International Journal of Recent Advances in Multidisciplinary Research
Vol. 02, Issue 08, pp.0608-0625, August, 2015

Research Article

TOWARDS A COMPUTER VISION MODEL OF RESTORATIVE SCENES

***[1]James Mountstephens and [2]Balvinder Kaur Kler**

[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
[2]Faculty of Business, Economics and Accounting, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

ARTICLE INFO

Article History:
Received 16th May 2015
Received in revised form
21st June, 2015
Accepted 28th July, 2015
Published online 31st August, 2015

Keywords:

Computer modeling,
Computer vision,
Interdisciplinary,
Attention restoration theory,
Restorative scenes.

ABSTRACT

This paper describes early progress in novel, interdisciplinary work that applies concepts and methods from Computer Vision to the development of a visual model of restorative scenes. Such a model has the potential to both enhance Attention Restoration Theory (ART) and to find numerous practical applications in the design and synthesis of living spaces and visual artefacts. To explore the feasibility of a visual model of restorative scenes, a comparison between known human gaze patterns and an exclusively bottom-up computational model of attention was performed. Similarities were found, providing evidence for a key claim of ART. Visual Models were then developed with 3 different motivations: i) biological plausibility, ii) a concern for model interpretability, and iii) a hypothesis that more abstract scene properties such as complexity and information content are responsible for fascination and restoration. Image datasets rated by humans for restorative potential were developed and used to construct and test these models and encouraging results were found. This work is the first to combine Computer Vision and Environmental Psychology and it is hoped that further collaborations are inspired.

INTRODUCTION

Interdisciplinary integration is often needed for the study of phenomena too broad or too complex for individual disciplines to adequately address (Repko, 2011; Klein and Newell, 1997). Towards integration, disciplines may contribute perspectives, concepts and theory that may yield a more comprehensive understanding of the phenomenon under study, as well as methodology and tools for both research and practical applications (Repko, 2011). Environmental Psychology is inherently interdisciplinary (Veith and Arkkelin, 1995). In its attempt to study the "molar relationships between behaviour and experience and the built and natural environments" (Bell *et al*, 2001:6) it has drawn upon Geography, Economics, Landscape Architecture, Sociology and Anthropology, among others. However, Environmental Psychology has had relatively little contact with computing and its related fields of study. Admittedly, some work has been done using computer graphics to *synthesise* visual environments (eg. de Koort *et al*, 2003) but the use of computers to automatically *analyse* visual environments (a vastly more difficult task) has, to the best of our knowledge, not been attempted.

Corresponding author: James Mountstephens,
Faculty of Computing & Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia.

The automatic analysis of images by machine is a key concern of Computer Vision. Humans are effortlessly able to extract a vast range of visual information from their environment and Computer Vision seeks to develop algorithms and hardware that allow machines to do likewise (Prince, 2012; Ponce and Forsythe, 2011; Szeliski, 2010). Also inherently interdisciplinary, Computer Vision draws upon Mathematics, Physics, Neurobiology and Cognitive Neuroscience, Signal Processing, Artificial Intelligence and Machine Learning, to build mathematical and algorithmic models of common processes in vision such as scene and object recognition, perceptual grouping, depth and motion perception (Ponce and Forsythe, 2011). Additionally, the sequential and selective aspects of scene perception have been studied with computational models of visual attention (Borji and Itti, 2013; Itti and Koch, 2001). Such models can enhance our understanding of human vision and help to reveal the structures and processes underlying our interpretation of scenes. At a practical level, implementations of models are intended to be able to receive a digital image of a scene as input and automatically extract and output information of interest contained therein. Although the main lesson of more than 40 years of work in Computer Vision is that achieving human-level performance is incredibly difficult (Szeliski, 2010), considerable progress has been made in the extraction of

objective image properties. However, little attention has been given to the automatic evaluation of visual scenes in terms of aesthetics, or the effect that certain images can have on a human viewer, which may be beneficial. Emprical work in Environmental Psychology has demonstrated that viewing specific visual scenes (even as photographs) can be beneficial, both subjectively and objectively. In comparison to most built scenes, viewing certain natural environments has been shown to improve cognitive performance, mood, the ability to plan, sensitivity to interpersonal cues, as well as physiological levels such as stress and arousal (Kaplan 1995; Kaplan and Talbot, 1983; Kaplan and Kaplan, 1981; Berto, 2005; Korpela and Hartig, 1996; Hartig et al, 2003; Hartig and Staats, 2005). One important way to understand these effects is as a form of *restoration*, the renewal of "physical, psychological and social capacities that have become depleted in meeting ordinary adaptational demands" (Hartig and Staats, 2005:281).

Kaplan's Attention Restoration Theory (ART) explains the benefits of viewing nature in terms of its effect on the crucial cognitive resource of attention (Kaplan, 1995). ART is predicated on a distinction between two main modes of attention, which is also recognised by cognitive science and by computational modelling. Involuntary (or exogenous) attention is driven largely bottom-up by sensory stimuli. It is responsible for effortless orientation to salient stimuli and is thought to be mediated subcortically in the superior colliculi. Voluntary (or endogenous, directed) attention involves top-down inhibition of involuntary attention and the neural excitation of task-relevant locations. It is mediated by a number of cortical areas forming a "dorsoparietal network" and is crucial for intentional action and concentrating on tasks. However, voluntary attention requires effort to sustain and long-term demands deplete this resource, leading to what ART calls directed attention fatigue (DAF). DAF leaves us unhappy, unable to plan, insensitive to interpersonal cues and increases our likelihood of errors in performance (Kaplan, 1995; Berto, 2005). ART claims that natural scenes contain stimuli that facilitate a move into involuntary mode, where "attention is typically captured in a bottom-up fashion by features of the environment itself" (Berman, Jonides and Kaplan, 2008:1207), and whereby directed attention is allowed to recover. By this account, natural scenes are more *fascinating*, containing patterns and objects that attract attention effortlessly but are not so stimulating as to require effortful focus and decision-making.

Although this explanation of the benefits of nature still requires further substantiation, the "black box" occurrence of attention restoration is well-supported using methods of assessment devised within ART for that purpose. In (Berto, 2005) for example, subjects were given a sustained cognitive task until fatigue, shown a slideshow of either nature or built scenes, and then asked to perform the task again. Only the subjects who viewed scenes of nature performed better on retest, demonstrating restored capabilities. If such restorative benefits exist it would be worthwhile to model the visual properties of the scenes that drive them. The details of such a model will be discussed in the next section but the basic idea is that it would describe the relationship between properties directly extractable from a digital scene image (such as colour, shape, texture) to the scene's restorative potential. If successful, the model would enhance our understanding of restoration and ART. It could, in effect, tell us *what makes a scene restorative* at a detailed visual level, which ART currently does not. The question is a profound one and addresses a deep relation between humans and the natural world.

This model could also be used practically to evaluate arbitrary scenes. Software systems able to automatically evaluate the restorative potential of a given image would literally add a new dimension to image analysis and selection; if a machine can reliably identify them, we might actively use restorative and fascinating images in our endeavours, be they website design or home decoration. Mobile phone apps presenting images predicted to restore might help us meet the challenges of day-to-day life. In architecture and urban planning, competing designs of living spaces might be selected based on restorative potential or the rating might even be incorporated into the design process itself. Knowing what makes a scene restorative could allow the automatic synthesis of fascinating scenes and images by computer too. If the relevant visual properties are sufficiently abstract, images that are not natural in a figurative sense might still restore. It might also be possible to objectively determine whether restorative scenes share common visual properties with other images considered fascinating, such as paintings and other works of art.

But what properties of a scene make it restorative? Its colour, shape, texture? The objects present? The scene's layout, organisation and viewing distance? More abstract notions of order and complexity? Or perhaps biological properties of our visual system? We do not know. ART provides little guidance in terms of directly measurable image properties and therefore the modelling task will largely be one of hypothesis and experiment. This paper will present exploratory work done so far in that capacity and is, to the best of our knowledge, the first work to combine Computer Vision with Environmental Psychology. As such, it would be worthwhile to cover some basic background material to allow the experiments described in the coming sections to be better appreciated.The model we seek will ideally be accurate, intelligible and have high explanatory power. It must also be practically feasible given the current capability of Computer Vision. These desiderata and constraints will inform the following presentation.

Background and Modelling Considerations

From a dataset of n scene images $\{\boldsymbol{I}_n\}$ labelled for restorative potential y by human subjects, we intend to build a regression model of the form shown below, which will predict y for a given image $\boldsymbol{I}$. We do not expect perfect determinism so a random error term ϵ is assumed.

$$y = g(f(\boldsymbol{I})) + \epsilon$$
$$= g(\boldsymbol{x}) + \epsilon \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The dependent variable y is normalised to the range [0, 1] and is fixed for a given set of subjects and scene images. The image properties$\boldsymbol{x}$ which drive restoration, and their exact relationship to y, are captured by the functions f and g respectively, which embody a combination of image processing and machine learning processes. Determining f (and therefore $\mathbf{x}$) and g will be the main modelling task here.

Images as Data

The function *f* extracts relevant visual information from an image ***I***. Being a rectangular array of pixels, a digital image can be considered a matrix, with a value for each pixel. In the simplest case of monochrome (intensity) images, the value for each pixel is a single number, often ranging from 0-255. For colour images, the value of each pixel is usually a vector of 3 numbers that represent colour component (for example red, green, blue).

