An Intelligent Model for Visual Scene Analysis and Compression

Amjad Rehman and Tanzila Saba
Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Malaysia

Abstract: *This paper presents an improved approach for indicating visually salient regions of an image based upon a known visual search task. The proposed approach employs a robust model of instantaneous visual attention (i.e., "bottom-up") combined with a pixel probability map derived from the automatic detection of a previously-seen object (task-dependent i.e., "top-down"). The objects to be recognized are parameterized quickly in advance by a viewpoint-invariant spatial distribution of Speeded Up Robust Features (SURF) interest-points. The bottom-up and top-down object probability images are fused to produce a task-dependent saliency map. The proposed approach is validated using observer eye-tracker data collected under object search-and-count tasking. Proposed approach shows 13% higher overlap with true attention areas under task compared to bottom-up saliency alone. The new combined saliency map is further used to develop a new intelligent compression technique which is an extension of Discrete Cosine Transform (DCT) encoding. The proposed approach is demonstrated on surveillance-style footage throughout.*

Keywords: *Visualization, DCT, image compression, scene analysis.*

Received May 27, 2010; accepted January 3, 2011

1. Introduction

Human vision is an active, dynamic process in which the viewer seeks out specific visual input as needed to support ongoing cognitive and behavioral activity [21]. Most vertebrates, including humans, can move their eyes. They use this ability to sample in detail the most relevant features of a scene, while spending only limited processing resources elsewhere. The ability to predict, given an image or video, where a human might fixate in a fixed-time free viewing scenario has long been of interest in the vision community. Besides the purely scientific goal of understanding this remarkable behavior of humans, and animals in general, to consistently fixate on important information, there is tremendous engineering application, e.g., in compression and recognition [39]. The standard approaches [26, 32] are based on biologically motivated feature selection, followed by center-surround operations which highlight local gradients, and finally a combination step leading to a "master map".

2. Related Work

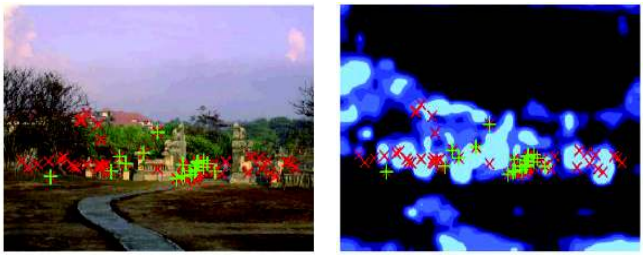
Recently, a few researchers have hypothesized that fundamental quantities such as "self-information" and "surprise" are at the heart of saliency/attention [6, 25]. However, ultimately, Bruce [6] computes a function which is additive in feature maps, with the main contribution materializing as a method of operating on a feature map in such a way to get an active, or saliency, map. Itti and Baldi [25] define "surprise" in general, but ultimately compute a saliency map in the classical sense for each of a number of feature channels, then operate on these maps using another function aimed at highlighting local variation. By organizing the topology of these varied approaches, we can compare them more rigorously: i.e., not just end to end, but also piecewise, removing some uncertainty about the origin of observed performance differences. Moreover, recent work referred in [19] is about the visual saliency for the particular task and therefore is task oriented only. In [12], authors present chaotic interleaving scheme for wireless image transmission with OFDM. The interleaving scheme is based on the chaotic baker map. The interleaving process is applied to the binary image data prior to the modulation step. Nonetheless, the scheme improves the performance of the OFDM system, where it generates permuted sequences with lower correlation between their samples. However, the process is complex and time consuming. Sundaram and Chang [40] proposed a two step process to condense scenes with respect to chromaticity, lighting and sound. Firstly, visual complexity of a shot is described and grammar of the film language is also analysed. However, the technique is suitable for scenes with text description. Authors claim good results on skims with compression rates between 60~80%. Thus, the leading models of visual saliency may be organized into three stages:

- *Stage 1. Extraction:* Extract feature vectors at locations over the image plane.

- *Stage 2. Activation:* Form an "activation map" (or maps) using the feature vectors.
- *Stage 3. Normalization/Combination:*To summarize, existing models of bottom-up saliency are reliable indicators of passive visual attention regions in an image [16, 26]. However, under task based viewing there is often a strong shift of attention away from the passive observation case [10]. This arises from the imposition of top-down processes by the observer under task in combination with the bottom up response [41]. Models have been constructed for the top-down case, but involve complicated prior learning of general object classes and their scene contextualization [13, 30, 33, 41]. Such models have the advantage of enhanced attention prediction power even in the absence of a target object being present in the image, but the complexity of the learning process and the specific scenarios makes these models hard to generalize.

In this paper, a task-oriented correction to bottom up models of visual attention is presented. The underlying premise of the proposed bottom-up correction is that if an object of interest is found to be present in the image, the general contextual information of the scene can be approximated [17]. The visual system under task-prioritized viewing is guided by prior experiences of associating task objects to likely scene location contexts [1, 3, 4]. Objects generally lie within semantically sensible parts of an image (e.g., pedestrians along pavements) and there is a known strong horizontal search bias in observers under task, based on the image context [17]. An illustration of this effect is presented in Figure 1 in which eye-fixation points from observers under task are imposed on an image. Detected objects can therefore be used to construct a horizontally-biased "object presences" attention map for combination with the attention maps from bottom-up/passive viewing models to give more accurate prediction of eye-fixations under task than the bottom up models alone.

