Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling

Hosnieh Sattar[1,2] Andreas Bulling[1] Mario Fritz[2]
[1]Perceptual User Interfaces Group, [2]Scalable Learning and Perception Group
Max Planck Institute for Informatics, Saarbrücken, Germany
{sattar,mfritz,bulling}@mpi-inf.mpg.de

Abstract

Predicting the target of visual search from eye fixation (gaze) data is a challenging problem with many applications in human-computer interaction. In contrast to previous work that has focused on individual instances as search target, we propose the first approach to predict categories and attributes of search targets based on gaze data. However, state of the art models for categorical recognition in general require large amounts of training data, which is prohibitive for gaze data. To address this challenge, we propose a novel Gaze Pooling Layer that integrates gaze information into CNN-based architectures as an attention mechanism – incorporating both spatial and temporal aspects of human gaze behavior. We show that our approach is effective even when the gaze pooling layer is added to an already trained CNN, thus eliminating the need for expensive joint data collection of visual and gaze data. We propose an experimental setup and data set and demonstrate the effectiveness of our method for search target prediction based on gaze behavior. We further study how to integrate temporal and spatial gaze information most effectively, and indicate directions for future research in gaze-based prediction of mental states.

1. Introduction

As eye tracking technology is beginning to mature, there is an increasing interest in exploring the type of information that can be extracted from human gaze data. Within the wider scope of eye-based activity recognition [4, 25], search target prediction [2, 23, 33] has recently received particular attention as it aims to recognise users' search intends without the need for them to verbally communicate these intends. Previous work on search target prediction from gaze data (e.g. [2, 23]) is limited to specific target instances that users searched for, e.g. a particular object. This excludes searches for broader classes of objects that share the same semantic category or certain object attributes. Such searches commonly occur if the user does not have a concrete target instance in mind but is only looking for an object from a certain category or with certain characteristic attributes.

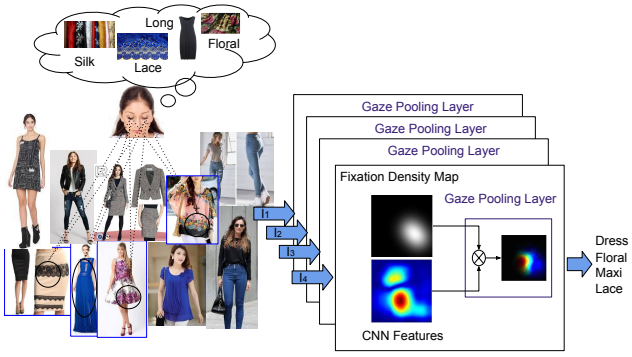

Figure 1. We propose a method to predict the target of visual search in terms of categories and attributes from users' gaze. We propose a *Gaze Pooling Layer* that leverages gaze data as an attention mechanism in a trained CNN architecture.

To address these limitations, we broaden the scope of search target prediction to categorical classes, such as object categories or attributes. One key difficulty towards achieving this goal is acquiring sufficient training data. We have to recall that object categorization only in the past decade has seen a breakthrough in performance by combining deep learning techniques with large training corpora. Collecting such large corpora is prohibitive for human gaze data, which poses a severe challenge to achieve our goal.

Therefore, we propose an approach for predicting categories and attributes of search targets that utilizes readily trained CNN architectures and combines them with gaze data in a novel *Gaze Pooling Layer* (see Figure 1). The gaze information is used as an attention mechanism that acts selectively on the visual features to predict users' search target. These design choices make our approach compatible and practical with current deep learning architectures.

Through extensive experiments we show that our method achieves accurate search target prediction for 10 category

and 10 attribute tasks on a new gaze data set that is based on the DeepFashion data set [17]. Furthermore, we evaluate different parameter settings and design choices of our approach, visualize internal representations and perform a robustness study w.r.t. noise in the eye tracking data. All code and data will be made publicly available upon acceptance .

2. Related Work

Predicting the target of visual search is a task studied both in computer vision [1, 2, 8, 23, 32, 33] and human perception [7, 5, 15, 21]. Existing approaches vary in the granularity of the predictions, either focusing on predicting specific object instances [2, 23] or operating at the coarser level and predicting target categories [8, 33]. The type of user feedback varies as well. While [2, 23, 33] solely use implicit information obtained from human gaze, [1, 8, 32] require the user to provide explicit relevance feedback. In the following we summarize previous works on gaze-supported computer vision, user feedback for image search and retrieval, as well as methods for search target prediction.

Gaze-Supported Computer Vision. Visual fixations have been used in [16, 30] to indicate object locations in the context of saliency predictions, and in [11, 22, 24] as a form of weak supervision for training of object detectors. Gaze information has been used to analyze pose estimation tasks in [18, 26] as well as for action detection [19]. Gaze data has also been employed for active segmentation [20], localizing important objects in egocentric videos [6, 28], image captioning and scene understanding [27], as well as zero-shot image classification [10]. While our work also combines visual representations (CNN) with gaze data, our task is user centric as we aim to predict search targets of the user and not aim for a computer vision task that is inherent in the image itself.