It is from these raw values that useful information must be determined, a task that is deceptively difficult. The average human is so effortlessly skilled at interpreting visual data that it is often hard to appreciate the sheer difficulty of moving from millions of simple retinal cell responses to the conscious perception of objects and scenes. Computer Vision researchers are faced with this task directly: how do we form any kind of high-level interpretation of a giant grid of numbers, especially when those numbers are inherently incomplete for the task of reconstructing the world (vision is an *inverse* problem: Ponce and Forsythe, 2011). The interested reader may find a convincing illustration of the difficulty of the problem in Hyvarinen *et al* (2009). Another challenge is the high dimensionality of image data: even a medium resolution (640x480) image in 24-bit colour provides almost a million values to deal with. This is in contrast to many other branches of science where variables are relatively few are and each have meaningful interpretations (eg the sampling unit is a person and variables are age, height, weight etc.).

Image Descriptors

Considered mathematically, a vast number of functions *f* could be calculated for this matrix of values ***I***, yielding an output ***x*** which might be scalar, vector or even another image. For example, the input image could be reduced to a scalar by taking its mean value over all pixels, its vector-valued intensity histogram could be calculated, or object edges could be identified as locations where intensity changes rapidly, yielding a new image consisting of edges alone. These examples are intentionally simple; a vast number of image processing operations have been developed with inspirations that include biological plausibility, a desire to capture visual information in a way that humans can understand, and by mathematical and more abstract considerations. The nature of *f* and ***x*** will depend on the task at hand. For example, if the task is to distinguish between certain very simple classes of image, intensity or colour histograms may be sufficiently distinctive image properties. In these cases, ***x*** will often be called an image *descriptor* or *feature*.

The present task will require the identification of an image descriptor that is related to restorative potential.Image descriptors are usually vectors and are notable in machine learning circles for their high dimensionality which is largely a result of the high dimensionality of the underlying image ***I***. It is also important to point out that although we would often like ***x*** to capture image properties with intuitive interpretations, this is not always possible. It may be tempting to speculate that, for instance, the presence of certain objects drives restoration but it may not be currently possible for Computer Vision to reliably identify those objects.

Model Forms

The function *g* describes how image properties ***x*** relate to restorative potential *y*. The relationship might be linear or nonlinear and may be embodied as a straightforward mathematical function or in more advanced machine learning models such as Artificial Neural Networks (ANNs; Haykin, 2009) or Support Vector Machines (SVMs; ibid). Although a considerable amount of scientific research attempts to find interpretable linear associations between variables, not all relationships of interest are linear or even an easily-interpretable function. SVMs and ANNs have proven remarkable in their ability to discover patterns in high dimensional data that are not obvious to either inspection or straightforward linear analysis but this comes at a price: although in theory SVMs and ANNs can be represented mathematically, their structure and the large number of free parameters involved make them very hard for people to interpret (Gershenfeld, 1998). They are usually treated as a black box.

Image Datasets

For a data-driven model, the construction of a dataset $\{\boldsymbol{I}_n\}$ of scene images labelled with restorative potential *y*, will be crucial. Ideally the dataset would constitute a reasonable sample of the relevant image population in both content and number. All images should be of high quality and should be labelled with a reliable measure of restorative potential by a reasonable sample of the relevant population of human subjects. Since the image and subject populations are arguably "all scenes" and "all people" the task is a huge one. No reference image datasets have been made publicly-available by any previous ART researchers –and this is a situation in strong contrast with Computer Vision research where a large number of datasets have been constructed and made available (eg. Russell *et al.*, 2005). But even if the images in those datasets are usable in terms of content, number and quality, they must still be labelled by human subjects with ratings of restorative potential andfor this task, the Perceived Restorative Scale (Hartig *et al.*, 1997) could be used.However, for manual labelling, the requirements of feasible timescale and large sample size are of course contradictory. The experiments described presently will construct and use datasets that fulfil some of the desiderata but as yet no ideal dataset exists.

Desiderata and Constraints

There are many options for ***x*** and *g* and little apriori to guide us. However, given our model desiderata, we can prioritise or at least evaluate the candidates. It is useful to consider the concreteness of the image properties under consideration and the explanation of restoration they would suggest. A descriptor ***x*** might be some measure of a scene's colour, texture or edges and would therefore be directly visual and easy to interpret. Or ***x*** might attempt to capture the scene's complexity or information content, which, although still derived from visual data, would describe more abstract properties and might be less easy to interpret. In the extreme, there could be mathematically-derived ***x*** that might successfully predict restoration but are opaque to interpretation; the model would be a black box, good for practical use but adding little to an explanation of attention restoration.

Intelligibility of the image property is a necessary condition for it to explain restoration, but it need not be sufficient. To take an obvious example of intelligibility: if it were found that a scene's overall level of the colour green caused restoration, what would that say about attention restoration as it is currently conceptualised? Natural environments often have high green content of course, but green in itself has no obvious causal power, beyond perhaps evolutionary hard wiring. On the other hand, finding that restorative scenes had a certain level of order, complexity or information content might give a reasonable explanation of why they produce fascination, the key component of attention restoration. For example, in scenes with a suitable ratio of order to disorder, effortless perception might be driven by order but sustained interest might be maintained by a certain level of disorder. This would arguably be a more satisfying type of explanation and will be addressed in more detail in section 6.

So, to summarise: a model with accuracy, intellibility, and explanatory power is sought, and is to be achievedwiith the current capabilities of Computer Vision. There is no guarantee that this will be possible and the endeavour could fail on any one of those counts. Before proceeding to the initial modelling experiments conducted here, it is first necessary to address a crucial assumption made about the whole enterprise. Namely, that a model of restorative scenes can be made directly in terms of image properties alone.

Assumption of Image Sufficiency

General scene perception is a complex task involving many levels of processing and may involve the top-down deployment of prior knowledge, task concerns, preferences and conscious intentions that go beyond bottom-up response to image information as stimulus. Modelling these factors would be an immeasurably more difficult task and possibly an insurmountable one. This issue relates directly to a key claim of ART: that states of fascination are driven bottom-up by visual features (Berman, Jonides and Kaplan, 2008). Evidence in favour of this claim would be a valuable finding in itself but for the purposes of developing a visual model it would serve as both direction and encouragement since it would suggest that the image contains sufficient information to account for attention restoration. This question motivates the first experiment presented here. It does not construct a visual model but attempts to validate the possibility of one and its findings are interesting in their own right.

Computational Attention Models

Initial work on combining Computer Vision and ART used computational attention modelling to gather evidence that states of fascination are driven bottom-up by visual features.The essential idea was to use a bottom-up computational model of visual attention to process scenes of both natural and built environments and to compare the resulting gaze patterns to known patterns measured in humans when viewing similar scenes. Since the model is free from both top-down task influences and higher-level processes such as object recognition, similar gaze patterns would suggest that attention restoration requires only a primitive level of processing. Marked differences might suggest that higher-level processes are at work.

Restorative Gaze Patterns

It is known how the gaze of humans differs when processing high and low restorative scenes.In (Berto *et al,* 2008) eye tracking technology was used to determine whether visual perception differs between scenes that are primarily natural or built. Eye movements are often taken to reflect both internal shifts of attention and the amount of effort employed in the viewing of a scene. Subjects were exposed to 50 scenes (25 natural, 25 built) and their gaze locations were recorded over 15s of free viewing. Three measures were used to characterise the differences in gaze patterns: distance covered in the total exploration of the scene, number of fixations and number of saccades. A fixation was defined as the gaze location remaining constant for >150ms. It was hypothesised that distance and number of fixations would be higher for built scenes as they are less fascinating and would engage directed rather than involuntary attention in a process of greater scrutiny. In contrast, participants were expected to scan the scenes of nature broadly, but not to attend carefully to any particular aspects. The findings of the study confirmed this hypothesis: in built scenes significantly greater distance was covered and more fixations were made.

Berto *et al*'s study provides the basic evaluation framework for our initial experiment. However, instead of tracking the sequence of gaze locations in human subjects, here a computer model was used to process the scene and determine locations of gaze. Computer modelling of attention is an active research area that overlaps with neuroscience and Computer Vision. A number of models have been devised ranging from top-down, task based to bottom-up entirely. A comprehensive survey can be found in Borji and Itti (2013) but here only the particular model used in this research will be described in detail.

Saliency-Based Computer Models

Building on the 'Feature Integration' theory of Treisman and Gelade (1980), Itti, Koch and Neibur (1998) developed a *saliency-based* computer model of selective visual attention (hereafter known as IKSM, for Itti and Koch Saliency Model) which, when given an image or image sequence, is designed to output a sequence of gaze fixation points. Saliency is a property related to the 'pop out' effect commonly found in visual search experiments where an object may be especially conspicuous relative to its neighbours because it differs in some property. For example, a circle found amongst squares or a green triangle amongst red triangles may 'pop out' of the scene.

In IKSM, saliency is a measure of the conspicuity of an image point based purely on local differences in low-level features and is embodied within a *saliency map*, an array of neural processing units analogous to the input image whose activity encodes saliency for each image point. During operation, a process of local competition amongst neurons determines that with the highest activation (the 'winner') which, as most salient, is taken to be the new location of gaze for the next time-step. Activation in the saliency map evolves over time in response to features of the inputimage sequence and an internal biasing mechanism of inhibition of return (IOR), which negatively weights the region in the saliency map centered on the current gaze location.