Figure 1-b shows the threshold Graph Based Visual Saliency (GBVS) [16] map generated from the Figure 1-a, the eye-fixations are still imposed. In Figure 1, an image is observed under task with eye-tracker points superimposed. x denotes all eye fixations of eight observers performing people count on image. + indicates the first three eye fixations across all participants (Figure 1-b) GBVS model of passive visual attention computed from the image. The intentional map is threshold to 10 - 50% by area. Nonetheless, overlap is generally good, but there is still substantial energy in the map lying away from the core region while eye-fixations lie outside the threshold area.



a) Map generated. b) Threshold graph based visual saliency.

Figure 1. Graph Based Visual Saliency map generated from the upper image.

In this research, Speeded up Robust Features (SURF) [2] interest-point matching to a reference image is employed to determine presence of object in a test image. We further propose an object confirmation technique based on comparing the distribution between the reference and test image matched points to introduce higher confidence to the object recognition process. This process is not general object recognition, but would apply to particular object retrieval, such as finding a particular vehicle in a database from a single or small number of stored reference images, allowing for the possibility of scale, viewpoint and illumination changes.

We present comprehensive statistics detailing the eye-fixation predictive power of our combined attention model in comparison to the pure bottom up models, showing an improvement in overlap. We further present data on our object detection scheme's reliability over different viewpoint. Finally, as an illustration of what our object-present-task model could be used for, we demonstrate a DCT-based task-targeted compression scheme that preserves regions of high saliency under task at high fidelity and non-task critical regions at lower fidelity to offer a notable increase in compression ratio compared to global application of DCT-based, JPEG-like compression.

3. Object Detection and Confirmation from Surf Fonts

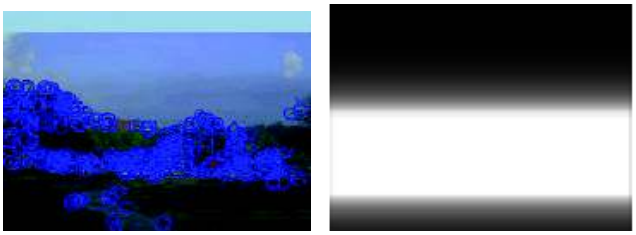
3.1. SURF Matching and Object Confirmation Refinement

SURF [2] is a robust feature-detector and descriptor combination that can be used for point to point matching between images. Generally, the SURF algorithm finds locally interesting points over many scales and stores these points into a set of point descriptors robust to rotation and scale transformations as well as skew anisotropic scaling and perspective effects, covered to some degree by the overall robustness of the description technique. The descriptor matching applies well over viewpoint change, scale and under different lighting conditions (see [2] for thorough performance measures) as well as being

naturally distributed towards visually salient information under different viewing conditions [18].

Due to the robustness of the matching technique under appropriate thresholding, the presence of descriptor matching between a reference image and a test image generally delivers a high confidence that the reference image content is present in the test image. For an example, SURF detection and matching process is exhibited in Figure 2. Circles in Figure 2 denote detected SURF points, "+" denotes matched points between reference and test images (image as shown in Figure 1). Figure 2-a manually extracted reference image with SURF points matched to test image. Figure 2-b test image showing SURF points matched to reference image. Figure 2-c, our object contextual correction map (unthresholded). In this paper we manually extract an object of interest from a larger image and store a squared-off copy of this object as a reference image along with a mask describing the object envelope within the reference image. We then "learn" such a reference image by running the SURF algorithm over it and storing the descriptors. Interest points outside the object envelope are excluded along with their descriptors. Of course, the thresholding for matching between images can vary and there could be mismatched point to point correspondences. For this reason, we propose a refinement in the object recognition technique based on the overlap between the matched points in the reference image and the matched points in the test image transformed homographically to the plane of the reference image. This allows greater confidence in the presence of the object as opposed to a series of unrelated but probably robust point to point correspondences.

a) Manually extracted reference image with SURF points matched.



b) Test image showing SURF points. c) Object contextual correction map.

Figure 2. Surf matching.

Although, there already exist good object classification and recognition techniques [14, 38, 15] but the technique proposed here is one of specific object recognition based on the distribution of interest points and not general object categorisation. In the process of object recognition a reference image (of an object of interest in this case) is "learned" by applying the SURF algorithm, transforming the matched interest-points into an invariant frame of reference then computing a spatial distribution of those points relative to one another. There is always the possibility of mismatching occurrence between points at poor thresholds and at substantially different viewing angles. We therefore use the corrected spatial distribution of the matches to parameterise the object. In brief there are two images, one is the Reference Image and the other is the Test Image. The matched points between these images have been calculated and the correspondence between these points is reliable. (This is achievable by choosing the appropriate thresholds at the SURF detection and matching phases.) It is assumed that the learned object in the reference image will have the features included in approximately the same plane. This should similarly constrain the matched points in the test image to be similarly planar and only 4 good matches are required. This is a reasonable assumption since surveillance objects are usually imaged in the medium to far field. Matched points are denoted (IMP) where I indicates image label, R or T for reference or Test, M denotes that the points are matched and P' is the plane. Therefore, (TMT) denotes the Matched Test Points in the Test Image plane. We first calculate the homograph matrix between test and reference image planes using matched points in each image, TMT and RMR. We then transform the matched points in the test image to the reference plane, TMR. This is done by using the homographic relationship standard in computer vision. We use the algorithm detailed in [20]. The spatial distribution of the TMR is then computed. This is the angular and radial displacement to all points from a zero point.

