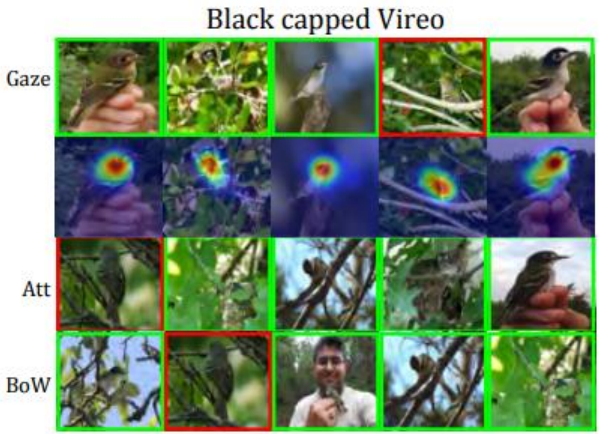
Gaze Embeddings for Zero-Shot Image Classification

Nour Karessli   Zeynep Akata   Bernt Schiele   Andreas Bulling

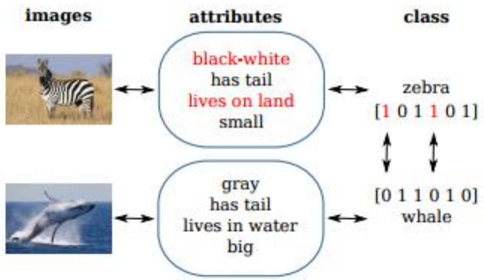


Presentation by Hsin-Ping Huang and Shubham Sharma

Introduction

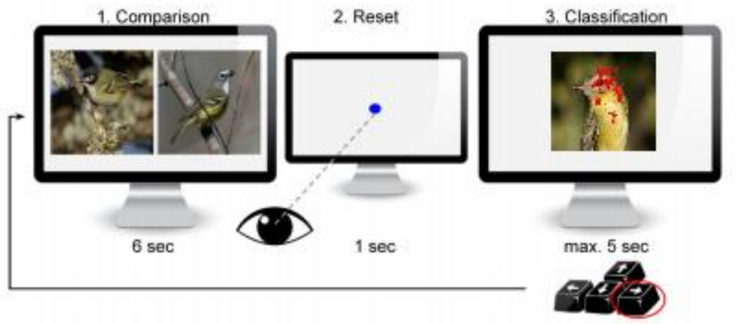
- Standard image classification models fail with the lack of labels.
- Zero-Shot Learning is a challenging task. Side information, e.g. attributes, is required.
- Several sources of side information exists: Attributes, detailed descriptions or gaze.
- Use gaze as the side information in this paper.

Attributes



Descriptions

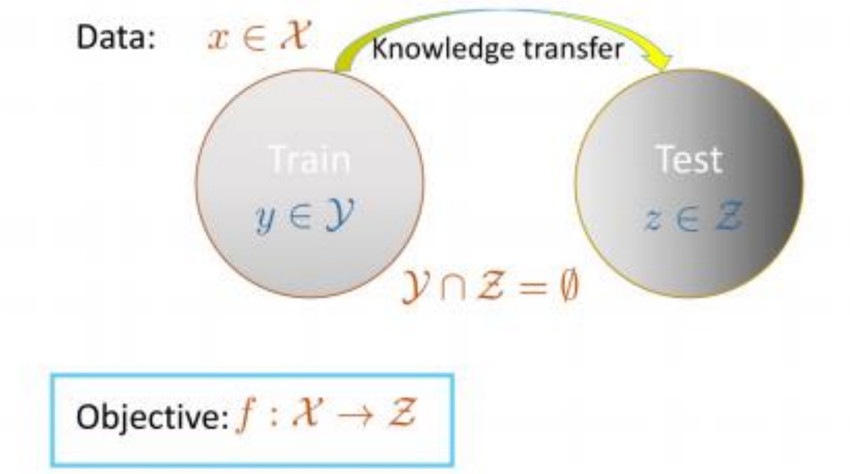
Gazes

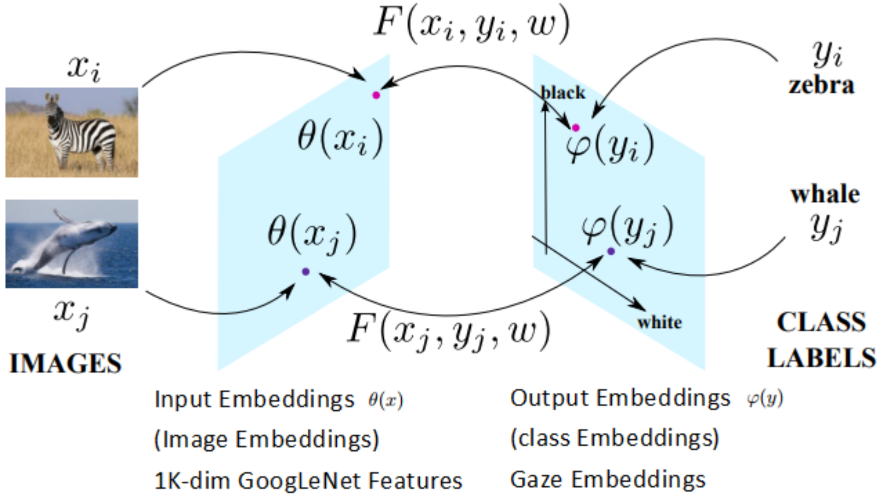


[Zero-shot learning tutorial, CVPR'17]

ZERO-SHOT LEARNING

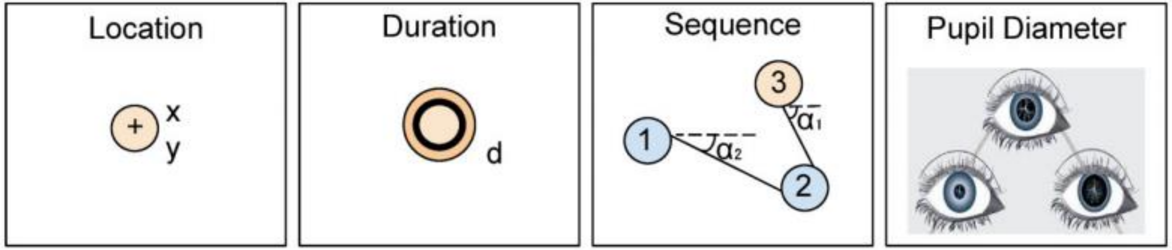
- Given training data and a disjoint test set, perform tasks such as object classification by mapping a function between the training data and test set.



