

INFERRING TASK-RELEVANT IMAGE REGIONS FROM GAZE DATA

Arto Klami

Aalto University School of Science and Engineering
Department of Information and Computer Science
Helsinki Institute for Information Technology HIIT
P.O.Box 15400, FI-00076 Aalto, Finland

ABSTRACT

A number of studies have recently used eye movements of a user inspecting the content as implicit relevance feedback for proactive retrieval systems. Typically binary feedback for images or text paragraphs is inferred from the gaze pattern. We seek to make such feedback richer for image retrieval, by inferring which parts of the image the user found relevant. For this purpose, we present a novel Bayesian mixture model for inferring possible target regions directly from gaze data alone, and show how the relevance of those regions can then be inferred using a simple classifier that is independent of the content or the task.

1. INTRODUCTION

Proactive systems analyze the behavior of the user and infer the users intentions and interest from measurements of their implicit activity. Eye movements are a natural information source for such systems, and they have been used for example in information retrieval systems where the goal is to infer implicit relevance feedback from gaze, to replace or complement explicit feedback. In text retrieval the task has been either to infer relevance for text blocks such as paragraphs or documents [1, 2], or even to learn a word-level relevance estimate that can directly be used as an implicit query string [3]. For images the approaches have been comparable to the former, i.e., the interest has been in inferring relevance of full-scale images [4, 5].

In this paper we take the first steps towards learning more finely grained relevance feedback in image perception. Instead of working at the level of full images, we study the problem of inferring which parts of an image are relevant for the user in a given task. In this work the relevance is inferred solely from the gaze pattern alone, not using the image content at all, and can be considered as the opposite of the bottom-up saliency models [6, 8, 9] that attempt to predict the gaze target from image content. Feedback given at the level of image regions is much richer than univariate (binary or continuous) implicit feedback on full images.

From the modeling point of view, we introduce a novel generative model for modeling a collection of fixation data of several users viewing the same image, with potentially different tasks. The model is a Gaussian mixture model (GMM) of the fixation data, with an interesting twist on the membership prior. A standard GMM assumes all data points (here fixations) are independent and identically distributed to the Gaussian image regions, and that the component membership variables are chosen from a fixed discrete distribution over the components. This means, for example, that a region inspected very carefully by a single user will have high weight, despite not necessarily being viewed by the other users at all.

We avoid such problems by a novel prior for the membership assignments. Instead of assuming them to be independent, we separate the notion of whether a component is used and how many samples it generates. This is done by having a discrete distribution for the former, separately for each potential task for the users, coupled by a truncated Poisson distribution for the number of samples. This allows finer control on the distribution of the samples over the components. Equivalently, the model can be viewed as a latent feature model [7] for the users, where each feature corresponds to inspection of a single image region and generates a set of fixations. The prior allows, for example, image regions that are unlikely to be observed but will be carefully investigated for several fixations when they are. We will show experimentally how the regions found by the model are more interpretable as potential gaze targets, compared to modeling the fixation collection with a standard GMM.

Based on the regions learned with the novel mixture model we then infer the relevance of those regions with logistic regression, using gaze-based features computed for each region. We show that we can infer the most relevant data-driven region as accurately as we can infer the most relevant full image. Hence, at least in a simplified retrieval setting we do not lose any accuracy by giving the feedback on sub-image level, but obviously get much richer feedback.

2. PSYCHO-VISUAL MOTIVATION

Image perception analysis is dominated by two major paradigms, bottom-up and top-down modeling. The bottom-up paradigm studies how the content influences gaze, while top-down processing is viewed as active control of attention. It is assumed that in a free-viewing task the bottom-up control dominates attention, which has led to development of models for visual saliency, models that infer likely attention targets from the image content, both based on psycho-visual motivation [6] as well as data-driven approaches [8, 9].

However, users with strong tasks are known to be able to override the bottom-up saliency. For example [10, 11] report that a viewers with clear tasks ignore low-level saliency almost completely and that they can do it already on image onset, during the very first fixations. These works hint towards the observation that under specific tasks (such as image retrieval when the user is focusing on the task) the low-level saliency is of not particular interest, but instead the focus should be on analyzing the gaze data itself.