This prevents gaze from becoming stuck in a single location and enforces a scan of the scene. An example of generated gaze locations and the saliency map that produced them is given in Figure 1.

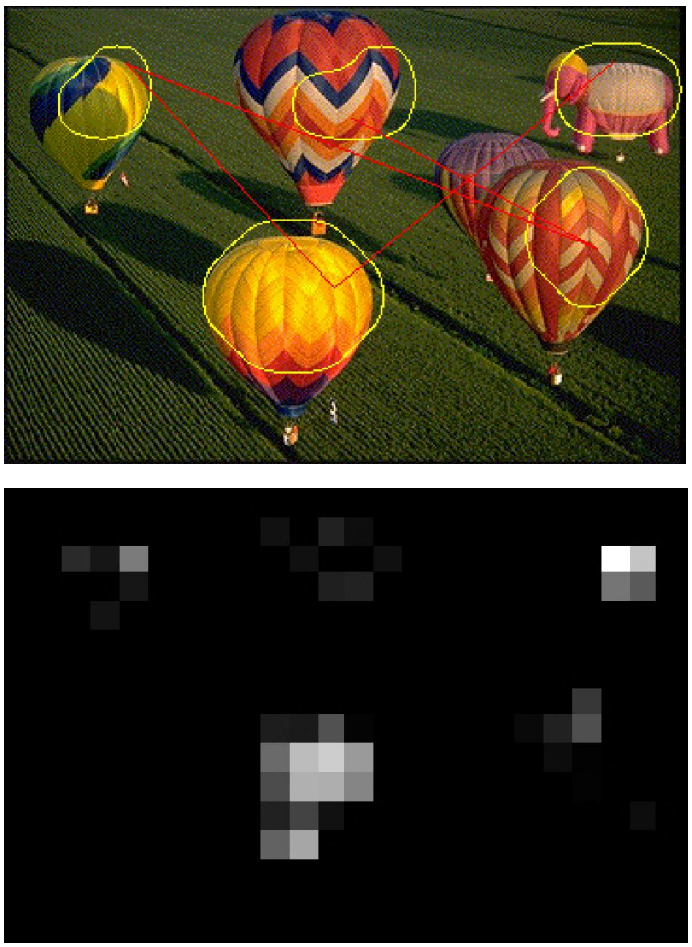

Figure 1. Example IKSM sequence of Gaze Locations (top) and Saliency Map (bottom)

Within IKSM, information is represented in maps which are analogous to the input image, or some function of it. These maps are produced by filtering and combining image features. The saliency map is the most important map since it ultimately determines the gaze fixation point. Inspired by biological visual receptors (Ittiand Koch, 2001), local differences are considered more important than absolute values and a filter's response at a given location depends on how the value there differs from its neighbours. Specifically, the values in each feature map are calculated in an approximation to a centre-surround response, produced by convolving a raw feature map with a Gaussian kernel at progressively larger standard deviations and differencing between this hierarchy, or 'Gaussian pyramid', of maps.

The features used in calculating saliency are inspired by those found in early regions of the human visual cortex (Hubel and Wiesel, 1959). Raw features can be either static or time-dependent. Colour opponency (red/yellow and green/blue), orientation and intensity are the most common static features and can be calculated from a single image whereas the dynamic features of motion and flicker require an image sequence for their calculation. Calculation of the centre-surround response for each of these raw feature mapsis the first stage in model execution and is conducted as above. Depending on the exact configuration of the visual cortex model there may be several feature maps for a feature type (eg. for motion, there may be motion left, motion right, up and down) so these are combined into a single *conspicuity map* to summarise the total response for the type of feature in question.

Conspicuity maps are combined in a similar fashion to form the saliency map for this time step and this provides enough information to decide the gaze location. A winner-takes-all process is used to find the point of highest saliency (or more accurately, the neuron in the saliency map array with the highest activation) which is the model's current output. Finally, a mechanism of inhibition of return(IOR) negatively weights the saliency map in an area centered at the current location of gaze so that this location becomes a very unlikely winner at the next time-step. Since gaze cannot return to the current location until the negative weighting subsides, a serial search of the image in order of decreasing saliency is enforced. A schematic showing the sequence of operations is found in Figure 2. IKSM has been used and cited in a large number of projects and publications and has shown to successfully replicate some aspects of human gaze allocation (Egner, Itti and Scheir, 2000).

Image Dataset

To allow a direct comparison with the results of Berto *et al* (2008), this experiment would ideally be performed on the same scenes used in that study. Unfortunately, said images have not been made publicly–available so a popular dataset used in Computer Vision research was used instead.

Images for this experiment were taken from the freely-available '8 Scene Categories Dataset' by the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT[1]. This dataset was compiled for work on determining the spatial envelope, or 'gist', of a scene (Oliva and Torralba, 2001) and has been successfully used to demonstrate objective differences between natural and built scenes, making it suitable for our current purposes. The images have since been made freely-available and have been used as a benchmark in various subsequent studies. The dataset consists of 2688 colour outdoor images, divided into natural and built scenes, and each further divided into 4 subcategories: {coast, mountain, forest, open country}, {street, inside city, tall buildings, highways}. There are approximately equal numbers of natural and built scenes: 1472 and 1216, respectively. This dataset can be considered ideal except for one thing: all images are only 256x256 pixels.Although upscaled to 800x600 for processing in this experiment, fine detail and texture are lacking. As stated earlier, dataset construction requires balancing competing criteria of content, size and quality and it was hoped that the available image detail and texture were sufficient.

Apparatus

The technical computing environment Matlab (release 2013b) provided the basic software infrastructure used in this experiment. Functions from the image processing, optimisation and statistics toolboxes were utilised at pre-and and post-processing stages.

[1]http://people.csail.mit.edu/torralba/code/spatialenvelope/

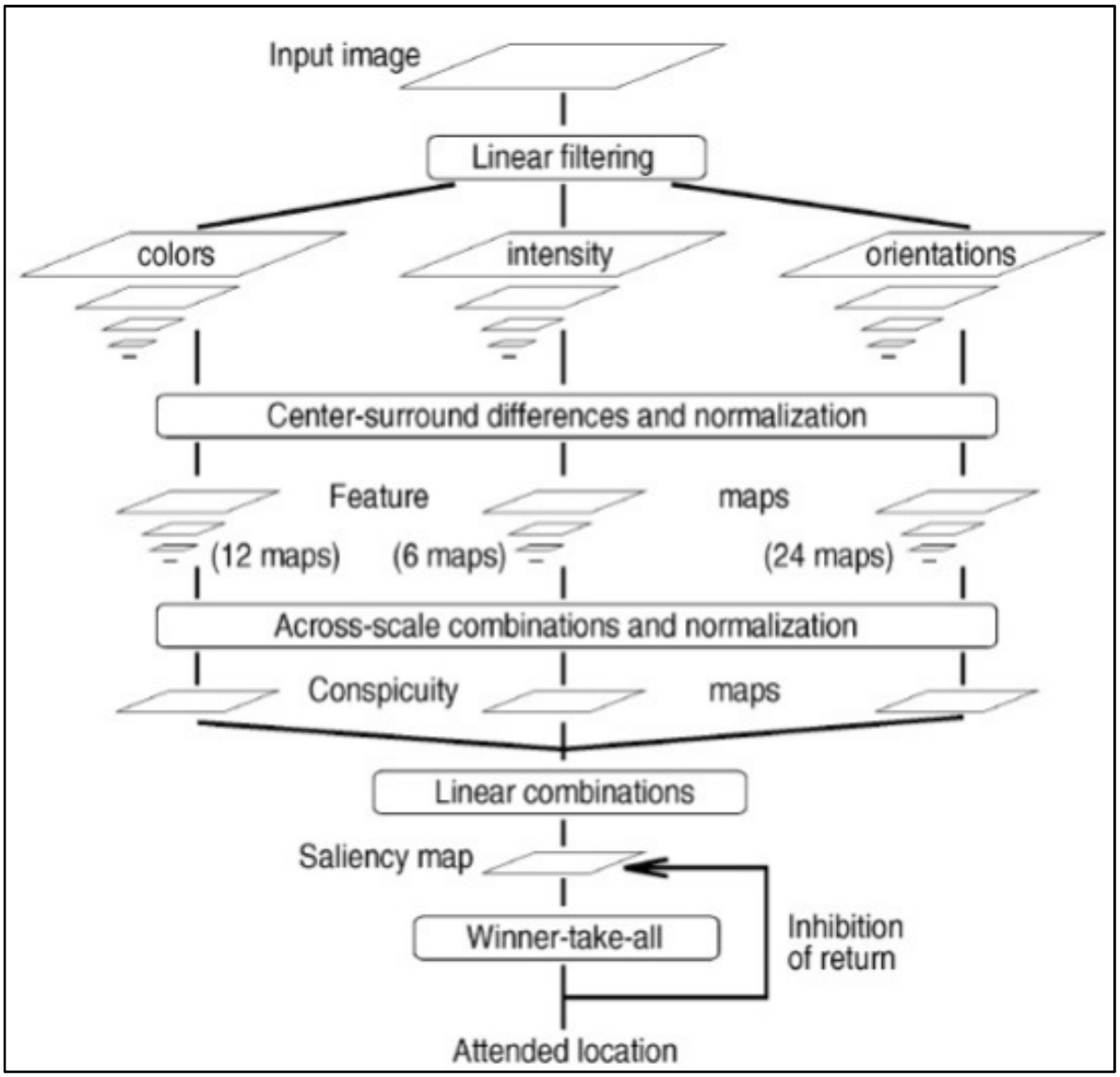

Figure 2. Schematic of Itti and Koch Saliency Model (Itti, Koch & Neibur, 1998)

