Task-Based Visual Saliency for Intelligent Compression

Patrick Harding
VISP group, School of Engineering and Physical Sciences
Heriot Watt University, Edinburgh
Email: pjh3@hw.ac.uk

Neil Roberston
VISP group, School of Engineering and Physical Sciences
Heriot Watt University, Edinburgh
Email: nmr3@hw.ac.uk

***Abstract*—In this paper we develop a new method for highlighting visually salient regions of an image based upon a known visual search task. The proposed method uses 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 recognised are parameterised quickly in advance by a viewpoint-invariant spatial distribution of SURF interest-points. The bottom-up and top-down object probability images are fused to produce a task-dependent saliency map. We validate our method using observer eye-tracker data collected under object search-and-count tasking. Our method shows 10% 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 DCT encoding. We demonstrate our technique on surveillance-style footage throughout.**

I. Introduction

Existing models of bottom-up saliency are reliable indicators of passive visual attention regions in an image [1], [2]. However, under task based viewing there is often a strong shift of attention away from the passive observation case [3]. This arises from the imposition of top-down processes by the observer under task in combination with the bottom up response [4]. Models have been constructed for the top-down case, but involve complicated prior learning of general object classes and their scene contextualisation [4], [5], [6], [7]. 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 generalise.

What we propose here is a task-oriented correction to bottom up models of visual attention in the case that a prior learned object is present. 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. The visual system under task-prioritised viewing is guided by prior experiences of associating task objects to likely scene location contexts [8], [9], [10]. 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 [11]. An illlustration of this effect is presented in Fig. 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 presence" 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. Fig. 1 (bottom) shows the thresholded Graph Based Visual Saliency [2] map generated from the upper image, with eye-fixations still imposed. 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 thresholded area.

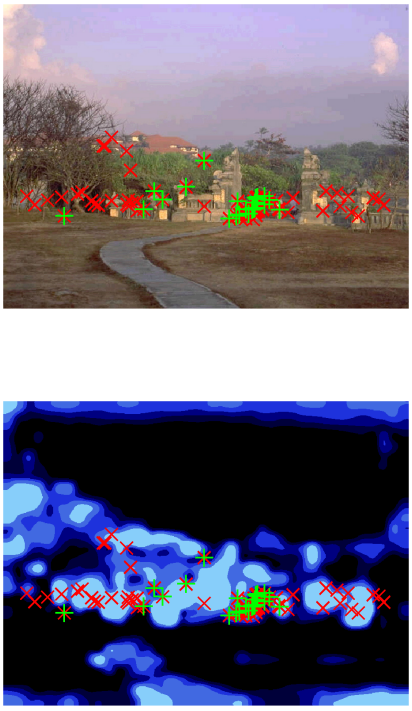

Fig. 1. (Top) An image observed under task with eye-tracker points superimposed. x denotes all eye fixations of eight observers performing people count on image. + denotes the first three eye fixations across all participants. (Bottom) Graph Based Visual Saliency (GBVS) model of passive visual attention computed from the image. The attentional map is thresholded to 10 - 50% by area.

We use SURF (*Speeded Up Robust Features* [12]) interest-

point matching to a reference image to determine object presence 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.

II. Object detection and confirmation from SURF points

A. SURF matching and and Object confirmation refinement

SURF (*Speeded Up Robust Features* [12]) 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 [12] for thorough performance measures) as well as being natually-distributed towards visually salient information under different viewing conditions [13].

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 of SURF detection and matching, see Fig. 2. 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 correspondances. 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

Fig. 2. Surf matching example. Circles denote detected SURF points, "+" denotes matched points between reference and test images. Image as shown in Fig. 1. Top: manually extracted reference image with SURF points matched to test image. Lower left: test image showing SURF points matched to reference image. Bottom right, our object contextual correction map (unthresholded).

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 correspondances.

There already exist good object classification and recognition techniques [14], [15], [16] 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 occurring 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 $(I_{MP'})$ where I is the image label, R or T for reference or Test, M denotes that the points are matched and P' is the plane. Therefore $(T_{MT'})$ denotes the Matched Test Points in the Test Image plane. We first calculate the homography matrix between Test and Reference Image planes using matched points in each image, $(T_{MT'})$ and $(R_{MR'})$. We then transform the matched points in the test image to the Reference plane, $(T_{MR'})$. This is done by using the homography relationship standard in computer vision. We use the algorithm detailed in [17]. The spatial distribution of the $(T_{MR'})$ is then computed. This is the angular and radial displacement to all points from a zero point.

