Gaze Embeddings for Zero-Shot Image Classification

Nour Karessli[1]* Zeynep Akata[1,2] Bernt Schiele[1] Andreas Bulling[1]

[1]Max Planck Institute for Informatics
Saarland Informatics Campus

[2]Amsterdam Machine Learning Lab
University of Amsterdam

Abstract

Zero-shot image classification using auxiliary information, such as attributes describing discriminative object properties, requires time-consuming annotation by domain experts. We instead propose a method that relies on human gaze as auxiliary information, exploiting that even non-expert users have a natural ability to judge class membership. We present a data collection paradigm that involves a discrimination task to increase the information content obtained from gaze data. Our method extracts discriminative descriptors from the data and learns a compatibility function between image and gaze using three novel gaze embeddings: Gaze Histograms (GH), Gaze Features with Grid (GFG) and Gaze Features with Sequence (GFS). We introduce two new gaze-annotated datasets for fine-grained image classification and show that human gaze data is indeed class discriminative, provides a competitive alternative to expert-annotated attributes, and outperforms other baselines for zero-shot image classification.

1. Introduction

Zero-shot learning is a challenging task given that some classes are not present at training time [1, 28, 35, 41]. State-of-the-art methods rely on auxiliary information to aid the classification, such as object attributes [5, 6, 16]. While image annotation using such attributes can be performed by naïve users, domain experts have to compile the initial list of discriminative attributes for a fixed set of classes and have to revise this list whenever new classes are added. Several recent works therefore evaluated alternatives, such as distributed text representations extracted from online text corpora such as Wikipedia [23, 32], web-search data [35] or object hierarchies, such as WordNet [24]. While such representations can be extracted automatically and are therefore less costly, they do not outperform attributes.

*Nour Karessli is currently with Eyeem, Berlin. The majority of this work was done in Max Planck Institute for Informatics.

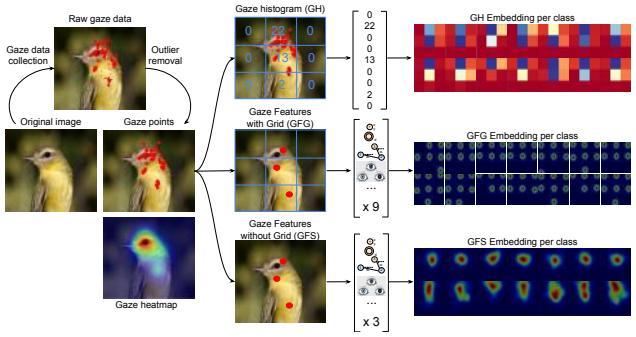

Figure 1: We encode gaze points into vectors using three different methods: gaze histogram (GH), gaze features with grid (GFG), and gaze features with sequence (GFS).

We instead propose to exploit human gaze data as auxiliary information for zero-shot image classification. Gaze has two advantages over attributes: 1) discrimination of objects from different classes can be performed by non-experts, i.e. we do not require domain knowledge, and 2) data collection only takes a few seconds per image and is implicit, i.e. does not involve explicitly picking class attributes but instead exploits our natural ability to tell objects apart based on their appearance. We further propose a novel data collection paradigm to encourage observers to focus on most discriminative parts of an object and thereby maximise the information content available for the classification task. The paradigm involves observers to first inspect exemplars from two different object classes shown to them side-by-side, and subsequently take a binary decision for class membership for another exemplar shown randomly from one of these classes. While human gaze data has previously been used to obtain bounding-box annotations for object detection [29] or approximated by mouse clicks to guide image feature extraction [4], this work is first to directly use human gaze data as auxiliary information for zero-shot learning.

The contributions of our work are three-fold. First, we propose human gaze data as auxiliary information for zero-shot image classification, being the first work to tackle this task using gaze. Second, we provide extensive human gaze

data of multiple observers for two fine-grained subsets of Caltech UCSD Birds 2010 (CUB) [48] and one subset of Oxford Pets (PET) [31] datasets. Third, we propose three novel class-discriminative gaze descriptors, namely Gaze Histograms (GH), Gaze Features with Grid (GFG), and Gaze Features with Sequence (GFS) and complement deep image features in a structured joint embedding framework [3]. Through extensive evaluations on our datasets, we show that human gaze of non-experts is indeed class-discriminative and that the proposed gaze embedding methods improve over several baselines and provide a competitive alternative to expert-provided attributes for zero-shot learning.

2. Related Work

Our work is related to previous works on zero-shot learning and gaze-supported computer vision.