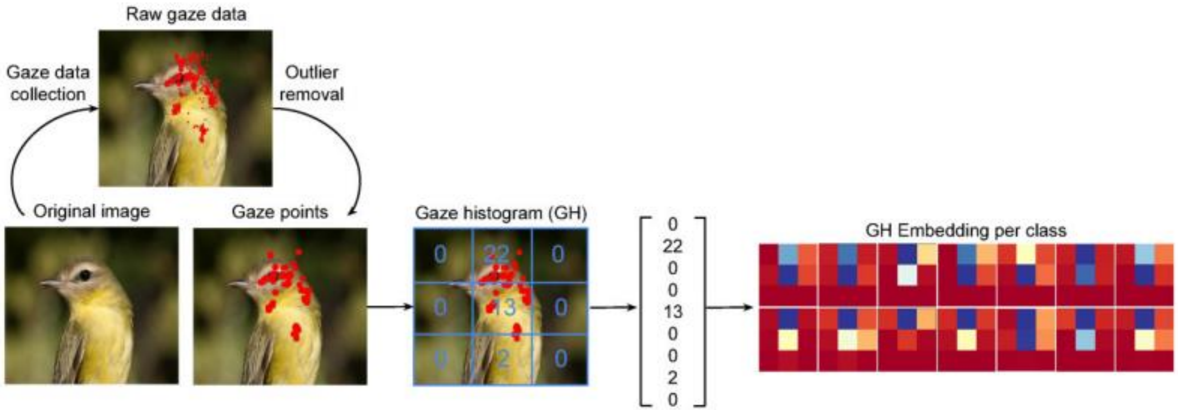


GAZE EMBEDDINGS

Gaze Features

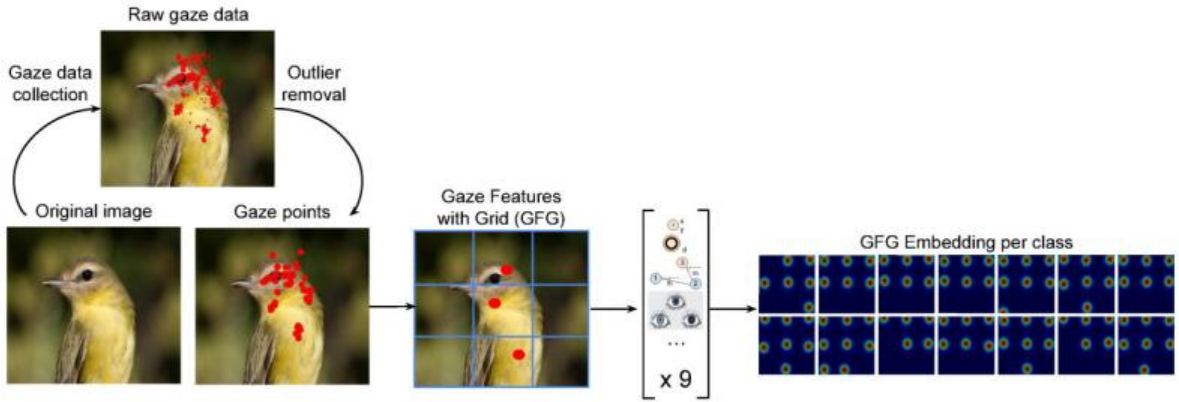


Gaze Histogram

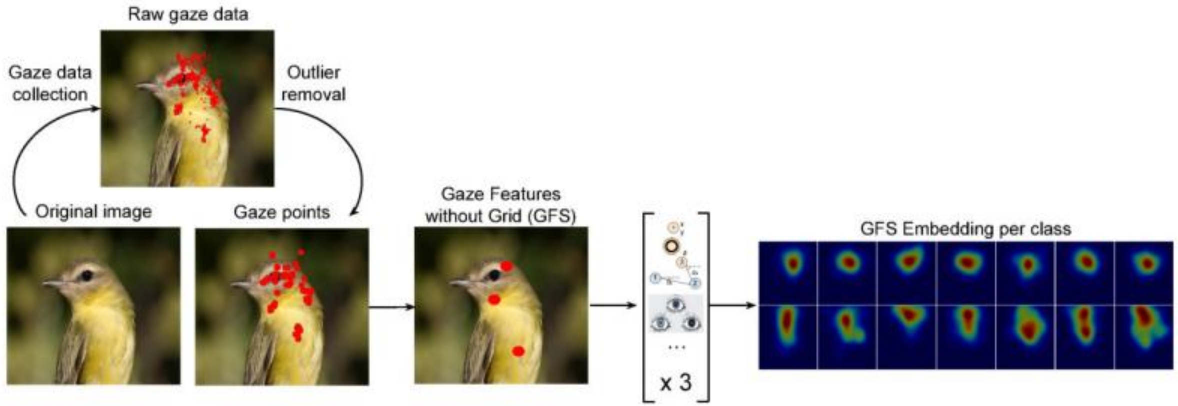


GAZE EMBEDDINGS

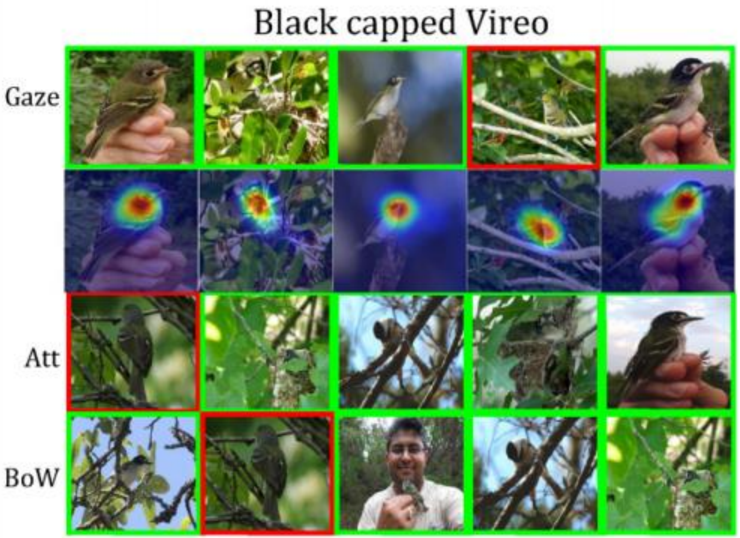
Gaze Features with Grid

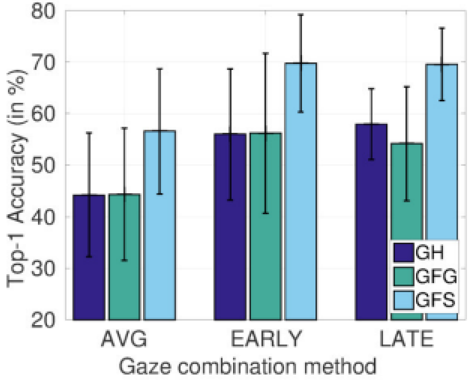


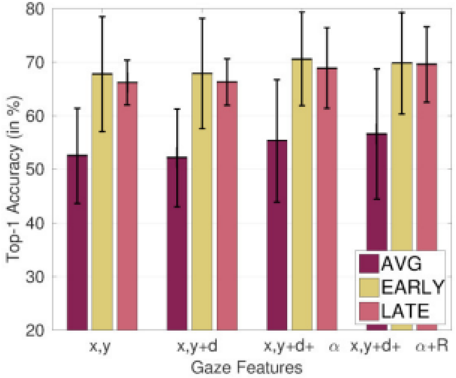
Gaze Features with Sequence



