What were we all looking at? Identifying objects of collective visual attention

We aim to identify the salient objects in an image by applying a model of visual attention. We automate the process by predicting those objects in an image that are most likely to be the focus of someone's visual attention. Concretely, we first generate fixation maps from the eye tracking data, which express the ground truth of people's visual attention for each training image. Then, we extract the high level features based on the bag-of-visual-words image representation as input attributes along with the fixation maps to train a support vector regression (SVR) model. With this model, we can predict a new query image's saliency. Our experiments show that the model is capable of providing a good estimate for human visual attention in test images sets with one salient object and multiple salient objects. In this way, we seek to reduce the redundant information within the scene, and thus provide a more accurate depiction of the scene.

Keywords: visual attention; bag of visual words; eye tracking; support vector regression

Subject classification codes: include these here if the journal requires them

1 INTRODUCTION

The massive amount of text and image data generated and spread by social media every day forms a publicly available global source of data representing the collective interest of all those who regularly use Facebook, Twitter and other sites. The data offers a stream of what people experience and what they are interested in, what they are thinking and doing in real time. The possibilities for data mining in this stream are many, varied and enormous. For example, it is possible to extract stories and items that are significantly more interesting than the rest automatically on the basis of assigned popularity ratings. Wu & Huberman (2007) provide a statistical model of collective attention based on novelty which describes the lifetime of stories on Digg.com. It is a simple proposition to tag and archive stories that are statistical outliers and which attract

exceptional attention. Hashtags in tweets on Twitter can be monitored automatically and analysed (Lehmann, Gonçalves, Ramasco, & Cattuto, 2012). In this work, the authors describe different profiles describing the lifetimes of hashtags. Some may increase in frequency, rise to a peak, and then decay, while others are triggered by a specific event causing a peak followed by a decay in interest. They apply lexical analyses to the hashtags in the clusters corresponding to the profiles and identify common themes. Subjects of collective interest and attention can be continuously monitored and recorded.

What then of applying the same ideas of automatic object extraction to images and video feeds? Together with GPS data, these provide very rich opportunities for describing topics of immediate collective interest and attention of those in particular locations. One data mining service offers to tag corporate branding appearing in images, so a company can monitor where its brand appears at anywhere and any time ("gazeMetrix"). Google has taken out a US patent (Zhao & Yagnik, 2012) for a means of automatic object identification in video streams. If successful this will enable automatic tagging, not just of specific corporate branding, but of any object the systems have been trained to recognise.

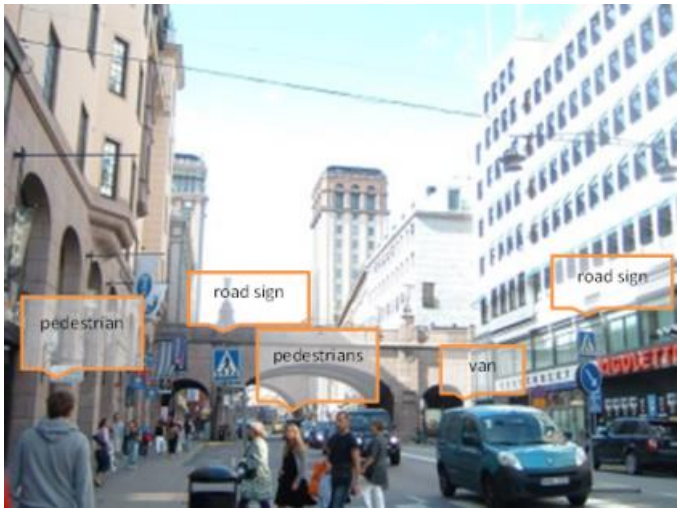


Figure 1. Google's patent describes a system that can automatically tag photos via object recognition.

This has the potential to mark up in close to real time, the content of all images and video feeds available to the system. The problem is now not being able to see the woods for the trees. Unless there is some way of identifying what the salient objects are in each image or video frame, the interesting objects of collective visual attention will be lost in a stream of redundant information about all of the other objects that happen to appear in the same shot.

Our work is aimed at identifying salient objects in an image by applying a model of visual attention, which is designed for subsequent object recognition in images. We aim to provide the means of making automatic object recognition sensitive to visual attention. We are able to predict those objects in an image that are most likely to be the focus of someone's visual attention on the basis of an analysis model that has been trained using eye tracking data obtained from many people viewing a large set of training images. The model is further enhanced by including automatic detection of high level salient features. The potential for this filter is very significant for being able to identify common objects likely to attract visual attention in images and video streams taken by different people in the same location and the same time.

In this paper, we describe the model and then show how well the predictions of saliency in a series of test images match with eye tracking data collected from participants viewing the same images.

2 RELATED WORK

The computational model of visual attention was first introduced by Itti, Koch, & Niebur (1998). Specifically, they proposed using a set of feature maps from three complementary channels, which were intensity, colour, and orientation. The normalized feature maps from each channel were then linearly combined to generate the overall saliency map. Based on this, many other researchers have since suggested

improvements. Meur, Callet, Barba, & Thoreau (2006) adapted the Itti model to include the features of contrast sensitivity functions, perceptual decomposition, visual masking, and centre-surround interactions. Privitera & Stark (2000) have improved the Itti model by adding symmetry as an additional feature. The above approaches are all based on low-level image features, such as intensity, colour, and orientation. We categorize these kinds of visual attention models as "bottom-up" models.

