Deliverable D2.2.1 + D2.2.2

Predicting relevance of parts of an image

Demonstrator for predicting relevance of image parts

Contract number: **FP7–216529** PinView

Personal Information Navigator Adapting Through Viewing

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under *grant agreement* n° 216529.

Identification sheet

Project ref. no.	**FP7–216529**
Project acronym	PinView
Status and version	Final, Revision: 1.00
Contractual date of delivery	31.12.2009
Actual date of delivery	31.12.2009
Deliverable number	D2.2.1 + D2.2.2
Deliverable title	Predicting relevance of parts of an image Demonstrator for predicting relevance of image parts
Nature	report + prototype
Dissemination level	Report: PU – Public Prototype: PP – Restricted to other programme participants
WP contributing to the deliverable	WP2 Learning relevance feedback from eye tracking
Task contributing to the deliverable	Task 2.2 Relevance of parts of an image from the viewing pattern
WP responsible	Teknillinen korkeakoulu
Task responsible	Teknillinen korkeakoulu
Editor	Arto Klami, <arto.klami@tkk.fi>
Editor address	P.O.BOX 5400, FI-02015 TKK, Finland
Authors in alphabetical order	Gabriela Csurka, Steve R. Gunn, David R. Hardoon, Samuel Kaski, Arto Klami, Kitsuchart Pasupa, Sandor Szedmak
EC Project Officer	Pierre-Paul Sondag
Keywords	gaze trajectory, image relevance, top-down saliency
Abstract	This report studies the task of inferring which parts of an image are relevant for the user viewing the image. The relevance is inferred from gaze trajectory of users viewing the images given a specific task. Novel computational models based on both Bayesian generative modeling and kernel methods are developed for inferring the regions of interest from raw fixation data, as well as from combination of eye-movements and image content features.

List of annexes

`pinview-d2-2-2-final.zip - Deliverable D2.2.2, a prototype Matlab implementation of a model for inferring task-relevant sub-image regions`

Contents

1 Overview

This deliverable that constitutes the output of Task 2.2 *Relevance of parts of an image from the viewing pattern* of the *Personal Information Navigator Adapting Through Viewing*, PinView, project funded by the European Community's Seventh Framework Programme under Grant Agreement n° 216529, consists of two parts. The first part is this report, Deliverable 2.2.1 *Predicting relevance of parts of an image*, describing three novel computational models for estimating relevance of sub-images from gaze, image content features, or combination of those. The second part is a prototype, Deliverable 2.2.2 *Demonstrator for predicting relevance of image parts*, that demonstrates the process of sub-image relevance estimation with one of the models.

The relevance of parts of images is studied from three perspectives. First, we present a method for inferring regional relevance estimates from collections of gaze measurements made in different tasks, and study to which degree it is possible to infer task-based relevance from gaze alone. The second model utilizes not only eye-movements but also image-level features. The model is based on the idea of multi-view learning, and predicts gaze target from both the eye-movements and image content descriptions. Further, we have developed a model for estimating visual saliency from image content, aiming to approximate the attention information even for cases where we do not have any gaze data for test images. All three models build on top of the concept of gaze or fixation heat map, a non-parametric method traditionally used for analyzing or visualizing attention.

The Deliverable 2.2.2, demonstrator for relevance prediction of parts of images, is described as a part of this report, in Section 6. It is an implementation of a method that takes an image and set of scan patterns as an input and produces local image regions estimated to be relevant as output.

The work presented in this report continues the work initiated in Task 2.1 *Relevance of an image from a scan pattern*. The main conclusion of Task 2.1 was that gaze pattern is useful in predicting image relevance in collage setups. The main findings in this report are similar: gaze is definitely informative of the sub-image relevance as well. Furthermore, the results suggest that image search tasks involve high degree of top-down control in gaze directing. The work will be continued primarily in Task 2.3 *Data fusion*, where means to complement the gaze information will be studied in order to improve the performance further. Data fusion will also be considered from the perspective of generalizing over queries and images.

This report describes contributions of four project partners, TKK, UCL, SOTON, and XRCE.

2 Introduction

Task 2.1 studied inferring the relevance of thumbnail images, making the assumption that only a subset of the images shown at a time will be relevant for the user. Gaze has been successfully used to predict that information in [15, 16, 18], showing that the approach is feasible. In this deliverable we look deeper in the issue of relevance by studying which parts or regions of larger images the user finds relevant. Relevance is here understood as something that needs to be examined in order to determine whether the image matches the needs of the user. That does not, however, necessarily match all the regions of the image the user looks at, since some fixations may emerge because of highly salient features of the image. Further, the relevant regions need not match relevant objects either; when searching for items kept on tables the user would typically need to look at all tables regardless of the presence of the search target.

Low-level visual saliency has been studied extensively in vision psychology, resulting in detailed models such as [12] designed based on how eyes and visual cognition work. Recently there has also been extensive research on more data-driven and computationally advanced models for the same task, such as [10, 14, 23]. These works, as well as the concept of visual saliency in general, are based on one strong assumption: the user is viewing the images without a particular task. It is assumed that in absence of a task the gaze will be directed by low-level image features, and hence gaze targets can be predicted from image content features alone.

In recent years there have been an increasing amount of studies on tasks other than free viewing, giving somewhat surprising results. For example [11] reports that a viewer with a clear task is able to ignore low-level saliency almost completely, indicating that the gaze control is primarily top-down instead of bottom-up. Studies like [6] go even further, noting that people can revert to top-down control immediately on image onset, already during the first fixation. These works hint towards the observation that under specific tasks (such as image retrieval when the user is focusing on the task) it may be sufficient to merely detect where the user is looking at, since everything they look at is relevant in some sense (either because it is the search target, or it tells about non-existence of the search target).

The existing knowledge on image perception is dominated by these two major paradigms. In free-viewing tasks the low-level saliency controls the gaze, while with clear tasks the gaze is explicitly controlled by the task. To which degree image search tasks fall into these extremes is, however, an open question that to our knowledge has not been studied extensively. Because of this uncertainty we worked on a number of models with different basic assumptions, while trying to figure out the degree of contribution for the two parts at the same time.

In this deliverable we report three different modeling approaches. First, we build a generative model for extracting task-relevant regions from fixation data alone, and use that to study a simple hypothesis: a region viewed by several users with the same task is likely to be relevant for that task. Second, we complement the pure gaze data with image-level features, and use multi-view learning to predict gaze direction of new users with the same task. Finally, we present a method for inferring top-down saliency maps from image features alone. The first approach seeks to provide an answer to the basic question of whether gaze alone can reveal information on relevance beyond the binary dichotomy on whether the region was viewed or not, while the latter two are useful also for scenarios where the eye-movements are not available on the test images.

3 Model for task-relevant regions

We build a two-stage model for predicting relevance of sub-images. The first stage of the model extracts possible target regions in a data-driven way, based on gaze data alone. The second stage then infers, for these targets, the relevance for the given task.