$$D(T_{MR'})=(r_{1\ldots n-1},\theta_{1\ldots n-1}) \quad (1)$$

The distribution of matched test points, TMR, with each reference point is overlapped and the displacement δ is calculated. Then for each reference point we apply a Euclidean distance threshold and count the inliers to δ. The best zero location is chosen based on greatest number of inliers. The object is judged to present or absent, based on the number of classified object points within the threshold for the best fit matches. We choose to set six confirmed object points as our threshold for object confirmation and this fix value is evaluated experimentally. In the case where there are no inliers, it is necessary to choose a different test point for constructing the distribution and to repeat the process. The full set of permutations could be explored but is not necessary if the confidence in the matching is high: after a few permutations, the method will find the overlap with the

points within a given tolerance if the object is present. We find that the method is robust for assessing the inliers of an object distribution. Since the object distribution is only derived from the matched points in the test image, the technique works in the presence of partial occlusion where the number of matches may fall but the distribution is still likely to overlap in the reference image.

3.2. Object Context Surface

We are trying to model visual attention under task by modulating the reliable bottom up maps with a contextual search surface based on object presence. Once a reference object of interest is detected in an image, we construct an "object context surface" for combination with the bottom up map of the raw test image. The premise behind the construction of this surface is that objects of similar class are generally horizontally distributed in an image and that under task there is consequently a strong horizontal bias in attention. Of course, there is a compromise in judgment required here. The horizontal search pattern is quite strong for scenes with some kind of horizon and starts to break down as the potential image area for search increases, e.g., as altitude is increased from eye-level observation towards aerial photography. The horizontal constraint is generally true for eye-level imagery. We choose to construct our surface using the following steps:

1. We find the centre height of the detected object from the matched SURF points in the test image.
2. We take the horizontal line through the centre point as our axis of object context.
3. We take two boundaries to define the core context of the image, one $1/6^{th}$ of the image height above and the other $1/6^{th}$ of the image height below the centre line. We saturate the map within this area. (In the case that the detected object is larger than the $1/6^{th}$ height either side of the centre, the size of the object is chosen instead of the $1/6^{th}$ image height.)
4. Outside the core context, the map is tailed-off in the vertical direction according to the formula:

$$distance\,(x,y)=(y-Cny)/2 \tag{2}$$

where, (x,y) is the current point in the map of the same dimensions as the image (excluding the core context) and cny is the closest point y dimension to the saturated mask. This equation weights the distance values so that there is a tail-off from the core context in the vertical direction. See Figure 2 for an object presence surface example.

3.3. Combination of Bottom up and Object-Presence Contextual Surfaces

Now we have models for the bottom up case and for the target present case. We wish to fuse these data maps together in a way that will preserve the core information. The bottom up map contains important contextual information likely to attract attention under passive observation. The top down map is based on the detection of a known object and is based on prior knowledge of search bias in observers in general naturalistic imagery. The combination depends strongly upon the degree of belief of the value of each component. In our case, by inspection on our data sets (see later-Validation), we set the surf detection and matching thresholds appropriately so that we have strong belief in the presence of our object based on SURF matching alone. In the case of further "object confirmation" by distribution as outlined above there is similarly a high belief in the plausibility of the top-down surface.

The human visual system deals with task by reading the bottom up information in a scene but imposing contextual constraint on the search. Therefore we seek to combine the surfaces in such a way that the object map dominates, but allows for strong bottom up areas to remain possible attention zones. Due to each being derived from different bases it is common to apply a power to the maps prior to combination in the general form of equations 1 and 2.

$$C\,(B(x,y) * O(x,y))\gamma \tag{3}$$

where, C is the combined map, BU is the GBVS saliency map [39], O is the "object" surface, either from SURF points or from object-classified points. The indices (x, y) are the pixel locations and are included to show that the above is elemental, not matrix multiplication. We choose γ=0.05 in this paper experimentally.

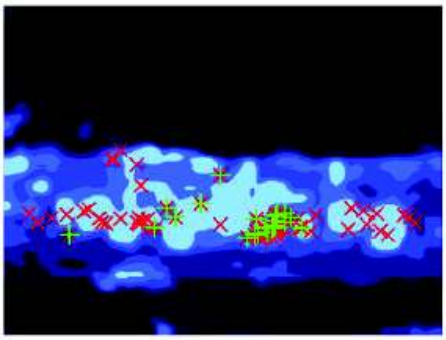


Figure 3. Thresholded combined bottom up GBVS map and object context correction.

This has the effect of flattening the object distribution somewhat. We also rescale the values before combination of the two different maps to reflect their degree of belief. We trust the bottom up map in the passive case, but it is less reliable in the task case. We therefore set the pixel values of the bottom up map between [0.32-0.95] and the values of the object map between [0.93-0.95] while retaining a large floating point value for each pixel, allowing for smoothness. These values are chosen out of many possible values to

produce maps that combine to offer domination of the attentional surface by the object context component with the possibility of diversional attention to the bottom up map. An example of the thresholded combined map is shown in Figure 3 against eye-tracker data. Eye fixation data overlaid from all participants. All fixations denoted by "x", first three only by "+". Note the shift in energy towards the known target area compared to the pure bottom up case in Figure 1. Note the overlap improvement compared to GBVS alone illustrated in Figure 3.

