Jitter-free Facial Landmark Tracking

Queried Learned Optimization for Tracking

Student

Dominik Gschwind

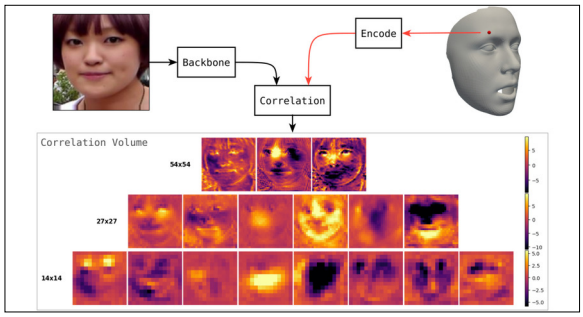
Introduction: Facial landmark detection is a vital component of many computer vision tasks. In the medical domain, specifically for smartphone-based eye gaze tracking, precision and temporal stability are paramount. However, existing mobile-oriented solutions often treat video as a stream of independent images, resulting in high-frequency signal jitter that degrades the reliability of subsequent gaze estimation.

Approach: To address this, we introduce Queried Learned Optimization for Tracking (QLOT), a new deep learning architecture tailored toward robust and stable facial landmark tracking on resource-constrained devices. QLOT bridges the gap between query-based detection and optical flow-inspired learned optimization. By adapting the recurrent update mechanism of RAFT (Recurrent All-Pairs Field Transforms), our model constructs a correlation volume from multiscale feature maps and iteratively refines landmark positions based on local and global semantics, as well as learned geometrical constraints.

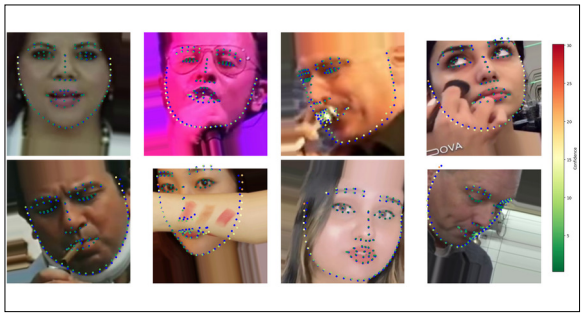
We design a lightweight feature pyramid network on top of HGNetV2 providing rich features. To target temporal stability, the model is trained on video sequences using various probabilistic losses to teach motion dynamics. Furthermore, we introduce an uncertainty aware gating mechanism, allowing the model to dynamically gauge confidence and modulate updates, thereby reducing jitter without most of the latency penalties associated with conventional filtering.

Result: As a first step we only test on accuracy and a specific temporal stability metric, where we are competitive despite the small model size. Although the initial results are promising, further work is needed to verify the temporal properties. Nonetheless, we can demonstrate that QLOT offers advanced keypoint detection and tracking capabilities within an end-to-end deep learning model.

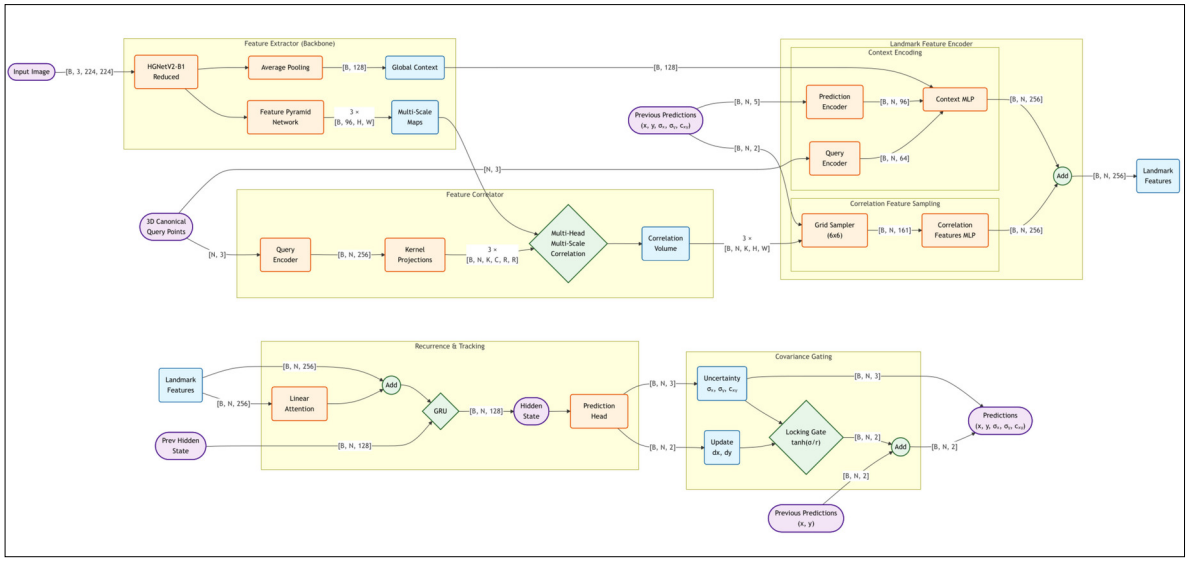
Conceptual flow for building the correlation volume from a single 3D query point.
Face image from Wider Facial Landmarks in the Wild dataset



Annotated faces, where blue is ground-truth, green-yellow was predicted.
Images from Wider Facial Landmark in the Wild Video Dataset



QLOT architecture where a single optimization iteration is shown.
Own presentment



Advisor
Prof. Dr. Martin Weisenhorn

Subject Area
Electrical Engineering, Data Science