3.1 Target detection

We propose a Bayesian statistical model for detection of target regions. The model is designed based on a number of observations from vision psychology, each of which is then converted to a probabilistic description fitting the model. The main properties can be summarized as:

1. Potential target regions in an image are typically localized into relatively small neighborhoods; even though there may be large objects of interest in the image, the users will typically focus on specific details of those.

2. The perceptive field of human eye is fairly wide. The highly accurate foveal vision covers roughly 1-2 degrees of visual field, but much wider area can be perceived with sufficient accuracy for determining the relevance.

3. In a search task a user will spend limited time observing a single image, and hence will only view a few target regions.

4. A user will typically observe a single target only for a few fixations.

5. The potential set of targets is not task-specific, assuming typical image search tasks, but instead a universal property of the image content. Users with a specific task will, however, look at the targets in different ways.

Based on the first two 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. The set of possible target regions for an image is hence parameterized as $T = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$, where $\boldsymbol{\mu}_k$ is a two-dimensional mean vector and $\boldsymbol{\Sigma}_k$ is the covariance matrix of the kth region.

The third observation points towards a factorial model [8] for the full gaze data. Given the set T, a single user will view a subset of the K regions. We use z_{ik} to denote whether the ith user viewed the kth target region, and can hence collect all the user-target allocations in a single matrix $\mathbf{Z}$. The assumption on each user viewing only a few regions is encoded by making the assumption that $m_i = \sum_k z_{ik} \sim \text{Poisson}(\lambda_1)$ with concentration parameter λ_1. This assumption is consistent with the assumption made by the Indian Buffet Process (IBP) models [9], the extension of factorial models to infinite latent feature models.

The fourth observation links the gaze data with the factorial model. The ith user has a total of N_i fixations on the image, which need to be distributed for the M_i target regions. Given the Gaussian regions, this can be formulated as Poisson-Gaussian likelihood for the factorial model. Each active user-target assignment (i.e., $z_{ik} = 1$) is assumed to generate $L_{ik} \sim \text{Poisson}(\lambda_2)$ fixation locations, each following the distribution $\text{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

The full model, identified as a finite variant of a IBP model with Poisson-Gaussian likelihood, for a collection $\mathbf{F} = \{\mathbf{F}^i\}_{i=1}^{U}$ of fixation data for U users is then given as

$$p(\mathbf{F}, \mathbf{m}, \mathbf{L}, \mathbf{z}, \mathbf{w} | \lambda_1, \lambda_2, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}) = \prod_{i=1}^{U} p(m_i|\lambda_1) \prod_{k=1}^{K} \left(p(L_{ik}|\lambda_2) \prod_{j=1}^{N_i} \left(p(\mathbf{f}_j^i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{w_{ijk}} \right)^{z_{ik}}.$$

Here w_{ijk} is an indicator variable telling for each fixation j of the ith user whether it belongs to the kth target region or not, and $\mathbf{f}_j^i$ is a two-dimensional vector of the fixation coordinates of the jth fixation of ith user. The probability distributions appearing above are

$$p(m_i|\lambda_1) = \text{Poisson}(\lambda_1), \quad p(L_{ik}|\lambda_2) = \text{Poisson}(\lambda_2), \quad p(f_{ij}|\mu_k, \Sigma_k) = \text{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The model is further complemented with prior distributions for the parameters λ_1, λ_2, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. This allows full posterior inference over the model parameters. For simplicity of computation, we adopt conjugate priors for all of the parameters, resulting in

$$\begin{aligned} p(\boldsymbol{\Sigma}_k|\nu, \mathbf{S}) &= \text{IW}(\nu, \mathbf{S}), \\ p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \boldsymbol{\mu}_0, \kappa) &= \text{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{\boldsymbol{k}}/\kappa), \\ p(\lambda_1|\alpha_1, \beta_1) &= \text{Gamma}(\alpha_1, \beta_1), \\ p(\lambda_2|\alpha_2, \beta_2) &= \text{Gamma}(\alpha_2, \beta_2), \end{aligned}$$

where $\text{IW}(\nu, \mathbf{S})$ denotes the Inverse-Wishart distribution with ν degrees of freedom and mean parameter $\mathbf{S}$.

The model relaxes the classical assumption of independence over data points (here the fixations). Instead, it places a prior directly on the assignment vectors $\mathbf{z}_i$ and $\mathbf{w}_{ij}$, only requiring that the sums of those follow the given Poisson distributions. This equals an implicit assumption that given the counts m_i and L_{ik} the actual contents of $\mathbf{z}_i$ and $\mathbf{w}_{ij}$, respectively, follow uniform distribution. The model can be modified to take the tasks into account by relaxing that assumption. In practice, we use the prior

$$p(\mathbf{z}_i|t_i) \propto p(m_i|\lambda_1) \prod_{k=1}^{K} p(z_{ik}|\theta_{t_i}),$$

where $p(z_{ik}|\theta_{t_i}) = \text{Bernoulli}(\theta_{t_i})$ measures the probability that the user with task t_i views the kth region. The model then has a set of these Bernoulli-distributions, one for each combination of region and possible task in the collection. The parameters θ_{t_i} of the Bernoulli distributions follow the conjugate prior

$$p(\theta_{t_i}|\alpha_{t_i}, \beta_{t_i}) = \text{Beta}(\alpha_{t_i}, \beta_{t_i}).$$

For inference of the full model we use posterior sampling, in particular a Gibbs-sampler. The sampling formulas for the mean and covariance parameters of the target regions, as well as the hyperparameters of those parameters, are identical to a Gaussian mixture model. The sampling formulas for the Poisson and Bernoulli parts are straightforward, while the target-region assignment vectors $\mathbf{z}_i$ and $\mathbf{w}_{ij}$ require explicitly computing the relative likelihoods of all possible assignments. The set of target regions is, however, finite, and hence these steps are also easy to derive and implement.

On a surface level, the output of the target detection model bears similarities with fixation heat maps, explained in more detail in Section 5, traditionally computed for pooled eye-tracking data (see, e.g., [21]). The task-specific regions of interest could alternatively be detected by computing heat maps for the fixation data of each task and comparing those. The proposed model, however, has several advantages. It gets rid of heuristic smoothing parameters of the fixation heat map, it takes into account the characteristic properties of gaze such as limited number of fixations within a single region of interest, it produces smoother results that also capture the uncertainty due to posterior modeling, and it is able to utilize the information from other tasks when detecting the regions. More importantly, it allows mapping fixations to well-defined regions instead of producing merely a density estimate of the data, which makes possible prediction of the relevance given the fixation features.

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

Table 1: List of features used for predicting the relevance of a target region.

3.2 Inferring target region relevance

The above model finds the regions and tells how frequently each of the regions is viewed by the different users. Given a collection of tasks it is sufficient for extracting the task-relevance for each of the tasks. However, it does not generalize to new images. 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.