RESULTS OF THE PAPER







	CUB-VW
Random points	39.5
Bubbles [Deng et al. CVPR'13]	43.2
Bag of Words from Wikipedia	55.2
Attributes	72.9
Gaze	73.9
Attributes + Gaze	**78.2**

EXPERIMENTS

Dataset: CUB-VW

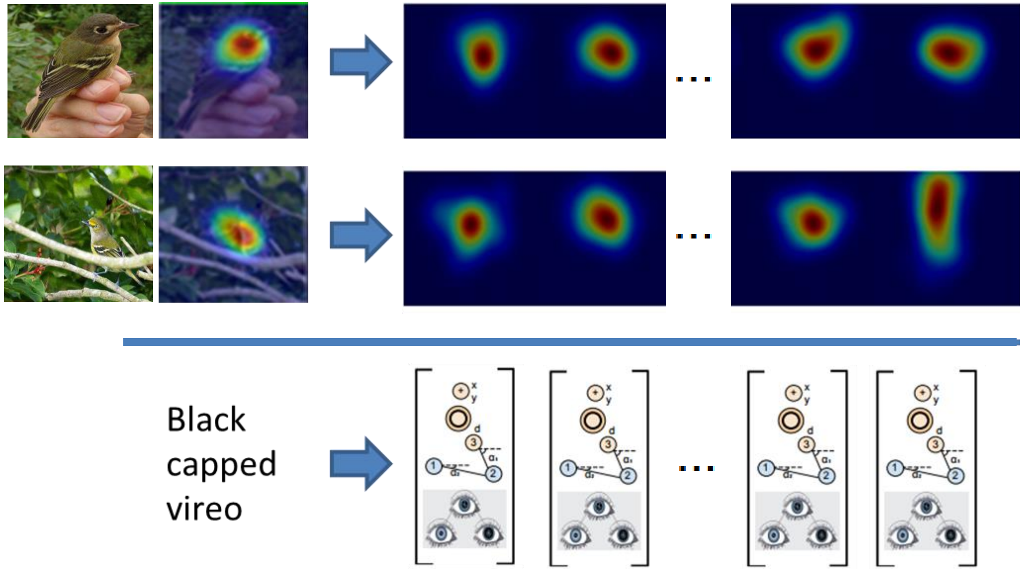
- 14 classes of Caltech-UCSD Birds 200-2010
- 10 different splits: 8/3/3 for train, validation and test classes
- Average **per-class top-1 accuracy**

