PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Visual attention: low-level and high-level viewpoints

Fred W. M. Stentiford

SPIE.

Visual Attention: low level and high level viewpoints

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

ABSTRACT

This paper provides a brief outline of the approaches to modeling human visual attention. Bottom-up and top-down mechanisms are described together with some of the problems that they face. It has been suggested in brain science that memory functions by trading measurement precision for associative power; sensory inputs from the environment are never identical on separate occasions, but the associations with memory compensate for the differences. A graphical representation for image similarity is described that relies on the size of maximally associative structures (cliques) that are found to reflect between pairs of images. This is applied to the recognition of movie posters, the location and recognition of characters, and the recognition of faces. The similarity mechanism is shown to model popout effects when constraints are placed on the physical separation of pixels that correspond to nodes in the maximal cliques. The effect extends to modeling human visual behaviour on the Poggendorff illusion.

Keywords: Visual attention, pattern recognition, top-down, bottom-up, popout, cliques

1. INTRODUCTION

Visual attention comes into play in the early stages of human vision often before any conscious recognition takes place, and provides a rich source of information on which efficient recognition can take place. The potential benefits that can arise from a model of attention are manifold and include applications to visual inspection in manufacturing processes, medical diagnosis, spotting security breaches, removing redundancy in data, various targeting applications, and many others.

Computer models of visual attention aim to imitate aspects of the behaviour of the human visual system. The models identify image regions that attract our attention either directly by our gaze or covertly in our peripheral vision. Many models are strictly bottom-up, that is, they rely totally on the information contained within the images in question. Others incorporate top-down methods that allow the statistics of related images to influence the parameters that determine local saliency values. In the extreme, top-down attention becomes recognition when attention is solely directed at a particular class of object and the features characterising those objects are used in the calculation of the saliency measure. In all cases feature measurements are selected that bear some relationship to the first stages of processing thought to take place in the human visual system.

This paper outlines some of the principal approaches taken towards modelling visual attention and indicates where some use of top-down factors is made. Some outstanding issues are described that pervade not just visual attention but the whole of the pattern recognition field. Recent research is reported that begins to address these problems.

2. APPROACHES TO BOTTOM-UP ATTENTION

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.

Optics, Photonics, and Digital Technologies for Multimedia Applications II, edited by
Peter Schelkens, Touradj Ebrahimi, Gabriel Cristóbal, Frédéric Truchetet, Pasi Saarikko,
Proc. of SPIE Vol. 8436, 84360L · © 2012 SPIE · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.923511

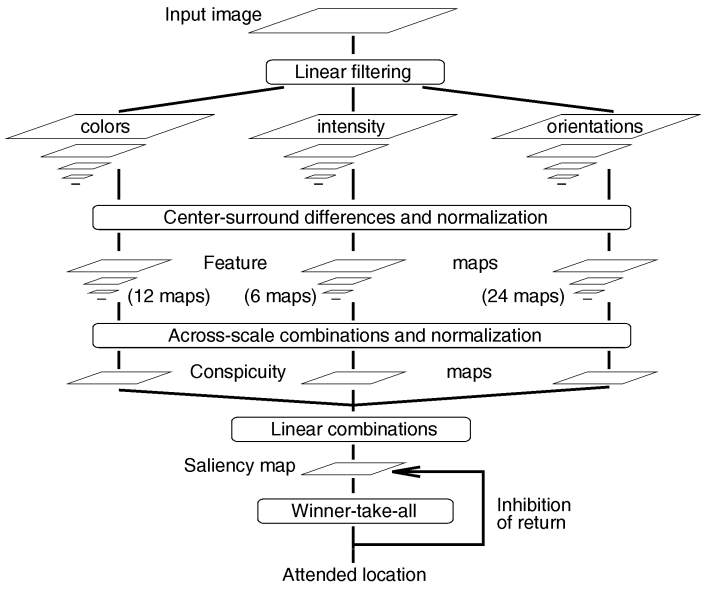


Figure 1. Model architecture from Itti[3]

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 (Figure 1). 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.