Assuming we know the true relevances for a training set, this is equivalent to the task performed for full images in Task 2.1. We know a set of regions the users could potentially have viewed, and we can infer a binary relevance for each of those. Building on the information learned on full-image prediction, we apply a logistic regression classifier used in [16] for a set of features computed from the gaze. The features considered in this work are fairly simple, borrowed largely from Deliverable D2.1.1 and [16], modified to work for image regions. The full set of features for relevance prediction is given in Table 1. All features are normalized with z-score transformation for each user+image combination before pooling all the data for learning the predictor. This makes the features relative to that particular user and image, making dependence on personal variation and image content smaller.

The relevance is predicted with logistic regression, the input being the features and all of their second order terms to enable non-linear decision boundaries. Better prediction models could be applied, and feature importance methods studied in WP5 can be used to further refine the representations.

3.3 Experiments and results

3.3.1 Detecting target regions from gaze

First, the results are evaluated visually, by inspecting the regions extracted by the algorithm. For this purpose a tool that highlights the relevant regions, averaging over the posterior samples to model uncertainty correctly, was developed. The visualization tool is included in the demonstrator deliverable D2.2.2 described in Section 6.

The method was applied on a data set with two different tasks for the users, obtained from collaborators at the Department of Psychology, University of Turku. The users were viewing full-screen outdoor and indoor images of houses, and the users 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. The full potential of the model lies in this kind of settings with more than one task for the users.

We applied the model for a set of 22 such images, using fixation data of 13+13 viewers.

Figure 1: Visualization of task-dependent relevant regions for two tasks, house-buying evaluation (blue) and burglary potential evaluation (red). The left image shows the results of the proposed model, with clearly extracted interpretable regions such as the cell-phone 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 giving overview on the room. The right image shows gaze heatmaps for the two tasks as a comparison. It is possible to make similar kinds of observations from that illustration, but the display is considerably more noisy and less clearly task-oriented.

Detailed analysis of this experiment is still ongoing, but the model immediately reveals strong dependence between the task and the target regions (Figure 1). Regions with high θ_1 and low θ_2 correspond to valuable-looking items, while the opposite is indicative of e.g. regions on empty walls or floors (the house-buyer is checking the condition of the house). Overall, the results point towards high top-down control in image viewing gaze patterns for clearly specified tasks.

3.3.2 Predicting relevance of regions

Next we turn our attention to predicting the relevance of regions given gaze properties of users viewing that region. If this can be done with sufficient accuracy then it is possible to extract not only the areas a single user viewed but also to which degree those were because of task-relevance.

We perform the experiments on the Sports data described in deliverable of Task 8.1 [20]. 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 users share the same task. We applied the model for 90 of the images, ignoring the first 10 for each user to ignore the time spent getting used with the system.

It is fairly difficult to collect explicit ground truth labeling for parts of images, and hence we create an artificial labeling for the sports image data set 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 will also have relevant regions that need to be checked in order to determine the relevance, but is sufficient for studying to which degree the relevance can be inferred.

We used 10-fold cross-validation on the data set, always using 9/10 of the images (and all 21 users) for training and the rest for testing. In total the data consists of 5983 non-empty user+region samples. If all cases are considered independent, we get on average roughly 5400 training samples and 600 test samples, which result in area-under-curve (AUC) score of

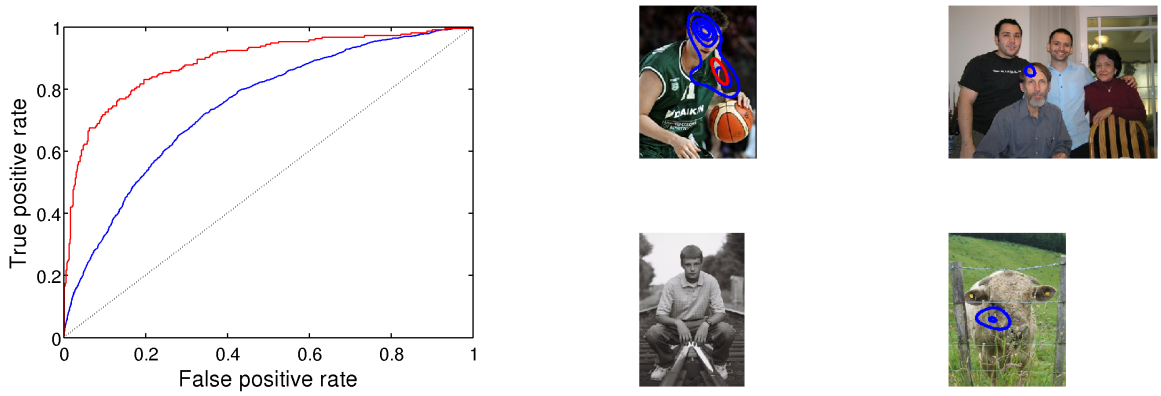

Figure 2: **Left**: ROC curves of classifying the image regions as relevant or non-relevant based on gaze-based features alone. A region is considered relevant if it was within a sports-relevant image. The blue line depicts the accuracy for predicting the relevance for each user+region separately (AUC=0.73), while the red line combines the independent predictions of different users for each region (AUC=0.88). The dashed black line indicates random classified. **Right**: An example page containing one of the top-ranked regions in the sports collection. The blue contours show the areas viewed by the users, while the red contour indicates the region predicted to be most relevant in the last posterior sample. As expected, the most relevant region contains only elements in the sports picture and not in any of the other pictures.

0.73. As we have multiple users for each image, we can also look at the combined prediction accuracy for each region separately. The predictions are still made independently for each user+region pair, but the predictions of all users viewing a particular region are combined by averaging the predictions. This increases the AUC score to 0.88. The results are illustrated in Figure 2. Overall, the prediction accuracy is reasonably high, despite the inaccurate ground truth. There is hence evidence that fixation-level statistics on the viewing patterns are informative of the relevance of the regions extracted by the model.

4 Multi-view learning approach

Multi-view learning describes how learning can be done in settings with multiple views of a problem. In the context of image relevance estimation the views can be various image content descriptions and eye-movements of users viewing the image. To make the problem more widely applicable, we assume it will not always be possible to observe all views. Therefore, the problem can be cast as a multi-view learning problem with missing data. The next figure summarises the scenario for four views where it can be seen that the training samples are complete, but the test samples contain missing data for at least one view:

	Views			
Source spaces:	$\mathcal{Z}_1$	$\mathcal{Z}_2$	$\mathcal{Z}_3$	$\mathcal{Z}_4$
	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Complete items:	$\mathbf{z}_1^1$	$\mathbf{z}_1^2$	$\mathbf{z}_1^3$	$\mathbf{z}_1^4$
(Training)	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\mathbf{z}_m^1$	$\mathbf{z}_m^2$	$\mathbf{z}_m^3$	$\mathbf{z}_m^4$
Incomplete items:	$\mathbf{z}_{m+1}^1$	$\mathbf{z}_{m+1}^2$	$\cdot$	$\mathbf{z}_{m+1}^4$
(Test)	$\cdot$	$\cdot$	$\mathbf{z}_{m+2}^3$	$\mathbf{z}_{m+2}^4$
	$\cdot$	$\mathbf{z}_{m+3}^2$	$\cdot$	$\cdot$
	$\mathbf{z}_{m+4}^1$	$\cdot$	$\cdot$	$\mathbf{z}_{m+4}^4$
	$\cdot$	$\mathbf{z}_{m+5}^2$	$\mathbf{z}_{m+5}^3$	$\cdot$
	$\cdot$	$\mathbf{z}_{m+6}^2$	$\mathbf{z}_{m+6}^3$	$\mathbf{z}_{m+6}^4$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$

The goal of the learning task is to estimate the values of the missing views for each sample. This scenario has wide application since it will often occur that data is incomplete, and it can be seen that the problem generalises classical supervised learning problems such as regression. For example, the approach could be used is face recognition when some parts (the views) of the faces are unknown, e.g. through occlusion. By using a training set of complete faces where all parts of the face are observed it is possible to learn the relationships among the parts of the faces. Consequently, on receiving a subsequent image containing occlusion of one part the technique is able to infer the missing part and reconstruct it.