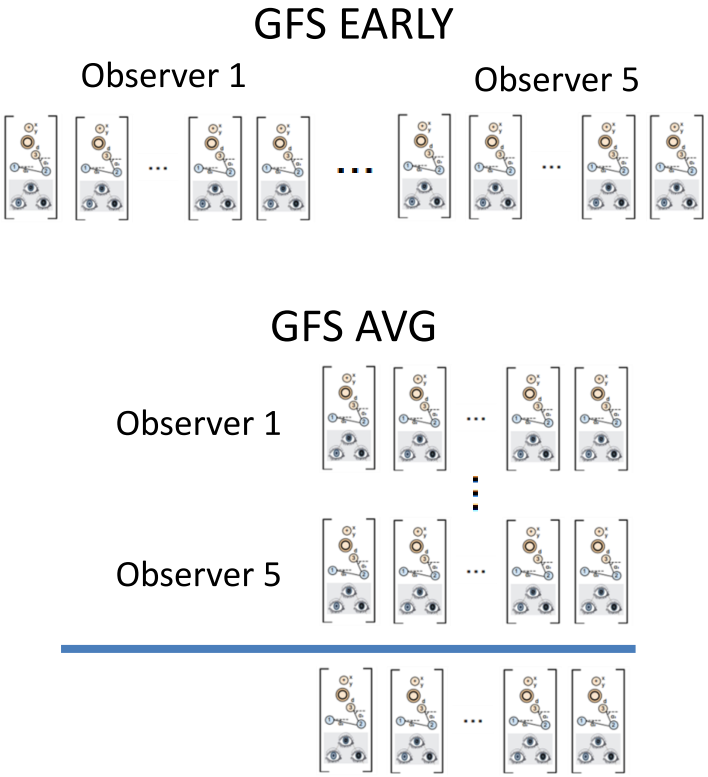
7 classes of Vireos

7 classes of Woodpeckers

Gaze Features with Sequence

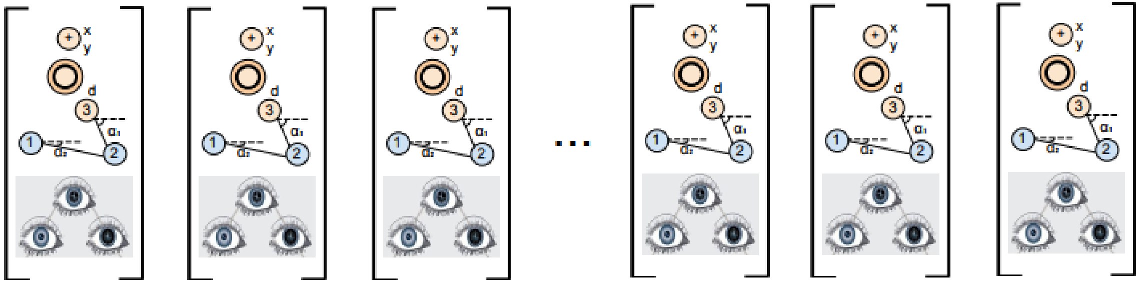
GFS of One Observer





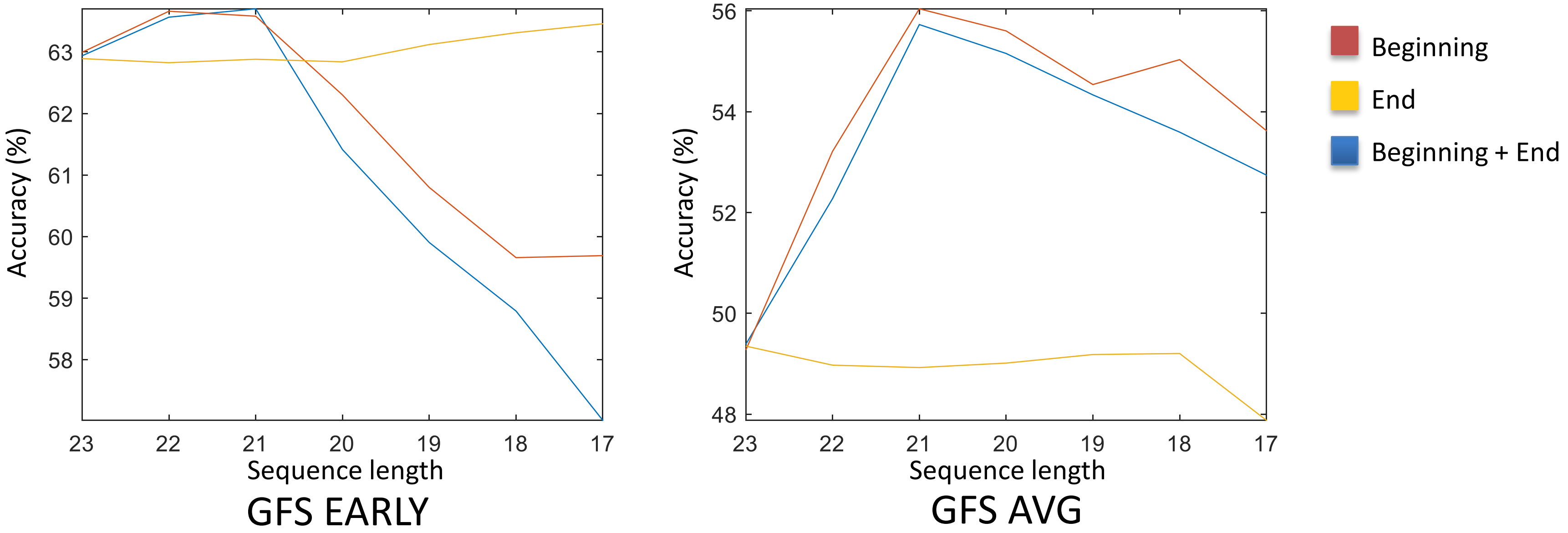
Experiment 1

- Gazes in the beginning contain less information because the observers just start viewing the image.
- Gazes in the end contain less information because the observers are tired or have done the observation.
- Ignore gazes in the beginning and the end.



Gaze Features with Sequence (GFS) of One Observer

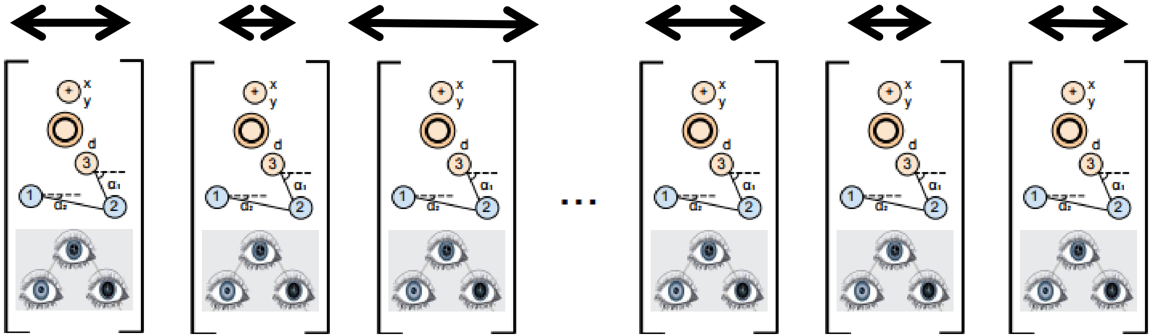
Experiment 1



- Ignoring gazes **in the beginning** yields better accuracy.
- Especially for AVG, the accuracy improves 6% when ignoring 2 gaze points.

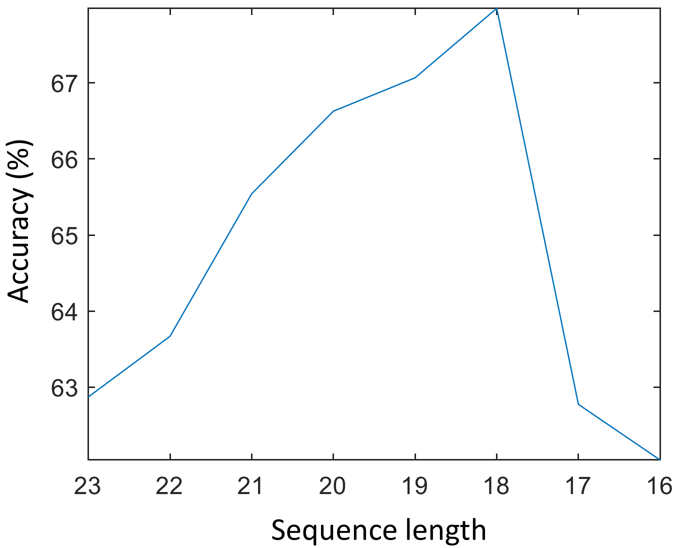
Experiment 2

- Gazes with shorter duration contain less information because those position are less salient in the image.
- Ignore gazes with shorter duration.