We take that approach to the extreme, excluding the image content from the analysis completely. This has the added benefit that we can learn models that are, at least to a degree, independent of the content. We will require a collection of gaze data for a particular image in order to learn the regions, but can then learn content-independent relevance predictors for those regions. For detecting the target regions we propose a Bayesian model learned from a collection of eye tracking data of several users viewing the same image. The model is designed based on a number of observations from vision research, each of which is then converted to a probabilistic description fitting the model:

1. Target regions are localized; even if there are large objects of interest the users will typically inspect specific details of those.
2. The perceptive field is fairly wide. The accurate foveal vision covers roughly 1-2 degrees of visual field, but relevance can be determined for wider area.
3. The set of potential targets is not task-specific, assuming typical image search tasks, but instead a property of the image content. Users with a specific task will, however, look at the targets in different ways.
4. In a search task a user only views a subset of possible targets.
5. A user will typically observe a single target only for a few fixations.

3. DATA-DRIVEN TARGET EXTRACTION

3.1. Mixture model

Based on the first three points we model each target region with a Gaussian distribution. A user viewing a particular

target is assumed to have a fixation that is localized around the mean of the region, while rather large variation is allowed because of the width of the perceptive field. Possible target regions for an image are hence parameterized as the set $T = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ containing the means and covariances of the Gaussians.

Our data consists of N users with N_i fixations each, denoted by the set $\{\mathbf{f}^i, t_i\}_{i=1}^N$ where $\mathbf{f}^i \in \mathbb{R}^{N_i \times 2}$ are the fixation coordinates and t_i is the task of the i th user. For assigning the fixations to the mixture components we use triple indexing. That is, binary \mathbf{w}_{ijk} tells whether the j th fixation of the i th user is assigned to the k th component. Plugging in the third observation of task influencing the region probabilities gives the Gaussian mixture model (GMM)

$$p(\{\mathbf{f}\}, \mathbf{w}|T) = \prod_i \prod_j \sum_k p(\mathbf{w}_{ijk}|t_i) \mathcal{N}(\mathbf{f}_j^i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The remaining task is specifying the prior $p(\mathbf{w}_{ijk}|t_i)$, which for GMM would typically be a multinomial distribution $p(\mathbf{w}_{ijk}|t_i) = \text{Mult}(\boldsymbol{\pi}_{t_i})$ with Dirichlet prior on the parameters, $\boldsymbol{\pi}_{t_i} \sim \text{Dir}(\alpha)$. Such a prior is, however, inconsistent with the last observations. Hence, we will next construct a novel prior for the assignment variables. The result is a prior over the whole matrix of assignment variables for a single user, that is $p(\mathbf{w}_i)$, which breaks the traditional assumption of independence between the samples.

Mathematically the fourth observation can be formulated as a single user viewing only a subset of the K available regions. We use \mathbf{z}_{ik} to denote whether the i th user viewed the k th target region, and encode the observation by making the assumption that $\mathbf{m}_i = \sum_k \mathbf{z}_{ik} \sim \text{Poisson}(\gamma_i)$ with user-specific concentration parameter γ_i . The probabilities for the components to be active are Bernoulli with task-dependent parameters θ_t . Finally, the fifth observation states that the user is likely to have only a few fixations within each region. Hence, each active user-target assignment (i.e., $\mathbf{z}_{ik} = 1$) is assumed to generate $\mathbf{l}_{ik} \sim \text{Trunc-Poisson}(\lambda_k)$ fixation locations, each following the distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The Poisson distribution is here truncated to exclude zero observations, since an active component by default generates at least one fixation

$$p(\mathbf{l}_{ik} | \lambda_k) = \frac{\lambda_k^{\mathbf{l}_{ik}} e^{-\lambda_k}}{(1 - e^{-\lambda_k}) \mathbf{l}_{ik}!} \mathbb{1}(\mathbf{l}_{ik} > 0).$$

Together these two assumptions result in the unnormalized prior

$$p(\mathbf{w}_i | t_i) \propto p(\mathbf{m}_i | \gamma_i) \prod_k (p(\mathbf{z}_{ik} | \theta_{t_i}) p(\mathbf{l}_{ik} | \lambda_k)^{\mathbf{z}_{ik}}),$$

where $\mathbf{l}_{ik} = \sum_j \mathbf{w}_{ijk}$, \mathbf{z}_{ik} is one if and only if $\mathbf{w}_{ijk} = 1$ for at least one j , and $\mathbf{m}_i = \sum_k \mathbf{z}_{ik}$. To summarize the dis-

tributions, $p(\mathbf{m}_i)$ is Poisson, $p(\mathbf{z}_{ik})$ is Bernoulli, and $p(\mathbf{l}_{ik})$ is truncated Poisson.