Studies from psychophysics and neurobiology indicate that, as well as bottom-up factors, top-down factors play an important role in attracting a person's attention (Frintrop, Rome, & Christensen, 2010). In order to obtain a better simulation of attention getting, the bottom-up and top-down approaches to saliency are fused to obtain a single focus of attention. Many factors may influence the visual attention. One of these is how attention is driven by current tasks. Wolfe, Cave, & Franzel (1989) proposed a model that takes this into consideration by modulating the weights of the feature maps depending on the task at hand. For example, if searching for a vertical green bottle, the model would increase the weights of the green and vertical orientation feature maps to allow those features to be attributed more saliency. Another important aspect of top-down factors is the people's knowledge of the outer world. That can be divided into two subcategories. One is the relationship between object and context. Oliva, Torralba, Castelhano, & Henderson (2003) proposed a method that regards context information, which means searching for a person in a street scene is restricted to the street region; the sky region is ignored. The contextual information is obtained from past search experiences in similar environments. Another subcategory is the prior knowledge about the target. In most cases, some particular types of objects are more likely to attract people's attention. Therefore, another way to add top-down components to a model is to use object detectors. Cerf, Frady, & Koch (2009) indicated that faces

and text strongly attract attention. They added a conspicuity map indicating the location of faces and text to the Itti model, and showed that it improves the ability to predict eye fixations in natural images. While adding object detectors improved the model, it's limited by the number of the object detectors, and hard to generalize to the generic category.

Gaze fixations provide the best indication in real-time of the focus of visual attention of a person, so one way to improve the predictive ability of the visual attention model is to exploit actual eye tracking data. Zhang et al. (2011) proposed a time delay neural network model that trained on the real eye tracking data, but their work aimed to simulate the time sequence of the eye gazes. Liang, Fu, Chi, & Feng (2010) used real eye tracking data as ground truth to refine a region based attention model with a Genetic Algorithm (GA). Their results showed that the refined model outperformed the original one. However, their refined model still only used low-level image features as the basis for optimization using the eye tracking data. This limited the extent to which improvement in performance was possible. Kienzle, Wichmann, Schölkopf, & Franz (2006) produced a visual saliency model directly from human eye movement data using a support vector machine (SVM). However, their training set only contained grey scale images. Judd, Ehinger, Durand, & Torralba (2009) trained a binary SVM classifier on a large colour database using both low-level and high-level image features. They treat the notion of visual attention as a binary classification problem, although they use some mathematical trick to enable the model to output continuous saliency value. However, since their models are binary classifiers, they won't be able to directly compare the saliency of different images or even compare salient regions within the same image.

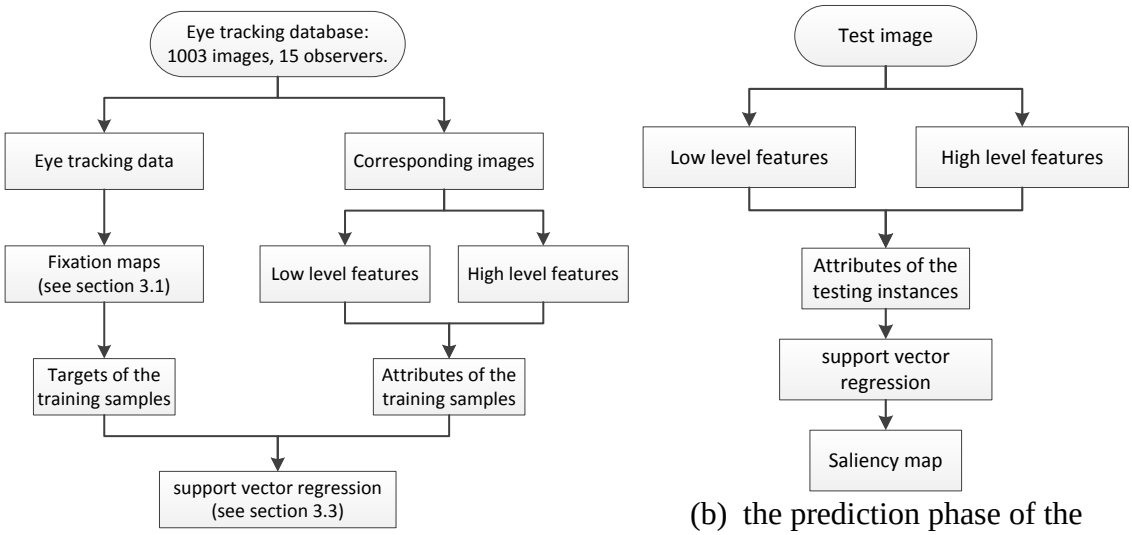
3 Our Model of Visual Attention

We try to simulate the collective visual attention of people. Intuitively, if most people

looked at the same place in an image, and the duration of attention is relatively long, then that area should be relatively salient, i.e. there may be something interesting in that area. On the contrary, if there isn't any interesting thing in an image, people might look around on the image, i.e. the fixation pattern should be scattered.

Based on that assumption, we propose a new visual attention model that learns from eye tracking data. The main differences with previous learning-based approach are: ~~one~~first, we generate a continuous fixation map as the ground truth of visual attention, and treat the learning problem as a regression problem. Thus, we hope to train a model that not only can determine salient areas in an image, but also can tell how salient this area is. Second, since the bag-of-visual-word image representation has successfully used in category recognition (Csurka, Dance, Fan, Willamowski, & Bray, 2004), we introduced a new type of high level features − the probabilities of each visual word occurs in the salient areas − into the visual attention model.~~Two, we introduced a new type of high level features that using the probabilities of each visual word occurs in the salient areas.~~ With this type of high level features, we can overcome the limitation of the object detectors, help us to find out any kind of objects that people are interested in.

