Asymmetry as a Measure of Visual Saliency

Ali Alsam, Puneet Sharma, and Anette Wrålsen

Department of Informatics & e-Learning (AITeL),
Sør-Trøndelag University College (HiST),
Trondheim, Norway
er.puneetsharma@gmail.com

Abstract. A salient feature is a part of the scene that stands out relative to neighboring items. By that we mean that a human observer would experience a salient feature as being more prominent. It is, however, important to quantify saliency in terms of a mathematical quantity that lends itself to measurements. Different metrics have been shown to correlate with human fixations data. These include contrast, brightness and orienting gradients calculated at different image scales.

In this paper, we show that these metrics can be grouped under transformations pertaining to the dihedral group D_4, which is the symmetry group of the square image grid. Our results show that salient features can be defined as the image features that are most asymmetric in their surrounds.

Keywords: Saliency, dihedral group D_4, asymmetry.

1 Introduction

We are frequently surprised by the difference between what we observe in our visual world and the observations of others around us. Commonly, we think of these differences as a product of our varying personalities or interests, i.e., we notice what we think of or like. The fact that we observe different visual realities can, however, be explained in a different manner—we are selective because our brains are limited. In other words, we are selective because our brains do not process all the visual information that surrounds us. In this view, which is supported by psychophysical experiments [1–4], visual selection, or attention, is an information reduction method.

Mathematically, information reduction methods start with a process of identifying the most important aspects of the data, i.e., the parts of the data that cannot be disregarded. As an example both factor analysis and principal component analysis are based on the idea that multi-dimensional data can be represented with a set of limited bases that account for them with limited information loss [5, 6]. Based on this mathematical analogy we might wonder how the reduction of visual information is achieved.

In the literature, two main methods have been proposed: Top-down, also know as attention, and bottom-up or pre-attention visual information reduction

J.-K. Kämäräinen and M. Koskela (Eds.): SCIA 2013, LNCS 7944, pp. 591–600, 2013.

[1, 7–13]. As an example of top-down we might consider the problem of locating an item such as the red book on the bookshelf. Here our visual system would be trying to quickly browse the scene, disregarding any other color. As such, top-down visual reduction is task-driven and voluntary, where we would be looking for an aspect in the scene that matches a mental representation. Bottom-up methods on the other hand are involuntary, faster than top-down and not task-driven. Instead they are driven by the identification of a new, unknown, visual experience. The question that arises in bottom-up approaches is: How do we reduce the visual data of an arbitrary scene?

Most of the bottom-up, pre-attention models share the same basic elements. The basic assumption is that the different regions of the visual information field differ in their visual content. Based on that, an area of the scene that is clearly different from its surround, salient, is thought to represent an anchor point for data reduction. In other words, the visual reduction task is similar to statistical methods such as principal component analysis, where the most salient features of the scene represent the set of bases around which the rest of the scene is arranged. To measure the difference between a center and its surround, a number of stimulus characteristics have been proposed. These include color difference, contrast and orientation. For a given scene, these differences are measured and the results stored in so-called feature maps which are then combined in a so-called saliency map.

While salient feature detection algorithms are progressively more efficient at predicting where a person might look under free viewing conditions, the actual definition of a salient feature and thereby the mechanism of selecting such regions is still debatable. Generally, a salient feature is defined as a region in the scene that is different from its surround. The nature of this difference is, however, loosely defined. As previously mentioned, the difference is measured in terms of a number of metrics pertaining to contrast and gradients or orientation at different spatial scales commonly implemented by means of image pyramid decomposition.

The question addressed in this paper is mathematical, namely, we ask if the differences used in estimating the level of saliency at a given scene location can be grouped in a unified mathematical definition. By examining the metrics used to construct the feature maps, we observe that all can be accounted for by transformations described by the dihedral group D_4. This is the symmetry group of the square image grid and includes two types of symmetries, i.e., rotation and reflection. The transformations defined by D_4 have exhibited immense power in image processing operations including image compression, denoising, and indexing [14–18].