In this work the views are users' eye movements and image content features. Several possible learning tasks handled by the unified framework can be derived for this setup, and hence we derive the framework in the general form and return to the practical application in the experimental section. For more application-oriented view, one can consider the following learning scenarios for the concrete case of image saliency/relevance estimation settings:

	Views					Scenarios
	Image parts			Users eye movement		
Source spaces:	$\mathcal{Z}_1$	$\mathcal{Z}_2$	$\mathcal{Z}_3$	$\mathcal{Z}_4$	$\mathcal{Z}_5$	
	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	
Complete items:	$\mathbf{z}_1^1$	$\mathbf{z}_1^2$	$\mathbf{z}_1^3$	$\mathbf{z}_1^4$	$\mathbf{z}_1^5$	
(Training)	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	$\mathbf{z}_m^1$	$\mathbf{z}_m^2$	$\mathbf{z}_m^3$	$\mathbf{z}_m^4$	$\mathbf{z}_m^5$	
Incomplete items:	$\mathbf{z}^1$	$\mathbf{z}^2$	$\mathbf{z}^3$	$\cdot$	$\cdot$	Predicting eye movements
(Test)	$\cdot$	$\cdot$	$\cdot$	$\mathbf{z}^4$	$\mathbf{z}^5$	Predicting image features

The preliminary results presented in this deliverable contain views only referring to eye movements. The features describing the images will be used in WP 5.3 (feature selection) and the results will be delivered in the corresponding deliverable.

4.1 Solution framework

In the learning framework development, we make the following two mild assumptions:

- There are a reasonably large number of observations where all views are known, that are made available to the learning procedure,
- In the incomplete observations at least one view is available, but no assumptions are made about which views are missing. However, any prior knowledge about the distribution of the missing data can be exploited to improve the estimation of their values,

We introduce a formulation which can be considered as a generalised regression problem whereby the missing values are estimated from the relationship amongst the views as well as the known views. Furthermore, assuming that the missing views of a sample item can be handled as output $\mathbf{y}$ and the known part as input $\mathbf{x}$, then we have $\mathbf{y} \Leftarrow \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is a linear operator learned from the complete data which describes the relationships between the different views. The difficulty of this kind of regression arises from the fact that the output and the input can vary among the sample items.

To solve the general missing value problem the next, a "Support Vector Machine"-style, maximum margin based optimisation problem is formulated for the regression task (see earlier application of the framework [2, 22]). The set $\mathcal{R}$ contains the indices of all views, the subsets $\mathcal{R}_X$ and $\mathcal{R}_Y$ of $\mathcal{R}$ refer to those views taken as inputs and outputs respectively, and furthermore we have $\mathcal{R}_X \cup \mathcal{R}_Y = \mathcal{R}$, $\mathcal{R}_X \cap \mathcal{R}_Y = \emptyset$. Now the optimisation problem reads as

$$\begin{array}{ll} \min & \frac{1}{2}\|\mathbf{W}\|_F^2 + C\sum_{i=1}^m \xi_i \\ \text{w.r.t.} & \mathbf{W} \text{ tensor} \in \mathcal{Z}^*,\ \boldsymbol{\xi} \in \mathbb{R}^m, \\ \text{s.t.} & \langle \underbrace{\bigotimes_{s\in\mathcal{R}_Y} \mathbf{z}_i^s}_{\text{Outputs}} , \mathbf{W} \underbrace{\bigotimes_{r\in\mathcal{R}_X} \mathbf{z}_i^r}_{\text{Inputs}} \rangle_F \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ i = 1, \ldots, m, \end{array} \tag{1}$$

where $C > 0$ is penalty constant and $\bigotimes$ denotes the tensor product.

The form is similar to the Support Vector Machine case with two notable exceptions:

- the outputs are no longer binary labels, $\{-1, +1\}$, but vectors of an arbitrary linear vector space,
- the normal vector of the separating hyperplane is reinterpreted as a linear operator projecting the inputs into the space of the outputs.

The regularisation term in the objective function forces the projections of the inputs and the outputs to be similar with respect to their inner products. When the inputs and the outputs are normalised they live on a sphere in both corresponding spaces, and then we solve a problem between spaces with structure of a Spherical rather then Euclidean geometry.

The following theorem, presented here without a proof, can illuminate the background of the use the optimisation approach displayed in (1):

Theorem 1 *For all partitions $\mathcal{R}_X, \mathcal{R}_Y$ of $\mathcal{R}$ the optimisation problem (1) is equivalent to the following one:*

$$\begin{array}{ll} \min & \frac{1}{2}\|\mathbf{W}\|_F^2 + C\sum_{i=1}^m \xi_i \\ \text{w.r.t.} & \mathbf{W} \text{ tensor} \in \mathcal{Z}^*,\ \boldsymbol{\xi} \in \mathbb{R}^m, \\ \text{s.t.} & \langle \mathbf{W}, \bigotimes_{r\in\mathcal{R}} \mathbf{z}_i^r \rangle_F \geq 1 - \xi_i,\ i = 1, \ldots, m, \\ & \xi_i \geq 0,\ i = 1, \ldots, m. \end{array} \tag{2}$$

This equivalence holds true if in every sample item the inputs and the outputs are partitioned independently.

This fact guarantees that the linear operator $\mathbf{W}$ has an universal property, namely *it is independent of how the views are grouped into inputs and outputs*, thus, it consistently characterises the underlying multi-view learning problem.