4. Validation of the Combined Surface under Task

Since here the aim is to build an improved attention map, the map is tested here against human observer eye-fixations, taken under tasking, to assess the valuable correction of the proposed object-surface to the bottom-up only case.

The eye-tracker data and image set from Torralba *et al.* [41] has been used to validate the model. The test image data set for this paper comprises 72 images and 108 search scenarios (3×36 tasks) performed by 8 observers, "count the people" on the first 36 images and "count the cups" and "count the paintings" on the other 36. Objects appropriate to the search-and-count task performed by the observers are manually extracted as reference images from the test images for all 108 search scenarios e.g., if task is to count the number of paintings, a painting would be extracted from the image, if present. Overall there are 61 object present cases and 61 appropriate objects for task are extracted out of the 108 tasks.

These objects are stored as reference images in a head-on, 0 object recognition test process. (A discussion of matching and object detection under different angular viewpoint follows.) Each reference image (i.e., 1 extracted "object" per task) is tested using one-pass surf descriptor matching against the descriptors from a one pass surf application to the other 108 search task images. Our surf matching thresholds are such that there is practically no mismatching between the reference and test images. The largest number of mismatches per image (i.e., matches from the wrong image) is 1 and statistically very rare event. 30 out of 61 objects are recognized using our object confirmation (This is > 6 matched surf points lying within overlap tolerance) and 47 examples has reliable SURF matching to at least three points in the test image. The reason for these low values can be attributed partially to our conservatively high matching threshold and partially to the object size in the image - often the area of the object pixels is very low(1% of image area of 800×600) and this does not allow for robust descriptor representation in quantity.

For each search scenario, three saliency maps are created: First, bottom up saliency; second, SURF combined map only from matched points; third, object combined map from object points. The construction mechanism for second and third saliency maps are identical, but there is a subtle change of the object centre line since not all matched points are classified as object points. Essentially the statistics from second and third saliency maps are identical within reasonable error, so below only cases first and second saliency maps are presented. Whereas, third saliency map is actually a refined subset of second saliency map allowing for higher confidence.

All 72 images have applied a bottom up map. SURF-only object maps are constructed when there are at least three matched points between the reference and test images (47/61 cases). Object surfaces are further constructed when there are more than 6 points classified as object points (30/61). Where the object is detected, the bottom up and top-down object contextual maps are combined as described above.

The attentional maps of each class are thresholded to different image areas representing the more salient half of the image. X=10, 20, 30, 40 and 50% of image area are chosen since these levels clearly represent the "more salient" half of the image to different degrees, as illustrated in Figures 1 and 4. The Figure 4 exhibits an overlap with the attentional maps at different threshold levels for all eye-points, gathered over 8 experimental participants under task. Left: eye-data vs. bottom-up only (72 images, 8 participants, 108 tasks). Right: eye-data vs. combined bottom-up and SURF-point object context attentional maps. (47 incidents of # SURF Matches > 3, 8 participants, 47 tasks). The bar indices 1 to 5 correspond to the 10 to 50% surface area coverage of the masks, as illustrated in Figures 1 and 3. The main axis is the percentage of interest points over the whole image set that lie within the saliency maps at the different threshold levels. The bars indicate average overlap at each threshold. Errors: standard deviation is plotted in red. There is a ten percent (or so) higher eye-fixation overlap when object context can be combined with the pure bottom up case.

For each search scenario, the eye tracker points lying within and without each threshold level of each mask are counted. We choose to use all eight participants and to process all of the eye points. This gives the exhaustive search case. The overlap is considerably higher if only the first three fixations are considered, but such fixations may contain elements of centre bias and so the statistics are not presented here. The comprehensive statistics for the overlap of the eye fixations under task are shown in Figure 4. On the left, the overlap of under-task eye fixations of all 8 observers over all 108 tasks vs. the 72 bottom up maps is shown. On the right, the overlap of the eye-fixations of all 8 observers in the 47 tasks where an object surface from atleast three SURF matches could be constructed. There is a substantial overlap improvement using our object-present surface, with

there being approximately a 10% higher attentional overlap relative to the bottom up models alone.

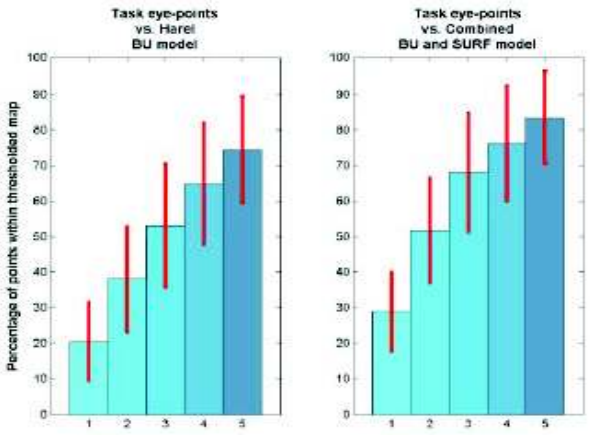


Figure 4. Overlapping with the attentional maps at different threshold levels.