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

The distribution of matched Test points, (T_{MR}), with each reference point is overlapped and the displacement δ 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 present or absent, based on the number of classified object points within the threshold for the best fit matches. We chose to set six confirmed object points as our threshold for object confirmation. In the case where there are no inliers, it is necessary to choose a different test point for constructing the distribution and to reapeat 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 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.

B. 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 judgement 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 1/6th of the image height above and the other 1/6th 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 1/6th height either side of the centre, the size of the object is chosen instead of the 1/6th image height.) 4. Outside the core context, the map is tailed-off in the vertical direction according to the formula:

$$distance(x, y) = \left(\frac{(y - c_{ny})}{2} \right) \tag{2}$$

where (x,y) is the current point in the map of the same dimensions as the image (excluding the core context) and c_{ny} 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 Fig. 2 for an object presence surface example.

C. 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 natualistic 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 equation (3) [4].

$$C(x, y) = (BU(x, y)) * (O(x, y))^{\gamma} \tag{3}$$

where C is the combined map, BU is the GBVS saliency map [2], 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 $\gamma = 0.05$

in this paper. This has the effect of flattenning 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 choose therefore to set the pixel values of the bottom up map to between [0.32 0.95] and the values of the Object map to between [0.93 0.95] while retaining a large floating point value for each pixel, allowing for smoothness. These values were 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 Fig. 3 against eye-tracker data. Note the overlap improvement compared to GBVS alone illustrated in Fig. 1.

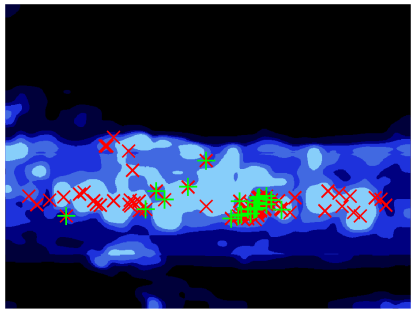

Fig. 3. Thresholded Combined Bottom up GBVS map and Object context correction. Eye fixation data overlaid from all participants. All fixations denoted by "x", first three only by "+". Note the shift in energy towards the known target area compared to the pure bottom up case in Fig. 1.

III. Validation of the Combined Surface Under Task

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

The eye-tracker data and image set from Torralba and Henderson has been used to validate the model (to view see URL in[4]). The test image data set for this paper comprises 72 images and 108 search scenarios (3x36 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 were manually extracted as reference images from the test images for all 108 search scenarios e.g. if task was to count the number of paintings, a painting would be extracted from the image, if present. Overall there were 61 object present cases and 61 appropriate objects for task were extracted out of the 108 tasks. These objects were 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) was 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 were such that there was practically no mismatching between the reference and test images. The largest number of mismatches per image (i.e. matches from the wrong image) was 1 and this was a statistically very rare event. 30 out of 61 objects were recognised using our object confirmation (This is > 6 matched surf points lying within overlap tolerance) and 47 examples had reliable SURF matching to at least three points in the test image. The reason for these low values can be attributed parially to our conservatively high matching threshold and partially to the object size in the image - often the area of the object pixels was very low(1% of image area of 800 x 600) and this did not allow for robust descriptor representation in quantity.

For each search scenario, three saliency maps were created: (i) Bottom Up saliency; (ii) SURF combined map only from matched points; (iii) Object combined map from Object points. The construction mechanism for (ii) and (iii) was identical, but there was a subtle change of the object centre line since not all matched points were classified as object points. Essentially the statistics from (ii) and (iii) were identical within reasonable error, so below only cases (i) and (ii) are presented (iii, is actually a refined subset of ii allowing for higher confidence).

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

The attentional maps of each class were thresholded to different image areas representing the more salient half of the image. X = 10, 20, 30, 40 and 50% of image area were chosen since these levels clearly represent the "more salient" half of the image to different degrees, as illustrated in Figs. 1 and 4.

For each search scenario, the eye tracker points lying within and without each threshold level of each mask were counted. We chose to use all eight participants and to process all of the eye points. This gives the exhaustive search case. The overlap was *considerably higher* if only the first three fixations were 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 Fig. 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 at least 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.

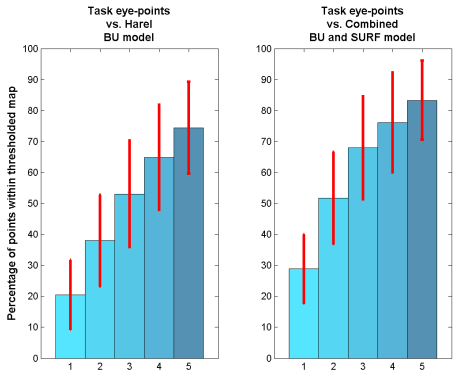

Fig. 4. The figure shows the overlap with the attentional maps at different threshold levels for *all* eye-points, gathered over 8 experimental participants under task. *left:* eye-data vs bottom-up only (72 images, 8 participants, 108 tasks). *right:* eye-data vs. combined bottom-up and SURF-point object 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 Fig. 1 and in Fig. 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.