Zero-Shot Learning. Zero-shot learning [1, 28, 35, 41] assumes disjoint sets of training and test classes. As no labeled visual data is available for some classes during training, typically some form of auxiliary information is used to associate different classes. Attributes [5, 6, 16] being human-annotated discriminative visual properties of objects are the most common type of auxiliary information. They have been shown to perform well in several tasks such as image classification [7, 30, 40, 46, 47], pedestrian detection [15, 20], and action recognition [19, 33, 51]. On the model side, multi-modal joint embedding methods [1, 3, 50] have been shown to provide a means to transfer knowledge from images to classes and vice versa through attributes. However as fine-grained objects [27, 48] are visually very similar to each other, a large number of attributes are required which is costly to obtain. Therefore, several alternatives have been proposed in the literature. Distributed text representations such as Word2Vec [23] or GloVe [23] are extracted automatically from online textual resources such as Wikipedia. Hierarchical class embeddings provide another alternative (e.g. using WordNet [24]) to learn semantic similarities between classes. On the other hand, search for alternative sources of auxiliary information has introduced the concept of fine-grained visual descriptions [34] which indicates that although novice users may not know the name of a fine-grained object, they have a natural way of determining discriminative properties of such objects.

Collecting labels from experts or attributes from novice users requires asking many yes/no questions for each image. We argue that, instead, it may be enough for them to look at an image to identify fine-grained differences between object classes. Although eye tracking equipment adds to the cost, recent advances suggest that eye tracking will soon become ubiquitous, e.g. in mobile phones [9]. Therefore, we propose to extract class-discriminative representations of human gaze and use them as auxiliary information for zero-shot learning.

Gaze-Supported Computer Vision. Gaze has been an increasingly popular cue to support various computer vision tasks. Gaze-tracking data has been used to perform weakly supervised training of object detectors [13, 39, 52], estimating human pose [21], inferring scene semantics [42], detecting actions [22], detecting salient objects in images [18] and video [14], segmenting images [25], image captioning [43] or predicting search targets during visual search [38]. Human gaze data is highly dependent on the task the annotators have to complete. While [12, 52] collected gaze tracking data for a free viewing task, [29] asked users to focus on a visual search task and built POET dataset. On the other hand, in [17] gaze has been used to evaluate saliency algorithms on video sequences. [4] imitated human gaze data with "bubbles" that they draw around mouse clicks where the annotators find distinguishing image regions. Others used saliency maps instead of real gaze data to improve object detection performance [26, 36]. Maybe the most closely related work to ours is [29], where fixations were used to generate object bounding boxes and thereby reduce the bounding-box annotation effort. On the other hand, to the best of our knowledge, we are first to collect real eye tracking data to extract class-discriminative representations and then in turn to use them as auxiliary information for the specific task of zero-shot image classification. Our technical novelty is in our design of effective gaze representations that provide a structure in class embedding space.

3. Gaze Tracking and Datasets

Here, we present our gaze data collection paradigm, detail our gaze datasets and our gaze embeddings.

3.1. Gaze Data Collection

We collect the eye tracking data with the Tobii TX300 remote eye tracker that records binocular gaze data at 300 Hz. We implement a custom data collection software in C# using the manufacturer-provided SDK which we will make publicly available. Our software logs a timestamp, users' on-screen gaze location, their pupil diameter, as well as a validity code for each eye that indicates the trackers confidence in correctly identifying the eye. We use gaze-points that are valid for both eyes. The participants are seated 67 cm from a 31.5 inch LCD screen. We use a chin rest to reduce head movement and consequently improve the eye tracking accuracy. The vertical extend of the image shown on the screen is ≈ 15 cm, thus the visual angle[1] is $\approx 25°$. We record 5 participants for every image which leads to 5 gaze streams for each image of three datasets. Almost 50% of our participants have impaired eye sight however, none of them wear glasses during the data collection although 30% of them wear contact lenses.

[1] Visual Angle $= 2 \times \arctan$ (Vertical Extend Stimuli / Distance)

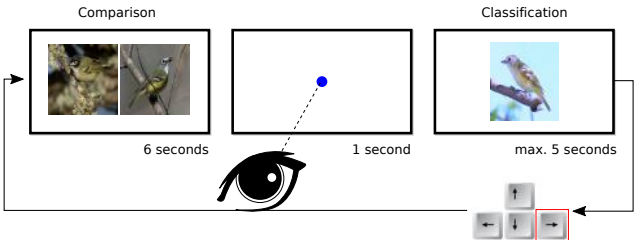

Figure 2: Participants first look at two images of two fine-grained classes (6 sec), then at the center of the screen to "reset" their gaze position (1 sec), finally they click the left or right arrow on the keyboard (max 5 sec) to select the class they think the image belongs to. We record their gaze only on classification screen.

Our data collection paradigm is illustrated in Figure 2. Participants first answer a short questionnaire on demographics, e.g. age, gender, eye sight, etc. and then we calibrate the eye tracker using the standard 5-point calibration routine. After calibration, participants follow a cycle of three steps, namely comparison, fixation, and classification. During the comparison step, we show two example images that we randomly sample from two fine-grained classes for six seconds where participants learn fine-grained differences between two classes. In the fixation step, we ask the participants to fixate on a dot in the center of the screen for one second to "reset" the gaze position. In the classification step, we show a new instance of one of the two classes which participants need to classify by clicking on right/left arrow of the keyboard in max five seconds. This step terminates before five seconds if the annotator decides earlier. A new cycle starts until all the images are annotated by the same user.