Calculation of the gaze locations for a given scene was performed by the freely-available Saliency Toolbox (STB)[2], a Matlab implementation of IKSM (Walther, 2006). The Saliency Toolbox accepts an input scene image and then evolves a saliency map for it, outputting a sequence of *n* gaze locations and the time in ms that each location is attended to, i.e. a list of triples $\{x_0\ y_0\ t_0;\ x_1\ y_1\ t_1,\ldots,x_n\ y_n\ t_n\}$. The number *n* of gaze locations, and the time between them, varies with each scene, according to its particular saliency map. The spatial coordinates in the output are in the coordinate frame of the image itself and therefore vary from 0-799 in the x-dimension and from 0-599 in the y-dimension. The machine used to conduct this experiment was a Dell Optiplex 990, with an Intel Core i3-2100 CPU running at 3.1 GHz and with 4GB RAM. For images of the resolution used here, the Saliency Toolbox does not run in realtime and total processing time for all 2688 images was approximately 30 hours.

[2] http://www.saliencytoolbox.net/

Procedures and Measures

To allow comparison between the results obtained by computer modelling and those from Berto *et al's* study of human attention patterns, the same presentation time of images and measures were used. The Saliency Toolbox was set to process each image in the dataset for 15s, and the sequence of gaze locations was extracted. After all images were processed, measures of total distance, number of fixations and number of saccades were calculated and their statistics analysed. Due to the differences in experimental design between Berto *et al's* work and the current study, certain modifications to the measures were needed. However, these differences are of measurement scale and would not affect any relative differences in observations between natural and built scenes.

Berto *et al's* distance measurement was calculated in pixels first and then converted to visual angles (degrees), which is a straightforward conversion when a real observer and screen are used and their relative distances measurable. In this work there is no real observer so distance was maintained in units of pixels.

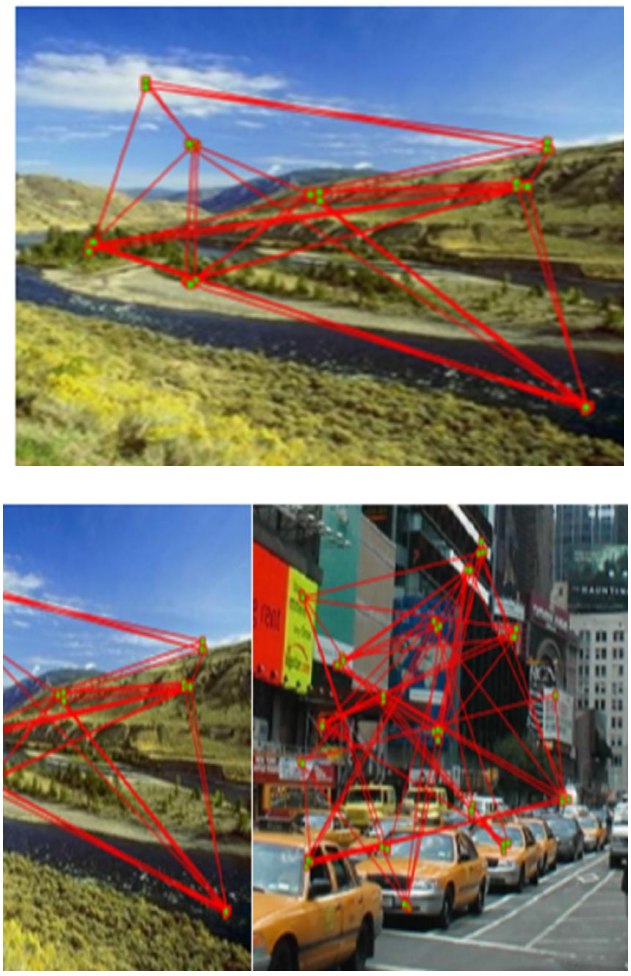

Figure 3. Example IKSM Gaze Sequences on Natural (top) and Built Scenes (bottom)

The Euclidean distance was used in the distance calculation. Fixations were considered to occur when gaze stayed at the same location for more than 150ms, as per Berto *et al's* study. Human saccades can be small, fast and frequent (Berto *et al* measured approximately 100 saccades in 15s). IKSM (and therefore the Saliency Toolbox) does not model such fine eye movements. Here, saccades are taken to occur when gaze changes location significantly and the absolute number of modelled saccades is likely to be lower than those found in humans. However, comparison of relative saccade counts was still expected to be informative.

RESULTS

Processing of scene images yielded sequences of gaze fixations, examples of which are shown in Figure 3. The mean distance and number of fixations and saccades were calculated for all 2688 scenes, across both the nature and built categories and are shown in Table 1 below.

Table 1. Mean Distance, Fixations and Saccades for Natural and Built Scenes

	Mean Distance (pixels)	Mean Fixations	Mean Saccades
Natural Scenes	10313.6 (SD 447.2)	31.4 (SD 8.1)	38.5 (SD 11.5)
Built Scenes	11125.1 (SD135.8)	33.1 (SD 7.5)	41.2 (SD 10.8)

More distance was covered when viewing the built scenes, with a mean distance of 11125.1 pixels (SD 3135.8), whereas viewing the natural scenes covered a mean distance of 10313.6 pixels (SD 3447.2). An independent samples *t*-test showed the differences in distance for natural and built scenes to be significant: $t(2686) = -6.32$, $p << 0.0001$. This is consistent with Berto *et al's* findings for human subjects. A greater number of fixations were measured in the built scenes, with a mean of 33.1 (SD 7.5) compared to 31.4 (SD 8.1) fixations in nature scenes. This difference in fixations for nature and built scenes was also found to be significant: $t(2686) = -5.63$, $p << 0.0001$. Again, this is consistent with Berto *et al's* study of human performance with the same types of stimuli.

Finally, a greater number of saccades were generated in the built scenes than the nature scenes: 41.2 (SD 10.8) ad 38.5 (SD 11.5) respectively. An independent samples *t*-test also showed this difference to be significant: $t(2686) = -6.07$, $p << 0.0001$. Berto *et al's* study found no significant difference in saccades during human exploration of scenes but the result here may be due to IKSM not modelling fine saccades. Berto *et al* do not appear to consider saccades to be as important as distance and fixation count in demonstrating fascination.

DISCUSSION

The result that a relatively simple bottom-up computer model of early stage vision and attention showed similar gaze statistics to humans when viewing fascinating scenes was surprisingbut is consistent with the ART claim that states of fascination associated with attention restoration are essentially stimulus-driven. Further investigation is necessary but the findings give encouragement to the idea that a model of restorative scenes based exclusively on image information might be possible.In the coming sections, a number of approaches to such a model will be described.The first is motivated by the experiment just presented in its use of biologically-inspired processing of visual scenes.

Biologically-Inspired Models

Attention restoration is a human phenomenon and, since our responses to stimuli are a product of both the properties of the stimulus itself and of the particular characteristics of our perceptual system, it is reasonable to first explore models inspired by biological visual systems.Being bottom-up and biologically-inspired, Itti and Koch's saliency-based model of visual attention (IKSM) just discussed,is appropriate for the task at hand.It will be explored here as a source of visual descriptorsable to characterise a scene's restorative potential. Specifically, the conspicuity maps for colour, intensity and orientation, described in 3.2 will be used. Another candidate model is HMAX by Riesenhuber and Poggio (1999),a physiologically-plausible computational model of object recognition in cortex, intended to explain cognitive phenomena in terms of simple and well-understood computational processes. HMAX is a purely feedforward model and has been shown capable of capturing the invariance properties and shape tuning of neurons in macaque inferotemporal cortex. Like IKSM, HMAX is explored for its potential to provide visual descriptors.