A. 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 $20°$ of offset from the reference, while the SURF matching alone generally started to collapse beyond $20°$.

IV. DCT COMPRESSION USING COMBINED SALIENCY MAPS

We have demonstrated a technique that can succesfully modulate attention maps from the bottom up model alone to adjust for task based viewing that relies on a 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 quantisation 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 quantisation is performed using a quantisation matrix derived from psychovisual tests and this matrix can be weighted to provide the required degree of compression in the block. The reverse process decodes the image [18], [19]. The heavier this quantisation, the larger the compression ratio achieved, however this is tempered by the fact that over-quantisation will produce blocking atifacts that significantly reduce image quantity and can damage real information within the image. In regular JPEG, the quantisation 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. We therefore threshold out half of the image for high and half for low compression. 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 derived from psychovisual experiments to give acceptable compression. The quantisation matrix we use is that specified in Annex K of the JPEG standard for the luminance component of images [18], appropriate for greyscale. We choose a low value of Q=3 for the outlying regions and weight the quantisation matrix according to the following relationship: $(50/Q) * Qmatrix$. We choose this exagerated 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 *greyscaled* 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*(quantised, linearised DCT image string) : *length*(huffman-encoded string)) for the Binary-Task-DCT-Huffman was 6 ± 0.2, while for the Global-DCT-Huffman the average was 5.5 ± 0.5. The final compressed output had a storage value of 0.1737 *bits per pixel (bpp)* for the binary, object-context compression and 0.1927 *bpp* for the global compression. The gain in compression outweighs the required storage for the reference descriptors and is valuable for large datasets. An effective illustration of object-context oriented compression is presented in Fig. 5.

Fig. 5. Task-oriented compression based on 9 matched and confirmed object points between test and reference. *Top left* our reference image containing an object of interest. *Top right* our test image, at a different viewing angle to reference. *Centre left* a zoom-in of the detected object in the full intelligent compression (seen in *Bottom*). *Centre right* 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. *Bottom* 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.)

V. Conclusion and Future Work

We have successfully demonstrated a new method of combining a reliable model of bottom up saliency with an object recognition scheme to contruct 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 this paper presents 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] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," pp. 1254–1259, Nov 1998.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 545–552.

[3] M. S. Castelhano, M. L. Mack, and J. M. Henderson, "Viewing task influences eye movement control during active scene perception," *Journal of Vision*, vol. 9, no. 3, pp. 1–15, 3 2009.

[4] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "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, October 2006. [Online]. Available: http://people.csail.mit.edu/torralba/GlobalFeaturesAndAttention/

[5] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally." *Neuron*, vol. 53, no. 4, pp. 605–617, February 2007.

[6] ——, "Modeling the influence of task on attention." *Vision Res*, vol. 45, no. 2, pp. 205–231, January 2005.

[7] R. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2007.

[8] J. R. Brockmole, M. S. Castelhano, and J. M. Henderson, "Contextual cueing in naturalistic scenes: Global and local contexts." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 32, no. 4, pp. 699–706, July 2006.

[9] J. R. Brockmole and J. M. Henderson, "Using real-world scenes as contextual cues for search." *Visual Cognition*, vol. 13, no. 1, pp. 99–108, 2006.

[10] ——, "Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements." *Quarterly Journal of Experimental Psychology*, vol. 59, no. 7, pp. 1177–1187, July 2006.

[11] B. C. Hansen and E. A. Essock, "A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes," *J. Vis.*, vol. 4, no. 12, pp. 1044–1060, 12 2004.

[12] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 4, pp. 346–359, 2006.

[13] P. Harding and N. M. Robertson, "A comparison of feature detectors with passive and task-based visual saliency," *LNCS*, vol. 5575, pp. 716–725, 2009.

[14] R. Fergus, "Visual object category recognition," Ph.D. dissertation, University of Oxford, 2005.

[15] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[16] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition," *International Journal of Computer Vision*, vol. 71, no. 3, pp. 273–303, March 2007.

[17] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004, p. 91.

[18] J. Committee, "Iso/iec 10918-1," ISO Standard, 1994.

[19] G. Wallace, "The jpeg still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, 1991.