4.1. Validation of SURF and Object Confirmation Over Angular Viewpoint Change

We performed a test on the matching performance and object classification over different viewpoint angles. Six sets of viewpoint shift images were collected, each based around a different object in the scene Viewing angle varied from head on 0° to 30° in steps of 5°. From each image in the set, an object region was manually extracted as a reference image. This object was between 10 and 20% of image area. The relative angles between reference plane and the other images in the set were known and the matching performance of both the SURF points alone and the Object Recognition refinement were tested over the different viewpoints. It was found that the object confirmation breaks down between 15 and 2 of offset from the reference, while the SURF matching alone generally started to collapse beyond 20°.

5. DCT Compression using Combined Saliency Maps

We have demonstrated a technique that can successfully modulate attention maps from the bottom up model alone to adjust for task based viewing that relies on simple object recognition. If we know the zone of an image that is of interest to an analyst, we can apply a selective compression targeted towards those areas of the image that are task critical. This could potentially save a lot of bandwidth. Many compression schemes are applied globally. This requires using some rule of thumb to maintain all potential information within an image, which means that the compression is not as strong as it could be, or it involves pushing the global compression further at the risk of destroying key information in an image. Here we propose a simple method of how our attention-enhanced map could be used to apply an intelligent compression to an image.

The JPEG algorithm is designed for good visual quality in photo-real images and so is appropriate in our examples. JPEG relies on quantization of the Discrete Cosine Transform applied to 8 by 8 pixel blocks of an image. This reduces the relatively unimportant high frequency components in each block, allowing for efficient Huffman or arithmetic coding. The quantization is performed using a quantization matrix derived from psycho visual tests and this matrix can be weighted to provide the required degree of compression in the block. The reverse process decodes the image [11, 43]. The heavier this quantization, the larger the compression ratio achieved, however this is tempered by the fact that over-quantization will produce blocking artifacts that significantly reduce image quantity and can damage real information within the image. In regular JPEG, the quantization is fixed across the whole image. In our case, however, we have a reliable method of selecting regions of contextual search interest under task. Our previous analysis leads us to "expect" 85% of eye fixations to lie within the top 50% of images by object contextual saliency. Therefore, half of the image for high and half for low compression are thresholded. However, this low information will not be lost altogether and will be available for contextual guidance.

We use a greyscale copy of the image and choose two quality factors to impose a high or low quality on the image region. The quality factor (Q) of 50 uses an unweighted matrix which is the original matrix derive from psycho visual experiments to give acceptable compression. The quantization matrix we use is that specified in Annex K of the JPEG standard for the luminance component of images [2], appropriate for greyscale. We choose a low value of Q=3 for the outlying regions and weight the quantization matrix according to the following relationship: $(50/Q)*Qmatrix$. We choose this exaggerated example of compression to illustrate the technique.

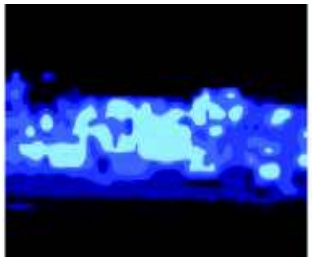
In practice, higher values of the non-core regions would be chosen which would be more visually pleasing and would not necessarily take up very much more storage. We ran the binary DCT technique over a set of 50 greyscale test images with object matches from reference images from differing angles. For comparison, we also applied a global DCT compression to the test images at Q=50. We found the average compression ratio (length (quantized, linearised DCT image string): length (Huffman-encoded string)) for the Binary- Task-DCT-Huffman is 6±0.2, while for the Global-DCT Huffman the average is 5.5±0.5. The final compressed output had a storage value of 0.1737 bits per pixel for the binary, object-context compression and 0.1927 bits per pixel for the global compression. The gain in compression outweighs the required storage for the reference

descriptors and is valuable for large datasets. Figure 5 presents a task-oriented compression based on 9 matched and confirmed object points between test and reference. Figure 5-a our reference image containing an object of interest. Figure 5-b our test image, at a different viewing angle to reference. Figure 5-c a zoom-in of the detected object in the full intelligent compression (seen in Figure 5-c). Figure 5-d the thresholded combined attentional map of the bottom up and object context in the test image. Black regions set to Q=3, others Q=50 in the attention based compression. Finally, Figure 5-e exhibits the full compressed image. Non task-core regions are heavily averaged while maintaining the background, core regions are preserved. (See Centre left for a zoom-in of the detected object.

a) Image containing an object of interest.

b) Test image.

c) Zoom-in of the detected object in the full intelligent compression.



d) The thresholded combined attentional map.

e) The full compressed image.

Figure 5. An effective illustration of object-context oriented compression.

6. Performance Comparison in the State of the Arts

The present study investigates how task instruction affects eye movement patterns during the viewing of a scene. The effect of task on eye movement patterns has been long established by pioneers of research into eye movement patterns [7, 44]. However, as convincing as they are, the accounts of these effects are descriptive, fixation data are depicted as images, and the results lack quantification. We sought to provide quantitative analyses, with a specific emphasis on investigating the nature of fixation durations in addition to their placement. We found that task effects are observed at both the scene and object level of analysis. It is also found that task affects both the placement and fixation duration patterns during scene viewing. Tasks effect on eye movements across the whole scene. At the level of the whole scene, fixations are more distributed in the memorization condition and are more focused on search-relevant regions in the search condition, which directly replicate the findings of [7, 44]. This is not a surprising finding when we consider the strategies involve for each task. In the Memorization task, participants are told that they would be tested on specific objects within the scene, and so to improve encoding of the different objects, it makes sense that they would try to fixate as many different objects as possible.