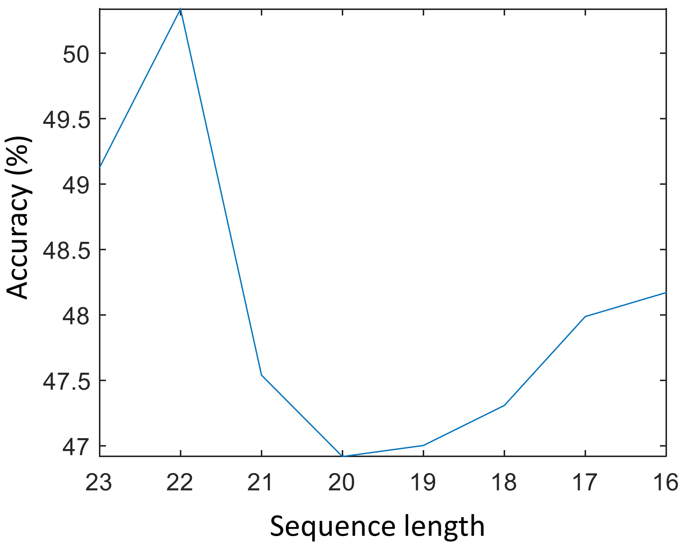


Gaze Features with Sequence (GFS) of One Observer

Experiment 2



GFS EARLY

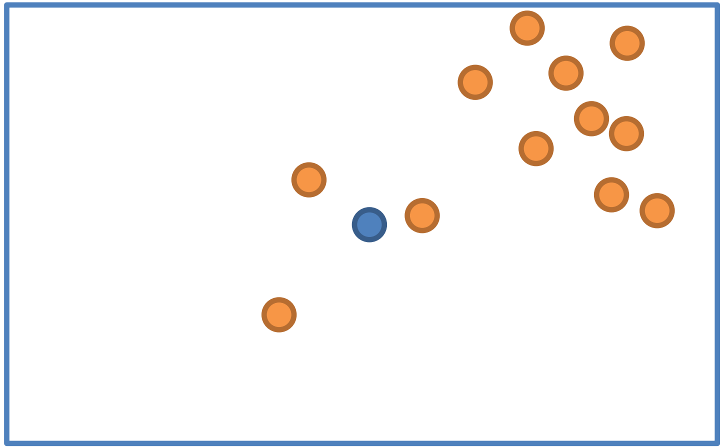


GFS AVG

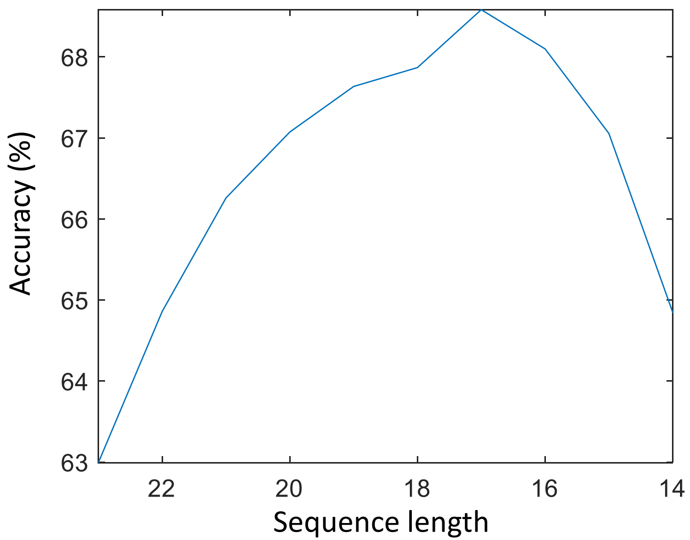
- Ignoring gazes **with shorter duration** yields better accuracy.
- Especially for EARLY, the accuracy improves 6% when ignoring 5 gaze points.

Experiment 3

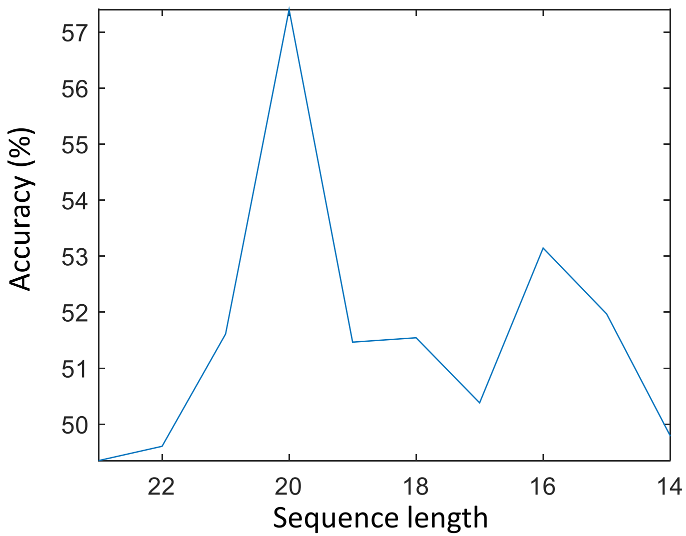
- Gazes close to the center contain less information because the observers have a tendency to look at the center.
- Ignore gazes close to the center of the image.



Experiment 3



GFS EARLY

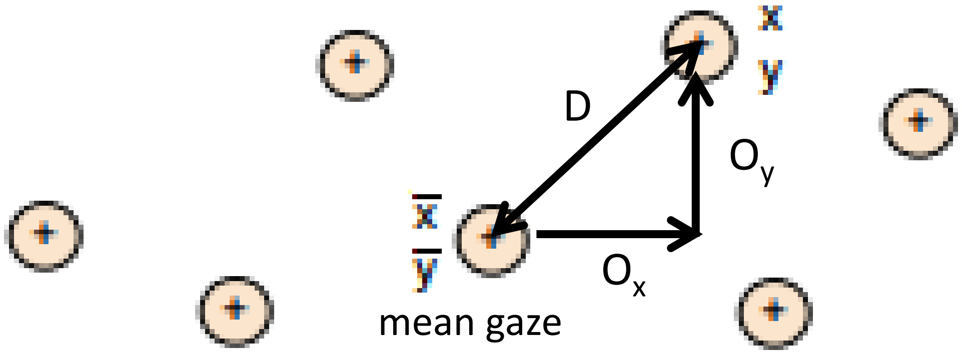


GFS AVG

- Ignoring gazes **close to the center** yields better accuracy.
- Especially for EARLY, the accuracy improves 5% when ignoring 6 gaze points.

