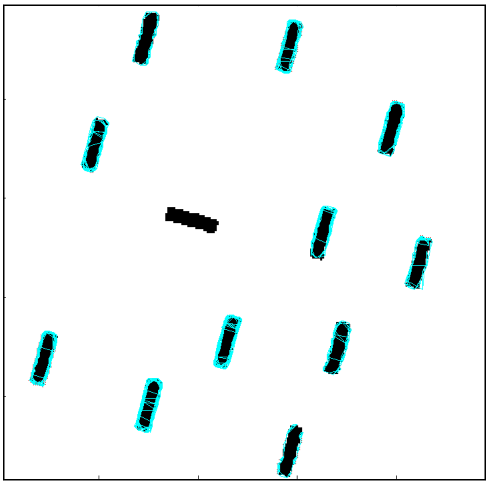
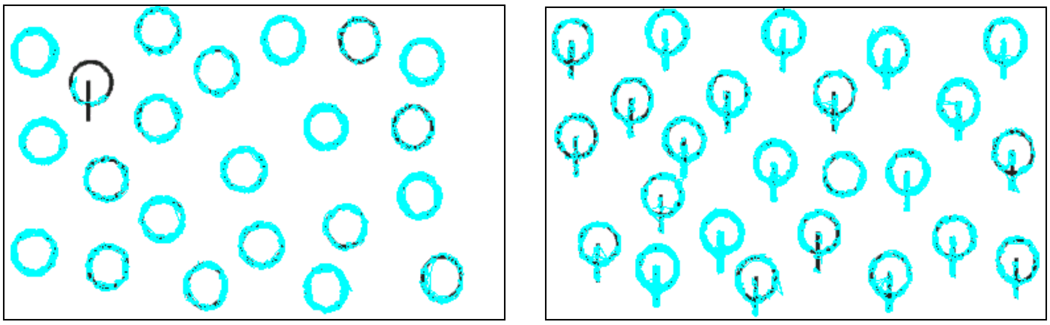
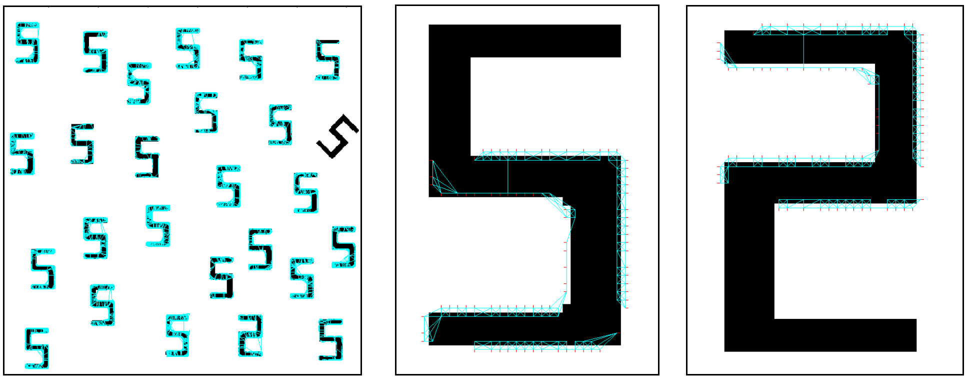


Figure 4. Composite set of matching cliques



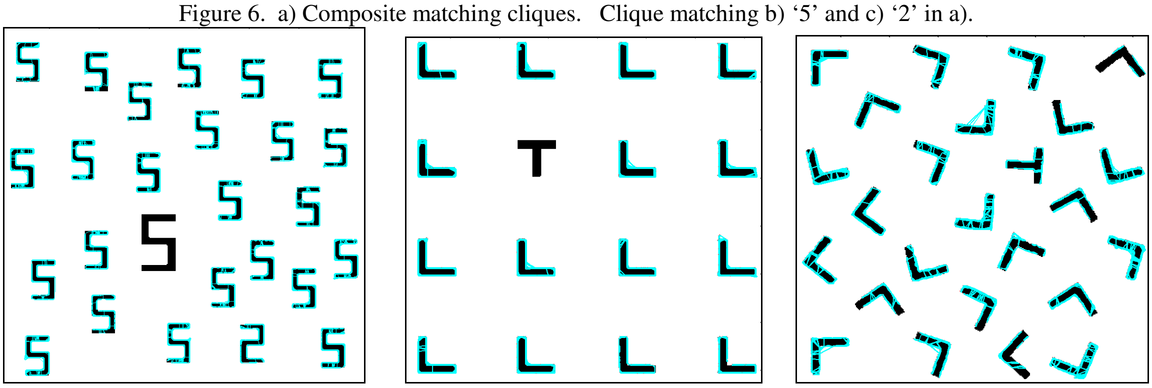
a) b)