The full model for the fixation data for all N users is then given as

$$p(\{\mathbf{f}\}, \mathbf{m}, \mathbf{L}, \mathbf{z}, \mathbf{w}) \propto \prod_{i=1}^N p(\mathbf{m}_i | \gamma_i) \prod_{k=1}^K \left(p(\mathbf{z}_{ik} | \theta_{t_i}) p(\mathbf{l}_{ik} | \lambda_k) \prod_{j=1}^{N_i} (N(\mathbf{f}_j^i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{\mathbf{w}_{ijk}} \right)^{\mathbf{z}_{ik}},$$

where conditioning on the parameters has been left out to simplify the formula.

3.2. Priors and inference

The model is complemented with prior distributions for the parameters, allowing posterior inference over the parameters. For efficiency, we choose conjugate priors for all of the parameters, resulting in

$$\begin{aligned} p(\boldsymbol{\Sigma}_k | \nu, \mathbf{S}) &= \text{IW}(\nu, \mathbf{S}), \\ p(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k, \boldsymbol{\mu}_0, \kappa) &= \text{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_k / \kappa), \\ p(\gamma_i | \alpha_\gamma, \beta_\gamma) &= \text{Gamma}(\alpha_\gamma, \beta_\gamma), \\ p(\lambda_k | \alpha_\lambda, \beta_\lambda) &= \text{Gamma}(\alpha_\lambda, \beta_\lambda), \\ p(\theta_t | \alpha_\theta, \beta_\theta) &= \text{Beta}(\alpha_\theta, \beta_\theta), \end{aligned}$$

where $\text{IW}(\nu, \mathbf{S})$ denotes the Inverse-Wishart distribution with ν degrees of freedom and mean \mathbf{S} . We choose $\boldsymbol{\mu}_0$ as the image centroid, $\kappa = 0.01$ to give noninformative distribution for the mean, and $\nu = 5$, $\mathbf{S} = [50^2, 0; 0, 50^2]$ to capture the prior information on preferred region size. The prior equals to already having 5 fixations for each region. For θ we use non-informative prior $\alpha_\theta = 1$, $\beta_\theta = 1$, whereas for γ we set $\alpha_\gamma = 10$, $\beta_\gamma = 2$ to favor users viewing on average around $\mathbb{E}[\gamma] = 5$ regions, and finally we set $\alpha_\lambda = 4$, $\beta_\lambda = 2$ to weakly prefer roughly $\mathbb{E}[\lambda] = 2$ fixations per region.

For inference we use Gibbs sampling. The sampling formulas for the set T are identical to a standard GMM. Also the formulas for sampling γ_i and θ_t are straightforward. The target-region assignment vectors \mathbf{w}_{ij} are sampled element by element, always computing the relative likelihoods of all K possible choices, conditioned on all other assignments. Finally, the prior for λ_k is not conjugate, but the posterior is log-concave and hence efficient inference can be done with adaptive rejection sampling [12].

3.3. Related work

Even though the model is above formulated as a mixture model for the fixation collection, it could alternatively be seen as a latent feature model [7] for the users. The \mathbf{z}_{ik} are the latent variables telling which features are active for each

Table 1. Features for predicting the relevance of a region.

| Feature | Type |
|---|------------|
| Number of fixations | discrete |
| Total fixation time | continuous |
| Length of the first fixation | continuous |
| How many times the region was visited | discrete |
| Time of first fixation since onset | continuous |
| Time of last fixation since onset | continuous |
| Index of the first fixation | discrete |
| Index of the last fixation | discrete |
| Standard deviation of fixation indices | continuous |
| Is the first fixation within this region? | binary |

of the users, and the task-dependent region weights and the distribution of the number of active features act as a prior for these variables. Contradictory to traditional latent feature models the model does not, however, generate a fixed observation vector but instead a set of data points associated with this user. Interestingly, the model also bears close relation to infinite latent feature models. Each user viewing a Poisson number of regions is consistent with the assumption made by the Indian Buffet Process (IBP) model [13], and the combination of a binary feature prior and Poisson for the number of fixations for each feature is close to the Gamma-Poisson prior for sparse count matrices [14]. These links suggest infinite extensions of the proposed model.