Gaze Datasets. We collect gaze tracking data for images of two publicly available datasets (See Table 1 for details). Following [4] we collect gaze tracking data of 14 classes (7 classes of Vireos and 7 classes of Woodpeckers: CUB-VW) for all the available 464 images. Each image is annotated by 5 participants. In addition, CUB-VWSW includes two more bird families of Sparrows and Warblers. CUB-VWSW contains $11,730$ gaze tracks of five participants for every image, i.e. 1882 images in total. Finally, we collect gaze tracking data for the images of a 24-class subset of the Oxford Pets dataset [31], where we take all 12 classes from Cats and a subset of 12 classes from Dogs. Following CUB setting, we collect $3,600$ gaze tracks of 720 images from five participants and name this dataset PET. We collect gaze data at the sub-species level, e.g. black-capped vireo vs red-eyed vireo. We observe that comparing birds at a higher level, e.g. woodpeckers vs vireos, is too easy and users take a decision instantly, while comparing birds at the sub-species level takes longer providing us more gaze points.

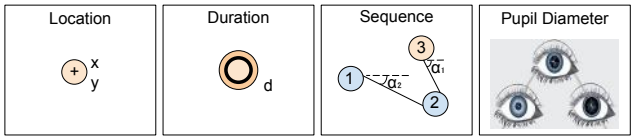

Figure 3: Gaze features include gaze point location (x, y), gaze point duration (d), angles to the previous and next gaze points in the sequence (α_1, α_2), and the pupil diameter (R).

3.2. Gaze Embeddings

We propose Gaze Histograms (GH), Gaze Features with Grid (GFG) and Gaze Features with Sequence (GFS) as three gaze embedding methods.

Gaze Histogram (GH). Gaze points are encoded into a $m \times n$-dimensional vector using a spatial grid overlayed over the image with m rows and n columns. The per-class gaze histogram embedding is the mean gaze histogram of a particular class. We encode gaze of each participant separately to evaluate the annotator bias.

Figure 1 (top) shows how we construct a 9-dimensional histogram using a spatial grid of 3×3. For simplicity, we show per-class histograms of seven vireos and seven woodpeckers with darker colors indicating higher number of occurrences. High attention points for Vireos (top row) fall in the middle of the image whereas for woodpeckers (bottom row) the top of the image seems to be more important. From visual inspection of the original images, we observe that vireos in CUB often sit on horizontal tree branches with their eyes being the most discriminative property. In contrast, woodpeckers often climb on large tree trunks with their head region being the most discriminative property. Gaze histograms capture these spatial, i.e. horizontal versus vertical, and class-specific differences.

Gaze Features (GFx). Counting the number of gaze points that fall into a grid cell encodes the location information coarsely and does not encode any information about the duration, sequence and the attention of the observer. Therefore, we build 6-dimensional gaze features, i.e. $[x, y, d, \alpha_1, \alpha_2, R]$, as shown in Figure 3. Our gaze features encode gaze location (x, y), gaze duration (d), angles (α_1, α_2) between the previous and subsequent gaze point in the scan path, and pupil diameter R that was shown to relate to processing load [10] of the observer. We embed these gaze features in two different ways, namely Gaze Features with Grid (GFG) and Gaze Features with Sequence (GFS).

Gaze Features with Grid (GFG) uses a spatial grid similar to gaze histograms (GH) to discretize the gaze space. Instead of counting the number of gaze-points per cell, we average the 6-dim gaze features of the points that fall in each cell. We then concatenate gaze features that fall inside each grid cell into a $6 \times m \times n$-dimensional vector with m and n being

the number of rows and columns of the spatial grid. The per-class GFG embedding is then the average of all GFG vectors from the same class. By encoding the spatial ordering of the gaze points, GFG captures information related to the typical behavioral patterns of birds such as sitting on horizontal tree branches vs climbing large tree trunks.

Gaze Features with Sequence (GFS) encodes the sequential order of gaze points. First we order gaze points with respect to time, i.e. first occurring gaze point to the last, then we sequentially select a fixed number (k) of gaze points from each gaze sequence and embed them as a $6 \times k$-dimensional vector. Here, k is typically the minimum number of gaze points extracted from the gaze-sequence of a certain observer. The GFS encodes the time sequence of the gaze points instead of focusing on their spatial layout. The per-class GFS embedding is the average GFS embeddings of the same class.