The whole procedure is as follow: First, we generate fixation maps from the eye tracking data of multiple users which express the ground truth output for each training image. Then, we extract the image features, including low-level and high-level features, and use these features as input attributes along with the gaze fixation map outputs from the eye tracking data to train a support vector regression (SVR) model. A diagram of the training phase for our model is presented in Figure 2 (a). After we train our SVR model, we can extract image features from a new image (not belonging to the training set) and use them to predict its saliency. A diagram of the prediction phase of the model is given



(a) the training phase of the attention model

(b) the prediction phase of the attention model

Figure 2. The diagram of the training phase of the computational attention model.

in Figure 2 (b). The training data, including images and corresponding eye tracking data, we use is from Judd et al. (2009).

3.1 Generating Fixation Maps

We need to convert the eye tracking data to a suitable fixation map, which is able to accurately reflect the gaze ground truth to train our SVR model. In order to extract the interesting objects and even events from a series of images, we need a fixation map with continuous saliency value that can indicate the different saliency value within different parts of an image, or across different images.

Intuitively, the area that the more people look at or the longer people look at should have higher saliency value. Thus, to produce the fixation maps, firstly, we generate a fixation sum map $\mathbf{F}^{(sum)} \in \mathbb{R}^{m \times n}$ from the eye tracking data, where m and n represent the size of the corresponding image. The value of $\mathbf{F}^{(sum)}$ is:

$$\mathbf{F}_{i,j}^{(sum)} = N_{i,j}, i \in [1, m], j \in [1, n]. \quad (1)$$

Where $N_{i,j}$ is the total number of the gaze fixations at the point (i, j) across all observers.

Then, the fixation sum map, which is a grid with discrete values, is smoothed by convolution with a Gaussian kernel to obtain a continuous fixation map $\mathbf{F}$:

$$\mathbf{F} = \mathbf{G} * \mathbf{F}^{(sum)}. \quad (2)$$

Here, $\mathbf{G}$ is the Gaussian kernel with a cut-off frequency of 8 cycles per image, about 1 degree of visual angle (Einhäuser, Spain, & Perona, 2008), to match the approximate area that an observer sees at high focus around the point of fixation.

3.2 Features Used for Support Vector Regression

We use image features that are associated with bottom up attractors of visual attention (low-level features) and top-down attractors (high-level features).

3.2.1 Low-level features

There are a variety of image features that are physiologically plausible and have been shown to correlate with visual attention. The features we use are:

- The local energy of the steerable pyramid filters (Simoncelli & Freeman, 1995);
- Intensity, orientation and colour centre-surround operators (Itti & Koch, 2000);
- The values of the red, green and blue channels as well as the probabilities of each of these channels. The probability of each colour is computed from 3D colour histograms of the image filtered with a median filter at 6 different scales (Judd et al., 2009);
- The saliency map generated by the model described by (Aude Oliva & Torralba, 2001).

3.2.2 High-level features

Some particular types of objects (such as face, person, text, etc.) are more likely to attract people's attention. To get a better prediction of the human's visual attention, we need to know the category on the images. The bag-of-word image representation (BoW) is successfully used in category recognition (Csurka et al., 2004)~~(Csurka, Dance, Fan, Willamowski, & Bray, 2004)~~. Based on this image representation, many methods were proposed to recognize generic category in the images (Tuytelaars, Lampert, Blaschko, & Buntine, 2010). We proposed a new type of high level feature, which based on the bag-of-visual-word image representation, to link the category information on the image with their saliency.

In the bag-of-visual-word image representation, the visual word in one image may be the components of different categories, such as face, person, etc. In our case, we try to link this category information with the saliency. We consider there are two classes in the images - the salient and the non-salient. Then we want to know the probability of each visual word occurs in each class. So, we can calculate the conditional probabilities of visual word w_i given class z_k directly from the training data. In order to avoid probabilities of zero, these estimates are computed with Laplace smoothing:

$$P(w_i|z_k) = \frac{1+\sum_{\{a_j \in z_k\}} N(w_i, a_j)}{|W|+\sum_{s=1}^{|W|}\sum_{\{a_j \in z_k\}} N(w_s, a_j)} \quad (3)$$

Where $|W|$ is the size of the vocabulary, videlicet, the number of visual words. $N(\mathrm{w}_i, a_j)$ is the number of times visual word w_i occurs in the image area a_j. This area could either be salient or non-salient.

Given one image, the $P(w_i|z_k)$ is computed at each visual word to get a discrete visual word map. We convolve it with a Gaussian kernel to get a continuous probability

map, which indicates the probabilities of the visual words in the corresponding image are salient. These probability maps are used as the high-level feature.

3.3 Training the SVR model

The database we used contains 1003 images. To speed up the training and prediction processes, once we have extracted the features from the image, we reduce the feature matrices to 160×160. However, there are still 160×160×1003=25,676,800 samples. If we use a kernel for our SVR training, the trained model will have a large number of support vectors, making the prediction quite slow. Therefore, to further decrease computational complexity we train the SVR model without a kernel (Ho & Lin, 2012).