User Feedback for Image Search and Retrieval. To close the semantic gap between user's envisioned search target and the images retrieved by search engines, Ferecatu and Geman [8] proposed a framework to discover the semantic category of user's mental image in unstructured data via explicit user input. Kovashka et al. [1] introduced a novel explicit feedback method to assess the mental models of users. Most recently Yu et al. [32] proposed to use free-hand human sketches as queries to perform instance-level retrieval of images. They considered these sketches to be manifestations of users' mental model of the target. The common theme in these approaches is that they require explicit user input as part of their search refinement loop. Mouse clicks were used as input in [8]. [1] used a set of attributes and required users to operate on a large attribute vocabulary to describe their mental images. In [32] the feedback was provided by sketching the target to convey concepts such as texture, color, material, and style, which is a non-trivial step for most users. In contrast, in our work, we do not rely on a feedback loop as in [1] or explicit user input or some form of initial description of a target as in [1, 8, 32]. We instead use fixation information that can be acquired implicitly during the search task itself, and demonstrate that such information allows us to predict categories as well as attributes of search targets in a single search session.

Visual Search Target Prediction. Human gaze behavior reflects cognitive processes of the mind, such as intentions [3, 12, 14], and is influenced by the user's task [31]. In the context of visual search, previous work typically focused on predicting targets corresponding to specific object instances [2, 23, 33]. For example, users were required to search for specific book covers [23] or specific binary patterns [2] among other distracting objects. In contrast, in this work we aim to infer the general properties of a search target represented by the object's category and attributes. In this scenario, the search task is guided by the mental model that the user has of the object class rather than a specific instance of an object [8, 29]. This presents additional challenges as mental models might differ substantially among subjects. Furthermore, [2, 23, 33] required gaze data for training, whereas our approach can be pre-trained on visual data alone, and then combined with gaze data at test time.

3. Data Collection

No existing data set provides image and gaze data that is suitable for our search target prediction task. We therefore collected our own gaze data set based on the DeepFashion data set [17]. DeepFashion is a clothes data set consisting of 289,222 images annotated with 46 different categories and 1,000 attributes. We used the top 10 categories and attributes in our data collection. The training set of DeepFashion was used to train our CNN image model for clothes category and attribute prediction; the validation set was used to train participants for each category and attribute (see below). Finally, the test set was used to build up image collages for which we recorded human gaze data of participants while searching for specific categories and attributes. In the following, we describe our data collection in more detail.

3.1. Participants and Apparatus

We collected data from 14 participants (six females), aged between 18 and 30 years and with different nationalities. All of them had normal or corrected-to-normal vision. For gaze data collection we used a stationary Tobii TX300 eye tracker that provides binocular gaze data at a sampling frequency of 300Hz. We calibrated the eye tracker using a standard 9-point calibration, followed by a validation of eye tracker accuracy. For gaze data processing we used the Tobii software with the parameters for fixation detection left at their defaults (fixation duration: 60ms, maximum time between

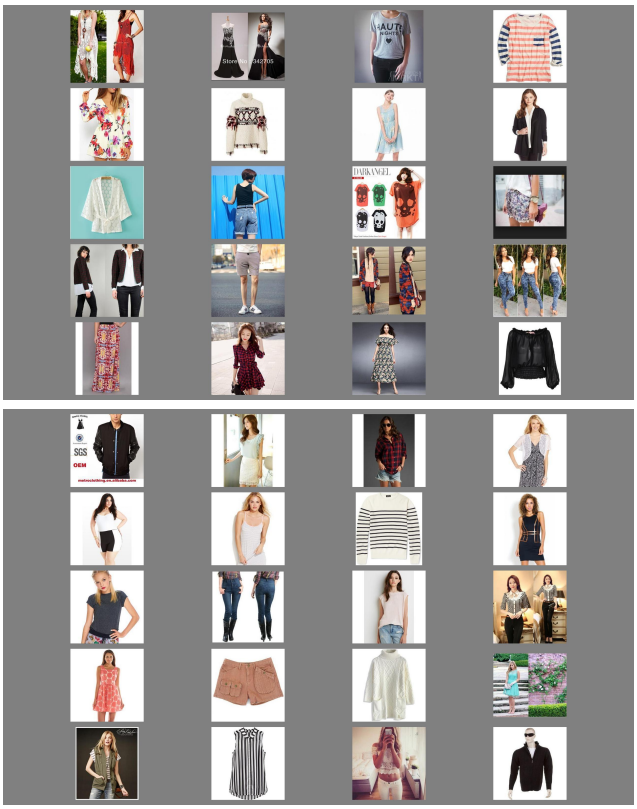

Figure 2. Sample image collages used for data collection: Attributes (top), Categories (bottom). Participants were asked to find different clothing attributes and categories within these collages.

fixations: 75ms). Image collages were shown on a 30-inch screen with a resolution of 2560x1600 pixels.