Combining Gaze Embeddings. As participants gaze at different regions of the same image, we argue that their gaze embeddings may contain complementary information. We thus propose three different methods to combine their gaze embeddings: First, we average the per-class gaze embeddings ($\phi(y)$) of each participant abbreviated by `AVG`. Second, we concatenate per-class gaze embedding of each participant through early fusion, i.e. `EARLY` and learn one single model. Third, we learn a model for each participant separately and then we average their classification scores before making the final prediction decision in the late fusion setting, i.e. `LATE`.

4. Gaze-Supported Zero-Shot Learning

In zero-shot learning, the set of training and test classes are disjoint. During training, the models have access to only the images and gaze embeddings of training classes but none of the images or gaze embeddings of test classes. The lack of labeled images from test classes is compensated by the use of auxiliary information that defines a structure in the label space [2, 3, 7, 49] and provides a means of associating training and test classes. In the following, we provide the details of the zero-shot learning model [3].

Zero-Shot Learning Model. Given image and class pairs $x_n \in \mathcal{X}$ and $y_n \in \mathcal{Y}$ from a training set $\mathcal{S} = \{(x_n, y_n), n = 1 \dots N\}$, we use the Structured Joint Embedding (SJE) model [3] to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ by minimizing the empirical risk

$$\frac{1}{N} \sum_{n=1}^{N} \Delta(y_n, f(x_n)) \tag{1}$$

where $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \{0, 1\}$ defines the cost of predicting $f(x)$ when the true label is y. The SJE model maximizes the compatibility function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ as follows:

$$f(x; W) = \arg\max_{y \in \mathcal{Y}} F(x, y; W). \tag{2}$$

Dataset	# img / class	Gaze	Bubbles [4]
CUB-VW	464 / 14	2320	210
CUB-VWSW	2346 / 60	11730	900
PET	720 / 24	3600	–

Table 1: Statistics for CUB-VW, CUB-VWSW datasets (images selected from CUB [48]) and PET dataset (images selected from Oxford PET [31]) w.r.t. number of images, classes, number of gaze and bubble [4] tracks.

that has the following bi-linear form:

$$F(x, y; W) = \theta(x)^\top W \varphi(y). \tag{3}$$

where the image embedding ($\theta(x)$), i.e. image features extracted from a Deep Neural Network (DNN) and the class embedding ($\varphi(y)$), i.e. gaze embeddings are provided as a pre-processing step. W is learned through structured SVM [45] by maximizing the ranking of the correct label:

$$\max_y (\Delta(y_n, y) + F(x_n, y; W)) - F(x_n, y_n; W) \tag{4}$$

and optimized through stochastic gradient descent (SGD). At test time, we search for the test class whose per-class gaze embedding yields the highest joint compatibility score.

5. Experiments

In this section, we first provide details on datasets, image embeddings and parameter setting that we use for zero-shot learning. We then present our detailed evaluation of gaze embeddings compared with various baselines both qualitatively and quantitatively.

Datasets. As shown in Table 1, [4] provide mouse-click data, i.e. bubble tracks, for 14 classes (7 classes of Vireos and 7 classes of Woodpeckers: CUB-VW) of CUB for a selection of 210 images. They collected bubble tracks for every image, however every annotator did not annotate every image. Therefore, unlike our 5 streams of gaze-tracks collected from 5 participants, there is only a single stream of bubble-tracks. On CUB-VW bubble-tracks, we build per-class bubble representations in three different ways, i.e. the same as gaze, and found out that Bubble Features with Sequence (BFS), encoding x, y location and radius of the bubble works the best, therefore we use these as bubble representations in all our experiments. We extensively evaluate our method on CUB-VW in the following section. Note that we validate all the gaze-data processing parameters on CUB-VW and use the same parameters for other datasets. Our CUB-VWSW dataset, i.e. including Vireos, Woodpeckers, Sparrows and Warblers, comes with 312 expert-annotated attributes for every class, i.e. embedded as 312-dimensional per-class attribute vectors. [11] extended [4] by collecting bubble-tracks for more bird species, therefore bubble-tracks of 900

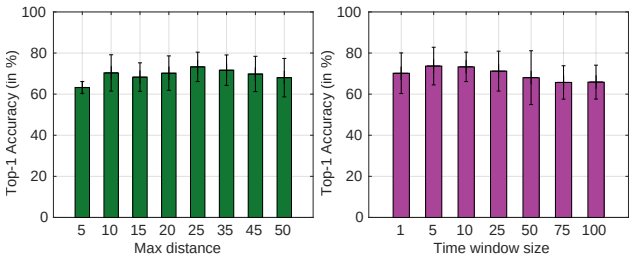

Figure 4: Raw gaze data processing: max distance between gaze points and time window size.

images that [11] selected for CUB-VWSW dataset are also available. The PET dataset neither contains attributes nor bubble-tracks. For both our CUB-VWSV and PET datasets, we further construct bag-of-words representations extracted from Wikipedia articles that describe a specific class to build per-class representations. Bag of words frequencies are produced by counting the occurrence of each vocabulary word that appears within a document. To obtain fixed-sized descriptors, we only consider the N-most frequent words across all classes after removing stop-words and stemming.