Mathematically, the model can be written like this:

$$s = b + \mathbf{w}^T \mathbf{f} \tag{4}$$

where s is the salient value at an area in the image, b is the bias of the model and $\mathbf{w}$ is the weight vector for each feature. Both b and $\mathbf{w}$ are learned from training data. The $\mathbf{f}$ is the features vector.

To fix the misclassification cost constant C and ϵ in the loss function of ϵ-SVR, we conducted an exhaustive grid search with the grid points equally spaced on a log scale. According to the results of a grid search of parameters, the constant C is set to 1, and ϵ is set to 0.0039.

4 Experiment to Evaluate the Ability of the Visual Attention Model to Predict Actual Fixations

We conducted an experiment to see how well predictions about which objects in a scene were predicted to be most salient aligned with actual eye tracking data. Two sets of test

images were used with two different groups of participants, one for each study. In one study, images consisted of scenes with one obvious object (Scene Type 1) and in the other study, images contain two or more obvious objects (Scene Type 2). We selected several scenes from around the De Montfort University campus and took photographs: Scene Type 1 contained 16 images and Scene Type 2 contained 21 images. Please note that the images we used for testing are not included in the original image dataset used for training our model.

In both scene types, the set of images was presented to participants for a total of 3 seconds each. No instructions were given about what the participant should look at. In such a way, we expect we can get the general idea of what the people think are interesting in these images.

All trials were carried out with the participant seated at a desk in front of a 20’ widescreen monitor. Eye movements were recorded with a Tobii X120 eye tracker located beneath the monitor. After the initial calibration sequence, all of the images were presented in sequence without any pause between them. After the trial, the purpose of the study was explained to the participant and their consent was obtained to use their data in the further analysis. An individual trial lasted approximately 5–7 minutes.

- Low-level features with eye tracking (LFET): here the SVR model trained with low level image features and eye tracking data was used to produce a saliency map based on low level features in the test image
- Low-level as well as high-level features with eye tracking together (LHFET): here the output from the SVR model that was learned with low level as well as high-level feature.

- Low-level features only (LF): as a control condition, a saliency map was generated using low level image features only. The map was generated by open source gaze data analysis software (Adrian, 2013).

We wished to establish whether our models (LFET and LHFET) produced better results in predicting the salient areas of the test images than a saliency map produced without exploiting eye tracking data and high level features (LF).

4.1 Measuring how well the saliency map can predict actual gaze positions

A saliency map is essentially a two-dimensional grid of probabilities. The elements in the grid are pixels. The probabilities refer to the likelihood that a gaze event will occur on just that pixel. For many pixels, the probabilities will be close to 0. Areas of the image where the probabilities are higher can be colour coded, resulting in a heat-map visualisation of the probabilities of a hit by a gaze event. The gaze events we have used are fixations rather than individual gaze points. The eye tracker delivers 60 measurements of gaze position each second, or one every 16 milliseconds (approximately). A fixation is a cluster of gaze points where the duration equated to the number of points in the cluster, and is represented by the average x and y position of the points within the cluster. The Actual Eye Tracking data is also represented as a map of locations in relation to the two-dimensional grid. People make very roughly about 3 to 4 fixations per second, giving 9 to 12 fixations over 3 seconds for each participant. The fixation data from all participants viewing a particular image is collected onto one map. There is no temporal sequence of fixations and consequently no scan path.

A measure of how close the saliency map is the map of actual fixations is needed for meaningful quantitative comparisons to be made. There are several candidate metrics, such as: Receiver Operating Characteristics (ROC), correlation-based

measures (Rajashekar, Linde, Bovik, & Cormack, 2008), Earth Mover's Distance (EMD), etc. Among these metrics, ROC is the most widely used in the community. Therefore, in this paper, we use the area under the ROC curve (AUC) to quantitatively evaluate the visual attention model. For a saliency map, one can convert it to only salient and non-salient areas depend on a threshold. For any particular value of the threshold, there is some fraction of the fixation points which are located in the salient areas (true positive rate), and some fraction of points which were not fixation points but labelled as salient anyway (false positive rate). Varying over all such thresholds yields a ROC curve and the area beneath it is generally regarded as an indication of the classifying power of the detector. The AUC score range between 0 and 1, the larger value indicates better prediction of the model.

4.2 Scene Type 1: Test images with one highly salient object only

The image set contained 16 images taken around the university campus, indoors and outdoors. All included just one prominent object that could be expected to attract the viewer's attention. In some cases the object was a person, in some an automobile or vehicle, and in others an object such a sign on a door. Participants recruited from staff and students in the faculty took part. We expected there to be no difference between the LFET maps compared with the LHFET map if there were no high level features (faces, figures or cars) present in the image. It is of course possible that features in the image might be mistakenly classified as high level features resulting in these areas being given higher saliency. We expected that both the LFET and LHFET maps would give significantly better predictions of actual gaze positions than the LF only map. Three sample images from this set and the maps associated with these are shown in Figure 3.

The ROC curves of these images are shown in Figure 4.

We conducted a t-test to test if there are significant differences between LHFET, LFET and LF. The p-values of the t-test are shown in Table 1. To see how big the differences are between these models, the average differences between them are also shown in the table.

From the Table 1, we can see that for Scene Type 1, the results of LHFET as well as LFET are both better than LF, the average improvement is about 6% for both of them, but the results of LHFET is not better than LFET.