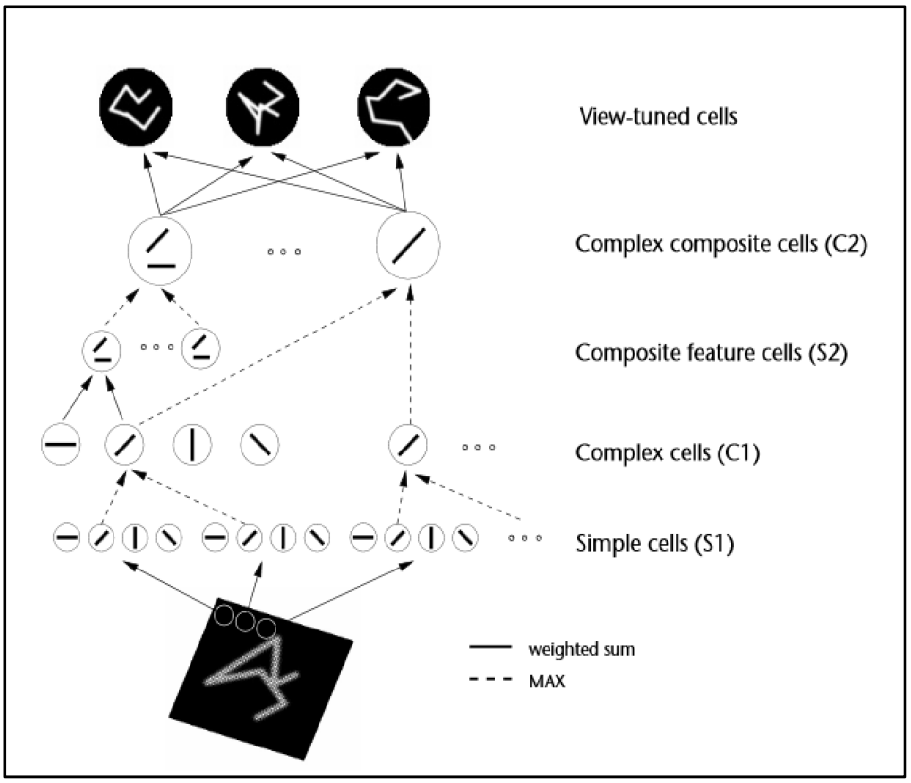

Fig. 4. HMAX General Architecture (Riesenhuber and Poggio, 1999)

The general approach used here will be to i) build a dataset of scenes rated for restorative potential by human subjects ii) validate the ratings by testing whether they actually do produce restoration, iii) extract features from the labelled scenes using HMAX and IKSM and train regression models using them, iv) predict the restorative potential of unseen scenes, and v) test whether the predictions do produce restoration.

Image Dataset

As in 3.3, the 2688 image '8 Scene Categories' dataset was used here. However for modelling, human rating for restorative potential is required. A subset of 72 scenes was manually selected from the master dataset in a way intended to provide coverage of natural vs built and the potential for restorative vs nonrestorative scenes. These 72 images were supplemented with a further 8 scenes of industrial views, taken from the web, which the source dataset was considered to lack. The 80 images were then rated and validedby humans online using a custom system built in PHP, Javascript and MySQL.

Image Rating using PRS

15 Malaysian undergraduates (8 male, 7 female, aged 23-25) were given the Perceived Restorativeness Scale (PRS) and asked to rate each of the 80 scenes. The results of the rating are shown in Figure 5overleaf. Ratings for images in each row of the figure increase from left to right and each row is a continuation of the previous one. Inspection demonstrates that the subjects did not simply distinguish between natural and built scenes. Some natural scenes have low ratings and some built scenes are rated as moderate-to-highly restorative. This is consistent with the ART literature and suggests that finding features to capture the distinction between scenes with high and low restorative potential may be challenging. Natural scenes with low ratings appear to lack openness and their texture and organisation is somewhat chaotic. Built scenes rated highest display a higher level of organisation. Symmetry and structure may be important and it is interesting to note that the subjects responded to the buildings even though the style of architecture is not familiar, suggesting that personal associations may be less important here.

Validation of Ratings using SART

To ensure that the ratings reflected genuine restorative potential, validation was carried out following the 'SART-slideshow-SART' protocol described in Berto (2005). The 25 top and 25 bottom rated images from the 80 were selected for the slideshow. Paired t-tests on each of the four performance measures across sessions 1 and 2 were conducted. Table 2 shows the means and standard deviations of the four measures and the significance level of differences between sessions.

Table 2. Mean and std dev (parentheses) for d', reaction time, numbers of correct and incorrect responses between sessions and across images rated high and low on the PRS

	Session	High Rated	Low Rated
d'	1	2.18 (.95)	2.46 (.48)
	2	2.61 (.79)	2.10 (.66)
p		< 0.5	< 0.5
RT (ms)	1	282.73 (48.24)	289.60 (32.95)
	2	302.36 (50.40)	286.30 (34.09)
p		< 0.5	ns
Correct	1	10.18 (5.62)	11.80 (3.77)
	2	13.36 (4.78)	9.20 (4.32)
p		< 0.5	< 0.5
Incorrect	1	1.82 (3.31)	1.10 (1.91)
	2	1.64 (3.35)	1.70 (2.06)
p		ns	ns

Figure 5: 80 scenes arranged row-wise in order of Perceived Restorativeness

Figure 6. Scenes predicted to have low (top half) and high (bottom half) restorative potential by (columns) i) HMAX with linear SVM kernel, ii) Conspicuity Maps with polynomial kernel and iii) HMAX with polynomial kernel

Subjects viewing images rated high on the PRS demonstrated a significant increase in d' and number of correct responses, and significant decreases were found in the same measures after viewing the low-rated scenes. This is consistent with (Berto, 2005) and, small sample size notwithstanding, we will consider the dataset validated for the initial experiments performed here.

Model Construction

The dataset was used to learn a regression model intended to be capable of predicting the restorative potential of a given image. Matlab provided the basic software infrastructure used in these experiments. Functions from the image processing, optimisation and statistics toolboxes were utilised at pre-and and post-processing stages. The freely-available Saliency Toolbox (STB), a Matlab implementation of IKSM (Walther and Koch, 2006) was used to calculate and extract conspicuity maps (CM) for colour, intensity and orientation, which were combined into a single vector of 768 components for each image. A basic implementation of HMAX, hmin (hmin, 2014) was used in the same manner to calculate feature vectors of 8150 components. The pretrained dictionary of S2 features included with hmin was used in the calculations but by default no colour information was used. Both types of feature vectors were fed into the freely-available SVM library libSVM (libSVM, 2014) training in nu-SVR regression mode. Linear, polynomial, RBF and sigmoid kernels were trained.

Model Predictions and Validation

After training, the full 2688 images (minus the 72 training images) of the '8 Scene Categories' dataset were processed and the trained models were used to predict the level of restorative potential for each scene. The following three feature/kernel combinations were manually chosen for study based on inspection of their output: i) HMAX with a linear kernel, ii) CM with a polynomial kernel and iii) HMAX with a polynomial kernel. The results for the top and bottom 25 ratings for each combination are shown in Figure 6 (overleaf).

Inspection shows that each feature/kernel combination has captured a different aspect of the groundtruth. HMAX with a linear kernel makes a clear distinction between built and natural scenes. It appears to have associated high levels of activity in neurons trained to respond to to bar-like stimuli with low levels of restorativeness and is therefore biased towards scenes of skyscrapers. However, it has not captured the low restorative natural scenes or the high restorative built scenes. CM with a polynomial kernel is mixedalthough there is a bias towards open nature in the high-rated scenes. Interestingly, although no sunsets were in the training dataset, it predicts them as highly restorative. This is not currently understood. HMAX with a polynomial kernel focusses on a distinction between spaciousness and detailed natural texture. Whether the emptiness of the low-rated scenes and the busy texture of the high-rated scenes is appropriate remains to be seen. Manual inspection may yield insight but the only way to verify how successful these predictions are is to test them for restorativeness.

The predictions made by each setup were validated using the same SART-slideshow-SART protocol used in constructing the groundtruth.Evidence of restoration,in terms of improved correctness and reaction time, was found only in the the HMAX/Linear SVM combination (shown in Table 3) for high-restorative scenes. The other two models showed no significant differences in any of the measures between sessions.

The HMAX/Linear SVM combination appears to have captured something and the results are based on predictions from a large dataset and actual testing on human subjects rather than statistical measures of model quality. However, understanding exactly what has been captured is challenging: although biologically-inspired features can have the advantage of greater plausibility when modelling human phenomena, high-dimensional feature vectors consisting of individual cell responses are harder to interpret than higher-level image processing descriptors such as colour, shape and texture, or mathematical properties such as entropy and complexity. This is especially true when advanced learning models like SVMs are used. Currently the model functions as a black box that provides little in the way of explanation. Another shortcoming of this experiment is that the images used were only of adequate rather than high quality, withinsufficient resolution to capture finer details of texture and pattern.This was commented on by subjects during the human rating of images. These deficiencies in interpretibility and image quality inspired the next experiment.

Table 3. Performance measures for images predicted high and low by HMAX/Linear SVM

	Session	Low Rated	High Rated
d'	1	2.63 (0.58)	2.53 (0.84)
	2	2.80 (0.62)	2.78 (0.73)
p		*ns*	*ns*
RT (ms)	1	322.6 (80.5)	330.4 (92.0)
	2	334.0 (82.4)	312.8 (89.0)
p		*ns*	*< 0.5*
Correct	1	12.73 (4.31)	11.60 (5.83)
	2	14.18 (4.33)	13.60 (5.66)
p		*ns*	*< 0.5*
Incorrect	1	0.73 (0.90)	0.80 (1.32)
	2	0.82 (1.54)	0.80 (0.79)
p		*ns*	*ns*

Higher-level Image Descriptors

The main goal of this experiment was to exploremodel interpretability; specifically, whether higher-level scene descriptors could capture restorative potential in more intuitive visual terms and whether a straightforward linear association would be discovered. In contrast with the previous experiment, a smaller dataset of superior visual quality was used and standard statistical measures were used to assess model quality.