To test the usefulness of the dihedral group in describing salient image features, we constructed a saliency map based on seven elements of D_4, namely, rotation by 90, 180 and 270 degrees and reflection about the horizontal, vertical and two diagonal axes. These transformations were performed on the blocks obtained by decomposing the image into square regions. The results at the higher and lower scales of image were calculated and stored in separate feature maps (details in the theory section). Finally, the feature maps were grouped into a

saliency map in linear manner, i.e., without the use of center surround operations. Having done that, we evaluated the correspondence between the proposed saliency map and human fixations data. Our results show that a saliency map derived based on the transformations of the dihedral group D_4 matches well with human fixation data, and has very high correspondence with the existing saliency map.

Based on these results and the knowledge that the D_4 transformations represent a mathematical measure of symmetry, we conclude with the hypothesis that a salient image feature is a part of the scene that is highly asymmetric compared to its surround and the more asymmetric a feature is the more salient it is. This hypothesis is strengthened by the knowledge that the transformations of D_4 are extremely fast. This latter aspect of the operations is in agreement with the knowledge that bottom-up operations are fast, in the order of 25 to 50 ms [10].

The rest of this article is organized as follows: In Section 2, we discuss the theory behind the dihedral group D_4 and the implementation of the proposed algorithm, in detail. In section, we examine the results obtained from the evaluation of saliency algorithms.

2 Theory

2.1 Mathematical Background

Mathematically, the symmetries of geometric objects can be defined by group theory, and in particular the symmetries of the square are encoded in the dihedral group D_4. In this section we briefly define and describe this group and then show how it can be applied to detect asymmetry in an image.

The Group D_4. A *group* is a set G together with a binary operation $*$ on its elements. This operation $*$ must behave in a very specific way:

i) G must be *closed* under $*$, that is, for every pair of elements g_1, g_2 in G we must have that $g_1 * g_2$ is again an element in G.

ii) The operation $*$ must be *associative*, that is, for all elements g_1, g_2, g_3 in G we must have that
$$g_1 * (g_2 * g_3) = (g_1 * g_2) * g_3.$$

iii) There is an element e in G, called the *identity element*, such that for all $g \in G$ we have that
$$e * g = g = g * e.$$

iv) For every element g in G there is an element g^{-1} in G, called *the inverse of* g, such that
$$g * g^{-1} = e = g^{-1} * g.$$

Groups appear in many places in mathematics. For instance, the integers form a group with the operation $+$, and the real numbers become a group under multiplication. We see that a group has just enough structure that every equation of the form $g_1 * x = g_2$, where g_1 and g_2 are elements of G, has a unique solution $x = g_2 * g_1^{-1}$ in G. For a good introduction to group theory, see [19].

In this paper we are interested in D_4, the symmetry group of the square. This group has eight elements, four rotational symmetries and four reflection symmetries. The rotations are 0°, 90°, 180° and 270°, and the reflections are defined along the four axes shown in Figure 1. We refer to these elements as $\sigma_0, \sigma_1, \ldots, \sigma_7$. Note that the identity element is rotation by 0°, and that for each element there is another element that has the opposite effect on the square, as required in the definition of a group. The group operation is composition of two such transformations. As an example of one of the group elements, consider Figure 2, where we demonstrate rotation by 90° counterclockwise on a square with labeled corners.

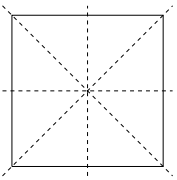


Fig. 1. The four axes of reflection symmetries of the square

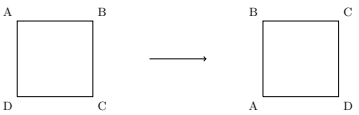


Fig. 2. Rotation of the square by 90° counterclockwise

The Average Asymmetry Matrix. The elements of D_4 can be viewed as transformations that act on a square. Such an action on a set which respects the group operation is called a *group action* on the set. We will not define this formally here, just note that this means that we can define the action of D_4 on the entries of a real square matrix in a natural way by letting the group elements rotate or reflect the entries according to the corresponding group elements. We will denote such an action by $\sigma_i M$, where σ_i is the element of D_4 acting on a square matrix M.

Let M be an $n \times n$-matrix and σ_i some element of D_4. We define the *asymmetry of M by σ_i*, denoted by $A_i(M)$, to be the matrix

$$A_i(M) = |M - \sigma_i M|. \tag{1}$$

We note that if M has a strong degree of the symmetry preserved by σ_i, the entries of this matrix will be close to zero.