The seemingly complex problem (2) leads to a simple Lagrangian dual:

$$\begin{array}{ll} \min & \frac{1}{2}\boldsymbol{\alpha}'\big(\mathbf{K}_1 \bullet \cdots \bullet \mathbf{K}_{n_R}\big)\boldsymbol{\alpha} - \mathbf{1}'\boldsymbol{\alpha} \\ \text{w.r.t.} & \boldsymbol{\alpha} \in \mathbb{R}^m \\ \text{s.t.} & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}, \end{array} \tag{3}$$

where

$$(\mathbf{K_r})_{ij} \quad = \langle \mathbf{z_i^r}, \mathbf{z_j^r} \rangle,\ r \in \mathcal{R},\ i,j \in \{1,\ldots,m\} \tag{4}$$

are kernels for each of the views, and $\bullet$ denotes the element-wise matrix product. This dual can be solved in a straightforward way for very large scale applications[1]. After computing the dual variables the optimum solution for the universal linear operator is given by

$$\mathbf{W} = \textstyle\sum_{i=1}^m \alpha_i \bigotimes_{r\in\mathcal{R}} \mathbf{z}_i^r.$$

If $\mathbf{W}$ is given then knowing only an arbitrary subset of the views considered as inputs the complement part of the views the outputs can be estimated in the following way

$$\begin{array}{l} \big(\bigotimes_{s\in\mathcal{R}_y} \mathbf{z}^s\big) \sim \mathbf{W} \bigotimes_{r\in\mathcal{R}_X} \mathbf{z}^r \\ = \sum_{i=1}^m \alpha_i [\bigotimes_{r\in\mathcal{R}} \mathbf{z}_i^r, \bigotimes_{r\in\mathcal{R}_X} \mathbf{z}^r] \\ = \sum_{i=1}^m \alpha_i \prod_{r\in\mathcal{R}_X} \langle \mathbf{z}_i^r, \mathbf{z}^r \rangle \bigotimes_{s\in\mathcal{R}_Y} \mathbf{z}_i^s \\ = \sum_{i=1}^m \beta_i \bigotimes_{s\in\mathcal{R}_Y} \mathbf{z}_i^s, \end{array} \tag{5}$$

where

$$\beta_i = \alpha_i \textstyle\prod_{r\in\mathcal{R}_X} \langle \mathbf{z}_i^r, \mathbf{z}^r \rangle,\ i = 1\ldots,m.$$

Thus the prediction is a linear combination of the corresponding outputs.

4.2 Experiments and results

We evaluate the performance of the proposed method on a real world dataset on eye movements [13]. In real scenarios, the eye tracker sometimes temporarily loses track of the subject's eyes due to the subject moving, etc. Furthermore, the data collection is expensive and laborious. Thus, we try to learn to predict which regions humans look based on available data. For focused search tasks with strong top-down control this corresponds also with the relevance.

In our experiment six users are shown a set of one thousand images (one image per page), hence, there are six views for an image in the learning problem. A view of each user's scanpath is represented by a heatmap of their eye movements assuming that *the higher density of eye movement on a part of an image implies the higher probability of the importance of that part*. We use 5-fold cross validation to evaluate the task and assume that at least one view, eye movement of a user, is available in the testing phase. A linear kernel function is used in the experiment. We evaluated the result on ten different training samples with sizes of the range $50, 100, \ldots, 500$, and on five different scenarios where the number of missing views was fixed to $1, \ldots, 5$ of the six given views. The performance is measured by the correlation and by the Kullback-Leibler (KL) divergence between the real heatmap of the test items and their prediction. The measures are

$$\text{corr} = \frac{\sum_{i\in\mathcal{S}_T, r\in\mathcal{R}_Y} \text{corr}(\mathbf{z}_i^r, \tilde{\mathbf{z}}_i^r)}{\sum_{i\in\mathcal{S}_T, r\in\mathcal{R}_Y} \text{corr}(\mathbf{z}_i^r, \mathbf{u}_i^r)}, \quad \text{KL} = \frac{\sum_{i\in\mathcal{S}_T, r\in\mathcal{R}_Y} \text{KL}(\mathbf{z}_i^r, \tilde{\mathbf{z}}_i^r)}{\sum_{i\in\mathcal{S}_T, r\in\mathcal{R}_Y} \text{KL}(\mathbf{z}_i^r, \mathbf{u}_i^r)},$$

[1] The web site of the authors provides an open source implementation to this problem

where $\mathcal{S}_T$ is the set of test items, $\mathbf{z}_i^r$ the real heatmap, $\tilde{\mathbf{z}}_i^r$ the predicted heatmap by the proposed method, $\mathbf{u}_i^r$ the random base line prediction in test item i for user r. Figure 3 illustrates the gain in these measures relative to the uniform random baseline predictions. The numerical results are complemented by illustrative examples in Figure 4.

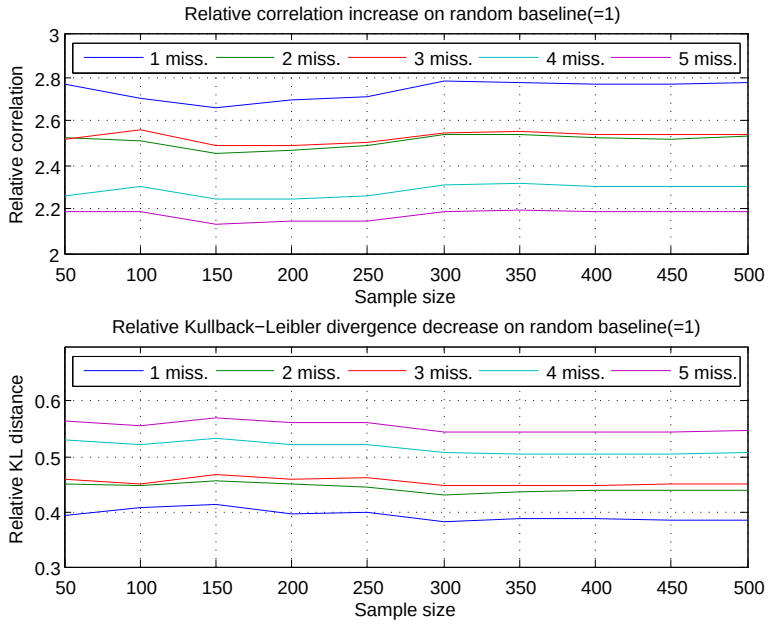

Figure 3: Illustration of the gain of the proposed method compared to the random baseline. The top figure measures the improvement with increase in correlation, and the bottom one with decrease in Kullback-Leibler divergence between the real and predicted heatmaps.

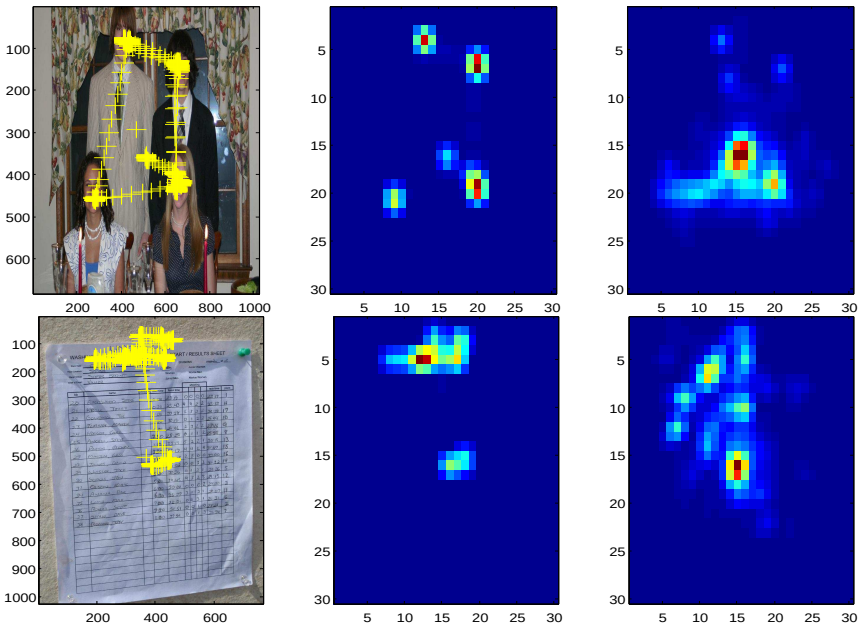

Figure 4: Illustrations of the prediction result. Left: Image + eye movement, Middle: real heatmap, Right: predicted heatmap