3.2. Procedure

We first trained participants by showing them exemplar images of all categories and attributes in a game like session to familiarize themselves with the categories and attributes. We did not collect any gaze data at this stage. For each category and attribute we then generated 10 image collages, each containing 20 images. Each target category or attribute appeared twice in each collage at a random location (see Figure 2 for an example). Participants were then asked to search for ten different categories and attributes on these image collages (see Figure 2) while their gaze was being tracked. We stress again that we did not show participants a specific target instance of a category or attribute that they should search for. Instead, we only instructed them to find a matching image from a certain category, i.e "dress", or with a certain attribute, i.e "floral". Consequently, search session guided by the mental image of participants from the specific category or attributes. Participants had a maximum of 10 seconds to find the asked target category or attribute in the collage that was shown full-screen. As soon as participants found a matching target, they were asked to press a key. Afterward they were asked whether they had found a matching target and how difficult the search had been. This procedure was repeated ten times for ten different categories or attributes, resulting in a total of 100 search tasks.

4. Prediction of Search Targets Using Gaze

In this work, we are interested in predicting the category and attributes of search targets from gaze data. We address this task by introducing the Gaze Pooling Layer (GPL) that combines CNN architectures with gaze data in a weighting mechanism. Figure 3 gives an overview of our approach. In the following, we describe the four major components of our method in detail: The image encoder, human gaze encoding, the Gaze Pooling Layer, and search target prediction. Finally, we also discuss different integration schemes across multiple images that allow us to utilize gaze information obtained from collages. As a mean of inspecting the internal representation of our Gaze Pooling Layer, we propose Attended Class Activation Maps (ACAM).

4.1. Image Encoder

We build on the recent success of deep learning and use a convolutional neural network (CNN) to encode image information [9, 13]. Given a raw image I, a CNN is used to extract image feature map $F(I)$.

$$F(I) = \textbf{CNN}(I) \tag{1}$$

The end-to-end training properties of these networks allows us to obtain domain-specific features. In our case, the network will be trained with data and labels relevant to the fashion domain. As we are interested in combining spatial gaze features with the image features, we use features $F(I)$ of the last convolutional layer that still has a spatial resolution. This results in a task-dependent representation with spatial resolution. In addition, to gain a higher spatial resolution we used same architecture as describe in [35]. We use their VGGnet-based model where layers after `conv5-3` are removed to gain a resolution of 14×14.

4.2. Human Gaze Encoding

Given a target category or attributes, participant $P \in \mathbb{P}$ look at image I and performs fixations $G(I,P) = (x_i, y_i), i = 1, ..., N$ in screen coordinates. We aggregate these fixations into fixation density maps $FDM(G)$ that capture the spatial density of fixations over the full image. Therefore, we represent the fixation density map $FDM(g)$ for a single fixation $g \in G(I, P)$ by a Gaussian:

$$FDM(g) = \mathcal{N}(g, \sigma_{\text{fix}}), \tag{2}$$

centered at the coordinates of the fixation, with a fixed standard deviation σ_{fix} – the only parameter of our representation.

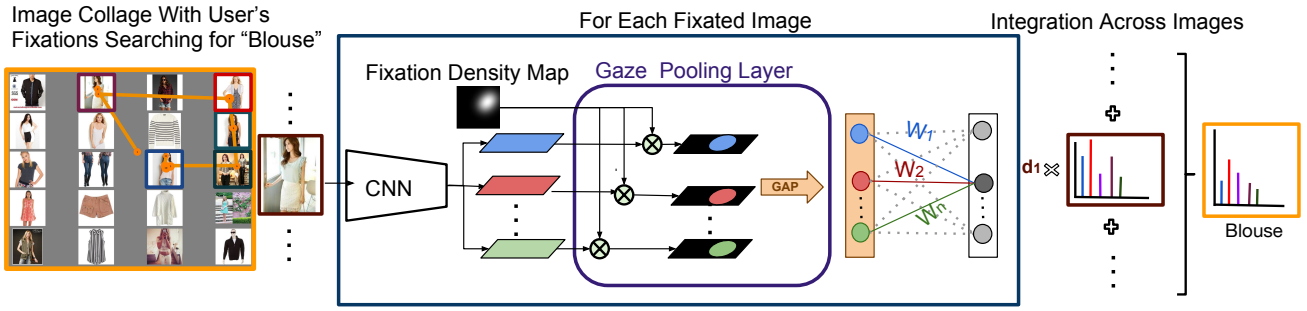

Figure 3. Overview of our approach. Given a search task (e.g. "Find a blouse"), participants fixate on multiple images in an image collage. Each fixated image is encoded into multiple spatial features using a pre-trained CNN. The proposed Gaze Pooling Layer combines visual features and fixation density maps in a feature-weighting scheme. The output is a prediction of the category or attributes of the search target. To obtain one final prediction over image collages, we integrate the class posteriors across all fixated images using average pooling.