Image Descriptors

The 9 image descriptors chosen here attempt to capture distinctive patterns in colour, texture, edge content, directionality, and layout. They each yield a histogram for an input image, which may be considered as a vector for machine learning purposes. Histograms have proved their worth in the context of content-based inexing and retrieval (CBIR), where they are a common descriptor used to represent global features

Figure 7. 32 Super high-resolution scenes arranged row-wise in order of Perceived Restorativeness

Of an image since they are invariant to translation and rotation and, when normalised, scale invariant. The desciptors used here are as follows:

- Colour histogram in RGB colour space
- Scalable Colour Descriptor (Chang, Sikora and Puri, 2001)
- Colour Layout Descriptor (Chang, Sikora and Puri, 2001)
- Edge Histogram (Chang, Sikora and Puri, 2001)
- Tamura Texture Features (Tamura, Mori and Yamawaki, 1978)
- Colour and Edge Directivity Descriptor (CEDD; Chatzichristofis and Boutali, 2008).
- Fuzzy Colour and Texture Histogram (FCTH; Chatzichristofis and Boutali, 2008).
- Auto Colour Correlation Feature. (Huang *et al*, 1997)
- Gist (Oliva and Torralba, 2001)

Descriptors i) – iv) are considered significant enough to be part of the worldwide MPEG-7 standard for representing visual content in digital video. Descriptor ix): the 'Gist', or 'spatial envelope' descriptor, was briefly introduced in the

section 3.3. Inspired by the human ability to quickly categorise scenes from coarse, low-resolution images (Oliva and Torralba, 2001), Gistwas developed to capture image energy across orientations, scales and locations in a relatively low-dimensional manner. It has been successfully used in human scene perception studies and in CBIR applications.

Image Dataset

The scene image dataset used consisted of 32 super high-resolution (≥ 1920x1080 pixels) scenes manually selected from the web. Nature scenes and built scenes were in the ratio 1:1and the scenes selected were intended to represent both high and low fascination environments in each category. The dataset was evaluated and labelled for restorative potential using the Perceived Restorativeness Scale (PRS) by 15 Malaysian students aged 20-22. Shown in figure 7 are the 32 scenes and their overall PRS score. They are arranged according to their PRS ranking: top-left is lowest and bottom-right is highest. As was the case for the image ratings in the previous experiment, natural scenes are generally highest-rated but some built scenes are also rated highly.

Model Construction and Results

Matlab and freely-available software libraries were used to process the 32 dataset images. The Java LIRE library (Lire, 2014) was used for descriptors i) – viii) and gist was calculated using its CSAIL implementation (Gist, 2014). A multiple linear regression model was then fit for each descriptor, with results shown in table 4 below. Before fitting the model, histogram components that were not significantly ($p < 0.05$) correlated with restorative potential were rejected, reducing the dimensionality of each descriptor for reasons of accuracy and the number of free parameters vs dataset size available. In Table 4, n is the original number of histogram components and n_{sub} is the remaining subset. It is clear that each descriptor contained mostly uncorrelated components. Two extreme cases were found: the Tamura texture descriptor had no correlated components and the ACC was found to have more than the available free parameters in the model, yielding no possible fit and a perfect fit, respectively.

Table 4. Model Fit Results for 9 Descriptors

Descriptor	n	n_{sub}	MSE	R^2	p
Colour Histogram	512	17	0.011	0.611	0.316
Scalable Colour	64	7	0.014	0.493	0.012
Colour Layout	120	3	0.017	0.412	0.001
Edge Histogram	80	23	0.003	0.875	0.096
Tamura	18	0	-	-	-
CEDD	144	14	0.012	0.583	0.147
FCTH	192	10	0.013	0.523	0.028
ACC	1024	125	0	1	0
Gist	512	16	0.008	0.732	0.038

Taking into account MSE, R^2 and p-values, it can be seen that only the Edge Histogram and Gist models givea good overall fit to the data, and this is in intuitive agreement with inspection when model predictions are plotted against human ratings in figure 8. To interpret these results in visually-meaningful terms it is necessary to considerthe edge histogram and gist descriptor in a little more detail.

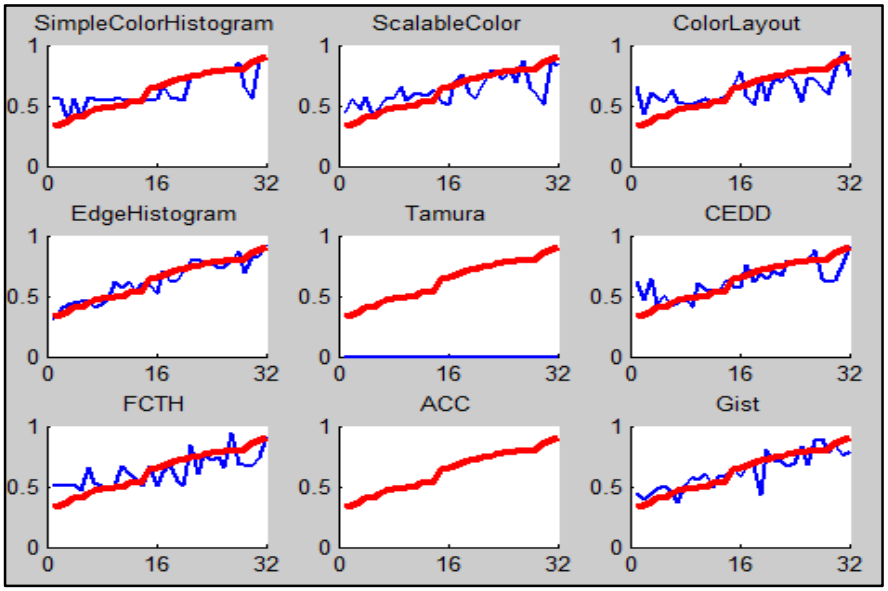

Figure 8. Predicted (blue) and Human-rated (red) Restorative Potential for 32 images

It has been long known that the detection of edges within an image is an important part of human vision, providing the raw material for shape and object perception (Hubel and Wesel, 1959). Essentially, the Edge Histogram descriptor describes the relative proportions, across image scales, of five types of edges, differentiated by their direction. The five edge types are shown below and can be broadly divided into vertical/horizontal (a-b) vs diagonal/non directional (c-e).

We find these results intriguing and encouraging but cannot consider them a convincing explanation yet. We conjecture that this result is actually a side effect of more profound and abstract properties that restorative environments possess which would also explain why it is possible for certain non-nature scenes to restore. This belief is explored in the final experiments performed in this paper.

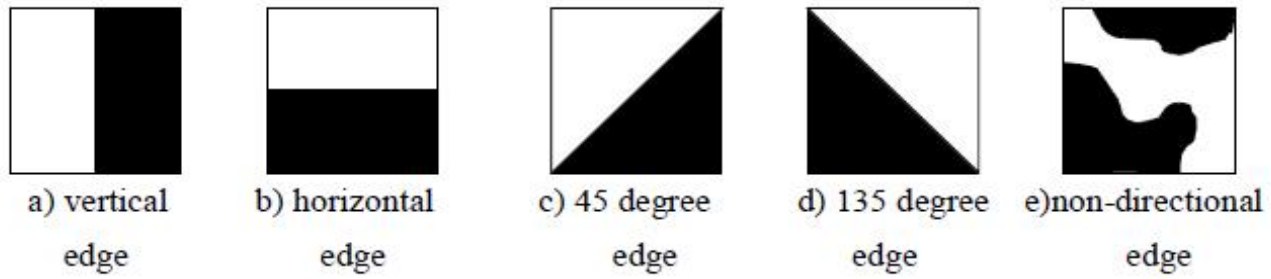

Figure 9. Five types of edges in the Edge Histogram descriptor (Park, Jeon & Won, 2000)

Analysis showed that the particular edge types correlated with restorative potential in this experiment were exclusively diagonal/non directional. These findingsare consistent with the everyday observation that modern city scenes are dominated by vertical and horizontal lines in a way that natural scenes generally are not. But the situation is perhaps not as simple as nature vs built: the dataset constructed here contained several highly-rated built scenes that are also not exclusively vertical and horizontal lines. The model was able to successfully predict their high restorative potential on this basis alone. It is worth noting that these buildings are in an older style of architecture with more curves, detail and complex patterns.

A shortcoming of the edge histogram is that variation in edge content at different locations in a scene is lost. The Gist descriptor does not measure edges directly but rather the image energy at various orientations, and analysis also showed that the orientations correlated with restoration were exclusively diagonal. Unlike the edge histogram, Gist has the advantage that it can also show the locations that distinguish restorative from non-restorative scenes. It was found that the presence of strong image energy oriented diagonally across the bottom quarter of the scene was most highly correlated with restorative potential.

These are currently the least developed and successful, but are, in our opinion, the most interesting and the best direction for future work.

Abstract Scene Properties

Order and Complexity

According to ART, fascinating scenes maintain sustained attentional engagement, but with minimal perceptual effort. The scene properties promoting this might be understood in terms of a balance between order and disorder. The element of order might allow minimal effort and element of disorder might provide uncertainty and continued interest. However, extreme order might be easy to perceive but quickly digested and unlikely to sustain attention andextreme disorder might require effort to make sense of and could be repulsive for that reason. Fascination would require a balance. Following (Grassberger, 1989), example extremes of order and disorder are illustrated in figure 10.The figure also shows a putative relationship between order and complexity;an appropriate balance between order and disorder is manifested as high complexity.