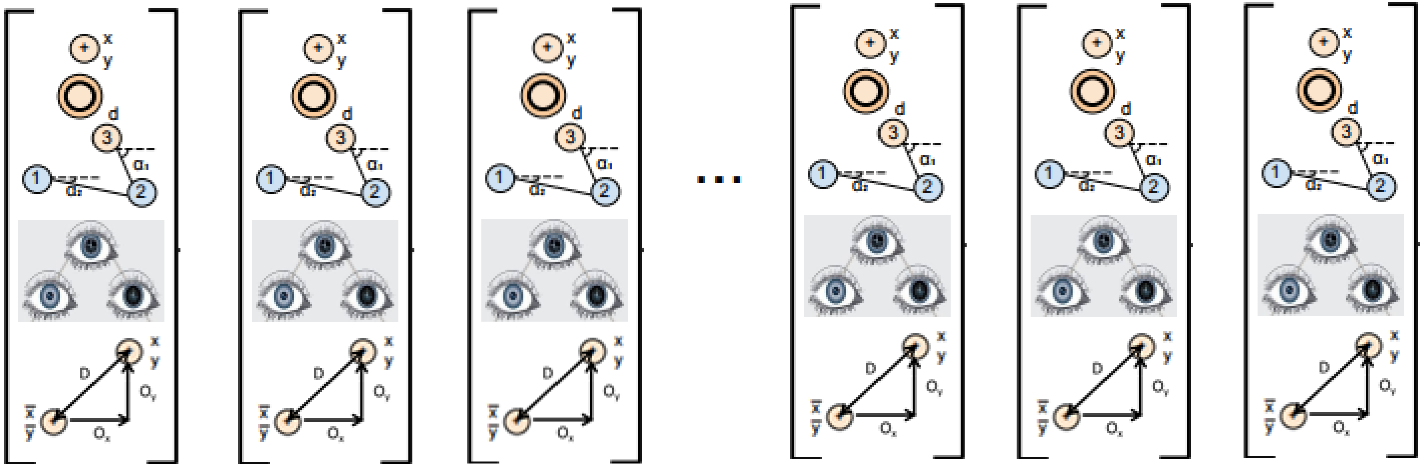
Experiment 4

- Not only the absolute positions, but also the offsets and distance between the mean gaze are informative.
  - Gazes have personal bias, each person have a different mean gaze.
  - The distribution of the gazes is important.
- Add the offsets and distance between the mean gaze as features.



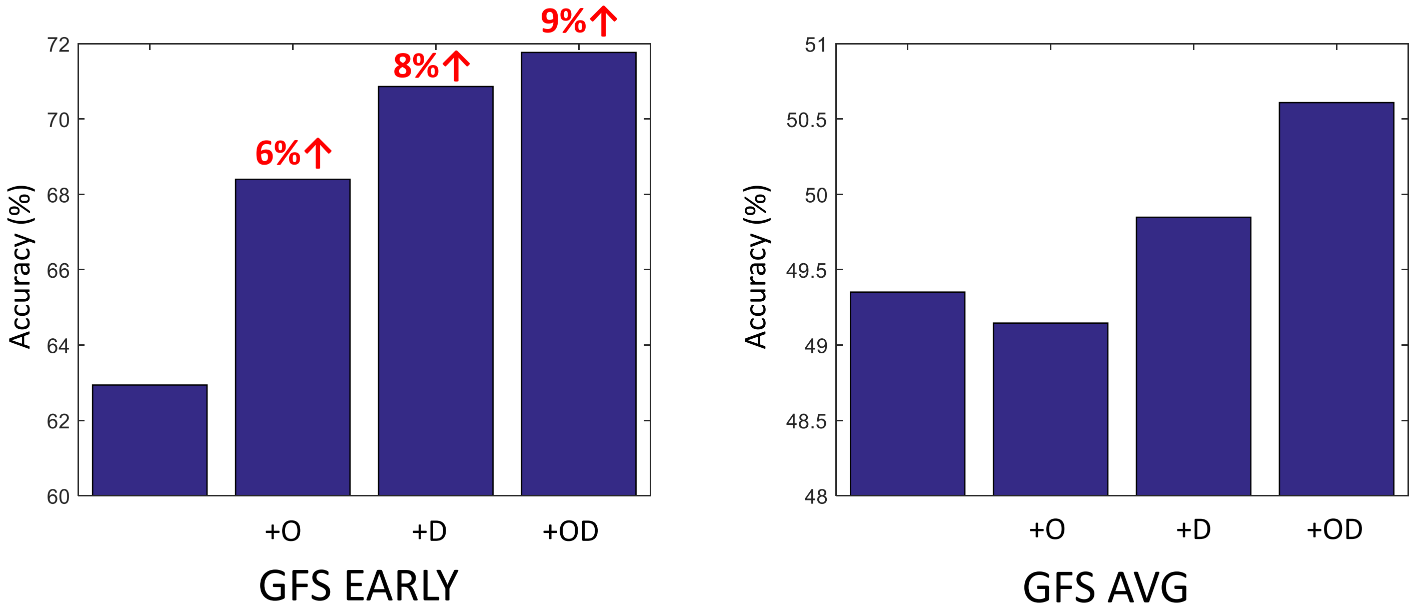
Experiment 4

- Add the offsets and distance between the mean gaze as features.



Gaze Features with Sequence (GFS) of One Observer

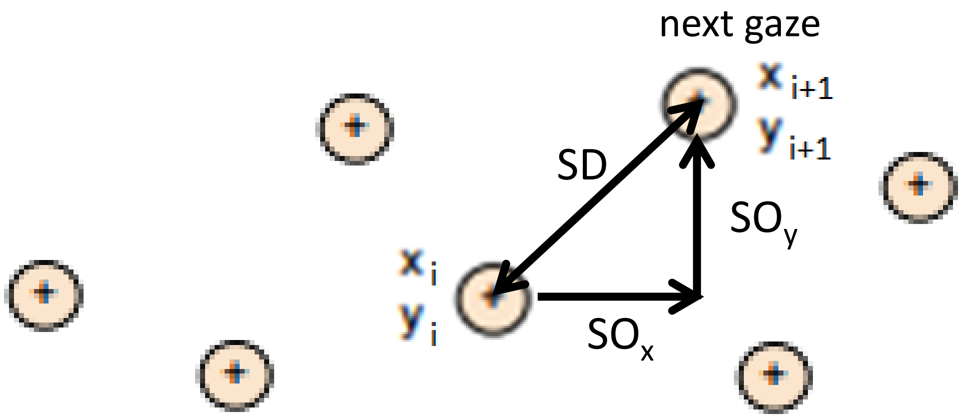
Experiment 4



- Adding ***the offsets and distance between the mean gaze*** yields better accuracy.