5 Learning salient regions

In this part we assume that when guided by a task the gaze is quickly attracted by the relevant region. That is, the scan path is primarily controlled by top-down processing. Hence, the relevance factor ψ_t can be estimated as a function of the proximity of $\mathbf{x}_t$ to gaze points (i_s, j_s). As suggested in [3] we model this function with an isotropic Gaussian, which is a plausible model for the foveation in human vision:

$$\psi(z) = \sum_{s=1}^{S} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(m-i_s)^2 + (n-j_s)^2}{2\sigma^2}\right) . \tag{6}$$

where $\psi(z)$ is the relevance of a pixel position $z = (m, n)$ and S denotes the total number of fixations. Note that a visualization of $\psi(z)$ over all pixels of the image is the heat-map of the gaze of a single user (see examples in the second line of Figure 6), and that a normalized heat map can be thought as the probability density function of the gaze positions in an image.

We have shown in PinView Deliverables 6.1 [5] and 6.2.1 [1] that using these heat maps in a weighted Fisher Kernel framework allow us to improve the categorization accuracy. However, gaze tracking systems are not yet ubiquitous and getting eye fixation data from a large set of images is impractical. In the absence of visual attention data obtained from gaze measurements, we can resort to automatic methods such as bottom-up and top-down saliency maps.

Therefore, in this section, we present a method [17] that aims to predict gaze heat-maps for new images without using the gaze tracker, by assuming that a set of images are available for which we have corresponding gaze trajectories (preferably obtained from several users) and hence gaze heat-maps. It is a learning based approach based on a simple principle: images sharing their global visual appearances are likely to share similar visual saliencies. If the users are assumed to share the same task, then this produces a top-down saliency map for that particular task.

The main steps of this approach, as illustrated in Figure 5, are (see more details in [17]):

- First we build (offline) a visual vocabulary [4] using a Gaussian mixture model $\lambda = \{w_i, \mu_i, \sigma_i, i = 1...N\}$:

$$p(x_t|\lambda) = \sum_{i=1}^{N} w_i p_i(x_t|\lambda) = \sum_{i=1}^{N} w_i \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \tag{7}$$

  where x_t are low level descriptors (in our case orientation histograms and local color statistics) associated with local patches extracted from images.

- Then we compute (similarly offline) image signatures for all images in the annotated[2] database using the Fisher kernel framework. The main idea described in more details in [19, 5] is to consider the gradient of the log-likelihood of (7) at each patch according to the parameters of the GMM :

$$\mathbf{f}_t = \nabla_\lambda \log p(x_t|\lambda) = \nabla_\lambda \sum_{i=1}^{N} w_i p_i(x_t|\lambda) \tag{8}$$

  Assuming independence between samples we can write that the Fisher Vector of the set of descriptors $X = \{x_t, t = 1...T\}$ (e.g. features extracted from an image) is the sum of the individual Fisher vectors:

$$\mathbf{f}_X = \sum_{t=1}^{T} \mathbf{f}_t \tag{9}$$

[2]Here annotated means that a gaze heat map is available for the image

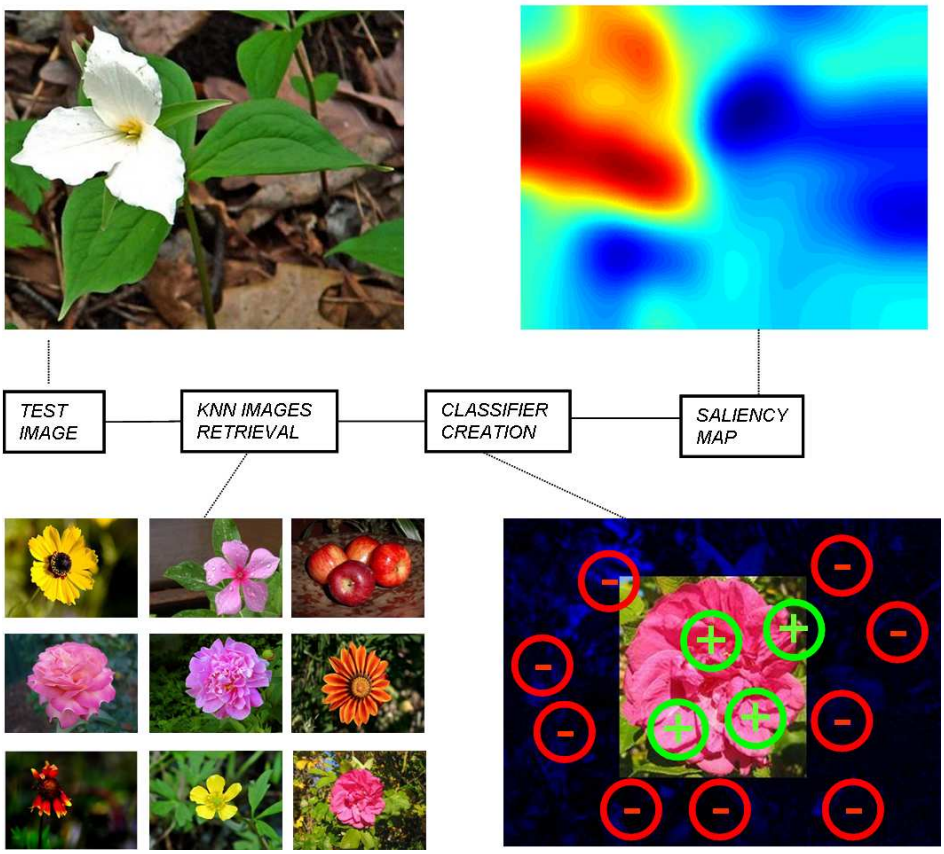

Figure 5: A schema illustrating the learning saliency from labeled nearest neighbors approach.

- Given a new image and using these image signatures, we can retrieve online the most similar images using the normalized L_1 similarity measure between Fisher vectors:

$$sim(I,J) = -||\hat{\mathbf{f}}_I - \hat{\mathbf{f}}_J||_1 = -\sum_i |\hat{f}_I^i - \hat{f}_J^i| \tag{10}$$

  where $\hat{\mathbf{f}}_\mathbf{I}$ is the Fisher vector $\mathbf{f}_I$ of image I normalized to norm L_1 equal to 1 ($||\hat{\mathbf{f}}||_1 = 1$).