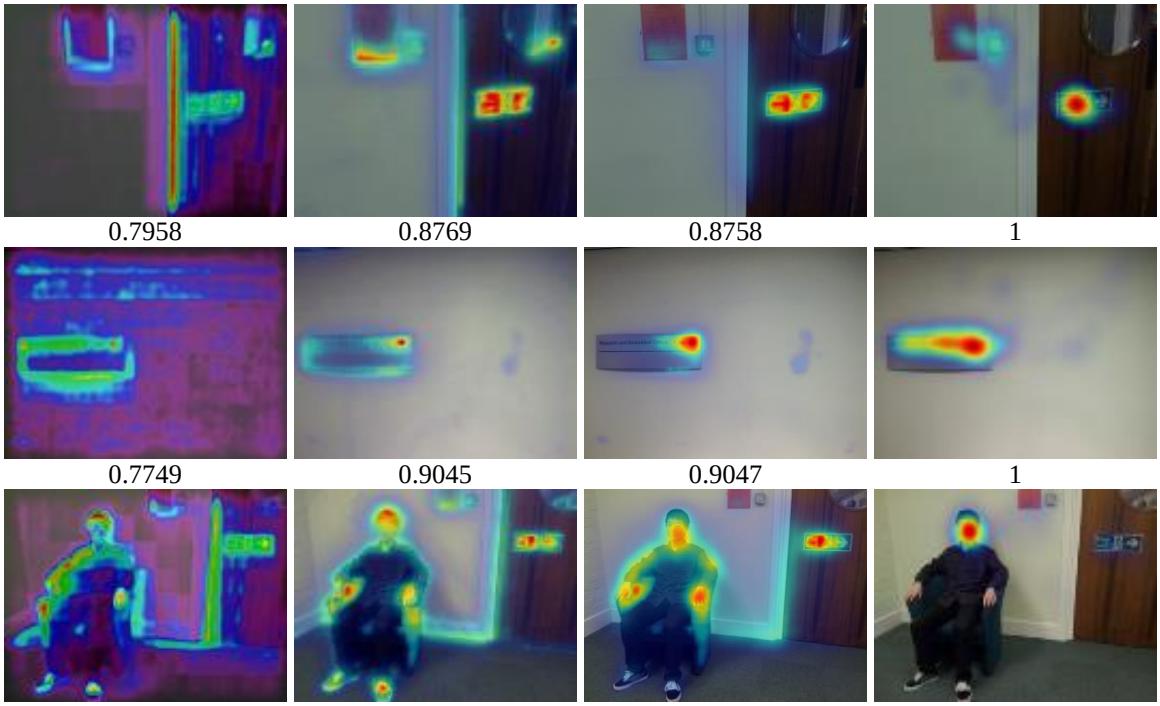
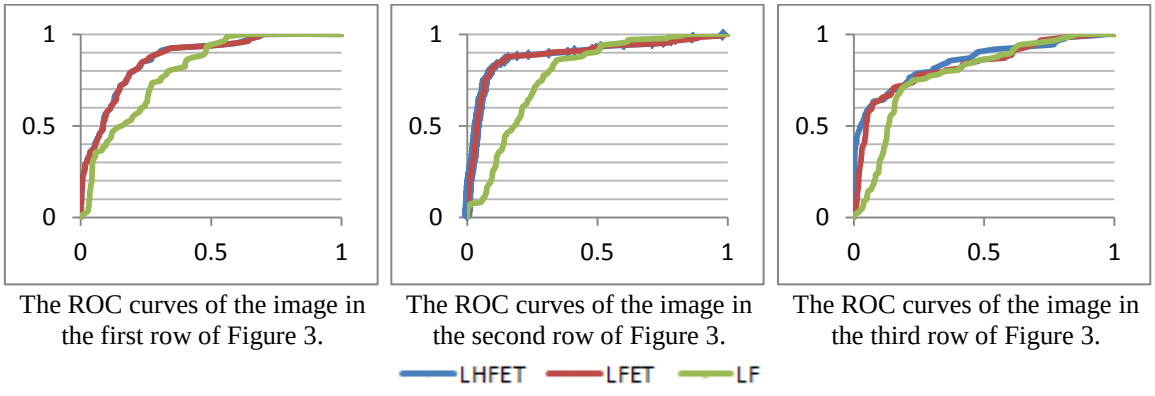


Figure 3. The sample of images and corresponding eye tracking data as well as results of visual attention models. The left three columns are the results of three visual attention models, the fourth column shows the actual eye tracking. Each row shows a sample of images, the number below each images is the AUC score of the model for the image.



The ROC curves of the image in the first row of Figure 3.

The ROC curves of the image in the second row of Figure 3.

The ROC curves of the image in the third row of Figure 3.

Figure 4. The ROC curves of the three example images shown in Figure 3. Different colours represent different models. The horizontal axis is the false positive rate, the true positive rate is on the vertical axis. Note that the curves of LHFET and LFET in the left two charts are too similar, so they look like there is only one curve left.

Table 1. The p-values of the t-test of whether there are significant differences between LHFET, LFET and LF for scene type 1.

Alternative hypothesis:	LHFET is larger than LF	LFET is larger than LF	LHFET is larger than LFET
The p-values of the null hypothesis:	<0.0001	<0.0001	0.3917
The average difference:	5.91%	5.81%	0.10%

4.3 Scene Type 2: Test images with many possible salient areas or objects

The study was similar to Study 1, although different participants were used. The images were of 'busy' scenes taken from the centre of the campus at lunchtime. We took 21 images in total. All images contained several high level features, either faces, figures or vehicles or a combination of these. We expected a clear difference between the LFET and LHFET maps. We also expected a number of different salient objects with different probabilities of attracting fixations. Three sample images from this set and the maps associated with these are shown in Figure 5.