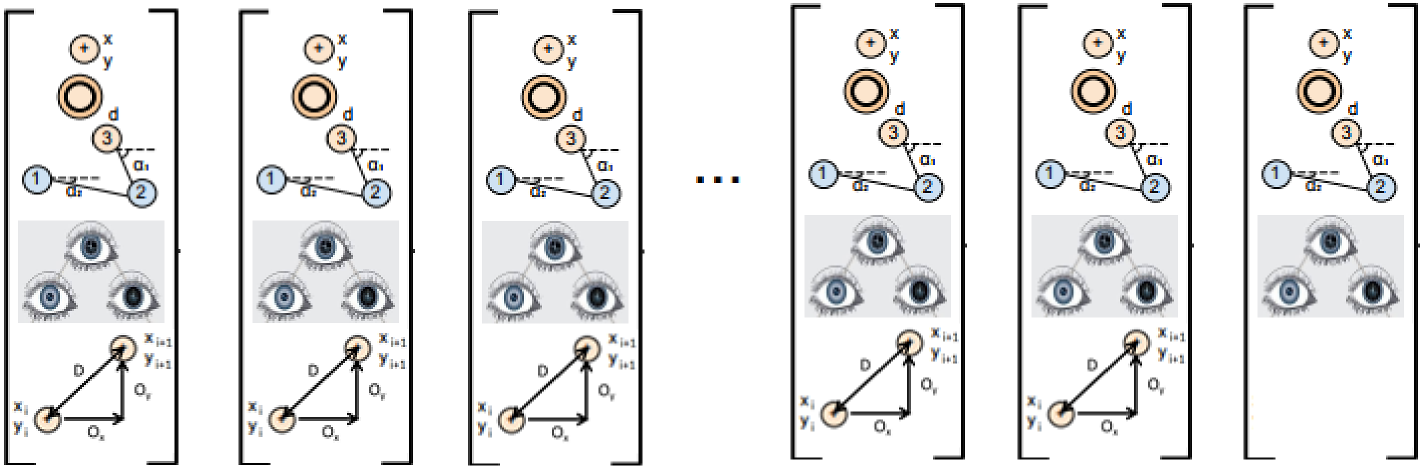
Experiment 5

- Not only the angles, but also the offsets and distance between two subsequent gazes are informative.
  - The saccade information is important.
- Add the offsets and distance between the subsequent gaze as features.



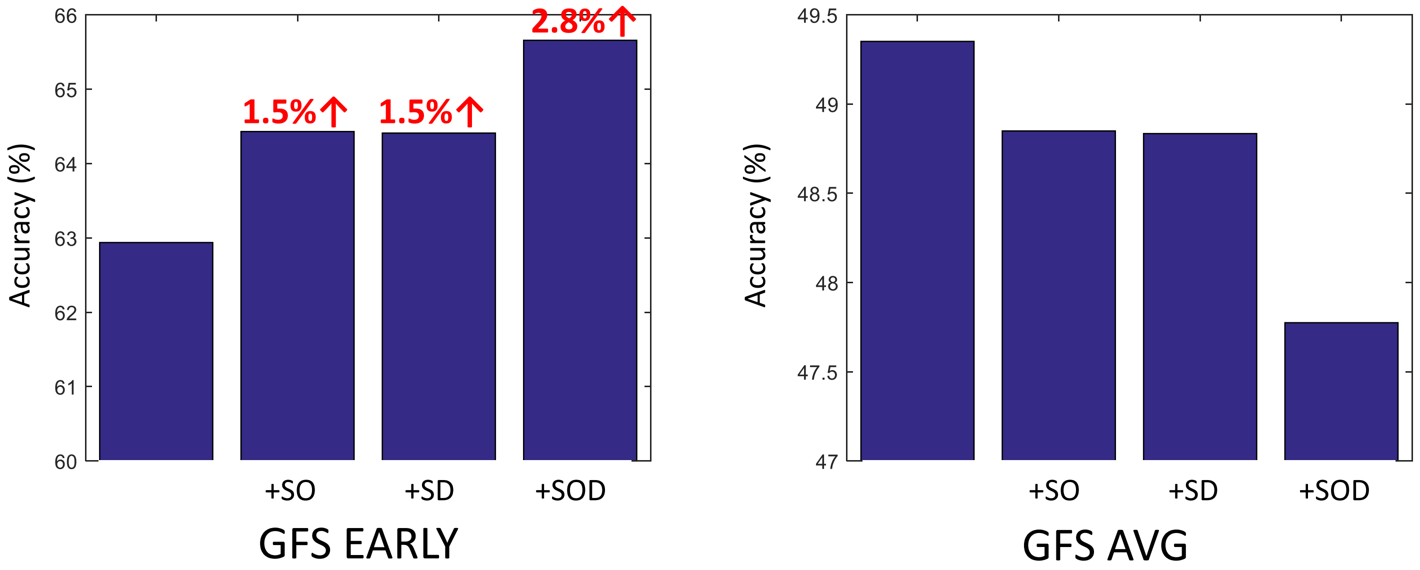
Experiment 5

- Add the offsets and distance between the subsequent gaze as features.



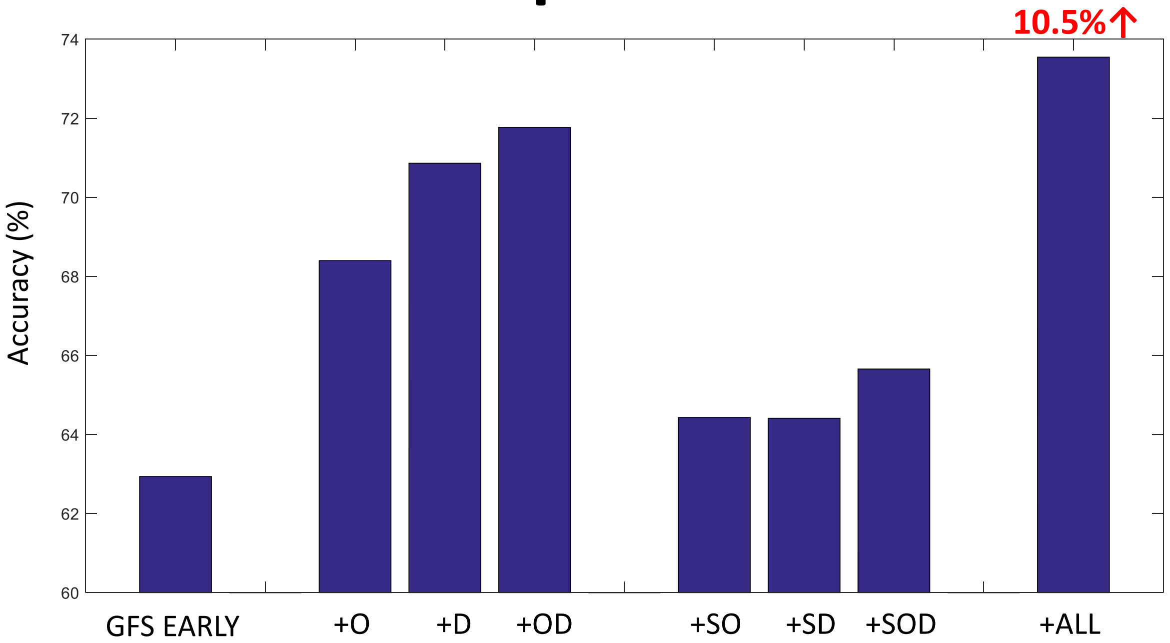
Gaze Features with Sequence (GFS) of One Observer