Image Embeddings and Parameter Setting. As image embeddings, we extract $1,024$-dim CNN features from an ImageNet pre-trained GoogLeNet [44] model. We neither do any task-specific image processing, such as image cropping, nor fine-tune the pre-trained network on our task. We cross-validate the zero-shot learning parameters, i.e. step size in SGD and the number of epochs, on 10 different zero-shot splits that construct by maintaining a ratio of $2/1/1$ for disjoint training, validation and test classes. We measure accuracy as average per-class top-1 accuracy.

5.1. Gaze Embeddings on CUB-VW

In this section, we first show how we pre-process our raw gaze data, and then extensively evaluate our gaze embeddings wrt. multiple criteria on the CUB-VW dataset.

Processing Raw Gaze Data. Raw gaze data is inherently noisy due to inaccuracies of the eye tracker. We reduce this noise using a dispersion-based method [37] which calculates the dispersion of gaze points using a sliding window approach with window size w_s and applies a threshold t_s on this dispersion value. All gaze points within the window are then set to the mean of all points below the threshold. In order to disentangle this raw-data pre-processing step from our end task of zero-shot learning, we train standard one-vs-rest SVM classifiers on stacked gaze features as training samples, and image label as supervision signal. We use $[x, y, d, \alpha_1, \alpha_2, R]$ as gaze features, GFS as gaze feature encoding and evaluate 10 random train and test splits.

Figure 4 (left) shows that the highest accuracy is obtained using $w_s = 25$ degrees (among $w_s = 5...50$). Time-window

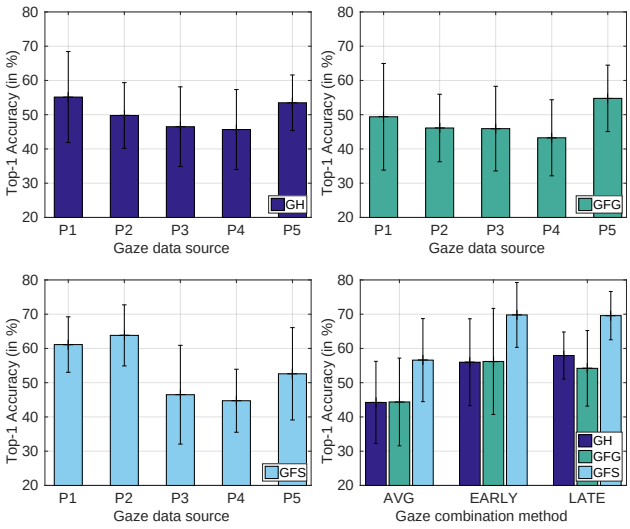

Figure 5: Comparing Gaze Histogram (GH), Gaze Features with Grid (GFG) and Gaze Features with Sequence (GFS). We evaluate 5 participants separately as well as their various combinations: Averaging each participant's gaze embeddings (AVG), Combining them through early fusion (EARLY) and through late fusion (LATE).

size (t_s) depends on how long the annotator needs to view the image before making a decision. As our users have significantly shorter viewing duration (≈ 0.5sec) compared to eye tracking studies that requires long viewing times, e.g. reading a textual document, we fix time window size on our data. By keeping $w_s = 25$, we evaluate $t_s = 1...100$ and observe that $t_s = 10$ms works the best (Figure 4, right). We observe that performance does vary across experiments, albeit not significantly. Therefore, at least for the datasets investigated in this work, gaze data can be processed in a generic fashion, i.e. does not have to be tailored for a particular user or object class.

Comparing Different Gaze Embeddings. We now compare the performance of Gaze Histograms (GH), Gaze Features with Grid (GFG) and Gaze Features with Sequence (GFS). We build GFx, i.e. GFG and GFS, with all gaze features, i.e. $[x, y, d, \alpha, R]$ for consistency. We first consider the gaze embeddings of our 5 participants separately and then combine the gaze embeddings of each participant by averaging them (`AVG`), concatenating them through early fusion (`EARLY`), and combining the classification scores obtained by each participant's gaze data through late fusion (`LATE`). We repeat these experiments on 10 different zero-shot splits to show a robustness estimate.