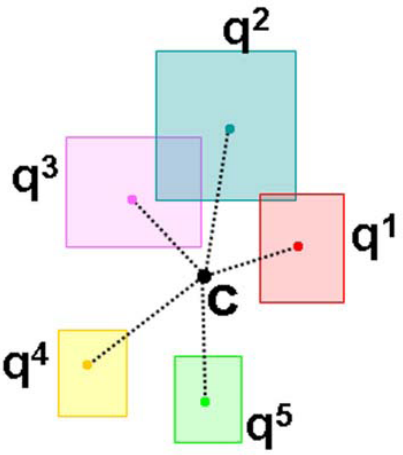


Figure 2 Ensembles of patches[16]

3. INCORPORATION OF FEATURES FOR TOP-DOWN ATTENTION

The use of top-down information is a strategy for improving performance in those cases where it is extremely difficult to deduce the salient regions where the content does not match the features chosen to characterise saliency. For instance, a complex textured background would be a problem for an approach based on entropy, and a method that relied heavily on centre weighting would be in danger of missing content near the edge of the image.

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 (Figure 2).

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.

3.1 Discussion

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. The problems faced in these studies however do represent those faced by the community as a whole. Many researchers [e.g.[18,19]] now feel that although much pre-attentive visual behaviour can be modelled fairly accurately human visual attention is also driven by other factors.

Human attentive behaviour in the first instance can be shown to focus on certain bottom-up features such as orientation, colour and intensity. But very quickly specific features of particular interest to the observer take over. It is very difficult to determine what these features are in any individual case. Sometimes the visual task and context are known and the features may be tuned to a representative image dataset. It is however, still unpredictable what memories and experience each individual possesses that are brought to bear on the attentive behaviour. This means that there may be a tradeoff between maintaining a strictly bottom-up model of visual attention that misses some objects of interest, and a strongly top-down system that highlights a limited class of objects to the exclusion of everything else.

The choice of features that characterise top-down attention is a problem that occurs more generally in pattern recognition where pattern classes are also distinguished by an optimal selection of feature measurements. Although a set of features may be found that obtain a good performance on a restricted dataset, unseen data frequently yield poor results because the features are no longer appropriate. Unless there is complete knowledge of the dataset being classified it is very difficult to obtain a representative subset from which a highly performing set of features can be extracted. In practice, class membership is not necessarily dependent on all patterns possessing a certain set of features. Quite often patterns belonging to a class contain structure in common with only some of the other members of the class and this structure in common might be different for all members of the class. Such relationships cannot be represented by linear

combinations of feature measurements in a feature space. So perhaps rather than finding similarity by reflecting features onto patterns we need to reflect patterns into each other.

Again the problems do not rest there! The search for solutions to pattern recognition problems can involve computation that grows exponentially and inevitably certain heuristics are introduced that restrict the search to practical levels. Such intuitive heuristics indicate the most promising avenues of exploration, but run the risk of unwittingly precluding fruitful parts of the search space. So on the one hand we need a very large space to encompass all the best solutions, and on the other hand we need a mechanism for search that is not exhaustive but does not use very restrictive heuristics.

3.2 Commonality in Vision

Comparing two patterns for their visual similarity in accordance with the human visual system has to take account of translation, scale, orientation to a certain extent, illumination and perspective distortion. Such variability is difficult to handle if feature measurements are used for reasons given above. However, the dependence on precise feature measurements can be minimised by making used of ideas taken from brain science. It is thought that memory functions by trading ultimate precision for associative power[20]. Sensory inputs from the environment are never identical on separate occasions, but the associations with separate details in memory compensate for the differences to achieve satisfactory recollection and recognition. Furthermore it has been established that subsets of neurons that discharge during a particular action also fire when that same action is observed in others[21]. These neurons are known as mirror neurons and show that patterns of sensory inputs are being reflected into those already memorized.

In terms of human vision such associations are spatial and the measurements are local. The dependence on measurement can be minimised by employing a local parameter such as local intensity gradient matched with a wide threshold. This will allow the measurement to be brightness independent. The spatial relationship can be angular again with a wide threshold for a match and gives scale independence and some rotational independence to the system. Associative power is obtained by seeking sets of points in each image that match in intensity gradient, but also reflect the same angular relationship with all others in both images. Such sets of points when viewed as nodes in a graph form cliques within a larger graph and provide evidence of structure common to both images with greater reliability the larger the clique.