This pattern of spreading fixations over many different objects is also reflected to some extent in the total scan pattern and to a greater extent in the greater total number of fixations; that is, there is a numerically longer scan pattern and a higher count of fixations in the Memorization than in Visual Search task. In the visual search task, fixations are more narrowly focused within the scene, and we can assume that participants limited their fixations to the areas that most likely contained the target. This finding is consistent with other studies showing that context information leads to more efficient searches [3, 4, 5, 9, 31]. Furthermore, this effect of context has been implemented in a recent computational model proposed in [41], which show that participants' fixations largely remain within scene areas that are statistically most likely to contain the target object. It is found that for both tasks average fixation durations increase as viewing time increased (for the first 5 fixations) and then remains stable in the later viewing period. This finding is consistent with earlier studies that are reported similar patterns [42].

It is also observed that the fixation durations stop increasing after only È2s. This steep increase during the first seconds of viewing is also found in other studies in which the task demands rely on the quality of the initial performance (e.g., È3.4s in [42]). One could conclude based on the fixation duration data that the initial scanning of the scenes does not differ between tasks. However, the saccade amplitude measure across the whole scene seems to point to a different pattern, which we will turn to now. When we examine average saccade amplitude, there are no systematic differences between tasks; however, there are differences in the saccade amplitude during the initial viewing of the scene. We find that participants make longer saccades during initial viewing in visual

search versus memorization, while later saccades do not differ across tasks. Again, this difference can be attributed to the strategies that participants are implementing as they examine the scene in each task condition. However, it is not clear whether the difference in saccade amplitude is due to the participants staying closer to the centre during the initial encoding for the memorization task, or whether they are simply scanning the whole scene more thoroughly in the visual search task. It is not clear what a proper baseline for these tasks would be, but in a preference rating task, Antes (1974) found that the average saccade amplitudes seem to decrease with increased viewing time. If we see this as the default (an initial wide scanning of the scene with the first few seconds of viewing), then it may be that when a memorization strategy is implemented the system can immediately start to examine details without the need for an initial wide scanning of the scene. This finding is in direct contrast to other studies that report that the first few fixations are controlled by stimulus factors alone [28].

For instance, Mannan *et al.* [28] measure eye movements while viewers examine greyscale photographs for 3s each. The photographs are either high-pass filtered, low-pass filtered, or unfiltered. Results show that fixation positions are similar on the unfiltered and lowpass filtered scenes during the first 1.5s of viewing. However, as noted by Henderson and Ferreira [23], even if eye movement control is largely determined by stimulus factors during the initial scanning of the scene, this does not prevent fixations from being influenced by task. In the present study, the immediate implementation of the memorization strategy is also seen with the elapsed time to the first saccade and is discussed further below.

The elapsed time to the execution of the first saccade is much longer for the memorization task than the visual search task. This elapsed time until the first fixation (or the initial fixation at scene onset) is theoretically different from other fixations made on the scene because it involves identifying the scene being presented [8, 9], as well as deciding and planning where to target the next fixation [9].

The additional 48 ms in the Memorization task, in addition to the shorter initial saccade amplitudes discussed above, suggest that the effect of task is immediate. That is, the largest differences between the tasks are seen within the first few seconds of viewing with both saccade amplitude and fixation duration becoming similar in the latter part of the viewing period. This immediate effect is interesting in light of other top-down influences, such as the effect of scene context on the examination of objects within the scene [22, 35], which seems to only emerge in later viewing. Top-down effects due to scene semantics seem to take a while to onset (relative to the whole viewing period), while top-down effects of task are seen immediately and seem to be more pronounced in the first few seconds of viewing. Task effects eye movements on objects. To better understand the effect of tasks on the examination of objects within the scenes, we also look at fixation patterns on objects. As would be expected from the effect of task on the distribution of fixations, we also find that participants tend to examine more objects in the memorization task condition. However, theoretically more interesting is the failure to observe an effect of task on the average fixation duration. The reason this is interesting or even surprising is that the lack of an effect of the task goes against the findings in the reading literature, in which effects of task, context, word difficulty, and word length are seen at the level of the average fixation duration [36]. Instead, we find that task affected gaze duration by modifying the number of fixations within a gaze on a given object.

The same pattern is also seen across other aggregate measures of eye movements on the objects viewed in the scenes. This finding is consistent with the failure to observe effects of other factors on individual fixation durations during scene viewing [23, 24]. In general, participants tend to spend more time fixating objects in the memorization task than in the visual search task. However, this is seen in the number of times that the objects are fixated, not in the average fixation duration.

This finding is consistent with an earlier study by Loftus [27] that reported memory for scene regions is not related to the average fixation duration but rather to the number of fixations made on the region. The finding that the number of fixations is greater for memory than the visual search task can be easily attributed as a system level strategy by which visual information in the memory task is encoded more thoroughly. However, because an equivalent effect is not found at the level of the fixation duration may indicate a limit in the architecture governing the decision of when to move the eyes. Rather than influencing when the eyes move, the effect of the task on scenes is observed in regards to where the eyes move. When to move the eyes during scene perception based on Morrison's [29] reading model, researchers in [22, 36] suggest that the decision of when to move the eyes during scene viewing is based on the processing of currently fixated visual information to a certain level. In reading, that level is thought to be lexical access of the word [37], whereas in scene viewing, it is proposed to be the recognition of the object at fixation [22].