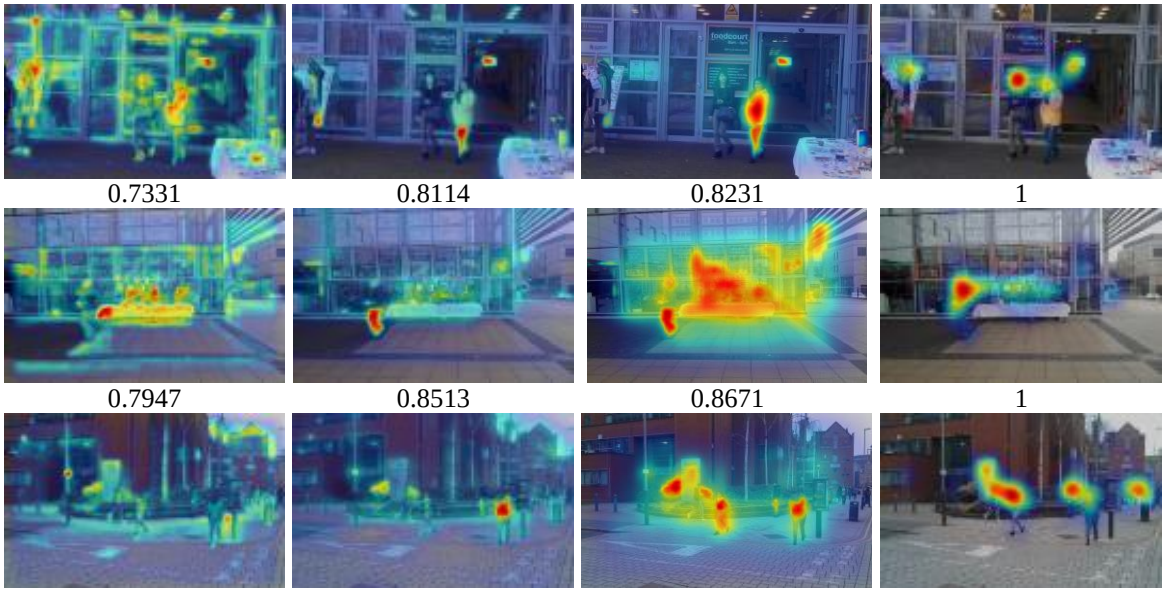
The p-values of the t-test of if there is a significant difference between LHFET, LFET and LF, as well as the average differences are shown in Table 2.

From the results of the t-test, we can see that, for Scene Type 2, the results of LHFET as well as LFET are both better than LF, and the results of LHFET is also better than LFET. The average improvement of the LHFET is about 4% in this scene type.

4.4 Evaluation based on the images' content

In addition, we divided the whole image set into two categories based on if there are high level features in the image. So we get two image sets: image set 1—image with high level features; image set 2—image without high level features.

The values of the areas under the ROC curves of the two image sets are showing in Figure 6.



0.7331	0.8114	0.8231	1
0.7947	0.8513	0.8671	1
0.7411	0.7802	0.8681	1
Low level features (LF)	low level features with eye tracking (LFET)	low level features with eye tracking together with high level features (LHFET)	Actual eye tracking (AET)

Figure 5. The sample of images and corresponding eye tracking data as well as results of visual attention models. The left three columns are the results of three visual attention models, the fourth column shows the actual eye tracking. Each row shows a sample of images, the number below each images is the AUC score of the model for the image.

Table 2. The p-values of the t-test of whether there are significant differences between LHFET, LFET and LF for scene type 2.

Alternative hypothesis:	LHFET is larger than LF	LFET is larger than LF	LHFET is larger than LFET
The p-values of the null hypothesis:	<0.0001	<0.0001	0.0031
The average difference:	4.28%	2.69%	1.59%

The p-values of the t-test of if there are significant differences between LHFET, LFET and LF, as well as the average differences are shown in Table 3.

From the results of the t-test, we can see that for both image sets, the results of LHFET and LFET are both better than LF, and the results of LHFET is also better than LFET for image set 1, but not better than LFET for image set 2.

