Automatic Acquisition of High-fidelity Facial Performances Using Monocular Videos: Supplementary Material

Fuhao Shi* Hsiang-Tao Wu† Xin Tong† Jinxiang Chai*
*Texas A&M University †Microsoft Research Asia

1. Evaluation and Experiment

We first demonstrate the accuracy and robustness of our facial feature tracking algorithm by comparing against alternative methods. We then validate our method by evaluating the importance of each key component in the system.

1.1. Comparisons Against Alternative Methods

We compare our facial tracking algorithm against Microsoft Kinect facial SDK [1] and CLM [2] running in both single frame detection mode and sequential tracking mode. We evaluate the algorithms based on seven video sequences taken from seven different subjects. The seven test sequences consist of 3341 frames in total and include variations caused by occlusions and differences in facial expressions, head poses, illuminations, and skin colors.

Figure 1 compares our facial detection/tracking method against alternative methods [1][2] in both single frame detection mode and sequential tracking mode. We also report the tracking errors corresponding to six facial regions (see Figure 2). The comparison clearly shows that our method with/without temporal coherence significantly outperforms other methods running in either single frame detection mode or sequential tracking mode.

The importance of temporal coherence. As described in Section 4.1.4, we utilize temporal coherence to further improve the accuracy and robustness of our detection/tracking process. Figure 2 shows that better accuracy is achieved by using temporal coherence.

Figure 3 confirms the benefit of adaptive AAMs modeling. It compares the registration errors of AAM term for a test sequence with 625 frames with/without adaptive AAMs. As shown in the figure, facial tracking using adaptive AAMs produces much smaller registration errors for the AAM term. The registration error drops rapidly after the first AAM model update and remains smaller across the entire sequence.

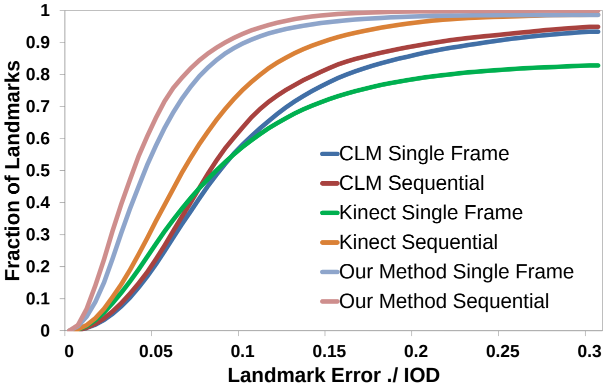

Figure 1. Comparison against alternative methods: cumulative error curves based on 7 testing video sequences.

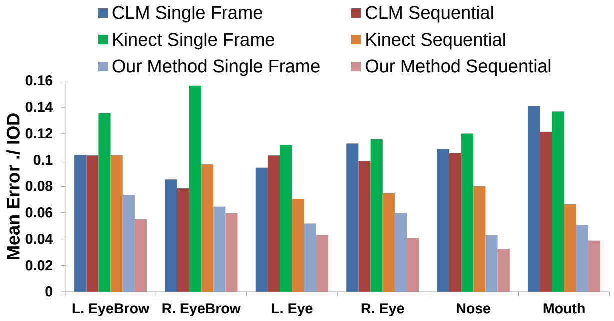

Figure 2. Comparison against alternative methods: mean error for each facial region.

1.2. Evaluation of Our Detection Method

We now evaluate the key components of our single frame detection process. We show the necessity of combining local detection with AAMs by dropping off each term in the cost function described in Equation 2. The evaluation is based on leave-one-subject-out cross validation on 317 sample images from 19 subjects. Figure 4 shows the cumulative error curves for each method. Per-frame AAM algorithm, even with ground truth global transformation, often fails to produce accurate results when actual feature locations are far away from their initial positions. In contrast, detection methods can often robustly detect feature locations in single images but often with less accurate results. Our detection method combines the power of AAMs and detection and can robustly identify feature locations with highest accuracy.

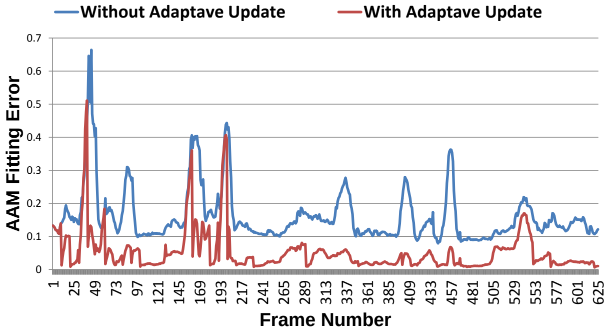

Figure 3. AAM fitting errors with/without adaptive updates on AAMs.

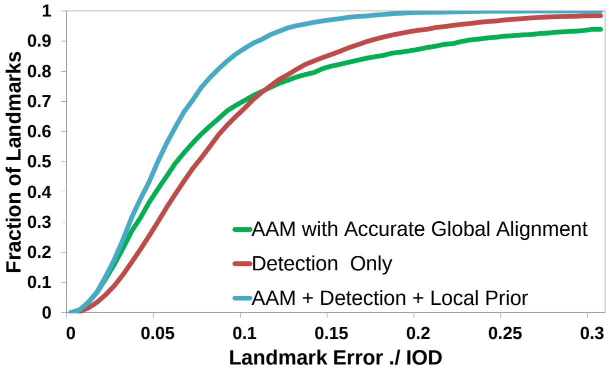

Figure 4. Evaluation of the importance of each component in our method. Cumulative error curves show the necessity of combining detection and AAMs.

References

[1] Kinect for Windows SDK. http://msdn.microsoft.com/en-us/library/jj130970.aspx/, 2013.

[2] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.