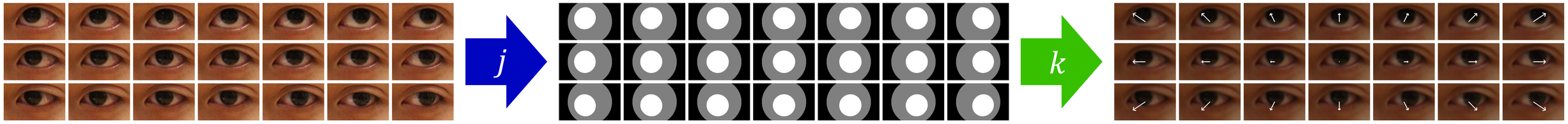
Deep Pictorial Gaze Estimation

Seonwook Park, Adrian Spurr, Otmar Hilliges

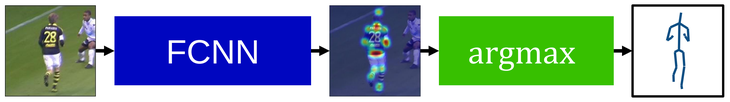
Advanced Interactive Technologies

ETH zürich

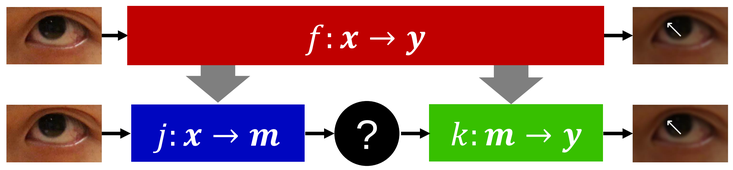


Motivation

Human Pose Estimation can be broken down into the easier task of heatmap regression and a simple argmax for coordinate extraction.



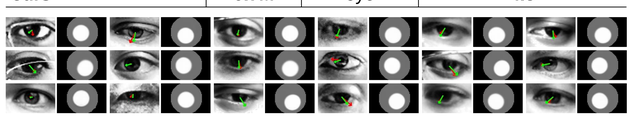
Similarly for **Eye Gaze Estimation** we would like to break down the task ($f: \boldsymbol{x} \rightarrow \boldsymbol{y}$) into two simpler mappings $j: \boldsymbol{x} \rightarrow \boldsymbol{m}$ and $k: \boldsymbol{m} \rightarrow \boldsymbol{y}$ such that $f = k \circ j$, where j and k are easier to learn than f.



Cross-person Gaze Estimation Results

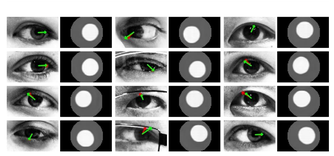
MPIIGaze (15 fold)

		# params	Inputs	Mean Test Error (deg)
kNN	[Zhang '15]	0	eye + head	7.2
RF	[Sugano'14]	-	eye + head	6.7
LeNet-5	[Zhang '15]	1.8M	eye + head	6.3
AlexNet		86M	eye	5.7
GazeNet	[Zhang '18]	90M	eye + head	5.5
VGG-16		158M	eye	5.4
ours		**0.7M**	eye	**4.5**



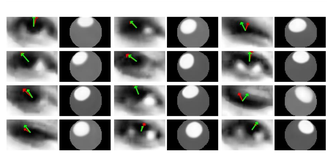
Columbia Gaze (5 fold)

	Mean Test Error (deg)
AlexNet	4.2
VGG-16	3.9
ours	**3.8**



EYEDIAP (5 fold)

	Mean Test Error (deg)
AlexNet	11.5
VGG-16	11.2
ours	**10.3**



(green: predicted, red: ground-truth)