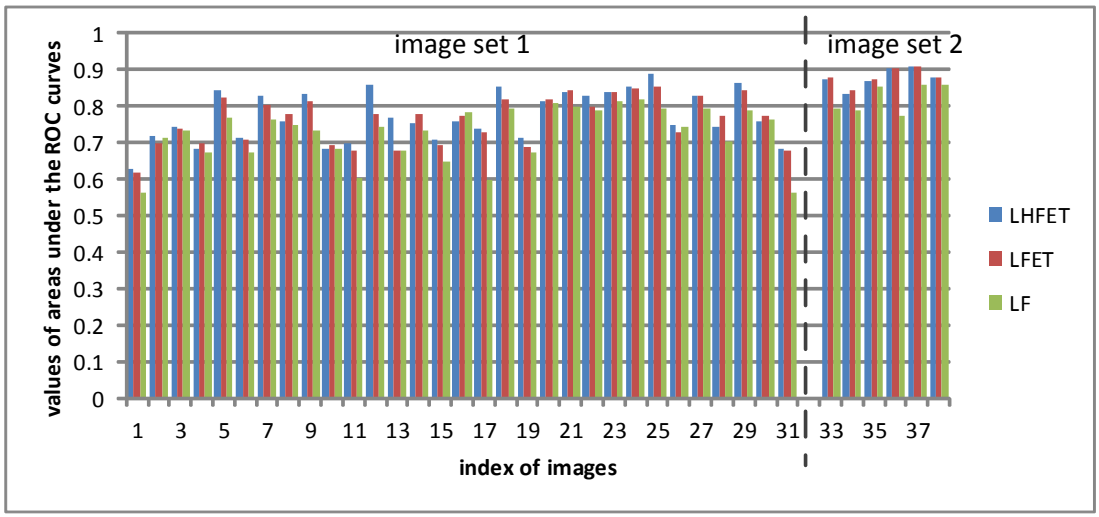


Figure 6. The values of the areas under the ROC curves of both image sets. The horizontal axis is index of images, the values of the areas under the ROC curve is on the vertical axis. On the left of the dotted line is the images belong to image set 1, on the right is the images belong to image set 2.

Table 3. The p-values of the t-test of whether there are significant differences between LHFET, LFET and LF, as well as the average differences for both image set.

	Alternative hypothesis:	LHFET is larger than LF	LFET is larger than LF	LHFET is larger than LFET
image set 1	The p-values of the null hypothesis:	<0.0001	<0.0001	0.0097
	The average difference:	4.84%	3.65%	1.19%
image set 2	The p-values of the null hypothesis:	0.0106	0.0074	0.9432
	The average difference:	4.28%	2.69%	1.59%

5 Discussion and Future Work

From the results of our experiments, we can see that it is apparent that people will subconsciously focus their attention on the salient object in the image, either for the images with only one salient object in it or the images with multiple salient objects in it, and ignore the background (see Figure 3 and Figure 5, the results of AET). Having a computational model to simulate this ability will be very useful for identifying interesting objects in a scene.

With respect to the visual attention model, using the actual eye tracking data to train the model improves the prediction of human fixations (for both scene types). Intuitively, training a model with real eye tracking data will allow us to effectively weight our image feature detectors, leading to a better result. Looking for the first row in Figure 3 for example, the edge of the door has a relatively large colour contrast, so the LF model marks this area as the most salient area. However, what people actually pay attention to is the sign on the door (see the AET image). The model trained with eye tracking data adjusts the weight of that type of feature, making the edge of the door less salient (see the LFET result), which better reflects real conditions.

As one might expect, the images with high-level features in it perform better with the model that seeks out high level features. Regarding the images without high-level features, there is no significant difference between the high-level and low-level trained models, however, using high-level features can perform slightly worse (see the average difference between LHFET and LFET for image set 2, in Table 3). This is because the high-level feature we used sometimes is not that accurate. And this is probably caused by the polysemy of the visual words that some features that look very similar but they actually belong to different categories. This also suggests one way to improve this model is to find a better way to make use of the high-level features in the images.

The model can be used to determine the most interesting parts of a new image. Part of our current work is the automatic tagging of the salient regions of images through known databases of objects in a bag-of-words model (Csurka et al., 2004). Here, there are several image representation methods, such as: dense sampling of image patches using a regular grid, scale-invariant features extracted by Hessian-Laplace or Harris-Laplace detectors, etc. However, these methods maybe not the most appropriate

for category recognition as it cannot differentiate the object from background. While others (such as Google (Zhao & Yagnik, 2012)) concentrate on automatically tagging everything it recognises in an image, we argue that a scene or image is better defined by its most salient parts, and therefore can provide an easier way of categorizing the image which is useful for modern social media purposes. Furthermore, by fusing image data acquired from a number of individuals from different viewpoints, we should be able to define a more sophisticated saliency model for the scene by consensus. Acquiring such data has become far easier with the advent of GPS data coupled with global image directories such as Facebook, Twitter and Instagram. With other techniques, like hashtags, it also can be used to track identified highly salient objects. So for example, 'elephant runs loose in the city centre of Penang' could be constructed automatically from lots of uploaded images where the most salient object is recognised as an elephant, all images suddenly occur at the same time, and all occur in the same small area identified by GPS location as being Penang.

A simple example of tagging only salient parts of images can be seen in Figure 7. Where traditional tagging may identify the lamppost and telephone box, these are not considered important by the saliency image. Additionally, where the tagging software may label these people as pedestrians, modelling through many images and GPS data may instead infer these people as students due to the location being known to be a university campus.

6 Conclusions

In this paper, we have proposed a computational visual attention model aimed at finding the most interesting objects in an image, and we discuss using it to identify objects of collective visual attention. We proposed a new type of high level feature that based on the bag-of-visual-word image representation, which can find the saliency on different