In a recent set of studies, Henderson and colleagues [24] investigate the degree to which the currently fixated visual information affects fixation durations on scenes. By masking the stimulus at the end of a saccade, the availability of the scene information is delayed. The rationale based on Rayner and Pollatsek [34] is that if fixation durations depend on the information currently being encoded and then the fixation durations should increase in proportion to the

delay of the stimulus onset. Results show that although there is a subpopulation of fixations whose durations are not affected by the delay, for a second population of fixations there is a substantial link between the availability of the fixated information and the duration of the fixations. The authors conclude that fixation duration in scene viewing is partially controlled by the immediately available information from the scene.

7. Conclusions and Future Work

This paper has successfully demonstrated an improved approach of combining a reliable model of bottom up saliency with an object recognition scheme to construct a combined bottom-up and object-present (i.e., task) attentional map for an image. This offers considerable advantages over previously reported methods. Specifically we do not require an intensive training phase and can remain viewpoint invariant. Testing of the resulting combined map against observer eye-fixations shows that the combined maps offer substantial improvement against the bottom-up only case at predicting the location of human observer eye-fixations under task, if an object is detected. Finally, we demonstrate the utility of this information, by proposing and demonstrating a DCT compression technique that uses the combined attentional maps to prioritise task salient information during compression.

Even though the object recognition technique that we use in this paper is based around the recognition of a specific object, the approach that is presented by this paper would still apply in the case of general object class recognition. This is one obvious future extension to this project but impact on performance needs to be considered carefully. The application of compression algorithms to images based on their task-salient regions can be extended by investigating alternative compression schemes.

References

[1] Abdelhamid A., Wang H., and Kulathuramaiyer N., "Spiral Bit-String Representation for Image Retrieval," *The International Arab Journal of Information Technology*, vol. 7, no. 3, pp. 223-230, 2010.

[2] Bay H., Tuytelaars T., and Gool L., "Surf: Speeded up Robust Features," *Computer Vision and Image Understanding*, Berlin, vol. 110, no. 4, pp. 346-359, 2006.

[3] Brockmole J., Castelhano M., and Henderson J., "Contextual Cueing in Naturalistic Scenes: Global and Local Contexts," *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, vol. 32, no. 4, pp. 699-706, 2006.

[4] Brockmole J. and Henderson J., "Using Real World Scenes as Contextual Cues for Search," *Visual Cognition*, vol. 13, no. 1, pp. 99-108, 2006.

[5] Brockmole J. and Henderson J., "Prioritizing New Objects for Eye Fixation in Real-World Scenes: Effects of Object-Scene Consistency," *Visual Cognition*, vol. 16, no. 2/3, pp. 375-390, 2008.

[6] Bruce N. and Tsotsos J., "Saliency Based on Information Maximization," *Neural Information Processing Systems*, vol. 18, pp. 155-162, 2006.

[7] Buswell G., *How People Look at Pictures*, University of Chicago Press, Chicago, 1935.

[8] Castelhano M. and Henderson J., "Stable Individual Differences Across Images in Human Saccadic Eye Movements," *Journal of Experimental Psychology*, vol. 62, no. 1, pp. 1-14, 2008.

[9] Castelhano M. and Henderson J., "Initial Scene Representations Facilitate Eye Movement Guidance in Visual Search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 4, pp. 753-763, 2007.

[10] Castelhano M., Mack M., and Henderson J., "Viewing Task Influences Eye Movement Control During Active Scene Perception," *Journal of Vision*, vol. 9, no. 3, pp. 1-15, 2009.

[11] Committee J., "ISO/IEC 10918-1," ISO Standard, 1994.

[12] El-Bakary E., Zahran O., and El-Dolil S., "A Tool to Enhance the Performance of OFDM Systems," *International Journal of Communication Networks and Information Security*, vol. 1, no. 2, pp. 54-58, 2009.

[13] Essack S., "Modeling the Influence of Task on Attention," *Vision Research*, vol. 45, no. 2, pp. 205-231, 2005.

[14] Fergus R., "Visual Object Category Recognition," *PhD Dissertation*, University of Oxford, 2005.

[15] Fergus R., Perona P., and Zisserman A., "Weakly Supervised Scale Invariant Learning of Models for Visual Recognition," *International Journal of Computer Vision*," vol. 71, no. 3, pp. 273-303, 2007.

[16] Harel J., Koch C., and Perona P., "Graph-Based Visual Saliency," *in Proceedings of Advances in Neural Information Processing Systems*, USA, vol. 19, pp. 545-552, 2007.

[17] Hansen B. and Essock E., "A Horizontal Bias in Human Visual Processing of Orientation and its Correspondence to the Structural Components of Natural Scenes," *Journal of Vision*, vol. 4, no. 12, pp. 1044-1060, 2004.

[18] Harding P. and Robertson N., "A Comparison of Feature Detectors with Passive and Task-Based Visual Saliency," *in Proceedings of the 16th*

Scandinavian Conference on Image Analysis, Berlin, pp. 716-725, 2009.

[19] Harding P. and Roberston N., "Task-Based Visual Saliency for Intelligent Compression," *in Proceedings of IEEE International Conference on Signal and Image Processing Applications*, Kuala Lumpur, pp. 480-485, 2009.