- Considering the top K most similar images, we compute for each of them 2 Fisher vectors splitting (9) into two parts: one considering only salient patches (those having a high average score in the gaze heat map) and one for the rest of the patches. We denote these Fisher vectors by $\mathbf{f}_{I_j^+}$ and $\mathbf{f}_{I_j^-}$ for the image I_j. Then we sum all positive respectively negative Fisher vectors associated to the K images for salient and non salient regions:

$$\mathbf{f}_{FG} = \sum_{j=1}^{K} \mathbf{f}_{X_j^+} \quad \text{and} \quad \mathbf{f}_{BG} = \sum_{j=1}^{K} \mathbf{f}_{X_j^-} \tag{11}$$

  and we call them (abusively) "foreground" and "background" Fisher models.

- Then for each patch x_t in the test image we compute a relevance score as follows:

$$s(x_t) = ||\hat{\mathbf{f}}_{x_t} - \hat{\mathbf{f}}_{FG}||_1 - ||\hat{\mathbf{f}}_{x_t} - \hat{\mathbf{f}}_{BG}||_1 \tag{12}$$

  However, generally we have overlapping patches. Therefore to compute the pixel scores $s(z)$ (in order to build a heat map) a weighted average of patch scores is computed as

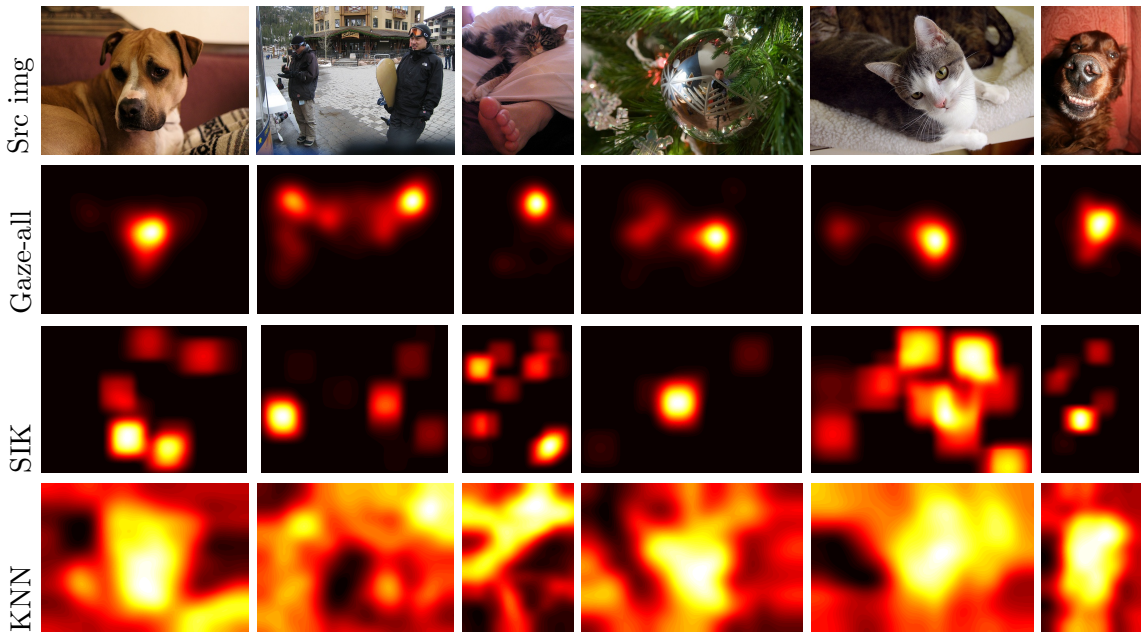

Figure 6: Saliency maps computed for some sample images using gaze trajectories (ground truth), smoothed Itti and Koch maps and the proposed method.

follows:

$$s(z) = \frac{\sum_{t=1}^{T} \omega_{t,z} s(x_t)}{\sum_{t=1}^{T} \omega_{t,z}}. \tag{13}$$

The weights $\omega_{t,z}$ are given by a Gaussian Kernel $\phi_{t,z} = \mathrm{N}(z|m_t, C_t)$, where m_t is the geometrical center of patch x_t and C_t is the isotropic covariance matrix with values $(\alpha s_t)^2$ on the diagonal with s_t being the size of patch x_t. The scaling parameter α is 0.6 in our experiments. Note that the isotropic covariance assumption corresponds to round patches, but elliptic patches can be easily accommodated by considering non-isotropic covariances C_t.

5.1 Experiments and results

For our experiments we used the Cats&Dogs dataset described in Pinview Deliverables 8.3 [20] and 6.1 [5]. It contains 262 images, including 105 images of cats, 105 images of dogs and 52 images of neither cats nor dogs. The gaze trajectories were obtained from 28 volunteers with an average of 147 gaze measurements per image (for full details please refer to the Deliverable 6.1 [5]). From these measurements we built the ground truth gaze heat maps using equation 6 with $\sigma = 5.6\%$ of the image height.

In order to evaluate the estimated gaze maps with the proposed method we used a 10-fold cross validation scheme. Each time we considered a fold as a test, we used the rest of the images with their ground truth gaze heat maps for training. From this subset we gathered for each image the K=20 most similar images and estimate the heat map as described above.

Figure 6 shows a few qualitative results. These images show that while the "hottest" positions are relatively well estimated, the method does have a tendency to over-estimate the relevant regions.

In order to get a quantitative evaluation we used two measures. The first is the intersection/union measure (used to measure the segmentation accuracy in Pascal VOC Challenge [7]) $V = \frac{TP}{TP+FP+FN}$ and the $F_\alpha = \frac{(1+\alpha)\,\mathrm{Pr} * Re}{\alpha * Pr + Re}$ measure with $\alpha = 0.5$ (to favour precision

over recall). To be able to compute these measures, we first need to binarize the heat maps. The best results (V=21%, $F_{0.5}$= 30%) of accuracy with the were obtained with a high threshold (0.75) when binarizing the estimated maps (the values in the estimated maps were first normalized to have values between 0 and 1) which reduces significantly the over-estimated region.

We compared this with smoothed maps obtained with the method proposed in [12] and smoothed with isotropic Gaussian where σ is 8% of the image size. The results with the best threshold are V=11%, $F_{0.5}$= 16% which shows that our method significantly outperforms the saliency maps obtained with the Itti & Koch's method. However, it needs to be remembered that our gaze trajectories and hence the gaze maps were explicitly controlled by the task (looking for cats and dogs), while the aim of Itti's maps is to estimate the free-viewing low-level saliency.