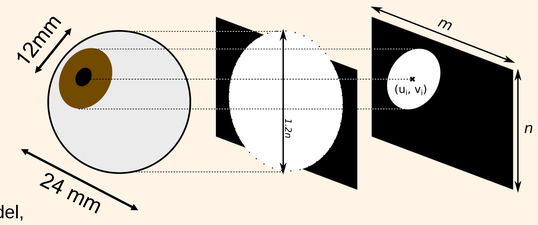
Method

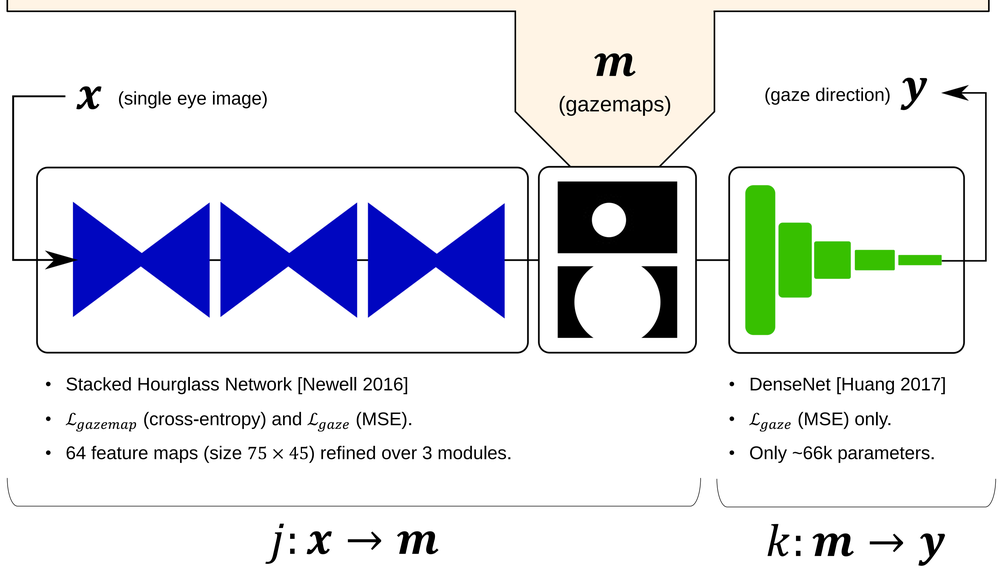
Pictorial Representation (gazemaps)

The human eyeball can be approximated by a perfect sphere (eyeball) and a circle intersecting its surface (iris).

A typical human has an eyeball diameter of 24mm, and iris diameter of 12mm. [Bekerman 2014, Forrester 2015]

Using this information, we form our eyeball model, rotate the eyeball to the required gaze direction, then project the eyeball and iris to yield two silhouettes. We call these binary maps "***gazemaps***".





- Stacked Hourglass Network [Newell 2016]
- $\mathcal{L}_{gazemap}$ (cross-entropy) and $\mathcal{L}_{gaze}$ (MSE).
- 64 feature maps (size 75×45) refined over 3 modules.

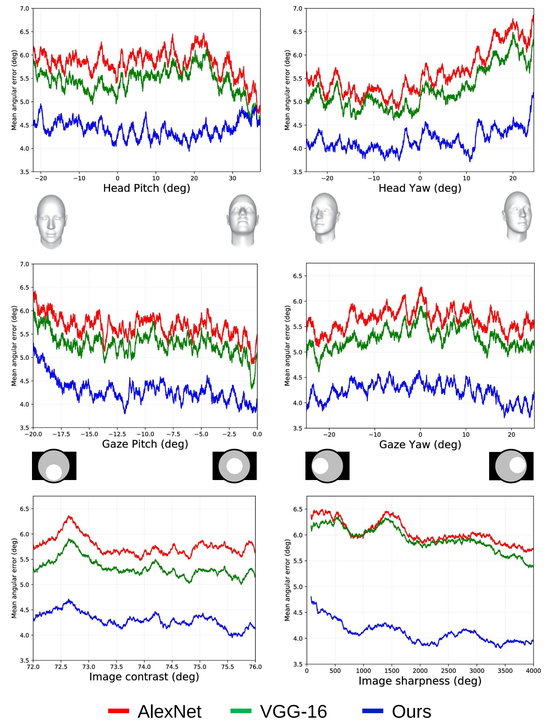
- DenseNet [Huang 2017]
- $\mathcal{L}_{gaze}$ (MSE) only.
- Only ~66k parameters.

$j: \boldsymbol{x} \rightarrow \boldsymbol{m}$ $\quad$ $k: \boldsymbol{m} \rightarrow \boldsymbol{y}$

Effect of Gazemap Supervision

Dataset	$\mathcal{L}_{gazemap}$	
	Without	With
MPIIGaze	4.67	**4.56**
Columbia	3.78	**3.59**
EYEDIAP	11.28	**10.63**

Robustness Evaluations



Conclusion and Future Work

- We can learn to predict gazemaps from challenging webcam-based images of single eyes.
- The first time in appearance-based gaze estimation that a fully-convolutional architecture is used.
- 1-degree (18%) improvement over [Zhang 2018] on the popular MPIIGaze cross-person gaze estimation benchmark.
- Alternative gazemaps and architectures should be explored.

Acknowledgements

This work was supported in part by ERC Grant OPTINT (StG-2016-717054). We thank NVIDIA for the donation of GPUs used in this work.

erc