[20] Hartley R. and Zisserman A., *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.

[21] Henderson J., "Regarding Scenes," *Current Directions in Psychological Science*, vol. 16, no. 4, pp. 219-222, 2007.

[22] Henderson J., "Eye Movement Control During Reading: Fixation Measures Reflect Foveal but not Parafoveal Processing Difficulty," *Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, vol. 47, no. 2, pp. 201-221, 1993.

[23] Henderson J. and Ferreira F., *Interface of Language, Vision, and Action, Eye Movements and the Visual World*, Psychology Press, New York, 2004.

[24] Henderson J. and Pierce G., "Eye Movements During Scene Viewing: Evidence for Mixed Control of Fixation Durations," *Psychonomic Bulletin and Review*, vol. 15, no. 3, pp. 566-573, 2008.

[25] Itti L. and Baldi P., "Bayesian Surprise Attracts Human Attention," *in Proceedings of Neural Information Processing Systems*, pp. 1-8, 2005.

[26] Itti L., Koch C., and Niebur E., "A Model of Saliency Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[27] Loftus G., "Eye Fixations and Recognition Memory for Pictures," *Cognitive Psychology*, vol. 3, no. 4, pp. 525-551, 1972.

[28] Mannan S., Ruddock K., and Wooding D., "Automatic Control of Saccadic Eye Movements Made in Visual Inspection of Briefly Presented 2-D images," *Spatial Vision*, vol. 9, no. 3, pp. 363-386, 1995.

[29] Morrison R., "Manipulation of Stimulus Onset Delay in Reading: Evidence for Parallel Programming of Saccades," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, no. 5, pp. 667-682, 1984.

[30] Navalpakkam V. and Itti L., "Search Goal Tunes Visual Features Optimally," *Neuron*, vol. 53, no. 4, pp. 605-617, 2007.

[31] Neider M. and Zelinsky G., "Scene Context Guides Eye Movements During Visual Search," *Vision Research*, vol. 46, pp. 614-621, 2006.

[32] Parkhurst D., Law K., and Niebur E., "Modelling the Role of Salience in the Allocation of Overt Visual Attention," *Vision Research*, vol. 42, no. 1, pp.107-123, 2002.

[33] Peters R. and Itti L., "Beyond Bottom-Up: Incorporating Task-Dependent Influences into a Computational Model of Spatial Attention," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, pp. 1-8, 2007.

[34] Rayner K. and Pollatsek A., "Eye Movement Control During Reading: Evidence for Direct Control," *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, vol. 33, no. 4, pp. 351-373, 1981.

[35] Rayner K., Castelhano M., and Yang J., "Eye Movements When Looking at Unusual/Weird Scenes: Are There Cultural Differences?," *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 35, no. 1, pp. 254-259, 2009.

[36] Rayner K., "Eye Movements in Reading and Information Processing: 20 Years of Research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372-422, 1998.

[37] Reichle E., Rayner K., and Pollatsek A., "The E-Z Reader Model of Eye Movement Control in Reading: Comparisons to other Models," *Behavioral and Brain Sciences*, vol. 26, no. 4, pp. 445-476, 2003.

[38] Russell B., Efros A., Sivic J., Freeman W., and Zisserman A., "Using Multiple Segmentations to Discover Objects and their Extent in Image Collections," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1605-1614, 2006.

[39] Rutishauser U., Walther D., Koch C., and Perona P., "Is bottom-up attention useful for object recognition?," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 11-37, 2004.

[40] Sundaram H. and Chang S., "Condensing Computable Scenes Using Visual Complexity and Film Syntax Analysis," *in Proceedings of the IEEE International Conference on Multimedia and Expo*, Japan, pp. 273-276, 2001.

[41] Torralba A., Oliva A., Castelhano M., and Henderson J., "Contextual Guidance of Eye Movements and Attention in Real-World Scenes: the Role of Global Features in Object Search," *Psychological Review*, vol. 113, no. 4, pp.766-786, 2006.

[42] Unema P., Pannasch S., Joos M., and Velichkovsky B., "Time Course of Information Processing During Scene Perception: The Relationship Between Saccade Amplitude and Fixation Duration," *Visual Cognition*, vol. 12, no. 3, pp. 473-494, 2005.

[43] Wallace G., "The Jpeg Still Picture Compression Standard," *Communication of ACM*, vol. 34, no. 4, pp. 30-44, 1991.
[44] Yarbus I., *Eye Movements and Vision*, New York, Plenum Press, 1967.

Amjad Rehman is a PhD in image processing and pattern recognition from Faculty of Computer Science and Information System Universiti Teknologi Malaysia. During his PhD, he proposed novel techniques for document analysis and recognition. He is author of dozens of papers published in international journals and conferences of high repute. Additionally, he is a reviewer and editor of five international journals. His research activities are in the area of image processing, visualization and graphics in particular machine learning, documents analysis and recognition. Currently, three PhD students are conducting research under his supervision.

Tanzila Saba is a PhD student in Graphics and Multimedia Department Faculty of Computer Science and Information System Universiti Teknologi Malaysia. She is at final stage of her PhD and the author of more than twenty papers of international journal and conferences. Her research activities are in the area of image processing and pattern recognition, in particular information security, machine learning and e-learning. She has excellent research record and awarded with international research awards of Keizo Obuchi and TWAS. She is expecting to complete her PhD in 2011.