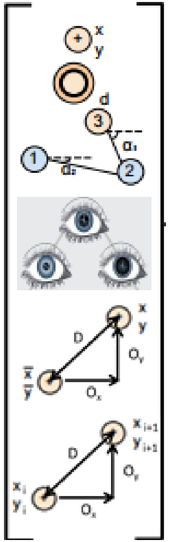
Experiment 5



- Adding ***the offsets and distance between the subsequent gaze*** yields better accuracy.

Experiment 5





- Adding ***the offsets and distance between the mean gaze*** and ***the subsequent gaze*** yields the best accuracy.

Experiment 6

- Use different zero-shot learning models.

Existing ZSL models can be grouped into 4:

1.Learning Linear Compatibility: ALE, DEVISE, **SJE**

2.Learning Nonlinear Compatibility: LATEM, CMT

3.Learning Intermediate Attribute Classifiers: DAP

4.Hybrid Models: SSE, CONSE, SYNC

Learning Linear Compatibility

Use bilinear compatibility function to associate

visual and auxiliary information

$$F(x, y; W) = \theta(x)^T W \phi(y)$$

SJE: Structured Joint Embedding

Gives full weight to the top of the ranked list

$$[\max_{y \in \mathcal{Y}^{tr}} (\Delta(y_n, y) + F(x_n, y; W)) - F(x_n, y_n; W)]_+$$

[Akata et al. CVPR’15 & Reed et al. CVPR’16]

Experiment 6

Hybrid Models

Express images and semantic class embeddings as a mixture of seen class proportions

CONSE: Convex Combination of Semantic Embeddings

Learns probability of a training image belonging to a class

Uses combination of semantic embeddings to classify

$$f(x,t) = \operatorname*{argmax}_{y \in \mathcal{Y}^{tr}} p_{tr}(y|x)$$

$$\frac{1}{Z}\sum_{i=1}^{T} p_{tr}(f(x,t)|x).s(f(x,t))$$

[Norouzi et al. ICLR'14]

SSE: Semantic Similarity Embedding

Leverages similar class relationships

Maps class and image into a common space

$$\operatorname*{argmax}_{u \in \mathcal{U}} \pi(\theta(x))^T \psi(\phi(y_u))$$

[Zhang et al. CVPR'16]

SYNC: Synthesized Classifiers

Maps the embedding space to a model space

Uses combination of phantom class classifiers to classify

$$\min_{w_c, v_r} \|w_c - \sum_{r=1}^{R} s_{cr} v_r\|_2^2.$$

[Changpinyo et al. CVPR'16]

Experiment 6

Gazes	
Method	Accuracy (%)
SJE	62.9
SSE	60.6
CONSE	63.7
SYNC	62.2

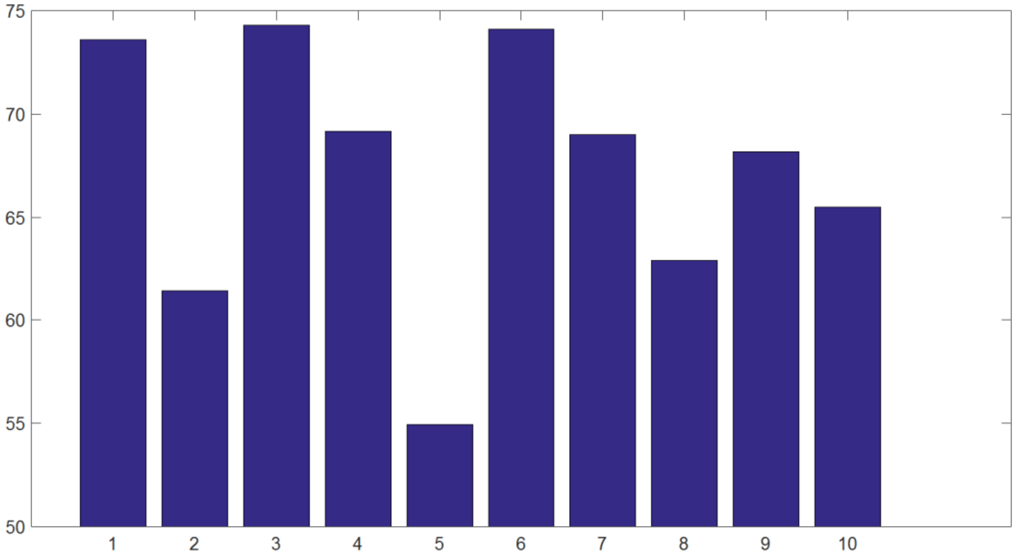
Attributes	
Method	Accuracy (%)
SJE	53.9
SSE	43.9
CONSE	34.3
SYNC	55.6

[Xian et al. CVPR’17]

- Using ***different zero-shot learning models*** yields similar accuracy for gaze embeddings.

Experiment 7

- Check the contribution of every participant to check if they contain complimentary information.



1: (1,2,3,4,5)
2: (4,5)
3: (1,2,3,4)
4: (1,2,3,5)
5: (5)
6: (1,2,4,5)
7. (1,2,3)
8. (1)
9. (1,2)
10. (1,3)

Failure Cases

- Birds are small or not salient in the pictures

- Birds have very different poses

CONCLUSIONS

- Using gaze embeddings for object recognition can be improved by processing the gaze data.
- The zero-shot model used in the paper works better when we think about either gaze or attributes.
- Not all participants necessarily contribute complimentary information.