Now we are ready to define the *average asymmetry* of M, denoted by $A(M)$. Let M be an $n \times n$ matrix. Then we define the average asymmetry matrix A of M, denoted $A(M)$, as the matrix

$$A(M) = \frac{1}{8} \sum_{i=0}^{7} A_i(M). \tag{2}$$

The more symmetries a matrix has, the smaller the entries of A(M) will be, and in this way we can say that $A(M)$ provides a way to measure the degree of asymmetry of M.

2.2 Proposed Group Based Asymmetry Algorithm

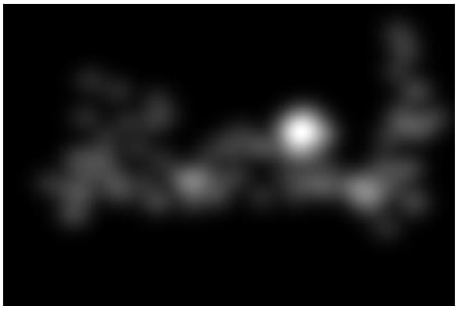
In this section, we outline the implementation of the proposed group based asymmetry algorithm. From the color image, we calculate three channels, i.e., luminance channel, red-green and blue-yellow color opponency channels as described by Walther and Koch [20]. In order to calculate a feature map, we decompose the channel into square blocks. In the general case when the image dimensions are not perfectly divisible by the selected block size we padd the image borders with neighboring pixels. For example, in our experiments we used a block size of 20 by 20 pixels for an image of size 1024 by 768 pixels, thus after padding the image size becomes 1040 by 780 pixels. For each block, we calculate the absolute difference between the block itself and the result of the D4 group element acting on the block. We take the mean of the absolute difference for each block, which is taken as a measure of asymmetry for the block and has a scalar value in the range [0,1]. The asymmetry values for all the blocks are then collected in an image matrix and scaled up to the size of original image using bilinear-interpolation. In the resultant feature map the saliency of a location is represented by its scalar value, where a greater value represents a higher saliency. From the the D4 group elements i.e., rotations by 90, 180 and 270 degrees, and reflections along the four axes of a square, we get seven feature maps. In order to capture both the local and the global salient details in a channel, we use three scales: the original, 1/2 and 1/4. This gives three scales which combined with the seven D4 group elements give 21 feature maps, i.e., from the three channels we get a total of 63 feature maps which are combined linearly to get a single saliency map.

2.3 Analysis Using ROC

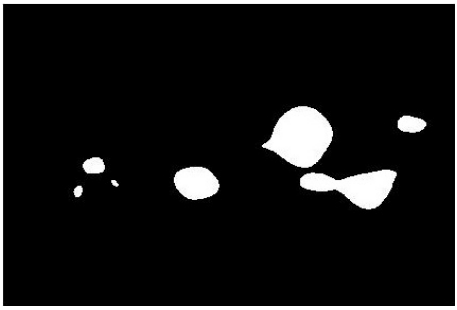
Approach. In this section, we discuss the approach taken for evaluating the performance of the visual saliency models. In keeping with published methods [21–23], we average all the fixations from different observers pertaining to a given image into a single two dimensional map, which is then convolved with a

two dimensional Gaussian filter. In the resultant fixations map, the intensity at a given location represents the density of fixations [24], where the more fixations a region receives the more salient its said to be. For example, figure 3(b) shows the fixations map for an image. Similar to the previous experiment, we calculated the fixations maps from the fixations data of 200 images and 15 observers.

(a) Image from database [26].



(b) Fixations map.



(c) Binary map.

Fig. 3. From the fixations map and the image, we can see that the region containing the road sign received a significant number of fixations. Figure 3(c) shows the binary map obtained by thresholding the fixations map by 20 percent.