The output of the model bears similarities with fixation heatmaps computed for pooled eye-tracking data [15]. The model has, however, several advantages over computing separate heatmaps for each task. It gets rid of heuristic smoothing parameters, takes into account properties such as limited number of fixations within a single region of interest, produces smoother results that also capture the uncertainty due to posterior modeling, and is able to utilize the information from other tasks when detecting the regions. More importantly, it maps fixations to well-defined regions, which makes possible prediction of the relevance given the fixation features.

4. INFERRING REGION RELEVANCE

The model finds the regions and tells how frequently they are viewed by users with different tasks. This is enough for understanding image perception under those tasks. However, it does not generalize to new images or new users with unknown tasks. In order to do that we need a classifier predicting for each region+user pair whether that region was relevant for that particular user, based on a feature representation of the gaze pattern.

This is, in fact, equivalent to the task studied e.g. by [5]. We know a set of regions the users could potentially

have viewed, and infer a binary relevance for each of those. Building on the earlier work on full-image prediction, we apply a logistic regression classifier

$$p(r|\mathbf{h}_k) = \frac{1}{1 + \exp(-\mathbf{a}^T \mathbf{h}_k - a_0)},$$

for a set of gaze-based features computed for each region. Here r_k is the estimated relevance, \mathbf{a} is a weight vector, a_0 is the bias term, and \mathbf{h}_k is the feature vector for the k th region, consisting of the 10 features in Table 1 and their second order products. All features are normalized with the z-score transformation for each user and image, making dependence on personal variation and image content weaker.

5. EXPERIMENTS AND RESULTS

To demonstrate two possible uses of the model, we perform experiments on two data sets consisting of voluntary users inspecting images under specific tasks. The first data set has 22 images, each viewed by 13+12 users having two different tasks. The users were viewing full-screen outdoor and indoor images of houses, and were asked to take the role of either burglar (task 1) or house-buyer (task 2), evaluating the potential of the particular house for their task. On this kind of setups the model can be used for interpreting how people with different tasks perceive images.

The other data set is the one used by [4] to evaluate implicit relevance feedback in image retrieval. The experiment consists of 100 pages of images, each showing 4 different images out of which at most one is relevant for the task of detecting sports images. All 21 users share the same task; they were asked to search for images related to sports. This data is used to show how the enhanced relevance feedback could be inferred in retrieval settings.

5.1. Detecting target regions from gaze

Figure 1 shows an example output of the model, averaged over the posterior samples, revealing clear differences between the tasks. The fixation heatmaps shown as comparison do not provide interpretable regions or basis for relevance prediction.

The advantage of the proposed model over the standard GMM is demonstrated by two quantitative measures, averaged over models trained for all 22 images. Ideal regions should be of roughly equal probability and size to represent meaningful gaze targets, yet it should be possible for the two tasks to have very different probabilities for some regions. We measure the first by the spread of the region probabilities (the difference between the most and least likely regions) and the latter by a measure of task-specificity: For each region we compute $\max_t \left| \frac{M_{kt}}{\sum_v M_{kv}} \right|$, where M_{kt} denotes the number of users with task t viewing the k th region. The

Table 2. Relevance prediction accuracies for two model complexities. The proposed model clearly outperforms the baselines of choosing the region with the most fixations and choosing a random region. For comparison, we also show the accuracy in predicting binary feedback for full images.

| | Predictor | Most fixated | Random |
|--------|-----------|--------------|--------|
| K=10 | 71.3% | 57.3% | 48.5% |
| K=15 | 73.1% | 55.5% | 47.9% |
| Images | 74.9% | 61.9% | 44.3% |

score is 0.5 for regions viewed equally often by users with both tasks and 1 for regions ignored by users of one task. A good model finds some regions with high specificity, indicating that it has captured task-specific viewing patterns.

