Few-Shot Adaptive Gaze Estimation

Seonwook Park\textsuperscript{1,2}, Shalini De Mello\textsuperscript{1*}, Pavlo Molchanov\textsuperscript{1}, Umar Iqbal\textsuperscript{1}, Otmar Hilliges\textsuperscript{2}, Jan Kautz\textsuperscript{1}

\textsuperscript{*}The first two authors contributed equally.

Overview

- We cast few-shot personalization as a meta-learning problem, where each person is a task in the meta-learning sense.
- We use MAML [Finn et al., ICML 2017] to yield a meta-learner (Adaptable Gaze Estimation Network - AdaGEN) via direct optimization of the within-person generalization error.
- We better leverage the subject-diversity of the large GazeCapture training set (993 subjects used in training).

Results

- MAML is better than naïve few-shot fine-tuning and does not suffer from over-fitting.
- MAML and DT-ED benefit with more training subjects (993 in GazeCapture vs 15 in MPIIGaze).
- Within-person consistency is important. Maximizing between-person differences is not beneficial.

- We do better than MAML applied to CNN features where the CNNs are trained directly for gaze estimation only.
- We out-perform state-of-the-art person-specific methods consistently and over all k values with lower variation in performance.

Overall, we show greater improvement compared to all prior art, and out-perform [Yu et al., CVPR 2019] even with 1 calibration sample.

Conclusions

- Our DT-ED learns a compact, rotation-equivariant representation of gaze.
- Learning a Few-Shot learner yields better performance than naïve fine-tuning or hand-designed personalization functions.
- FAZE can apply to other personalization problems such as gesture recognition and affective state estimation.

Acknowledgements

Seonwook Park carried out this work during his internship at Nvidia. This work was supported in part by the ERC Grant OPTINT (StG-2016-717054).

Source Code

github.com/NVLabs/few_shot_gaze

Motivation

- Large performance gap between empirical lower bound and state-of-the-art cross-person gaze estimation methods.
- We need to consider person-specific factors (below) while requiring as few calibration samples as possible.

Few-Shot Adaptive Gaze Estimation (FAZE) Framework

- Via a novel disentangling transforming encoder-decoder (DT-ED) architecture.
- Using novel loss terms for a) embedding consistency within a subject, (b) gaze estimation, and (c) image reconstruction with transformed gaze/pose.
- The learned gaze direction and head orientation representations are:
  - Rotationally equivariant to eyeball / head rotation
  - Disentangled from head / eyeball rotations respectively
  - Compact & task-specific

REPRESENTATION LEARNING

- We cast few-shot personalization as a meta-learning problem, where each person is a task in the meta-learning sense.
- We use MAML [Finn et al., ICML 2017] to yield a meta-learner (Adaptable Gaze Estimation Network - AdaGEN) via direct optimization of the within-person generalization error.
- We better leverage the subject-diversity of the large GazeCapture training set (993 subjects used in training).

META LEARNING

- We cast few-shot personalization as a meta-learning problem, where each person is a task in the meta-learning sense.
- We use MAML [Finn et al., ICML 2017] to yield a meta-learner (Adaptable Gaze Estimation Network - AdaGEN) via direct optimization of the within-person generalization error.
- We better leverage the subject-diversity of the large GazeCapture training set (993 subjects used in training).

Few-Shot Adaptive Gaze Estimation (FAZE) Framework

- Via a novel disentangling transforming encoder-decoder (DT-ED) architecture.
- Using novel loss terms for a) embedding consistency within a subject, (b) gaze estimation, and (c) image reconstruction with transformed gaze/pose.
- The learned gaze direction and head orientation representations are:
  - Rotationally equivariant to eyeball / head rotation
  - Disentangled from head / eyeball rotations respectively
  - Compact & task-specific

ACKNOWLEDGEMENTS

Seonwook Park carried out this work during his internship at Nvidia. This work was supported in part by the ERC Grant OPTINT (StG-2016-717054).