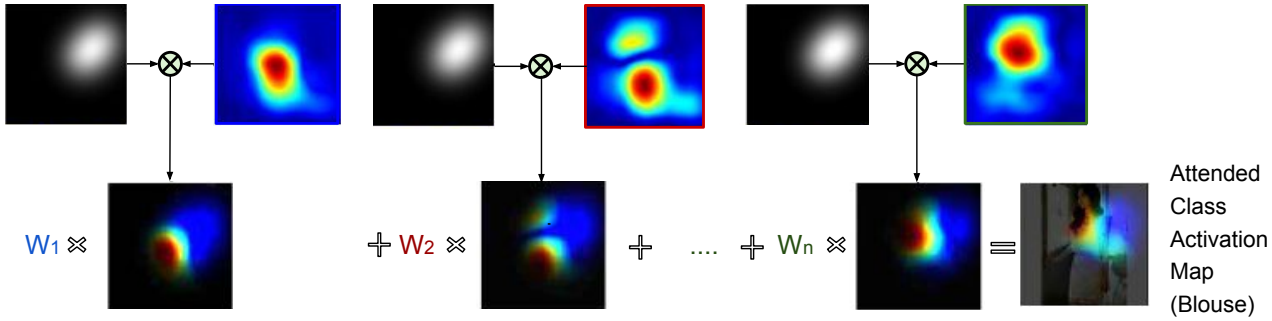

Figure 4. The proposed Gaze Pooling Layer combines fixation density maps with CNN feature maps via a spatial re-weighting (top row). Attended class activation maps are shown in the bottom row, which the predicted class scores are mapped back to the previous convolutional layer. The attended class activation maps highlight the class-specific discriminative image regions.

The fixation density map for all fixations $FDM(G)$ is obtained by coordinate-wise summation:

$$FDM(G) = \sum_{g \in G} FDM(g) \tag{3}$$

This corresponds to an average pooling integration. We also propose a max pooling version as follows:

$$FDM(G) = \max_{g \in G} FDM(g) \tag{4}$$

4.3. Gaze Pooling Layer

We combine the visual features $F(I)$ with fixation density map $FDM(G)$ in a Gaze Pooling Layer. The integration is performed by element-wise multiplication between both to obtain a gaze-weighted feature map (GWFM)

$$GWFM(I, G) = F(I) \otimes FDM(G). \tag{5}$$

In spirit of [35], we then perform Global Average Pooling (GAP) on each feature channel separately in order to yield a vector-valued feature representation.

$$GAP_{\text{GWFM}}(I, G) = \sum_{x,y} GWFM(I, G) \tag{6}$$

We finish our pipeline by classification with a fully connected layer and a soft-max layer.

$$p(C|I, G) = softmax(W\ GAP_{\text{GWFM}}(I, G) + b), \tag{7}$$

where W are the learned weights and b is the bias and C are the considered classes. The classes represent either categories or attributes depending on the experiment and we decide for the class with the highest class posterior (see Figure 3).

4.4. Integration Across Images

In our study, a stimulus is a collage with a set of images $I_i \in \mathbb{I}$. During the search task, participants fixate on multiple images in the collage, which generates fixations $G_i \in \mathbb{G}$ for each image I_i. Hence, we need a mechanism to aggregate information across images. To do this, we propose a weighted average scheme of the computed posteriors per image:

$$p(C|\mathbb{I}, \mathbb{G}) = \sum_i \frac{d_j}{\sum_j d_j} p(C|I_i, G_i). \tag{8}$$

We consider for the weights d_i the total fixations duration on image I_i as well as fixed d_i (see Figure 3). The latter

corresponds to plain averaging.

4.5. Attended Class Activation Mapping

In order to inspect the internal representation of our Gaze Pooling Layer, we propose the attended class activation map visualization. It highlights discriminative image regions for a hypothesized search target based on CNN features combined with the weights from the gaze data. In this vein, it shares similarities to the CAM of [35] but incorporates the gaze information as attention scheme. The key idea is to delay the average pooling, which allows us to show spatial maps as also illustrated in Figure 3. In more detail, our network consists of several convolutional layers which the features of last convolutional layer is weighted by our fixation density map (GWFM). We do global average pooling over the GWFM and use those features for a fully connected layer to get the user attended categories or attributes. Given that our features maps are weighted by gaze data of users, it represents their attended classes. We can identify the importance of the image region for attended categories by projecting back the weights of the output layer onto a gaze-weighted convolutional feature map, which we call Attended Class Activation Map (ACAM):

$$ACAM_c(x,y) = \sum_k w_k^c \, GWFM_k(I,G) \qquad (9)$$

where w_k^c indicates the importance $GWFM_k(I,G)$ of unit k for class c. The procedure for generating the class activation map are shown in Figure 4.