4. TOP-DOWN APPLICATIONS

4.1 Movie Posters

Detecting cliques of interest points in photos of movie posters is used as a method of recognition against a reference set[22]. Interest points are located in regions of changing colour values in reference images and the candidate. Pairs of matching points that bear the same relative angle in the reference and candidate images are extracted to define a graph representing the relationships. Software for extracting cliques is applied to find the maximal clique although heuristics are invoked to reduce computation. Recognition is achieved providing the number of nodes in the clique exceeds a threshold (Figure 3). The approach is shown obtain good performance on low quality images where it is not possible to extract SIFT key points.

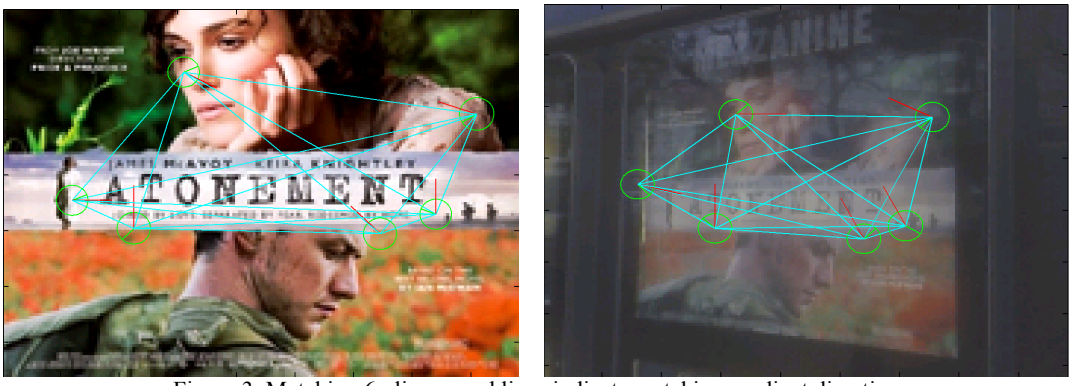


Figure 3. Matching 6-cliques; red lines indicate matching gradient direction

4.2 Text Location and Recognition

The problem of recognising characters in general scenes is difficult if the characters are not already located. It has been shown that characters may be located by extracting matching maximal cliques between a generic character and a window scanned across the image. When the generic character is replaced with particular references recognition as well as location is obtained[23]. Figure 4 shows a 117 point clique matching a larger candidate character illustrating some elastic deformation without breaking any node relationships. Local gradients are quantized into 4 directions without loss of performance.

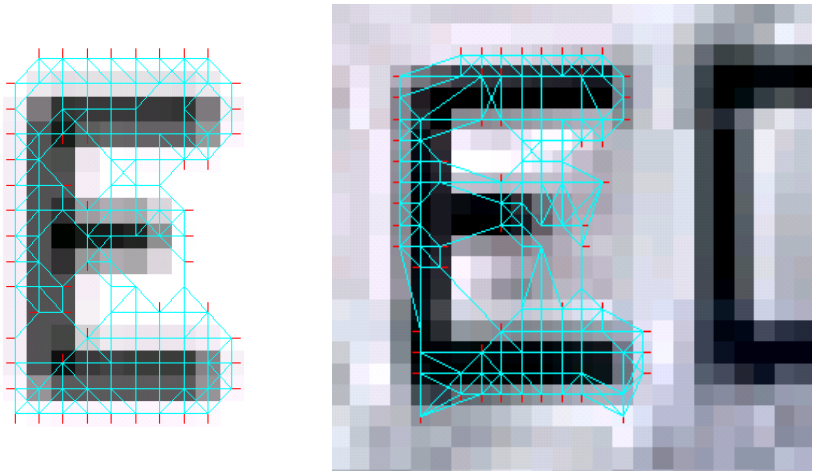


Figure 4. 117 node clique present in small reference and larger scale image.