In order to evaluate, how well the visual saliency models predict the fixations maps for different images, we use a receiver operating characteristic(ROC) curve [25] which requires that a fixations map is thresholded to yield a binary map with two classes – the positive class consisting of fixated regions, and the negative class consisting of non-fixated regions. As an example, figure 3 shows the binary map obtained by thresholding the fixations map by 20 percent. This procedure is in keeping with the study by Judd et al. [26]. The ROC curve evaluates how well the visual saliency algorithm predicts the two classes [25]. For plotting the ROC curve we randomly select 500 pixels from the positive class and an equal number of pixels from the negative class. The area under the ROC curve (AUC) is used as a measure of the performance of a classifier. AUC gives a scalar value in the interval [0,1] where larger the area, better is the performance [25].

Results. We plot the ROC curves for the visual saliency algorithm proposed by Itti et al. [3], and the proposed group based asymmetry algorithm(GBA). Results in figure 4 show that the GBA algorithm results in an AUC value of 0.81 which is better than that achieved with the visual saliency algorithm by Itti et al. [3] which gives AUC of 0. 77.

In order to measure the similarity between the proposed group based asymmetry algorithm and the visual saliency algorithm by Itti et al. [3] we calculated another ROC curve. In this case, we use the saliency maps from the visual saliency algorithm [3] as the ground truth maps. By following the procedure described in section 2.3, we evaluated how well the maps obtained from the GBA algorithm predict the maps obtained from the visual saliency algorithm [3]. Figure 5 shows the ROC curve for the proposed GBA algorithm which gives an AUC of 0.88 indicating that the prediction of the saliency values obtained by the proposed algorithm is indeed close to that of the visual saliency model.

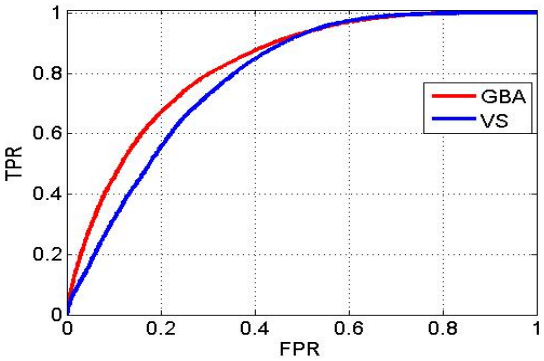


Fig. 4. Figure shows the ROC curves for the visual saliency(VS) model by Itti et al. [3](AUC = 0.77), and the proposed group based asymmetry (GBA) model (AUC = 0.81). The x-axis shows the false positive rate(FPR) and the y-axis shows the true positive rate(TPR).

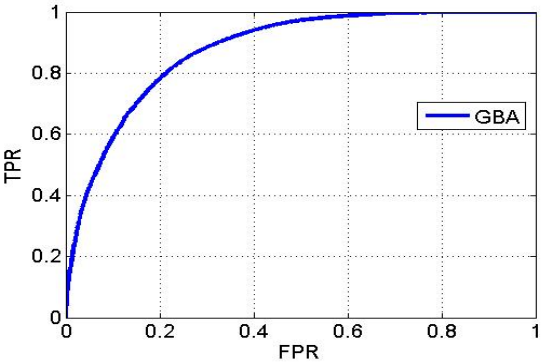
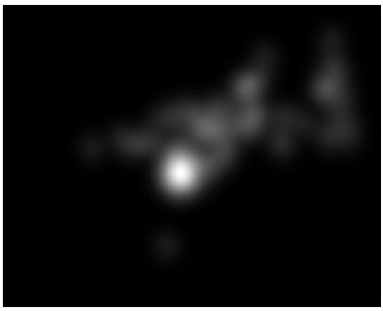


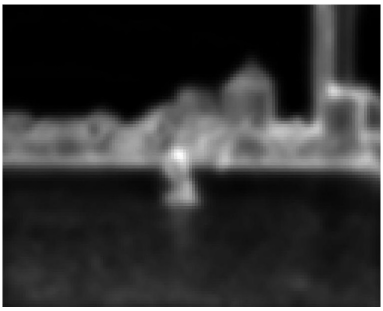
Fig. 5. Figure shows the ROC curves for the proposed group based asymmetry (GBA) model, AUC = 0.88. The x-axis shows the false positive rate(FPR) and the y-axis shows the true positive rate(TPR). Here we use the maps from the visual saliency algorithm [3] as the ground truth.