4.6. Implementations Details

In order to obtain the CNN features maps, we follow [35] and build on the recent VGGnet-GAP model. For our categorization experiments, we fine-tune on a 10 class classification problem on the DeepFashion data set [17]. For attribute prediction, we fine-tune a model with 10 times 2-way classification in the final layer. We perform a validation of the VGGnet image classification performance model in the same setting as [17] and obtained comparable results ($\pm 5\%$) for category and attribute classification. To ensure that the images and collages are not informative of the category or attribute search tasks, we have performed a sanity check by using only the CNN prediction on the images of our collages. The resulting performance is at chance level, which validates our setup as search task information cannot be derived from the images or collages and therefore can only come from the gaze data.

5. Experiments

To evaluate our method for search target prediction of categories and attributes, we performed a series of experiments.

Global vs.		——— Category ———			Attribute
Local	◔	Top1	Top2	Top3	Accuracy
Global		31%±5	48% ±8	62% ±8	20%±1
Local		49%±7	68%±6	78%±6	26 %±1
Global	✓	52%±6	68%±6	78%±6	25%±1
Local	✓	**57%±8**	**74%±7**	**84%±4**	**34%±1**

Table 1. Evaluation of global vs. local gaze pooling with and without weighting based on the fixation duration ◔.

We first evaluated the effectiveness of our Gaze Pooling Layer, the effect of using a local vs global representation, and of using a weighting by fixation duration. We then evaluated the gaze encoding that encompasses the pooling scheme of the individual fixation as well as the σ_{fix} parameter to represent a single fixation. Finally, we evaluated the robustness of our method to noise in the eye tracking data, which sheds light on different possible deployment scenarios and hardware that our approach is amendable to. Additionally, we provide visualization of the internal representations in the Gaze Pooling Layer. Across the results, we present Top-N accuracies denoting correct predictions if the correct answer is among the top N predictions.

5.1. Evaluation of the Gaze Pooling Layer

Fixation information enters our method in two places: The fixation density maps in the Gaze Pooling Layer(subsection 4.3) as well as the weighted average across images in the form of fixation duration (see subsection 4.4 and Figure 3). In order to evaluate the effectiveness of our Gaze Pooling Layer, we evaluate two conditions: "*local*" makes full use of the gaze data and generates fixation density maps using the fixation location as described in our method section. "*global*" also generates a fixation density map, but does not use the fixation location information and therefore generates for each fixation a uniform weight across the whole fixated image. In addition, we evaluate two more conditions, where we either used the fixation duration (◔) as a weight to the average class posterior of each fixated image (see subsection 4.4) or ignore the duration.

Table 1 shows the result of all 4 combinations of these conditions, with the first column denoting if local or global information was used and the second column ◔ whether fixation duration was used. Absolute performance of our best model using local information and fixation duration were 57%, 74%, and 84% on top1-3 accuracy respectively for the categorization task and 34% accuracy for attributes. The results show a consistent improvement (16 to 18 pp for categories, 6 pp for attributes) across all measures and tasks

$\sigma_{\text{fix}} \rightarrow$	1	1.2	1.4	1.6	1.8	2
Top1	55%	54%	56%	56%	57%	57%
Top2	74%	74%	74%	74%	74%	75%
Top3	83%	84%	84%	85%	85%	84%

Table 2. Evaluation of different gaze encoding schemes using different per-fixation σ_{fix}.

going from a global to a local representation (first to second row). Adding the weighting by fixation duration yields another consistent improvement for both local and global approach (another 6 to 5 pp for categories). Our best method, improves overall by 22 to 26 pp on the categorization task and 14 pp on the attributes. The global method without fixation duration (first row) is in spirit similar to [23] – although the specific application differs. All further experiments will consider our best model (last row) as the reference and justify the parameter choices (average pooling, $sigma_{\text{fix}}$) by varying each parameters one by one.

5.2. Evaluation of the Gaze Encoding

We then evaluated the gaze encoding that takes individual fixations as input and produces a fixation density map. We first evaluated the representation of a single fixation that depends on the parameterσ_{fix}, followed by the pooling scheme that combines multiple fixations into fixation density maps.

Effects of Fixation Representation Parameter f_σ. The parameter σ_{fix} controls the spatial extend of a single fixation in the fixation density maps as described in subsection 4.2. We determined an appropriate setting of this parameter to be $\sigma_{\text{fix}} = 1.6$ in a pilot study to roughly match the eye tracker accuracy and analyzed here the influence on the overall performance by varying this parameter in a sensible range (given eye tracker accuracy and coarseness of feature map) from 1 to 2 as shown in Table 2. As can be seen from the Table, our method is largely insensitive to the investigated range of reasonable choices of this parameter and our choice of 1.6 is on average a valid choice within that range.