Figure 5. Composite matching cliques illustrating pop-out asymmetry



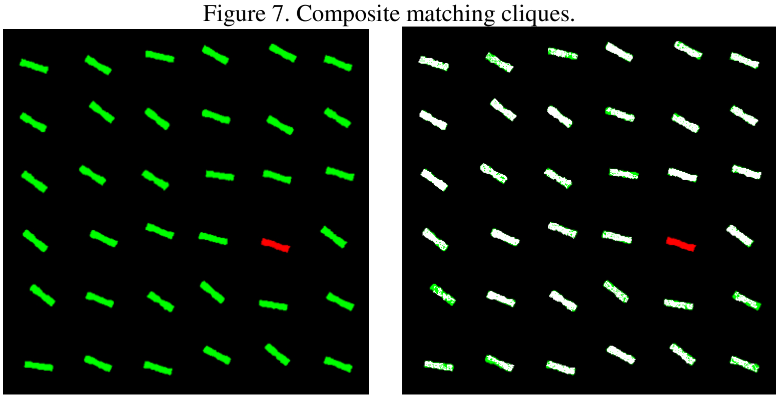
a) b) c)

Figure 6. a) Composite matching cliques. Clique matching b) '5' and c) '2' in a).



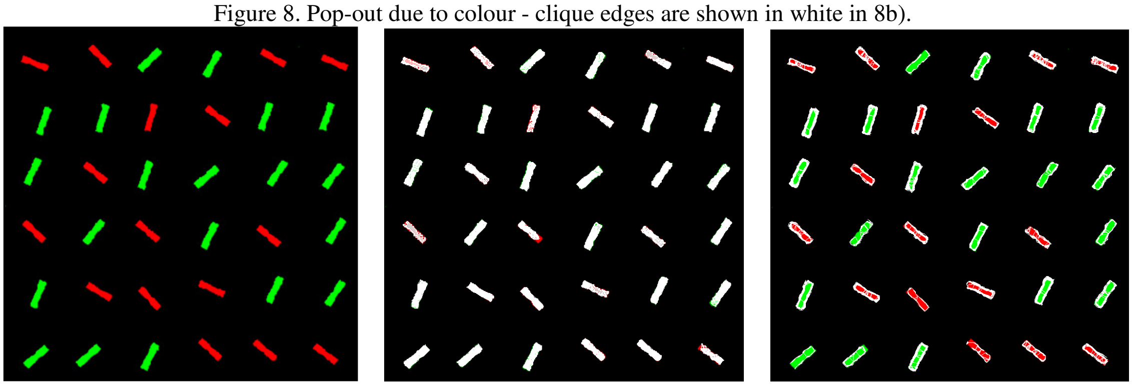
a) b) c)

Figure 7. Composite matching cliques.



a) b)

Figure 8. Pop-out due to colour - clique edges are shown in white in 8b).



a) b) c)

Figure 9. a) Conjunction of colour and orientation. b) colour only. c) gradient orientation only

Figures taken from Gao *et al* [24] also show that the absence of cliques matching a unique feature in the target produces pop-out (Figure 3a). However, the reverse is not the case because there is plenty of opportunity for the circular shape to fit substructure amongst the distractors that also possess the additional feature (Figure 3b). This example illustrates a common asymmetry in which the presence of a feature in the target absent from the distractors produces pop-out, whereas the reverse, i.e. pop-out due to absence in the target, does not hold [25].

Figure 6a) from Wolfe *et al* [26] is another example of pop-out rising from features present in the target but not in the background. However, the '2' in Figure 6a) is a conjunction of features from the background '5's and is not salient without top-down processing [27]. Figures 6b) and 6c) show corresponding cliques that match the bottom part of a '5' and the top part of a '2'. No such cliques are found that represent a conjunction of features in the tilted '5' which remains salient in terms of this model. A similar result is obtained where size is a distinguishing feature in the target (Figure 7) while conjunctions of matching cliques keep the '2' in the background.

Figure 7b) shows that the background 'L' shapes have more structure in common with each other than with the 'T' in that large matching cliques are only found amongst the 'L's. However, when the orientation is disturbed in 7c) there is more opportunity for matches to be found and the saliency of the 'T' is lost [26].