(a) Image from database [26]



(b) Fixations Map

(c) Saliency Map [3]



(d) Group based Asymmetry Map(GBA)

Fig. 6. Comparison of visual saliency algorithms, both algorithms return the region containing the boat at the center as salient, which is also in agreement with the fixations map obtained from the eye fixations data

To offer a visual comparison between the two methods we show the fixations map, and the saliency maps obtained from the GBA algorithm and the visual saliency algorithm [3] for an example image. In figure 6, we can see that the maps from both the algorithms are quite similar. In fact both of them return the region containing the boat at the center as salient, which is also in agreement with the fixations map.

3 Discussion

In this study, we set about unifying the mathematical description of saliency in a single metric. Backed by the knowledge gained from research in image processing where it has been shown that the dihedral group D_4 can be used to encode edges and contrast which are the main current descriptions of saliency we chose to devise an algorithm that represents the level of saliency in an image region by virtue of the transformations of D_4. D_4 is the symmetry group of the square image grid and includes two types of symmetries, i.e., rotation and reflection.

In our implementation, we chose to describe the symmetry of an image region at three different scale, however, we didn't perform any center surround operations by taking the differences between the scales. In this view, what we have presented in this study is a new unified metric together with a new description of saliency where we define saliency as the combined level of asymmetry at different image scales.

In our experiment, we a used a receiver operating characteristic(ROC) curve to compare the performance of the proposed method with that of Itti et al. [3] . Here we used 200 images and fifteen observers and found that the new method results in a predication of fixations that is better than that achieved with the saliency algorithm. We thus concluded that the transformations of the dihedral group D_4 are a good metric to estimate salient image regions which if backed by further studies can represent a mathematically sound method to define a salient image region.

References

1. Suder, K., Worgotter, F.: The control of low-level information flow in the visual system. Reviews in the Neurosciences 11, 127–146 (2000)
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4, 219–227 (1985)
3. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 1254–1259 (1998)
4. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40, 1489–1506 (2000)
5. Gorsuch, R.L.: Factor Analysis. Lawrence Erlbaum Associates, LEA (1983)
6. Jolliffe, I.T.: Principal component analysis. Springer (2002)
7. Braun, J., Sagi, D.: Vision outside the focus of attention. Perception and Psychophysics 48, 45–58 (1990)

8. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual Reviews in the Neurosciences 18, 193–222 (1995)
9. Steinman, S.B., Steinman, B.A.: Vision and attention. i: Current models of visual attention. Optometry and Vision Science 75, 146–155 (1998)
10. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience 2, 194–203 (2001)
11. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2049–2056 (2006)
12. Mozer, M.C., Sitton, M.: 9. In: Computational modeling of spatial attention, pp. 341–393. Psychology Press (1998)
13. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 185–207 (2013)
14. Lenz, R.: Using representations of the dihedral groups in the design of early vision filters. In: ICAASP, pp. 165–168 (1993)
15. Lenz, R.: Investigation of receptive fields using representations of the dihedral groups. Journal of Visual Communication and Image Representation 6, 209–227 (1995)
16. Foote, R., Mirchandani, G., Rockmore, D.N., Healy, D., Olson, T.: A wreath product group approach to signal and image processing. i. multiresolution analysis. IEEE Transactions on Signal Processing 48, 102–132 (2000)
17. Chang, W.Y.: Image processing with wreath products. Master's thesis, Harvey Mudd College (2004)
18. Lenz, R., Bui, T.H., Takase, K.: A group theoretical toolbox for color image operators. In: IEEE International Conference on Image Processing, ICIP 2005, vol. 3, pp. 557–560 (2005)
19. Dummit, D.S., Foote, R.M.: Abstract Algebra. John Wiley & Sons (2004)
20. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks 19, 1395–1407 (2006)
21. Cerf, M., Harel, J., Einhauser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: Advances in Neural Information Processing Systems (NIPS), vol. 20, pp. 241–248 (2007)
22. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proceedings of Neural Information Processing Systems (NIPS) (2006)
23. Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision 9, 1–15 (2009)
24. Duchowski, A.T.: Eye Tracking Methodology: Theory and Practice. Springer, Heidelberg (2007)
25. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. Pattern Recognition Letters 27, 882–891 (2004)
26. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: International Conference on Computer Vision (ICCV) (2009)