Fixation Pooling Strategies. We evaluated two options for how to integrate single fixations into an fixation density map: Either using average or max pooling. The results are shown in Table 3. As the Table shows, while both options perform well, average pooling consistently improves over the max pooling option.

5.3. Noise Robustness Analysis

While our gaze data is recorded with a highly-accurate stationary eye tracker, there are different modalities and types of eye trackers available. One key characteristic in which they differ is the error at which they can record gaze data – typically measured in degrees of visual angle. While our controlled setup provides us with an accuracy of about 0.7 degrees of error, state-of-the-art eye trackers based on web-cams, tablets or integrated into glasses can have up to 4 degrees depending also on the deployment scenario [34]. Therefore, we finally investigated the robustness of our approach w.r.t. different levels of (simulated) noise in the eye tracker. To this end, we sampled noise from a normal distribution with $\sigma = 1, 3, 5$. This corresponds roughly to $60, 120$ and 200 pixels and to $1.2, 2.5$ and 4.2 degrees of visual angles and hence covers a realistic range of errors. The results of this evaluation are shown in Figure 5. As can be seen, our method is quite robust to noise with only a drop of 5 to 10pp for Top3 to Top1 accuracy, respectively – even at the highest noise level. In particular, all the results are consistently above the performance of the corresponding global methods shown as dashed lines in the plot.

Fixation	——— Category ———			Attribute
Pooling	Top1	Top2	Top3	Accuracy
Max	54%±8	73%±9	83%±6	32%±1
Average	**57%±8**	**74%±7**	**84%±4**	**34%±1**

Table 3. Evaluation of different fixation pooling strategies using average or max pooling.

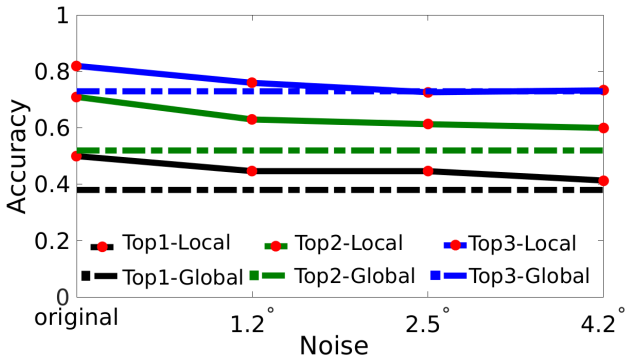

Figure 5. Accuracy for different amounts of noise added to the eye tracking data. Our method is robust to this error which suggests that it can also be used with head-mounted eye trackers or learning-based methods that leverage RGB cameras integrated into phones, laptops, or public displays.

5.4. Visualization and Analysis of Gaze Pooling Layer on Single Images

We provide further insights into the working of our Gaze Pooling Layer by showing visual examples of the attended class activation maps, associated fixation density map and

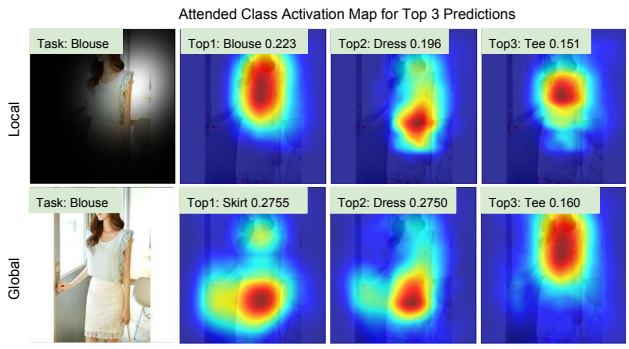

Figure 6. Attended class activation maps of top 3 predictions in local and global method for a given image. Participants were searching for target category "Blouse". The maps shows the discriminative image regions used for for this search task.

search target prediction results. While the quantitative evaluation was conducted on full collages, this is impracticable for inspection. Therefore, we show in the following visualizations and analysis on single images.

Predictions. Figure 7 shows results for the categorization task and Figure 8 for the attribute task. Each of these figures shows the output of the "global" method that use uniform fixation density map as well as the "local" method that makes full use of the gaze data. We observe that for the "local" method a relevant part of the images is fixated on which in turn leads to correct prediction of the intended search task.

Attended Class Activation Map (ACAM) Visualization. Figure 6 shows the attended class activation map (ACAM) of top 3 predictions, for "local" as well as "global" approach. The "global" method exploits that this image was fixed on - but does not exploit the location information of the fixations. Therefore it reduces in the case of a single image to a standard CAM. E.g the lower part of image is activated for "skirt" and the upper part is activated for "Tee". One can see that highlighted regions vary across predicted class. The first row shows the ACAM for the "local" method. It can be seen how the local weighting due to the fixation is selective to the relevant features of the search target, e.g. eliminating the "skirt" responses and retaining the "blouse" responses.