We compared the model to Bayesian GMM (with all other priors identical to our model), running the latter with two different hyperparameters α to demonstrate how it can only achieve one out of two desirable properties at a time. With small α GMM tends to find a few large regions that are not interpretable, whereas encouraging more equal region sizes by setting larger α results in loss of task-specificity; each region is viewed by users of both tasks (Figure 2).

5.2. Predicting relevance of regions

It is fairly difficult to collect explicit ground truth labeling for parts of images, and hence we create an artificial labeling for the sports data using the actual image relevances as labels. In brief, every region within the sports-relevant image is considered as relevant, while the rest are marked non-relevant. This will not be a perfect labeling as non-relevant images also have regions that need to be checked in order to determine the relevance, but is sufficient for our purposes.

We apply the model in a leave-one-page fashion, pooling the data for all users, pages and regions for training a predictor for one left-out-page. We then test the model on the remaining page, reporting the percentage of users with the most relevant region within the sports image, and average the results over all pages. We compare the method with two baselines. One chooses a random region out of the ones the user viewed, measuring the gain from having the gaze data (without gaze random guessing with 25% accuracy is the only option). The other baseline chooses the region with the most fixations, approximating selection based on the mode of the fixation heatmap. Table 2 reports the scores for two model complexities, showing clear gain compared to both baselines, and Figure 3 shows two examples.

We also applied the same relevance predictor for the whole images instead of the extracted regions. Each image is considered a region, and the model produces as an out-



Fig. 1. Visualization of relevant regions for two tasks, house-buying (blue) and burglary potential (red) evaluation. **Left:** Proposed model extracts interpretable regions such as the cell-phone on the table and the contents of the shelf on the right being relevant for burglars. The house-buyers, instead, focused on the window on the left and the central area showing access to other rooms. **Right:** Gaze heatmaps for the two tasks are noisier and less clearly task-oriented, but reveal similar aspects.

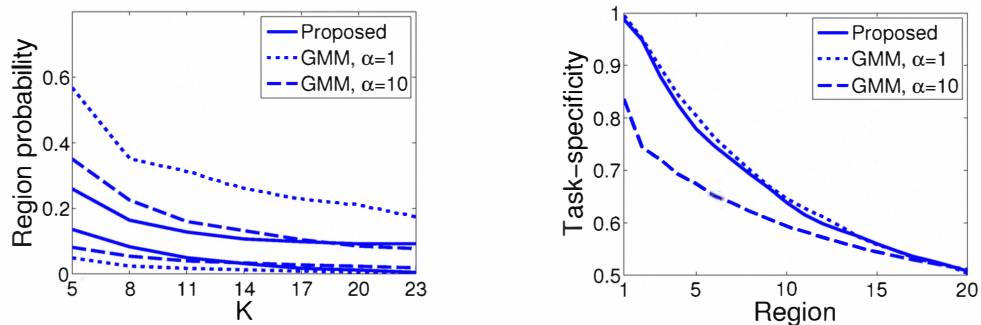


Fig. 2. **Left:** Spread of component sizes (the lines are the frequencies of the smallest and largest regions) as function of the number of regions (model complexities). **Right:** Regions of the models (with $K = 20$) ordered according to the task-specificity (higher is better). The proposed model achieves both roughly uniform component sizes and good specificity at the same time, while GMM only achieves the former with large α and the latter with small α .

put a probability of relevance for each image. Interestingly, the accuracy for predicting the relevant regions is comparable to that of full image prediction, meaning that the richer sub-image feedback comes with no decrease in accuracy.

6. DISCUSSION

Gaze is informative of the interest of the user, and can be used for inferring relevance feedback in information retrieval setups [2]. We studied the feasibility of making image retrieval feedback richer by complementing or replacing full image relevance [5] by extracting image regions estimated to be relevant. The results are not yet integrated in an actual retrieval system, but instead we studied merely the tasks of extracting the regions and inferring their relevance.