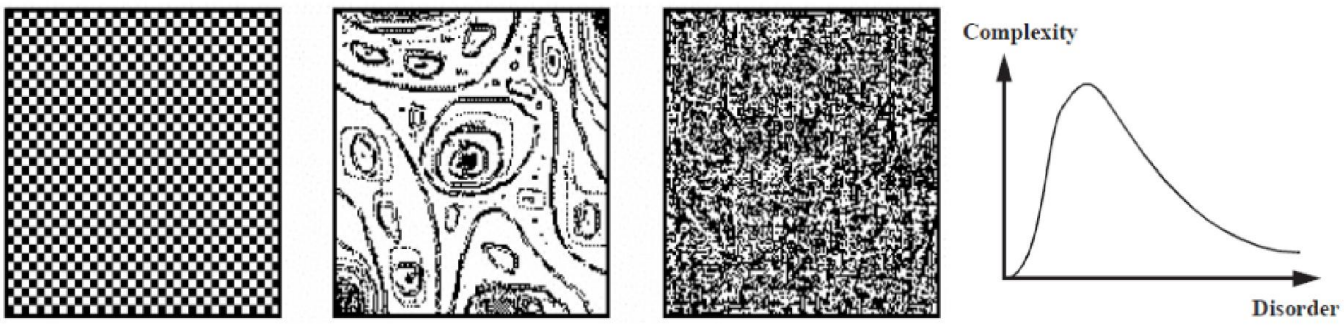

Figure 10. Examples of Ordered and Disordered Stimuli (Grassberger, 1989). Plus graph of complexity vs disorder

We believe that restorative scenes may have this balance, and therefore high complexity. To explore this hypothesis, it will be necessary to measure the order of an image, which is no simple task. In all that follows, the image dataset and ratings constructed in 5.2 will be used.

Symmetry

Order is a relationship between parts and a key manifestation of order is the presence of *symmetry* (Badii and Politi, 1999). Parts of a system are related to each other through simple transformations – they are similar to each other by reflectionor rotation, demonstrating harmony rather than disorder. The automatic detection of symmetries in images is a currently active problem inComputer Vision, though highly-challenging when applied to real-world images. Loy and Eklundh (2006) presents analgorithm for detecting local symmetry based on properties of the influential SIFT (Scale-Invariant Feature Transform) descriptor. The details of the algorithm are beyond the scope of this paper but its example output is illustrated in Figure 11.

The algorithm accepts an arbitrary image as input and attempts to identify the presence of both rotational and reflectional symmetry, showing lines of reflection and centres of rotation, each with an associated confidence level. A Matlab implementation (Symmetry, 2014) of this algorithm is freely-available and was used here to process the 32 images. A simple descriptor based on the algorithm output was the average of all detected instances of symmetry, weighted by their confidence level. No significant association with restorative potential was found. Inspection revealed that the current performance of symmetry detectors is simply inadequate for the task here.

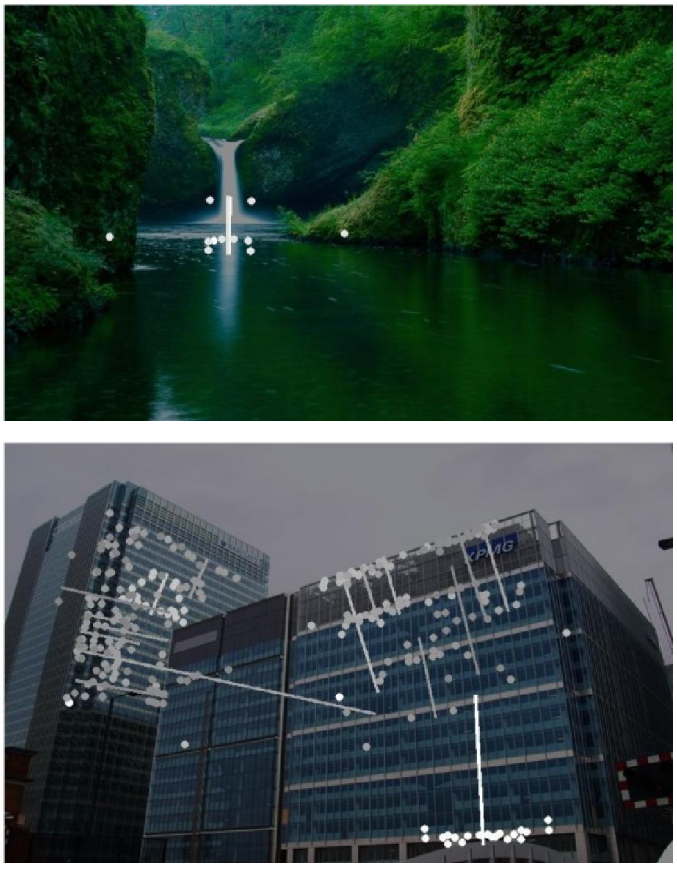

Figure 11. Examples of reflectional symmetry detected in the dataset

Self-Similarity

Related to symmetry is *self-similarity*, another manifestation of order (Brown and Liebovitch, 2010). In mathematics, a self-similar object is exactly or approximately similar to a part of itself (i.e. the whole has the same shape as one or more of the parts). Many objects in the real world, such as coastlines, are statistically self-similar: parts of them show the same statistical properties at many scales and self-similarity is a defining property of *fractals* (ibid). A commonly-used measure of the self-similarity of a shape is the Hausdorff dimension. Straight lines have Hausdorff dimension 1 but fractals and more interestingand complex shapes are associated with higher values. The dataset images were first segmented by a popular algorithm (Shi and Malik, 2000) to reveal some approximation to the shapes therein. See below for an example contour identified for the bottom scene in Figure 11.

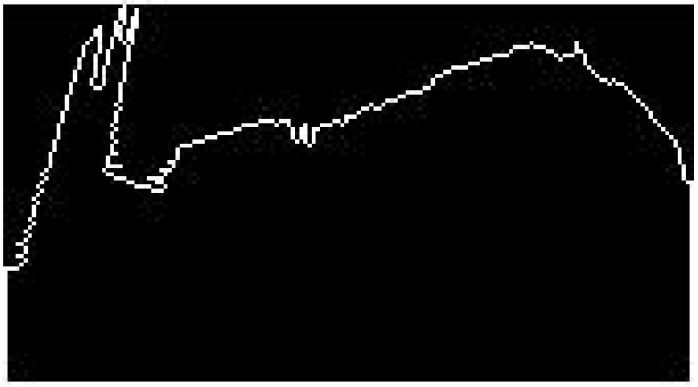

Figure 12. Example of scene shape contours determined by segmentation. This one has Hausdorff dimension 1.17

Then a Matlab implementation of the Hausdorff dimension for images was used to process the contour of each segment, and The average taken over all segments in the scene.Again, no correlation with restorative potential was found. This is also likely to be due to current deficiencies in segmentation algorithms.

Perception, Uncertainty and Information

A complementary way to interpret ideas of order and complexity is in terms of uncertainty and information. States of a totally ordered system are completely certain and therefore observation yields no information since its state could be completely predicted beforehand. In contrast, observing systems with high disorder yields high information since their state cannot be predicted before time. Situations between these two extremes can reflect different levels of perceptual effort and interest. Appropriate levels of uncertainty (and therefore information) in a scene can be mysterious and engage sustained attention if not too much effort is required. The change of perspective from order to information is useful because it lends itself to the consideration of scene perceptionand attention as a process of reducing uncertainty about what is there in a scene. At the start, we may know nothing about the scene and our uncertainty is total, but each glance provides new information and uncertainty is reduced to a level acceptable for our current purposes. Among other things, the content and the structure of the scene determine how much information is available and how it is revealed to an observer.

We believe that fascinating scenes possess moderate levels of information and reveal it at a steady rate. In other words, the more one looks, the more is seen. Mathematical information theory gives us a means to formalise an approximation to these ideas and explore this hypothesis. Although mathematical information does not capture everything one would intuitively mean by the term, it does reflect some important aspects of perception and communication in terms of probability distributions. In mathematical terms, the information content of a stimulus (signal, image) received by an observer increases with its uncertainty.

This final experimentis based on work by Rigau *et al.* (2008) in the computational characterisation of fine art aesthetics. It will capture scene structure and diversity in information-theoretic terms. For some scene property of interest such as colour or texture, we may ask how much information a given local region provides about that property for the scene as a whole. If the scene is not very diverse then most regions will predict the global property well, but if diversity is high then most regions will not. The technical details of the method used here can be found in (Rigau *et al,* 2008) but the basic idea is that the a visual property such as colour or texture is chosen and a scene image is be successively divided in half in such a way as to maximise the information the segments created provide about the property for the scene as a whole. If a scene is not diverse, it contains little information, and this process will finish quickly since all regions are similar. If a scene is diverse it will have much information to give and the segmentation process will take a long time. The key point is that associated with each stage in the segmentation process is the proportion of total information known so far to total information available. Therefore, the rate of information acquisition can be found for any given image and we could therefore ask two complementary questions about the segmentation process: how much of the total image information has been acquired after the nth segmentation or how many segmentations does it take to acquire x% of it? The answers will be different for different images and therefore offer a means to compare how different images might be perceived in terms of information, and potentially their restorative potential.

We custom implemented the method of Rigau *et al* in Matlab and processed the image dataset to segment each image on histograms for each of the 9 image properties described in section 5. Some examples of the segmentation and the varying information rate revealed are shown in Figure 13 below.