6. Discussion

In this work we studied the problem of predicting categories and attributes of search targets from gaze data. Table 1 shows strong performance for both tasks. Our Gaze Pooling Layer represents a modular and effective integration of visual and gaze features that is compatible with modern deep learning architectures. Therefore, we would like to highlight three features that are of particular practical importance.

Parameter Free Integration Scheme. First, our proposed integration scheme is basically parameter-free. We introduce a single parameter σ_{fix} but the gaze encoding is only input to the integration scheme and, in addition, the method turns out to be not sensitive to the choice (see experiments in subsection 4.2).

Training from Visual Data. Second, fixing the fixation density maps to uniform maps yields a deep architecture similar to a GAP network that is well-suited for various classification tasks. While this no longer addresses the task of predicting categories and attributes intended by the human in the loop, it allows us to train the remaining architecture for the task at hand and on visual data, which is typically easier to obtain in larger quantities than gaze data. This type of training results in a domain-specific image encoding as well as task-specific classifier.

Training Free Gaze Deployment. Gaze data is time consuming to acquire – which makes it rather incompatible with today's data hungry deep learning models. In our model, however, the fixations density maps computed from the gaze data can be understood as spatially localized feature importance that are used to weight feature importance in the spatial image feature maps Figure 6. Ours results demonstrate that strong performance can be obtained with this re-weighting scheme without the need to re-train with gaze data. As a result, our approach can be deployed without any gaze-specific training. This result is surprising, in particular as the visual model on its own is completely uninformative without gaze data on the task of search target prediction (as we have validated in subsection 5.1. We believe this simplicity of deployment is a key feature that makes the use of gaze information in deep learning practical.

Biases in Mental Model of Attributes and Categories Among Users. In order to illustrate the challenges our Gaze Pooling Layer has to deal with in terms of the variations in the observed gaze data, we show example fixation data in Figure 9. In each image, fixation data of two participants (red and green dots) is overlaid over a presented collage. Although both participants had the same search target (top: attribute 'Floral'; bottom: category 'Cardigan'), we observe a drastically different fixation behaviour. One possible explanation is that the mental models of the same target category or attribute can vary widely depending on personal biases [8]. Despite these strong variations in the gaze information, our Gaze Pooling Layer allows to predict the correct answer in all 4 cases. The key to this success is aggregating relevant local visual feature across all images in the collage, that in turn represent one consistent search target in terms of categories and attributes.

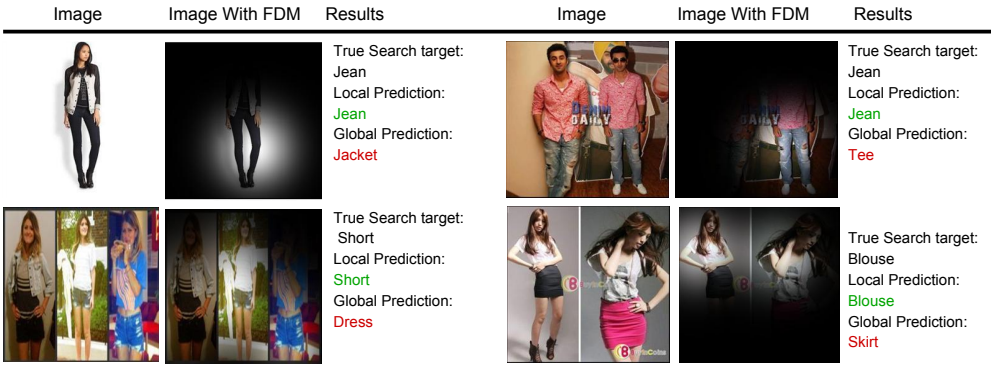

Figure 7. Example category responses of local and global method. Green means correct and red means wrong target prediction.

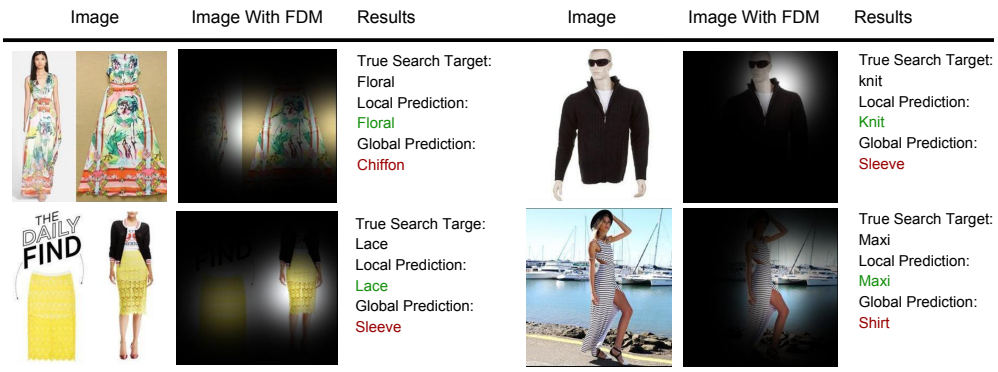

Figure 8. Example attribute responses of local and global method. Green means correct and red means wrong target prediction.