4.3 Face Recognition

Successful face recognition must handle varying expressions, pose, illumination and other factors that affect image quality. Facial structure that is in common between a reference and variants of that face must be greater that any present between the reference and the faces of other people to be useful for recognition. Shape driven graphical approaches[24, 25] assign fiducial points to nodes and maximise a similarity function to obtain recognition of candidate images. The methods however, require a training set of manually annotated images and it is not clear how different backgrounds affect performance.

Large maximal cliques have been extracted from pairs of faces from the Yale face dataset[26] that allow elastic deformation across varying expressions whilst maintaining discrimination between different people. Figure 5 shows a 50x50 reference face being compared with an image containing four faces at double the definition of the reference. Figure 6 shows the sizes of maximal cliques extracted between the same reference and all 15 normal faces (100x76) in the database and indicates an important score difference between the correct image and the others.

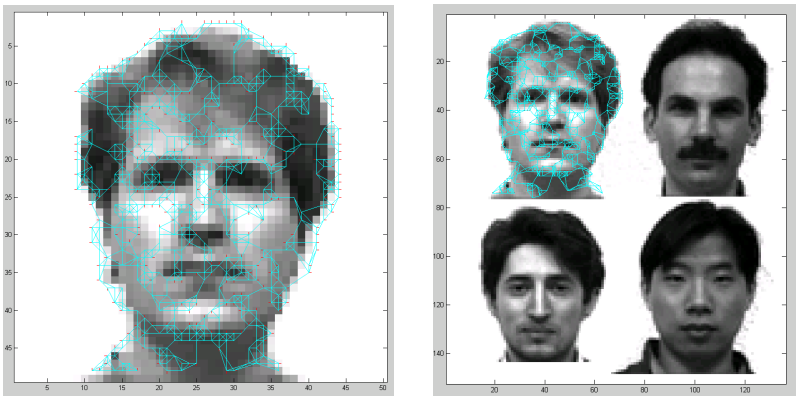


Figure 5. Maximal 585 node clique present in small scale reference and larger scale 4 face image.

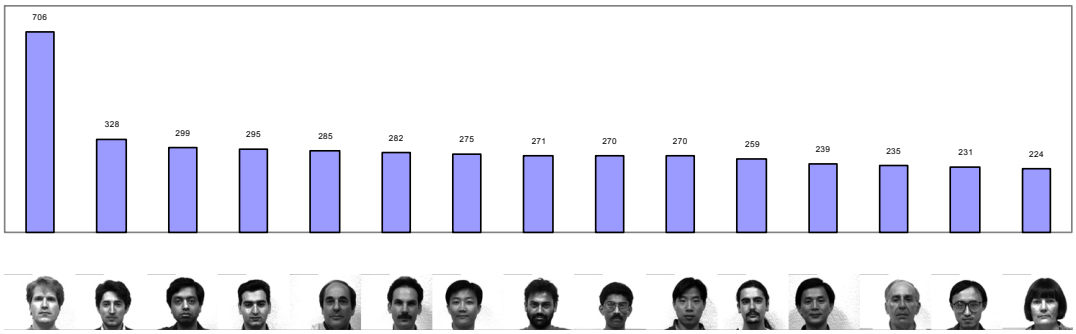


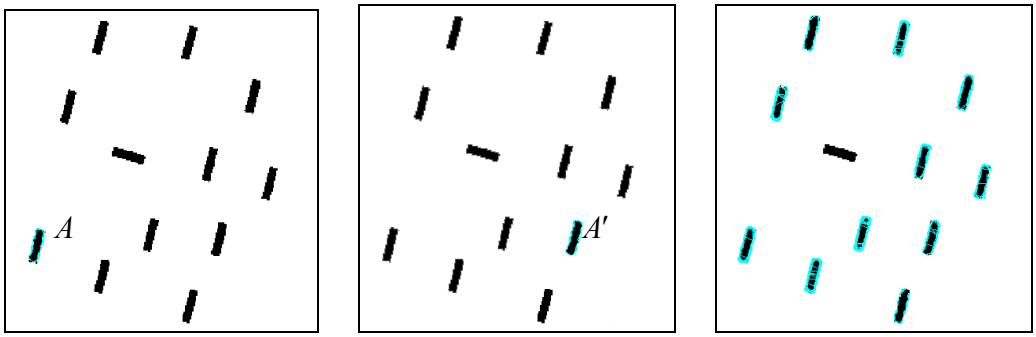
Figure 6. Maximal clique sizes obtained when comparing subject 1 against all in the set.

