Gaze Embeddings for Zero-Shot Image Classification

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

1 Max-Planck Institute for Informatics 2 University of Amsterdam

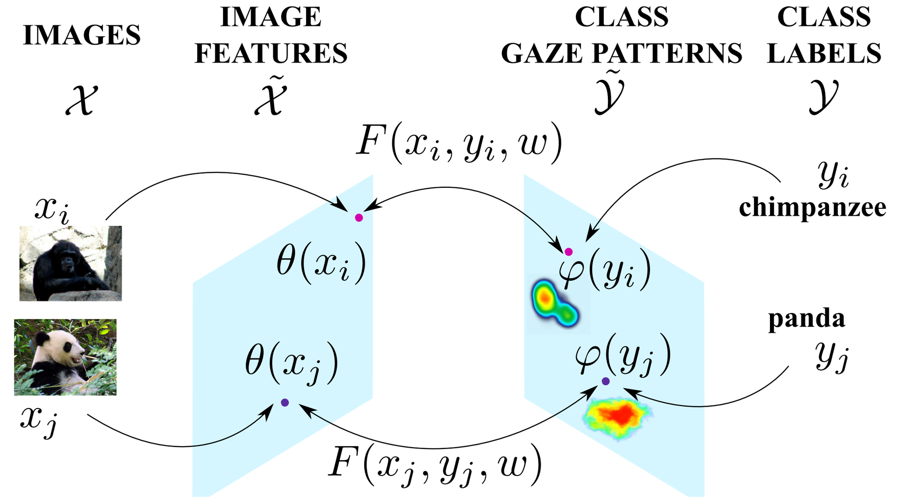
Zero-shot Learning

Lack of training data → zero-shot learning
Attributes aid classification → requires expensive domain expert annotation
Gaze information → novice users implicit annotation

Compatibility function (SJE [1])

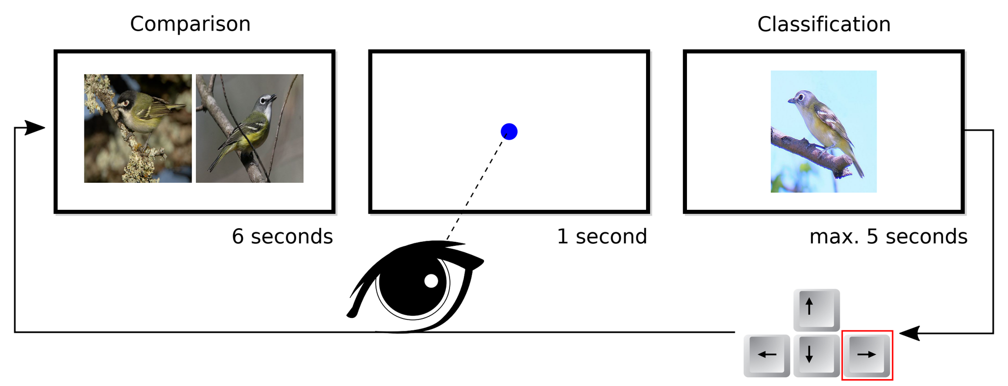
$$F(x, y; W) = \theta(x)^\top W \varphi(y)$$

θ DNN image features, φ class gaze embedding

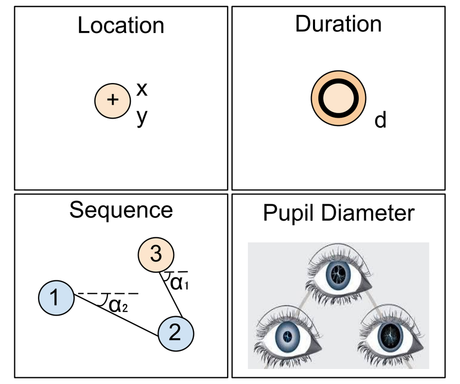


Gaze Collection and Datasets

Using Tobii TX300 remote eye tracker
- Example images of two classes (6 sec)
- Reset gaze position to center (1 sec)
- Classify class instance (max. 5 sec)



Gaze features
- Location x, y
- Duration d
- Scan path angles α_1, α_2
- Pupil diameter R