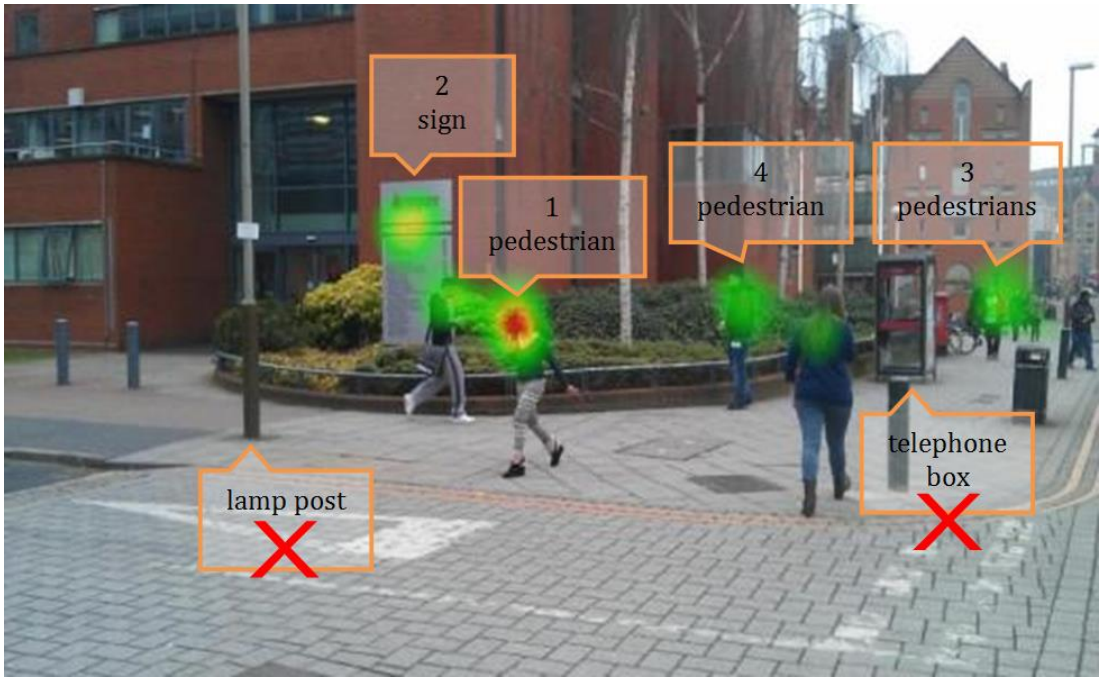


Figure 7. A simple example of tagging only salient image regions. Where conventional tagging software may tag the lamppost and telephone box, these are not considered salient and so are ignored in the new model.

categories. The visual attention model takes advantage of the low-level and high-level features of the images, using the actual eye tracking data as the ground truth of visual attention to train a SVR model. We have also conducted the experiments to evaluate the model. The results of the experiments show that for both image set—images with one salient object and images with many salient objects—using the actual eye tracking data to train the model improves predicting the human fixations, and it is more helpful for the images with many salient objects. For the images with high-level features, adding the BoW based high-level feature to the model will result in a better prediction of human fixations, whereas for the images without high-level features in it, adding the BoW based high-level feature to the model will not get a better prediction.

References

Adrian, V. (2013). OGAMA (Version 4.3). Retrieved from http://www.ogama.net/

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12), 10–10. doi:10.1167/9.12.10

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 1, p. 22).

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14).

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception*, *7*(1), 1–39. doi:10.1145/1658349.1658355

gazeMetrix. Retrieved September 10, 2013, from https://www.gazemetrix.com/

Ho, C.-H., & Lin, C.-J. (2012). Large-scale linear support vector regression. *Journal of Machine Learning Research*, *13*, 3323–3348.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506.

Itti, L., Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. doi:10.1109/34.730558

Judd, T., Ehinger, K. A., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *International Conference on Computer Vision* (pp. 2106–2113). doi:10.1109/ICCV.2009.5459462

Kienzle, W., Wichmann, F. A., Schölkopf, B., & Franz, M. O. (2006). A Nonparametric Approach to Bottom-Up Visual Saliency. In *Neural Information Processing Systems* (pp. 689–696).

Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web* (pp. 251–260). ACM.

Liang, Z., Fu, H., Chi, Z., & Feng, D. D. (2010). Refining a region based attention model using eye tracking data. In *Image Processing, IEEE International Conference* (pp. 1105–1108). doi:10.1109/ICIP.2010.5651804

Meur, O. L., Callet, P. L., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817. doi:10.1109/TPAMI.2006.86

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

Oliva, A., Torralba, A. B., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *International Conference on Image Processing* (Vol. 1, pp. 253–256). doi:10.1109/ICIP.2003.1246946

Privitera, C. M., & Stark, L. W. (2000). Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(9), 970–982. doi:10.1109/34.877520

Rajashekar, U., Linde, I. V. D., Bovik, A. C., & Cormack, L. K. (2008). GAFFE: A Gaze-Attentive Fixation Finding Engine. *IEEE Transactions on Image Processing*, *17*(4), 564–573. doi:10.1109/TIP.2008.917218

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *Image Processing, IEEE International Conference*. doi:10.1109/ICIP.1995.537667

Tuytelaars, T., Lampert, C. H., Blaschko, M. B., & Buntine, W. (2010). Unsupervised Object Discovery: A Comparison. *International Journal of Computer Vision*, *88*(2), 284–302. doi:10.1007/s11263-009-0271-8

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology-Human Perception and Performance*, *15*(3), 419–433. doi:10.1037//0096-1523.15.3.419

Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences*, *104*(45), 17599–17601.

Zhang, Y., Zhao, X., Fu, H., Liang, Z., Chi, Z., Zhao, X., & Feng, D. (2011). A Time Delay Neural Network model for simulating eye gaze data. *Journal of Experimental & Theoretical Artificial Intelligence*, *23*(1), 111–126. doi:10.1080/0952813X.2010.506298

Zhao, M., & Yagnik, J. (2012, August 28). Automatic large scale video object recognition.