Figure 13. Example of information-theoretic segmentation for 20 segments on intensity histogram. The amounts of total information revealed at this stage are 40% and 53% respectively

Unfortunately, no significant associations between the percentages of information revealed with this method and restorative potential were found.

We are currently in the process of extending Rigau *et al*'s method with IKSM to only include actual locations of attention instead of the whole scene, since humans often do not attend to all available information. It is hoped that improvements will be found.

Conclusions and Future Work

This paper has decribed the early stages of work in a novel,but potentially significant,area.The results presented here are encouraging but suggestive rather than definitive. There are manymore options for developing a visual model of restorative scenes and work to improve all the preceding in terms of method and analysis is ongoing.

There is exploration too of other ways in which computing and technology can contribute to ART. The state of fascination which is so central to attention restoration is little-understood and no method of directly measuring it has been developed. So far, only the relevant subscale of the PRS provides any means of measuring fascination but this is from self-report about scene properties and not the subject's inner state. We are developing a method to directly measure the three components of fascination, as distilled by Joye *et al* (2013): attentional engagement, positive affect, and effortlessness. Eye tracking technology will be used to measure attentional engagement (by scanpath analysis) and effort (through pupil size), and EEG analysis for emotional valence and arousal. If successful, the PRS could be supplemented or superseded and faster rating of images might be possible.

Work is also underway to extend Itti and Koch's system to directly model the fatigue of directed attention, a feature missing from all known computational models of attention. It is hoped that this will allow a grounded extension of ART which may be of value to both computational modelling and Environmental Psychology. As a pilot study the work presented here may be considered successful and the main hope is that it will stimulate other researchers to join us and improve it. This is only the beginning, but in the future Environmental Psychology might seek to use Computer Vision techniques more and Computer Vision researchers will be challenged to develop methods to meet these new demands.

REFERENCES

Badii, R. and Politi, A. 1999. Complexity : hierarchical structures and scaling in physics. New York : Cambridge.

Bell, P. A., Greene, T. C., Fisher, J. D. and Baum, A. 2001. Environmental Psychology (5th edition). Toronto: Harcourt Brace. College Publishers

Berman, M. G., Jonides, J. and Kaplan, S. 2008. The Cognitive Benefits of Interacting With Nature. Psychological Science, 19(12), 1207-1212.

Berto, R. 2005. Exposure to restorative environments helps restore attentional capacity. *Journal of Environmental Psychology,* 25, 249–259.

Berto, R., Massaccesi, S. and Pasini, M. 2008. Do eye movements measured across high and low fascination photographs differ? Addressing Kaplan's fascination hypothesis. *Journal of Environmental Psychology*, 28(2), 185-191.

Borji, A. and Itti, L. 2013. State-of-the-art in Visual Attention Modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 1, pp. 185-207.

Brown, C.T. and Liebovitch, L.S, 2010. Fractal analysis Los Angeles : SAGE.

Chang, S.F., Sikora, T. and Puri, A. 2001. Overview of the mpeg-7 standard. IEEE Transactions on Circuits and Systems for Video Technology, 11(6):688–695.

Chatzichristofis, S. A. and Boutalis, Y. S. 2008. Cedd: Color and edge directivity descriptor. a compact descriptor for image indexing and retrieval. In A. Gasteratos, M. Vincze, and J. Tsotsos, editors, Proceedings of the 6th International Conference on Computer Vision Systems, ICVS 2008, Volume 5008 of LNCS, pages 312–322, Santorini, Greece, May 2008. Springer

Chatzichristofis, S. A. and Boutalis, Y. S. 2008. Fcth: Fuzzy color and texture histogram a low level feature for accurate image retrieval. In Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2008, pages 191–196, Klagenfurt, Austria, May 2008. IEEE

De Kort, Y.A.W., IJsselsteijn, W.A., Kooijman, J. and Schuurmans, Y. 2003. Virtual Laboratories: Comparability of Real and Virtual Environments for Environmental Psychology. In Presence Vol. 12, No. 4, Pages 360-373.

Egner, S., Itti, L. and Scheier, C. R. 2000. Comparing attention models with different types of behavior data. In Proceedings of Investigative Ophthalmology and Visual Science 41(4), 39.

Forsyth, D.A. and Ponce, J. 2011. Computer Vision: A Modern Approach. Prentice-Hall.

Gershenfeld, N. 1998. The nature of mathematical modelling. Cambridge University Press.

Gist (2014)

Grassberger, P. 1989. Problems in Quantifying Self-organized complexity. Helvetica Physica Acta 62, 498-511.

Hartig, T. and Staats, H. 2005. Linking preference for environments with their restorative quality. In B. Tress, G. Tress, G. Fry and P. Opdam. (Eds.). From Landscape Research to Landscape Planning. Aspects of Integration, Education and Application. Dordrecht: Springer: 279-292.

Hartig, T., Evans, G.W., Jamner, L.J., Davis, D.S. and Gärling, T. 2003. Tracking restoration in natural and urban field settings. *Journal of Environmental Psychology*, 23:109-123.

Hartig, T., Korpela, K., Evans, G. W. and Gärling, T. 1997. A measure of restorative quality in environments. Scandinavian Housing and Planning Research, 14, 175-194.

Haykin, S. 2009. Neural Networks and Learning Machines (3rd Edition), Prentice-Hall.

hmin, (2014) http://cbcl.mit.edu/jmutch/hmin/. Accessed 30th August 2014.

http://people.csail.mit.edu/torralba/code/spatialenvelope/ Accessed 30th August 2014.

http://www.nada.kth.se/~gareth/homepage/local_site/code.html . Accessed 30th August 2014.

Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J. and Zabih, R. 1997. Image indexing using color correlograms. In Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, CVPR '97, volume 00, pages 762–768, San Juan, Puerto Rico, June 1997. IEEE

Hubel, D. H. and Wiesel, T. N. 1959. Receptive Fields Of Single Neurones In The Cat's Striate Cortex, *Journal of Physiology*, I48, 574-59I.

Hyvarinen, A., Hurri, J. and Hoyer, P. O. 2009. Natural Image Statistics. Springer-Verlag.

Itti, L. and Koch, C. 2001. Computational Modeling of Visual Attention. Nature Reviews Neuroscience. 2(3), 194-203.

Itti, L., Koch, C. and Niebur, E. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. In IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(11), 1254-1259.

Joye, Y., Pals, R., Steg, L., and Lewis Evans, B. 2013. New methods for assessing the fascinating nature of nature experiences. PLOS ONE, 8(7)

Kaplan, S. 1995. The Restorative Benefits of Nature: Toward an Integrative Framework. *Journal of Environmental Psychology*, 15: 169-182.

Kaplan, S. and Talbot, J. F. 1983. Psychological Benefits of a Wilderness Experience. In I. Altman and J. F. Wohlwill. (Eds.). Behaviour and the Natural Environment. New York and London: Plenum Press. 6: 163- 203.

Kaplan, S., and Kaplan, R. 1981. Cognition and the environment. Functioning in an uncer-tain world. Ann Arbor, MI: Ulrich.

Klein, J.T., and Newell, W.H. 1997. Advancing interdisciplinary studies. In J. Gaff and J. Ratcliff (Eds.), Handbook of the undergraduate curriculum: A comprehensive guide to purposes, structures, practices, and change (pp. 393-415). San Francisco: Jossey-Bass

Korpela, K. and Hartig, T. 1996. Restorative qualities of favorite places. *Journal of Envi-ronmental Psychology,* 16, 221-233.

Libsvm, 2014. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ Accessed 30th August 2014.

Lire, 2014. http://www.semanticmetadata.net/lire/. Accessed 30th August 2014.

Loy, G. and Eklundh, J.O. 2006. Detecting Symmetry and Symmetric Constellations of Features. Proc European Conference on Computer Vision (ECCV).

Oliva, A. and Torralba, A. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. Journal of Computer Vision,* 42(3):145–175.

Park, D.K., Jeon, Y.S. and Won, C.S. 2000. Efficient use of local edge histogram descriptor. Proceedings of the 2000 ACM workshops on Multimedia. pp51-54

Prince, S.J.D. 2012. Computer Vision: Models, Learning, and Inference. Cambridge University Press.

Repko, A.F. 2011. Interdisciplinary Research Process and Theory. Sage Publications.

Riesenhuber, M. and Poggio, T. 1999. Hierarchical Models of Object Recognition in Cortex. Nature Neuroscience 2: 1019-1025.

Rigau, J., Feixas, M. and Sbert, M. 2008. Informational Aesthetics Measures.IEEE Computer Graphics and Applications, 28(2):24-34.

Russell, B. C., Torralba, A., Murphy, K. P. and Freeman, W. T. 2005. LabelMe: a database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025.

Shi, J. and Malik, J. 2000. Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997

Symmetry (2014)

Szeliski, R. 2010. Computer Vision: Algorithms and Applications. Springer Texts in Computer Science.

Tamura, H., Mori, S. and Yamawaki, T. 1978. Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics, 8(6): 460–472.

Treisman, A. and Gelade, G. 1980. A feature-integration theory of attention. Cognitive Psychology, Vol. 12, No. 1, pp. 97–136.

Veith, R. and Arkkelin, D. 1995. Environmental psychology: an interdisciplinary perspective. New Jersey: Prentice-Hall.

Walther, D. and Koch, C. 2006. Modeling attention to salient proto-objects. Neural Networks 19, 1395-1407.

****