By matching colour in equation (1) rather than brightness and ignoring the requirement to match local gradient orientation in (2), the model can be applied to images where saliency arises from the single feature of colour (Figure 8). When colour and orientation are in conjunction (Figure 9a) none of the shapes are indicated as salient either by colour alone (Figure 9b) or orientation alone (Figure 9c).

4. DISCUSSION

The popout mechanism suggested here identifies attentive structure by reflecting similarity amongst the distractors. This is in contrast to many models of attention that prescribe features that are likely to be rare thereby attributing something special and out of the ordinary to the attentive object. The problem with this approach is that it is never possible to guarantee in advance that the preselected features will produce the required result. Or put another way it is always possible given a particular set of features, to devise an image configuration that will cause the process to fail to extract the attentive object.

The inability of human vision to resolve the red target pre-attentively in Figure 9a can be explained by a deliberate disregard for the conjunction of features. It could be argued that detecting conjunctions would improve human vision in some circumstances. However, there are potentially a huge number of combinations and the additional complexity could be an evolutionary disadvantage [24].

It is necessary to restrict the separation (4) of the pixels used to determine the similarity of regions to obtain the results described in this paper and also to model the Poggendorff illusion [23]. The separation parameter is chosen so that regions and associated cliques obtain approximately the spacing displayed in the figures. This factor is analogous to the classical receptive field in human vision.

An approach that identifies commonality amongst a group of shapes is also consistent with the law of similarity that states that similar objects appear to belong together in human vision. Indeed the pop-out effects above show that a dissimilar object is not grouped together with the distractors and the results are consistent with human visual behaviour reported in the literature. The process of detecting similarity by searching for and detecting maximal cliques that represent structure common to pairs of images could also be seen as analogous to groups of neurons firing simultaneously when visual inputs match those experienced and assimilated earlier.

Some investigations have been carried out on natural images where matching pairs of cliques occupy the background thereby isolating the attentive object. The result is obtained without any prior assumptions about the object or the background (Figure 10). In this case node properties include grey level as well as gradient direction.

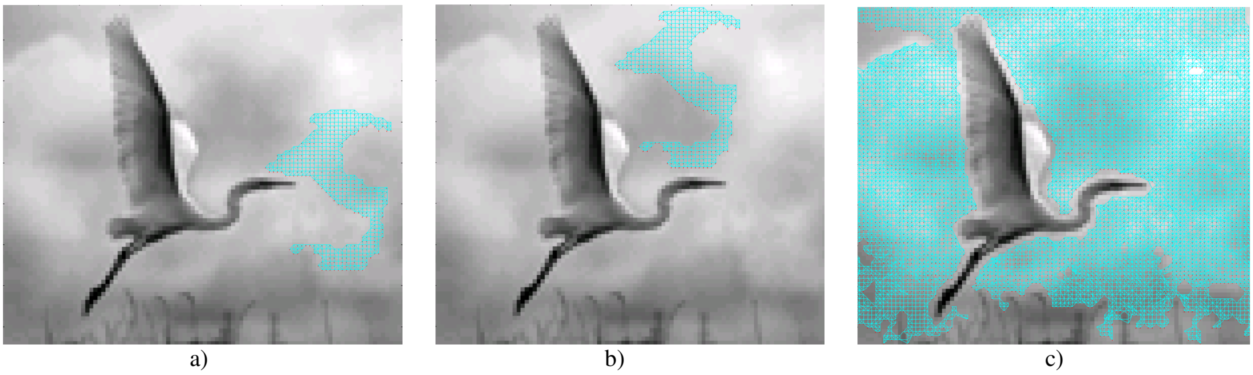


Figure 10. a), b) Matching maximal cliques c) Composite matching cliques

5. CONCLUSIONS

A new similarity measure has been applied to the problem of detecting saliency. It has been shown that large matching cliques exist and can be extracted to reflect similarity between regions within images. When region sizes are restricted this mechanism provides an illustrative model for several pop-out effects in which objects become salient by virtue of the absence of structure that is common in the background. The models illustrated here need further investigation on larger sets of data to establish greater significance, but the measure of similarity does suggest an interesting area for further research into visual attention and certain optical illusions.

REFERENCES

[1] Osberger, W. and Maeder, A. J., "Automatic identification of perceptually important regions in an image," 14th IEEE Int. Conference on Pattern Recognition, 16-20th August, (1998).