Figure 9. Example fixation data of 2 participants (red and green dots) with search target attribute='Floral' on top and category='Cardigan' below.

7. Conclusion

In this work we proposed the first method to predict the category and attributes of visual search targets from human gaze data. To this end, we proposed a novel Gaze Pooling Layer that allows us to seamlessly integrate semantic and localized fixation information into deep image representations. Our model does not require gaze information at training time, which makes it practical and easy to deploy. We believe that the ease of preparation and compatibility of our Gaze Pooling Layer with existing models will stimulate further research on predicting mental states from gaze data.

Acknowledgment

This research was supported by the German Research Foundation (DFG CRC 1223) and the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University. We also thank Mykhaylo Andriluka for helpful comments on the paper.

References

[1] K. G. A. Kovashka, D. Parikh. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2), 2012.

[2] A. Borji, A. Lennartz, and M. Pomplun. What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. *Neurocomputing*, 2014.

[3] F. J. Brigham, E. Zaimi, J. J. Matkins, J. Shields, J. McDonnouugh, and J. J. Jakubecy. *The eyes may have it:Reconsidering eye-movement research in human cognition*, volume 15. Advances in Learning and Behavioral Disabilities, 2001.

[4] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster. Eye movement analysis for activity recognition using electrooculography. *IEEE TPAMI*, 33(4), 2011.

[5] X. Chen and G. J. Zelinsky. Real-world visual search is dominated by top-down guidance. *Vision Research*, 46(24):4118 – 4133, 2006.

[6] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014.

[7] M. P. Eckstein. Visual search: A retrospective. *Journal of Vision*, 11(5):14, 2011.

[8] M. Ferecatu and D. Geman. A statistical framework for image category search from a mental picture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1087–1101, June 2009.

[9] A. V. K. Simonyan, Karen and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

[10] N. Karessli, Z. Akata, B. Schiele, and A. Bulling. Gaze embeddings for zero-shot image classification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. Manjunath. From where and how to what we see. In *ICCV*, 2013.

[12] C. L. Kleinke. Gaze and Eye Contact: A Research Review. *Psychological Bulletin*, 100(1), 1986.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[14] M. F. Land and S. Furneaux. The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 352(1358), 1997.

[15] D. T. Levin, Y. Takarae, A. G. Miner, and F. Keil. Efficient visual search by category: Specifying the features that mark the difference between artifacts and animals in preattentive vision. *Perception & Psychophysics*, 63(4):676–697, 2001.

[16] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2015.

[17] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[18] E. Marinoiu, D. Papava, and C. Sminchisescu. Pictorial human spaces. how well do humans perceive a 3d articulated pose? In *ICCV*, 2013.

[19] S. Mathe and C. Sminchisescu. Multiple instance reinforcement learning for efficient weakly-supervised detection in images. *arXiv:1412.0100*, 2014.

[20] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *ICCV*, 2009.

[21] M. B. Neider and G. J. Zelinsky. Searching for camouflaged targets: Effects of target-background similarity on visual search. *Vision Research*, 46(14):2217 – 2235, 2006.

[22] G. Papadopoulos, K. Apostolakis, and P. Daras. Gaze-based relevance feedback for realizing region-based image retrieval. *ACM-MM*, 16(2), Feb 2014.

[23] H. Sattar, S. Müller, M. Fritz, and A. Bulling. Prediction of search targets from fixations in open-world settings. In *CVPR*, 2015.

[24] I. Shcherbatyi, A. Bulling, and M. Fritz. Gazedpm: Early integration of gaze information in deformable part models. arxiv:1505.05753 [cs.cv], 2015.

[25] J. Steil and A. Bulling. Discovery of everyday human activities from long-term visual behaviour using topic models. In *UbiComp*, 2015.

[26] R. Subramanian, V. Yanulevskaya, and N. Sebe. Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements. In *ACM-MM*, 2011.

[27] Y. Sugano and A. Bulling. Seeing with humans: Gaze-assisted neural image captioning. arxiv:1608.05203, 2016.

[28] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel. Gaze guided object recognition using a head-mounted eye tracker. In *ETRA*, 2012.

[29] S. P. Wilson, J. Fauqueur, and N. Boujemaa. *Mental Search in Image Databases: Implicit Versus Explicit Content Query*. Springer Berlin Heidelberg, 2008.

[30] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1), 2014.

[31] A. L. Yarbus. *Eye movements and vision*. Springer, 1967.

[32] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M.Hospedales, and C. C. Loy. Sketch me that shoe. In *CVPR*, 2016.

[33] G. J. Zelinsky, Y. Peng, and D. Samaras. Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision*, 13(14), 2013.

[34] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015.

[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.