5. BOTTOM-UP PROPERTIES

Only information contained within an image can come into play when an image is compared with itself. Furthermore restrictions on the separation of pixels reduce the size of pixel cliques that can be matched and top-down influences are thereby reduced. Under these circumstances when a typical image illustrating popout is compared with itself, global structure is not reflected in the maximal matching cliques. Instead smaller local shapes (*A, A′*) are matched that satisfy the restrictions on size (Figures 7a and 7b). The result is that a set of maximal cliques is created that embraces all the similarly shaped objects, but not the dissimilar object that pops out (Figure 7c).

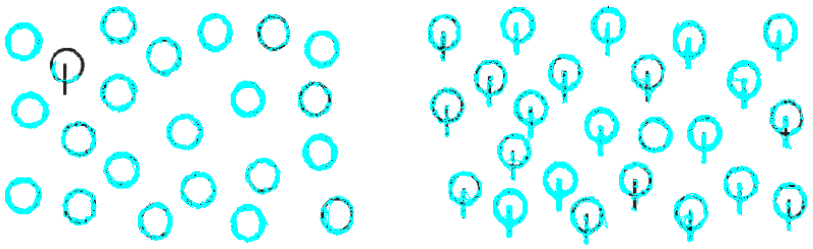
In a similar fashion the absence of matching cliques produces popout (Figure 8a), whereas the reverse is not the case because there is plenty of opportunity for the circular shape to fit substructure in the distractors that possess the additional feature (Figure 8b). In Figure 9c the attentive tilted "5" contains no matching cliques, but the "2" is matched by cliques that match different parts of the distractor shapes (Figures 9a, 9b) and does not popout. Figure 10 illustrates the analysis applied to a natural image 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. In this case node properties include grey level as well as gradient direction.

A further effect arising from the restriction on separation is observed in the comparison of a diagonal line with a series of figures containing the same diagonal but distorted by varying amounts (Figure 11). The right-hand section of the transversal is shifted through a series of vertical steps passing through the collinear position at step 4. The sizes of maximal cliques peaks in the collinear position, but also at a shifted position[27]. This is consistent with psychophysical experiments conducted with the Poggendorff illusion[28].



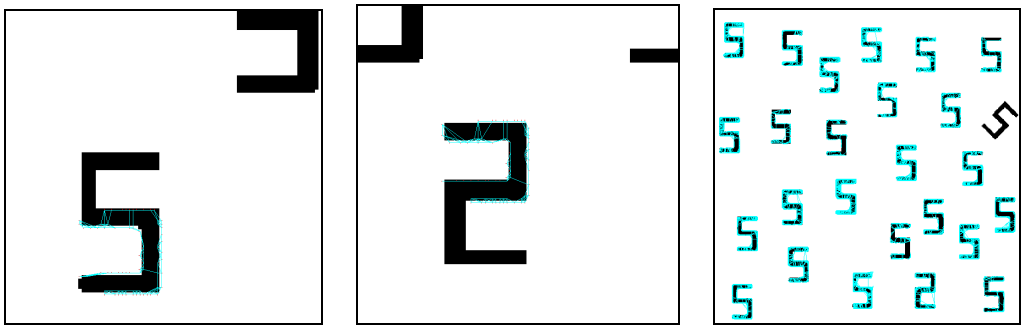
a) b) c)

Figure 7. a),b) Locally constrained matching cliques, c) Composite set of matching cliques.



a) b)

Figure 8. Composite matching cliques illustrating popout asymmetry



a) Matching part of "5" b) Matching part of "2" c) Composite

Figure 9. Matching cliques illustrating suppression of "2" popout

a) b) c)