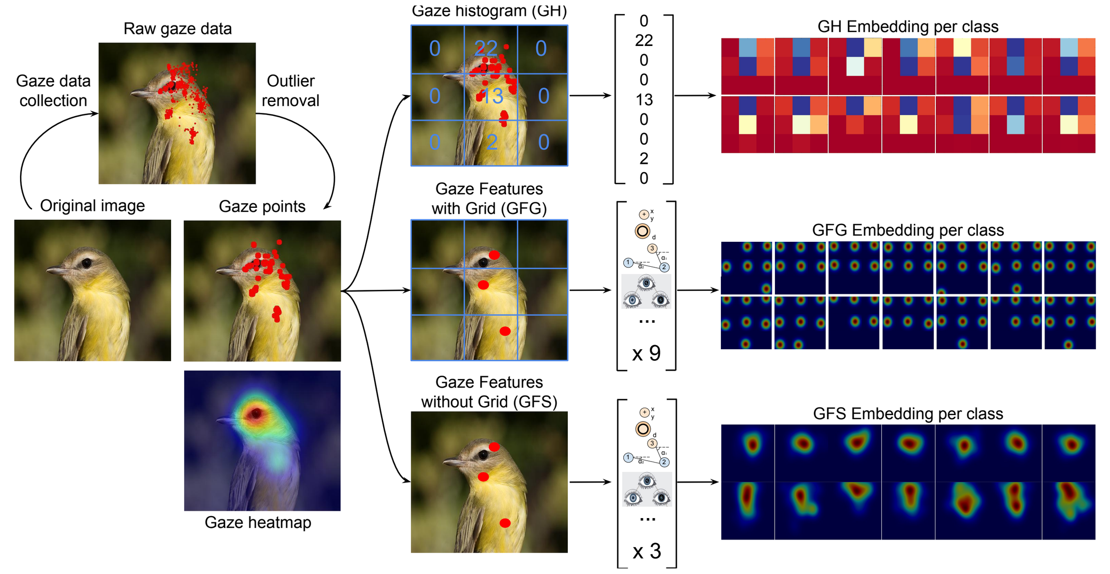


Dataset	# img/class	Gaze	Bubbles [2]
CUB-VW [4]	464/14	2320	210
CUB-VWSW [4]	2346/60	11730	900
PET [3]	720/24	3600	–

[1] Z. Akata et al. Evaluation of output embeddings for fine-grained image classification. In CVPR, 2015.
[2] J. Deng et al. Fine-grained crowdsourcing for fine-grained recognition. In CVPR, 2013.
[3] O. M. Parkhi et al. Cats and dogs. In CVPR, 2012.
[4] P. Welinder et al. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, Caltech, 2010.

* Nour Karessli currently works with EyeEm, Berlin.

Gaze Embeddings



Gaze Histogram (GH)
Count gaze points within spatial grid

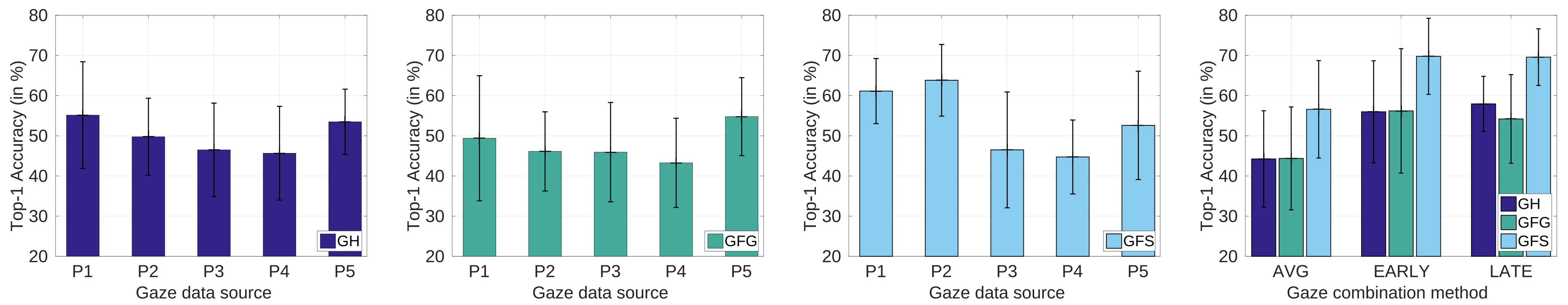
Gaze Features with Grid (GFG)
Average gaze features in each grid cell

Gaze Features with Sequence (GFS)
Concatenate gaze features of (k) points

Multiple participants → complementary information

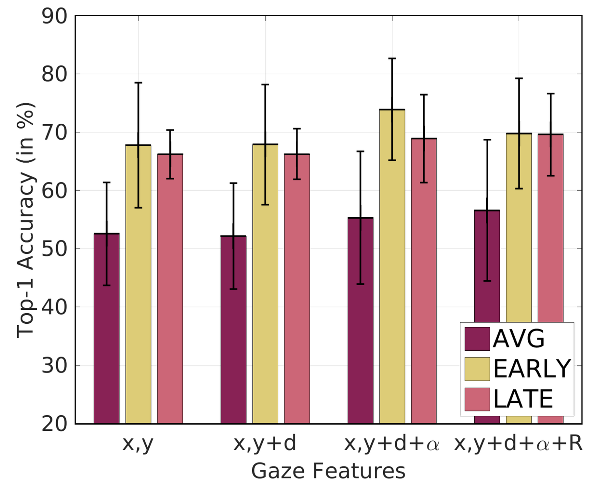
- AVG average participants per-class gaze embeddings
- $EARLY$ and $LATE$ fusion

Comparing Gaze Embeddings GFS = best→ Sequence helps more than spatial grid
Annotator bias → combining pariticipants embeddings improves performance



Analyzing Gaze Features
$EARLY$ performs the best using $[x, y, d, \alpha_1, \alpha_2]$

- As annotator becomes familiar with categories
- Pupil diameter R not helpful: Concentration drops



Comparing Gaze and Baselines
Bubble → novice users find distinctive regions
Gaze outperforms SoA → **class specific**

	Method	Accuracy
Baselines	Random points	39.5
	Bubbles [2]	43.2
	BoW from Wiki	55.2
SoA	Attributes	72.9
Ours	Gaze	73.9
	Attributes + Gaze	**78.2**

Ablation from Gaze to Bubbles
Gap is due gaze features+images quantity

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

Gaze Embeddings on Other Datasets
Gaze can be generalized to other domains

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