Holistically Constrained Local Model: Going Beyond Frontal Poses for Facial Landmark Detection

KangGeon Kim[1]
kanggeon.kim@usc.edu

Tadas Baltrušaitis[2]
tbaltrus@cs.cmu.edu

Amir Zadeh[2]
abagherz@cs.cmu.edu

Louis-Philippe Morency[2]
morency@cs.cmu.edu

Gérard Medioni[1]
medioni@usc.edu

[1] Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA, USA

[2] Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Facial landmark detection is an essential initial step for a number of facial analysis research areas such as expression analysis, 3D face modeling, facial attribute analysis, and person recognition. It is a well researched problem that has seen a surge of interest in the past couple of years.

However, most state-of-the-art methods still struggle in the presence of extreme head pose, especially in challenging in-the-wild images. Furthermore, as most methods operate in a local manner [1, 2], they rely on good and consistent initialization, which is often very difficult to achieve. While some images attempt to combat this by evaluating a number of proposals and initializations, this comes at a computational cost.

In our work, we present a new model – Holistically Constrained Local Model (HCLM), which unifies local and holistic facial landmark detection by integrating head pose estimation, sparse-holistic landmark detection and dense-local landmark detection. Our method's main advantage is the ability to handle very large pose variations, including profile faces. Furthermore, our model integrates local and holistic facial landmark detectors in a joint framework, with a holistic approach narrowing down the search space for the local one.

For a given set of k facial landmark positions $\mathbf{x} = \{x_1, x_2, ..., x_k\}$, our HCLM model defines the likelihood of the facial landmark positions conditioned on a set of sparse landmark positions $X_s = \{x_s,\ s \in S\}$ ($|S| \ll k$) and image $\mathcal{I}$ as follows:

$$p(\mathbf{x}|I, X_s, \mathcal{I}) \propto p(\mathbf{x}) \prod_{i=1}^{k} p(x_i|X_s, \mathcal{I}). \qquad (1)$$

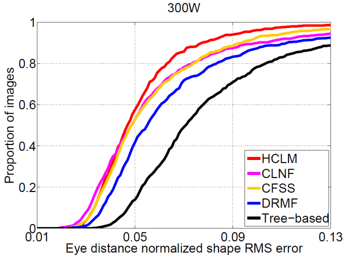

Figure 1: *Cumulative error curves on 300-W dataset.* Measured as the mean Euclidean distance from ground truth normalized by the inter-ocular distance. Note that we use 68 points for this comparison.

In Equation 1, $p(\mathbf{x})$ is prior distribution over set of landmarks $\mathbf{x}$ following a 3D point distribution model (PDM) with orthographic camera projection.

Some of the results comparing our HCLM model to state-of-the-art baselines can be seen in Figure 1. Our model demonstrates competitive or better performance to most of the baselines. Furthermore, HCLM demonstrates superior performance in especially difficult images, such as profile ones. This is due to the both better initializations and combination of *holistic* and *local* approaches of our model.

[1] Xuehan Xiong and Fernando Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[2] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.