[2] Luo, J. and Singhal, A., "On measuring low-level saliency in photographic images," IEEE Conf. On Computer Vision and Pattern Recognition, June, (2000).

[3] Itti, L. and Koch, C., "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention," Vision Research, 40(10), 1489-1506 (2000).

[4] Stentiford, F. W. M. "An estimator for visual attention through competitive novelty with application to image compression," Proc. Picture Coding Symposium, Seoul, 101-104 (2001).

[5] Kadir, T. and Brady, M., "Saliency, scale and image description," International Journal of Computer Vision, 45, 83-105 (2001).

[6] Le Meur, O., Le Callet, P, Barba, D. and Thoreau, D., "A coherent computational approach to model bottom-up visual attention," IEEE Trans. Pattern Analysis and Machine Intelligence, 8(5), 802-817 (2006).

[7] Harel, J., Koch, C. and Perona, P., "Graph-based visual saliency," Proc. Neural Information Processing Systems, (2006).

[8] Gao D. and Vasconcelos, "Bottom-up saliency is a discriminant process," Int. Conf. on Computer Vision, Rio de Janeiro, Brazil, (2007).

[9] Gopalakrishnan, V., Hu, Y. and Rajan, D., "Salient region detection by modelling distributions of color and orientation," IEEE Trans. Multimedia, 11, 892-905 (2009).

[10] Valenti, R., Sebe, N. and Gevers, T., "Isocentric color saliency in images," Proc. IEEE Int. Conf. Image Processing, (2009).

[11] Fang, Y., Lin, W., Lee, B-S., Lau, C-T., Chen, Z. and Lin, C-W., "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," IEEE Trans. Mutimedia, 14, 187-198 (2012).

[12] Liu, T., Sun, J., Zheng, N., Tang, X. and Shum, H. Y., "Learning to detect a salient object," in Proc. Int. Conf. Computer Vision and Pattern Recognition, (2007).

[13] Zhang, L., Tong, M. H., Marks, T. K., Shan, H. and Cottrell, G. W., "SUN: A Baysian framework for saliency using natural statistics," J of Vision, 8(7), 1-20 (2008).

[14] Bruce, N. D. B. and Tsotsos, J. K., “Saliency, attention and visual search: an information theoretic approach,” J. of Vision, 9(3), 1-24 (2009).

[15] Oliva, A., Torralba, A., Castelhano, M. S. And Henderson, J. M., “Top-down control of visual attention in object detection,” International Conference on Image Processing, (2003).

[16] Boiman, O. and Irani, M., “Detecting irregularities in images and in video,” Int. Conf. on Computer Vision, (2005).

[17] Hou, X. and Zhang, L., “Saliency detection: a spectral residual approach,” Proc. CVPR, (2007).

[18] Borji, A. and Itti, L., “State-of-the-art in visual attention modelling,” PP(99), IEEE Trans. PAMI, (2012).

[19] Frintrop, S., Rome, E. and Christensen, H. I., “Computational visual attention systems and their cognitive foundations: a survey,” ACM Trans. Appl. Percept. 7(1), 1-39 (2010).

[20] Mancas, M., “Relative influence of bottom-up and top-down attention,” Attention in Cognitive Systems, Springer, 5395, 212-226 (2009).

[21] Follet, B., Le Meur, O. and Baccino, T., “Modeling visual attention on scenes,” Studia Informatica Universalis 8(4), 150-167 (2010).

[22] Stentiford, F. W. M. and Bamidele, A, “Image recognition using maximal cliques of interest points,” Proc. Int. Conf. on Image Processing, Sept. 26 - 29, Hong Kong, (2010).

[23] Stentiford, F. W. M., “Interest point analysis as a model for the Poggendorff illusion,” Proc. Human Vision and Electronic Imaging XVII, SPIE Conf., San Francisco, (2012).

[24] Gao, D. and Vasconcelos, N., “Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics,” Neural Computation 21(1), 239-271 (2009).

[25] Treisman, A and Souther, J., “Search asymmetry: a diagnostic for preattentive processing of separable features,” J. of Experimental Psychology: General, 114, 285-310 (1985).

[26] Wolfe, J. M. and Horowitz, T., “What attributes guide the deployment of visual attention and how do they do it?,” Nature Reviews Neuroscience, 5, 495-501 (2004).

[27] Treisman, A. and Sata, S., “Conjunction search revisited,” J. of Experimental Perception and Performance, 16, 459-478 (1990).