The regions were extracted by a novel Bayesian mixture model. Instead of assuming all samples independent, we developed a component assignment prior that better captures the process underlying image perception. The most impor-

tant detail is that the amount of data and the probability of the component being active are separated from each other, using a representation akin to latent feature models [7]. We then showed on real image perception data how the new prior results in components that can better be interpreted as potential attention targets.

We also showed how the task-relevance of such regions can be inferred for a new user given a simple classifier based on only the gaze data, clearly outperforming the alternative of picking the region with the most fixations. The classifier is content- and task-independent, and readily applicable to new images without heavy computation.

7. ACKNOWLEDGMENTS

The author belongs to AIRC, a CoE of the Academy of Finland. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under *grant agreement*



Fig. 3. Left: Demonstration of relevant region detection. The blue contours show the areas viewed by the users, while the red contours indicate the regions predicted to be most relevant for each of the users. Note how the blue contours cover regions in all images, yet the most relevant region is outside the sports image for only one user. **Right:** A closeup of the relevant image on one of the pages, showing how different regions of the image are chosen most relevant by different users. In both cases the relevant regions of different users have been drawn with jitter, to reveal multiple choices of the same region.

n° 216529 and from the Academy of Finland decision number 133818, and was in part supported by the PASCAL2 EU NoE. We would like to thank Dr. Johanna Kaakinen (Department of behavioural sciences and philosophy, Division of Psychology, University of Turku) for providing the data used in the first experiment.

8. REFERENCES

- [1] G. Buscher, A. Dengel, and L. van Elst, “Eye movements as implicit relevance feedback,” in *CHI '08: CHI '08 extended abstracts on Human Factors in Computing Systems*, 2008, pp. 2991–2996.
- [2] K. Puolamäki, J. Salojärvi, E. Savia, J. Simola, and S. Kaski, “Combining eye movements and collaborative filtering for proactive information retrieval,” in *Proceedings of SIGIR'05*, 2005, pp. 146–153.
- [3] K. Puolamäki, A. Ajanki, and S. Kaski, “Learning to learn implicit queries from gaze patterns,” in *Proceedings of ICML*, 2008, pp. 760–767.
- [4] A. Klami, C. Saunders, T. de Campos, and S. Kaski, “Can relevance of images be inferred from eye movements?,” in *Proc. of the ACM Conference on Multimedia Information Retrieval*, 2008, pp. 134–140.
- [5] L. Kozma, A. Klami, and S. Kaski, “GaZIR: Gaze-based zooming interface for image retrieval,” in *Proc. ICM-MLMI 2009, International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction*, 2009, pp. 305–312.
- [6] L. Itti and C. Koch, “A saliency-based search mechanism for overt and covert shifts of visual attention,” *Vision Research*, vol. 40, 2000.
- [7] Z. Ghahramani, “Factorial learning and the EM algorithm,” in *Advances in Neural Information Processing Systems 7*, 1995.
- [8] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems*, 2006.
- [9] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M.O. Franz, “A nonparametric approach to bottom-up visual saliency,” in *Advances in Neural Information Processing Systems 19*. 2007, pp. 689–698.
- [10] J. Henderson, J. R. Brockmole, M. S. Castelhana, and M. Mack, “Visual saliency does not account for eye movements during visual search in real-world scenes,” in *Eye movements: A window on mind and brain*, pp. 538–562. Elsevier, 2007.
- [11] W. Eihäuser, U. Rutishauser, and C. Koch, “Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli,” *Journal of Vision*, vol. 8, pp. 1–19, 2008.
- [12] W.R. Gilks and P. Wild, “Adaptive rejection sampling for gibbs sampling,” *Applied Statistics*, vol. 41, pp. 337–348, 1992.
- [13] T. Griffiths and Z. Ghahramani, “Infinite latent feature models and the Indian buffet process,” in *Advances in Neural Information Processing Systems 18*, 2006.
- [14] M.K. Titsias, “The infinite gamma-poisson feature model,” in *Advances in Neural Information Processing Systems 19*, 2007.
- [15] O. Spakov and D. Miniotas, “Visualization of eye gaze data using heat maps,” *Electronics and electrical engineering*, vol. 2, no. 74, 2007.