Full-Face Appearance-Based Gaze

Links

- Website: https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/gaze-based-human-computer-interaction/its-written-all-over-your-face-full-face-appearance-based-gaze-estimation/
- Paper: https://perceptual.mpi-inf.mpg.de/files/2017/11/zhang_cvprw2017.pdf

Notes

- Goal: full-face approach for 2D and 3D appearance-based gaze estimation (as opposed to only eyes)
- Encodes the face image using a CNN with spatial weights applied on the feature maps to flexibly suppress or enhance information in different facial regions
    - Spatial weights mechanism -- learns spatial weights on the activation maps of the convolutional layers, reflecting the information contained in different facial regions
    - Spatial weights network facilitates the learning of estimators that are robust to significant variation in illumination conditions as well as head pose and gaze directions
- Appearance-based
- Other regions of the face beyond the eyes may contain valuable information for gaze estimation
    - Uses additional layers that learn spatial weights for the activation of the last convolution layer
        - There could be some image regions that do not contribute to the gaze estimation task such as background regions, and activations from such regions have to be suppressed for better performance
        - Activations from other facial regions (besides the eyes) are expected to be subtle
    - Introduced a mechanism that forces the network explicitly to learn and understand that different regions of the face can have different importance for estimating gze for a given test sample
- Spatial weights mechanism: 3 additional convolutional layers with filter size 1x1 followed by a rectified linear unit layer
- Baseline CNN architecture: AlexNet (5 convolution layers and 2 fully connected layers); trained an additional linear regression layer on top of the last fully connected layer
- 2D gaze estimation - input face images were cropped according to the six facial landmark locations (4 eye corners and 2 mouth corners)

- 3D gaze estimation: fit the generic 3D face model provided with MPIIGaze to the landmark locations to estimate the 3D head pose
- Results
    - All methods that take full-face information as input significantly outperformed the single eye baseline
    - Performance was further improved by incorporating the spatial weights network
    - Performance gap greater for 3D than 2D
    - More robust to facial appearance variation caused by extreme head pose, gaze directions, and illumination

OpenGaze

Links

- GitHub: https://github.molgen.mpg.de/perceptual/opengaze
- Paper: https://arxiv.org/pdf/1901.10906.pdf

Notes

- First software toolkit for appearance-based gaze estimation and interaction
- Potential of appearance-based gaze estimation: attentive user interfaces, mobile gaze interaction, social signal processing
- Why appearance-based gaze estimation not widely used
    - Remains unclear how they perform compared to dominant, special-purpose eye tracking equipment
        - Haven't been compared with other families of gaze estimation methods in a principled way due to different requirements in terms of hardware and deployment setting
    - Methods remain challenging for user interface and interaction designers
        - Available code written by computer vision experts for evaluation purposes -- not optimized for real-time use, doesn't implement all functionality required for interactive applications in a single pipeline, or cannot be easily extended or integrated into other software or user interface frameworks
- Comparisons between different families of gaze estimation methods
    - Varying distance between user and camera (best to worst): feature-based, appearance-based, model-based
        - Appearance-based has a much larger operational range, so has more practical usefulness (feature-based works best within a certain range)

    - Number of calibration samples vs gaze estimation accuracy -- calibration with a large number of samples can be time consuming and prohibitive for certain applications where spontaneous interaction is crucial
        - Current appearance-based methods can achieve accuracy competitive with feature-based (even for few calibration samples -- e.g. 4)
    - Indoor and outdoor settings (different illumination conditions) (best to worst): feature-based, appearance-based, model-based
        - Better accuracy tends to be achieved for the indoor environment
    - With and without glasses (due to strong reflections and distortions they may cause)
        - Have a stronger effect on gaze estimation accuracy than illumination conditions
        - Estimation results differences are larger for appearance-based methods than for feature-based
            - Possible causes: training data does not contain a sufficient number of images of people wearing glasses (for this experiment), or since Tobii EyeX (feature-based) uses infrared light, which filters out most reflections on the glasses
- OpenGaze Software Toolkit
    - Input can be single RGB images
    - First detects faces and facial landmarks, which are used to estimate 3D head pose and data normalization. Cropped face image will be input to the appearance-based gaze estimation method
        - Data normalization: crops face image with a normalised camera to cancel out some of appearance variations caused by head pose
    - Output: gaze direction in the camera coordinate system, which can be further projected to the screen coordinate system
    - Face and facial landmark detection
        - Integrates OpenFace for facial landmark detection that, in turn, relies on dlib to detect faces
        -
    - Data normalization
        - Data normalization - used to efficiently train appearance-based gaze estimators, and cancels out geometric variability by warping input images to a normalized space
        - First crops the face image after rotating the camera so the x-axis of the camera coordinate system is perpendicular to the y-axis of the head coordinate system. Then it scales the image so that the normalized camera is located at a fixed distance away from the face center.
        - → input image has only 2 degrees of freedom in head pose
    - Gaze estimation
        - Appearance-based
        - Whole face image fed into the CNN to output 3D gaze directions

        - Neural network architecture -- based on the AlexNet architecture, pretrained on two commonly-used gaze datasets with full-face images
    - Projection on screen
        - Projects gaze direction back to the original camera coordinate system
        - Provides APIs to project 3D gaze direction from camera coordinate system to the 2D screen coordinate system, and vice versa
            - Requires camera intrinsic parameters and camera-screen relationship
    - Personal calibration
        - Comes with pre-trained generic gaze estimator which works across users, environments, and cameras without any personal calibration
        - Further provides a personal calibration scheme to make corrections to raw gaze estimates from the appearance-based gaze estimation model
        - Provides a GUI to collect the personal calibration data from users
    - Implementation and speed
        - Components implemented as separate classes written in C++ with interfaces to communicate between each other
        - Most time-consuming process: face and facial landmark detection