As shown in Figure 5, GFS embeddings outperform GH and GFG embeddings. This implies that the sequence information is more helpful than the spatial discretization by using the grid. Therefore, we argue that in fine-grained zero-

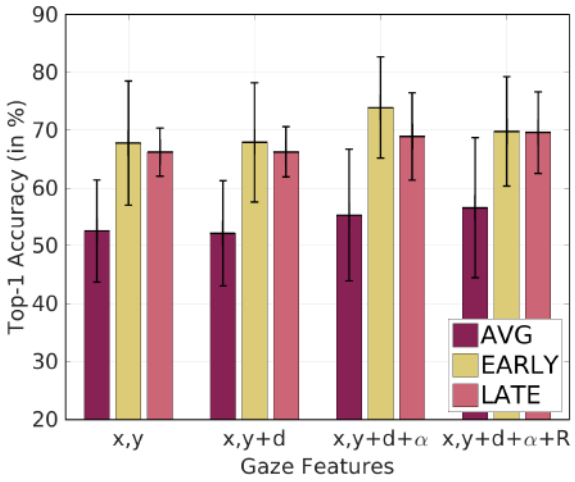

Figure 6: Effect of gaze features: Location (x, y), Duration (d), Sequence (α_1, α_2), Pupil diameter (R). We start with x, y and concatenate d, α_1, α_2 and R cumulatively.

shot learning task, the sequence of gaze points is important to obtain best performance. Our second observation is that indeed each participant's gaze embedding lead to different results, therefore considering the annotator bias is important. For GH, the best performing participant is the first, while for GFG it is the fifth and for GFS it is the second. We argue that gaze embeddings of each participant is complementary, therefore we propose to combine gaze embeddings of different participants. We obtain 56.6% using AVG, 69.8% using EARLY and 69.6% with LATE. These results support our intuition that there is complementary information between gaze embeddings of different participants.

Analyzing Gaze Features. We evaluate the effects of encoding gaze location (x, y), duration (d), sequence (α_1, α_2) and the annotator's pupil diameter (R) that measures concentration. We build gaze features cumulatively by starting with the parameter x, y, followed by d, α_1, α_2 and R sequentially in this order. Being the best performing method from Figure 5, we evaluate the effect of gaze features with respect to participants combined GFS with AVG, EARLY and LATE. We observe from Figure 6 that EARLY achieves the highest 73.9% accuracy when we use $[x, y, d, \alpha_1, \alpha_2]$ features. The $[x, y]$ features already achieves high accuracy, adding duration, i.e. d slightly improves results and adding the sequence information, i.e. α, adds further improvement. However, the pupil diameter parameter does not bring further improvements. As our annotators go through all the images which requires a total of one hour of constant concentration. Although they take a break after half an hour, their concentration drops towards the end of the task while they become familiar with the fine-grained bird species.

Comparing Gaze and Baselines. Table 2 shows a performance comparison of our gaze embeddings with several baselines. Saliency histogram (35.8%) is a discretization of a saliency map [8] using a spatial grid over the image. As a second baseline, we randomly sample points in the image and obtain 39.5% accuracy. Another baseline is taking the location of the central point as an embedding, which leads to 41.5%, indicating a certain center bias in CUB-VW images. Bubbles [4], mouse-click locations of visually distinguishing object properties, are the closest alternative to our gaze data. Bubbles achieve 43.2% accuracy, which supports the hypothesis that non-expert users are able to determine distinguishing properties of fine-grained objects. As the final baseline, we evaluate class embeddings extracted from Wikipedia articles (55.2%). Our best performing gaze embeddings, i.e. GFS EARLY with $[x, y, d, \alpha]$ from Figure 6, achieve 73.9% accuracy and outperform all these baselines. Moreover, they outperform expert annotated attributes with 72.9% being the current state-of-the-art. This result shows that human gaze data is indeed class discriminative while being more efficient than attributes to collect. Finally, we combine our gaze embeddings with attributes and show by obtaining 78.2% accuracy that human gaze data contains complementary information to attributes.

	Method	Accuracy
Baselines	Saliency histogram	35.8
	Random points in the image	39.5
	Central gaze point	41.5
	Bubbles [4]	43.2
	Bag-of-Words from Wiki	55.2
SoA	Human annotated attributes	72.9
Ours	Gaze embeddings	73.9
	Attributes + Gaze	**78.2**

Table 2: Comparing random points, mean gaze point, saliency histogram using [8], bubbles [4], Bag of Words and expert annotated attributes on CUB-VW.

Method	Accuracy
Gaze	**73.9**
Gaze: same images as bubbles	69.7
Gaze: same location as bubbles	64.0
Gaze: same number as bubbles (avg)	55.0
Gaze: same number as bubbles (rnd)	49.2
Bubbles (mouse-clicks)	43.2

Table 3: Ablation from gaze to bubbles: using our full gaze data with GFS EARLY embedding, using same images as bubbles, concatenating gaze points located inside bubbles, averaging those gaze points and using one among those gaze points vs bubbles.