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

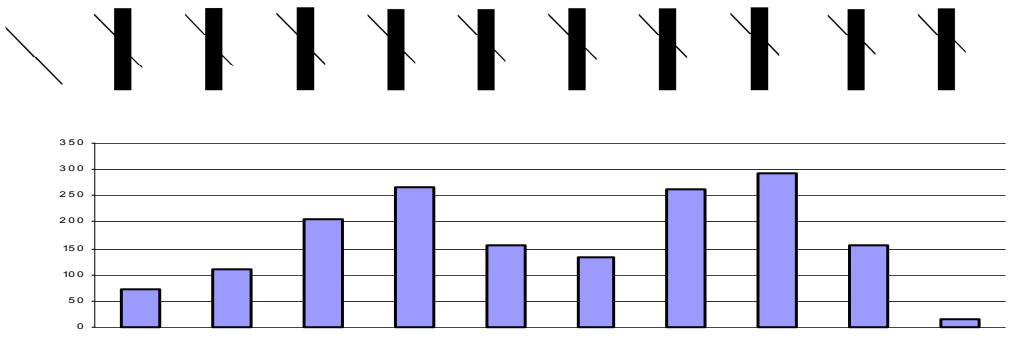


Figure 11. Poggendorff figures and corresponding maximal clique sizes.

5.1 Discussion

The popout mechanism suggested here identifies attentive structure by recognising similarity amongst the distractors. This is in contrast to many models of attention that look for features that are 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 effect. 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.

An approach that identifies commonality amongst a group of shapes is also consistent with the law of similarity that states that similar objects seem to belong together in human vision. Indeed the popout effects above show that a dissimilar object is not grouped together with the distractors. The law of proximity that states that objects near one another appear to form a group would also be consistent with the same mechanism that also restricted the separation of matching cliques as well as the nodes within them.

It might be conjectured that the restrictions on node and clique separations that yield these effects could offer an analogous explanation for the speed of pre-attentive behaviour in human vision. Indeed if a process of recognition can take place among a small set of locally connected neurons, it will be more likely to be completed before one which involves many more neurons at possibly greater physical distances from each other.

6. CONCLUSIONS

It can be concluded that both top-down and bottom-up influences play major parts in the process of visual attention in human vision. It has been shown that large matching cliques exist and can be extracted that reflect similarity between pairs of images. The similarity mechanism also provides an illustrative model for some popout effects in which objects become attentive by virtue of the absence of structure that is common in the background. The models illustrated here need further investigation on much larger sets of data to establish greater significance, but the measure of similarity does suggest an interesting area for further research into visual attention.

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, 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, 28, 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):32, 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):5, 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," In Conf. on Computer Vision, (2005).
[17] Hou, X. and Zhang, L., "Saliency detection: a spectral residual approach," Proc. CVPR, (2007).
[18] Mancas, M., "Relative influence of bottom-up and top-down attention," Attention in Cognitive Systems, Springer, 5395, 212-226 (2009).
[19] Follet, B., Le Meur, O. and Baccino, T., "Modeling visual attention on scenes," Studia Informatica Universalis, 8(4), 150-167 (2010).
[20] Edelman, G. M., [Second nature: brain science and human knowledge], Yale Univ. Press, (2006).
[21] Rizzolatti, G. and Craighero, L., "The mirror-neuron system," Annu. Rev. Neurosci., 27, 169-192 (2004).
[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. and Bamidele, A. "Text detection in natural scenes using cliques of interest points for mobile visual search," International Classification Conference, St Andrews University, (2011).
[24] Wiskott, L., Fellous, J-M., Kruger, N. and von der Malsburg, C., "Face recognition by elastic bunch graph matching," IEEE Trans. Pattern Anal. Machine Intell., 16, 775-779, (1997).
[25] Cootes, T. F., Edwards, G. J. and Taylor, C. J., "Active Appearance Models," IEEE Trans. Pattern Anal. Machine Intell., 23, 681-685 (2001).
[26] http://cvc.yale.edu/projects/yalefaces/yalefaces.html
[27] 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).
[28] Day, R. H., "The Poggendorff illusion and apparent interparallel extents," Perception, 21, 599-610 (1992).