Nevertheless, the results are encouraging because the show that eye track information really helps (as we have shown in the deliverables 6.1 [5] and 6.2.1 [1]) and therefore there is a benefit to using this type of data. We have seen also that when applied to visual saliency with a bigger training datasets obtained with explicit feedback [17] we obtain much higher accuracies in estimation. We fully expect the same trend for gaze heat maps if more eye-tracking data is used.

learnForImage()	Draws a collection of posterior samples from a model learned for the fixation data of several users.
summarizePosterior()	Summarizes a subset of posterior samples for the image, creating as an output an estimate for viewing density for each of the tasks.
createHeatmap()	Creates standard gaze heatmaps from the raw fixation data for comparison.
collectFeatures()	Finds the feature representation for the regions viewed by a given user, to be used for predicting the relevance of each of those.
example1	Demonstrates the estimation of task-relevance for a set of users viewing the same image
example2	Demonstrates inferring task-relevance for a single user viewing a new image

Table 2: Overview of the functionality in Deliverable 2.2.2 *Demonstrator for predicting relevance of image parts*.

6 Demonstrator for predicting relevance of image parts

The Deliverable 2.2.2 *Demonstrator for predicting relevance of image parts* implements the model presented in Section 3. It is provided as a Matlab package, included as an Annex of this report, that can be used for two different tasks. First, given an input image and a set of eye fixation data matrices for users viewing the image with given tasks, it detects regions considered relevant for each of the tasks. The result can be analysed e.g. with illustrations like the one in Figure 1. Second, the package has functionality for predicting task-relevant regions given a single user with unknown task.

Table 2 shows a brief overview of the interface of the package, listing the names and purposes of the functions. More details, such as the list of parameters given for the functions, are described in the package. The package also contains a small fixation data set for testing the functionality.

7 Conclusions

The Task 2.2 studied prediction of relevance of sub-images in search tasks. Three complementary novel computational models were developed to address various questions within the general scope. Since the degree of top-down versus bottom-up control in image search tasks was not known before the task, different kind of approaches were tried. Two of the models assume strong top-down control and learn gaze targets for new users, while one of the models can be used to study the degree of top-down vs bottom-up saliency, as well as to infer which attention targets are relevant. The computational models also span a range of different kind of machine learning approaches, including both Bayesian generative modeling and advanced kernel methods.

The main results of the work can be summarized as:

1. There is strong top-down control in image search tasks, which indicates that pure detection of gaze targets is already highly useful in determining the relevant sub-images.

2. Low-level gaze features will still help in predicting the relevance of a region viewed by the user, which means information like viewing time and temporal pattern are indicative of relevance.

3. Gaze and low-level features can be combined in multi-view setting to improve prediction of gaze targets, and it is possible to build top-down saliency models based on image content features.

The work done in this task was largely fundamental basic research on a difficult open problem of how images are viewed under search tasks. Still, the results have clear implications for the PinView project in general. First, the strong top-down control means that gaze-targets can directly be used as crude estimates of sub-image relevance, and the accuracy provided by such methods may be sufficient for the final system. Second, the promising results on top-down saliency models indicate that it is possible to utilize gaze-based information even when not assuming gaze-tracking for actual use scenarios. Third, it was found out that in order to focus on sub-image parts the images need to be fairly large. On traditional thumbnail displays one or two nearby fixations are typically enough for near-complete understanding of the image. Hence, to fully utilize sub-image relevance more advanced interfaces would be needed, such as displays that interactively present larger versions of potentially interesting images.

The results also provide a good starting point for Task 2.3 *Data fusion* of WP2, starting M25. The multi-view approach is already a crude data fusion method, and work on similar models can be continued. Data fusion is in general expected to be more useful when moving away from highly controlled tasks where gaze direction alone is a strong indication of relevance, which opens up possibilities for fundamental research on relative importance of top-down and bottom-up saliency in different kinds of search setups.

Acknowledgements

We thank Dr. Johanna Kaakinen for providing the data set used in the experiments in Section 3 and Dr. Craig Saunders for his comments on the draft of this report.

References

[1] H. Ali, M. Antenreiter, G. Csurka P. Auer, T. de Campos, Z. Hussain, J. Laaksonen, R. Ortner, K. Pasupa, F. Perronnin, C. Saunders, J. Shawe-Taylor, and V. Viitaniemi. Description, analysis and evaluation of confidence estimation procedures for sub-categorisation. Technical report, PinView, April, 9th 2009. Deliverable 6.2.1 available from http://www.pinview.eu/.

[2] K. Astikainen, L. Holm, E. Pitkänen, S. Szedmak, and J. Rousu. Towards structured output prediction of enzyme function. In *BMC Proceedings, 2(Suppl 4):S2*. 2008.

[3] Ee-chien Chang, Stéphane Mallat, and Chee Yap. Wavelet foveation. *Journal of Applied and Computational Harmonic Analysis*, 9:312–335, 2000.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.

[5] Teofilo de Campos, Florent Perronnin, Ville Viitaniemi, Jorma Laaksonen, and Marco Bressan. Description and evaluation of novel local features with usable sub-categorisation performance. Technical report, PinView, October, 1st 2008. Deliverable 6.1, available from http://www.pinview.eu/.

[6] 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*, 8:1–19, 2008.

[7] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool. The pascal visual object classes challenge. http://pascallin.ecs.soton.ac.uk/challenges/VOC/.

[8] Zoubin Ghahramani. Factorial learning and the EM algorithm. In *Advances in Neural Information Processing Systems 7*, 1995.

[9] Tom Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, 2006.

[10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, 2006.

[11] J. Henderson, J. R. Brockmole, M. S. Castelhano, 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*, pages 538–562. Elsevier, 2007.

[12] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 2000.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where people look. In *International Conference on Computer Vision, ICCV*. 2009. http://people.csail.mit.edu/tjudd/wherepeoplelook.html.

[14] 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*, pages 689–698. MIT Press, 2007.

[15] Arto Klami, Craig Saunders, Teófilo de Campos, and Samuel Kaski. Can relevance of images be inferred from eye movements? In *MIR '08: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 134–140. ACM, New York, NY, USA, 2008.

[16] László Kozma, Arto Klami, and Samuel Kaski. GaZIR: Gaze-based zooming interface for image retrieval. In *Proc. ICMI-MLMI 2009, The Eleventh International Conference on Multimodal Interfaces and The Sixth Workshop on Machine Learning for Multimodal Interaction*, pages 305–312, New York, NY, USA, 2009. ACM.

[17] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, 2009.

[18] Kitsuchart Pasupa, Craig Saunders, Sandor Szedmak, Arto Klami, Samuel Kaski, and Steve Gunn. Learning to rank images from eye movements. In *Proceedings of ICCV'2009 Workshop on Human-Computer Interaction (HCI'2009)*, pages 2009–2016, 2009.

[19] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[20] Craig Saunders and Arto Klami. Database of eye-movement recordings. Technical report, PinView European Comunity project FP7-216529, July 01 2008. D8.3, available at http://www.pinview.eu/.

[21] O. Spakov and D. Miniotas. Visualization of eye gaze data using heat maps. *Electronics and electrical engineering*, 2(74), 2007.

[22] S. Szedmak, T. De Bie, and D.R. Hardoon. A metamorphosis of canonical correlation analysis into multivariate maximum margin learning. In *ESANN Proceedings, Brugge*. 2007.

[23] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 12 2008.