Ablation from Gaze to Bubbles. As we observed a large

		CUB		
Method	**Side-Info**	**VW**	**VWSW**	**PET**
Random points	Image	39.5	9.0	21.0
Bubbles	Novice	43.2	10.3	N/A
Bag of Words	Wikipedia	55.2	24.0	33.5
Human Gaze	Novice	**73.9**	26.0	**46.6**
Attributes	Expert	72.9	**42.7**	N/A

Table 4: Comparing random points, bubbles [4], bag of words, attributes, and our gaze embeddings (GFS EARLY), on CUB-VW = CUB with Vireos and Woodpeckers, CUB-VWSW = CUB with Vireos, Warblers, Sparrows, Woodpeckers and PET=Oxford Pets with Cats and Dogs.

accuracy gap between gaze and bubble embeddings previously, we now investigate the reason for this gap through an ablation study. We gradually decrease the information content of gaze embeddings in the following way. We first use the same images as bubbles and observe from Table 3 that the accuracy decreases from 73.9% to 69.7%. We then concatenate the gaze features of the gaze points that fall inside the bubbles, i.e. use gaze points at the same location as bubbles, and observe the accuracy decline to 64.0%. Instead of concatenating, averaging the gaze points or taking one random point inside bubbles decreases the accuracy to 55.0% and 49.2% respectively. We attribute the accuracy difference between 49.2% and 43.2 (bubbles) to the gaze features, i.e. $[x, y, d, \alpha, R]$. We conclude from this experiment that the images that the annotators viewed while we recorded their gaze as well as their attention and the quantity, the location, the duration of the gaze-tracks are all important to obtain good zero-shot learning results.

5.2. Gaze Embeddings on Other Datasets

In this section, we first evaluate gaze embeddings on CUB with 60 species of Vireos, Woodpeckers, Sparrows and Warblers (CUB-VWSW) [48]. To show the generalizability of our idea to other domains, we also evaluate results on Oxford PET [31] dataset with 24 types of cats and dogs (PET). Note that we set the parameters based on experiments on CUB-VW and used those across all datasets.

Experiments on CUB-VWSW dataset. We use GFS-EARLY embedding for being the best performing method in our previous evaluation. We compare it with random points, bubbles, bag-of-words and attributes. Results on CUB-VWSW dataset show that gaze performs significantly better than random points that come from the image itself, bubble embeddings extracted from mouse-click locations, and BOW embeddings extracted from Wikipedia articles. On the other hand, expert annotated attributes outperform non-expert annotated gaze data. This is expected since our novice annotators did not compare different vireo, woodpecker, sparrow and warbler species and especially vireos, sparrows and warblers looks very similar to each other, i.e. having similar size, shape and colors. On the other hand, the fact that gaze embeddings perform better than BoW by itself is an interesting result. We suspect that allowing the annotators to explore differences between bird species at sub-species level and also super-species level, e.g. our annotators never compared woodpeckers and vireos but only two different woodpecker species, or annotating images using bird expert opinion would improve our results. Additionally, an improved zero-shot learning model that takes into account the hierarchical relationships between classes may lead to better results. We will explore these options in future work. Finally, fine-tuning the parameters of gaze embedding, which we intentionally avoid, may improve results.

Experiments on PET dataset. Here, as attributes and bubbles are not available, we use random points in the images and bag-of-words extracted from Wikipedia articles as baselines. The random chance is 16% as we sample 6 test classes, we repeat our experiments on 10 different zero-shot splits and report the average in the last column of Table 4. We observe that Wikipedia articles of PET classes include more information than random points in the image (21.0% vs 33.5% with bag of words). Whereas our gaze embeddings obtain 46.6% accuracy that significantly outperforms the results obtained with bag-of-words. As all the images show cat and dog breeds, our annotators are more familiar with these classes which makes this dataset less challenging than CUB. Note that by fine-tuning raw-gaze processing or gaze embedding parameters such as gaze features, gaze embedding type, etc. on this dataset, these results may potentially get higher. We conclude from PET results that our proposed gaze embeddings indeed capture class discriminative information and they can be generalized to other domains.

5.3. Qualitative Results

Qualitative results of birds, cats and dogs on Figure 7 shows five highest ranked images for three different test classes comparing gaze embeddings with competing methods. We additionally visualize the gaze heatmaps extracted from gaze tracks corresponding to that particular test image. Although we do not use the gaze embeddings for these test classes while training, we include these visualizations as give an intuition of how gaze-tracks look like.

For birds, we compare gaze with both human annotated attributes and bag-of-words. Gaze ranks "Black capped Vireo" images correctly in the first three positions whereas attributes and bow makes mistakes. The misclassified "Black capped Vireo" is a "White eyed Vireo" which also has its distinguishing property on the head region. The misclassified image in expert annotated attributes belongs to "Blue headed Vireo" whose embedding is similar to "Black capped Vireo".

Figure 7: Qualitative Results: Five highest ranked images for unseen classes of birds, cats and dogs. We compare gaze with attributes (when available) and with bag of word representations and show gaze heat-maps of selected images.

On the other hand, the word 'head' is highly frequent in both Wikipedia articles which makes the BoW embedding for these two classes similar and thus, it leads to a mismatch. For other examples, gaze embedding ranks correct images the highest. These results also illustrate the difficulty of the annotation on fine-grained datasets.

For cats and dogs, we observe that qualitative results follow a similar trend as quantitative results. Qualitatively, gaze performs better than bag-of-words representations. Comparing gaze and bag-of-words results shows that gaze never confuses cats and dogs whereas such confusion occurs for bag-of-words. As a failure case of gaze embeddings, gaze retrieves "Abyssinian Cat" images for "Russian blue" query as these two cats have a similar form but can be distinguished only with color information, not encoded with gaze.

6. Conclusion

In this work, we proposed to use gaze data as auxiliary information to learn a compatibility between image and label space for zero-shot learning. In addition to a novel eye tracking data collection that captures humans' natural ability to distinguish between two objects we proposed three gaze embedding methods that 1) use spatial layout of the gaze points and employ first order statistics, 2) integrate location, duration, sequential ordering and user's concentration features to spatial ordering information, and 3) sequentially sample gaze features. Through extensive quantitative and qualitative experiments on the CUB-VW dataset we showed that human gaze is indeed class-discriminative and improves over both expert-annotated attributes and mouse-click data (bubbles). Our qualitative and quantitative results on the PET dataset showed that gaze can be generalized to other domains. On the other hand, our results on larger fine-grained datasets, e.g. CUB-VWSW might indicate that the results would benefit from alternative data collection paradigms that allow annotators to view super-species as well as sub-species. In future work we will investigate the gaze behavior by focusing on over two fine-grained images.

Acknowledgements. This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, Germany. We would like to thank Semih Korkmaz for his helpful insights.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label embedding for attribute-based classification. In *CVPR*, 2013.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *TPAMI*, 2015.

[3] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[4] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. *CVPR*, 2009.

[6] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.

[7] Y. Gong and S. Lazebnik. Comparing data-dependent and data-independent embeddings for classification and ranking of internet images. In *CVPR*, 2011.

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 19*, pages 545–552. MIT Press, 2007.

[9] M. X. Huang, J. Li, G. Ngai, and H. V. Leong. Screenglint: Practical, in-situ gaze estimation on smartphones. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2017.

[10] J. Hyönä, J. Tommola, and A.-M. Alaja. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, 48(3):598–612, 1995.

[11] M. S. L. F.-F. Jia Deng, Jonathan Krause. Leveraging the wisdom of the crowd for fine-grained recognition. *TPAMI*, in press.

[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009.

[13] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. Manjunath. From where and how to what we see. In *ICCV*, pages 625–632, 2013.

[14] S. Karthikeyan, T. Ngo, M. Eckstein, and B. Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *Proc. CVPR*, pages 3241–3250, 2015.

[15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[16] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. In *TPAMI*, 2013.

[17] J. Li, Y. Tian, T. Huang, and W. Gao. A dataset and evaluation methodology for visual saliency in video. In *ICME*, pages 442–445, 2009.

[18] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, Washington, DC, USA, 2014. IEEE Computer Society.

[19] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[20] M. Livne, L. Sigal, N. F. Troje, and D. J. Fleet. Human attributes from 3d pose tracking. *Computer Vision and Image Understanding*, 116(5), 2012.

[21] E. Marinoiu, D. Papava, and C. Sminchisescu. Pictorial Human Spaces. How Well do Humans Perceive a 3D Articulated Pose? In *ICCV*, 2013.

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

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[24] G. A. Miller. Wordnet: a lexical database for english. *CACM*, 38:39–41, 1995.

[25] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *Proc. ICCV*, pages 468–475, 2009.

[26] F. Moosmann, D. Larlus, and F. Jurie. Learning saliency maps for object categorization. 2006.

[27] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICCVGI*, 2008.

[28] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[29] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *ECCV*, 2014.

[30] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[31] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[32] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[33] Q. Qiu, Z. Jiang, and R. Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *ICCV*, 2011.

[34] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] M. Rohrbach, M. Stark, and B.Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[36] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *CVPR*, volume 2, 2004.

[37] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

[38] H. Sattar, S. Müller, M. Fritz, and A. Bulling. Prediction of search targets from fixations in open-world settings. In *CVPR*, pages 981–990, 2015.

[39] I. Shcherbatyi, A. Bulling, and M. Fritz. GazeDPM: Early Integration of Gaze Information in Deformable Part Models. arxiv:1505.05753, 2015.

[40] B. Siddiquie, R. Feris, and L. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.

[41] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[42] R. Subramanian, V. Yanulevskaya, and N. Sebe. Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements. In *MM*, pages 33–42, 2011.

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

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[45] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005.

[46] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009.

[47] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010.

[48] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, Caltech, 2010.

[49] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[50] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[51] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and F.-F. Li. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[52] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *CVPR*, pages 739–746, 2013.