Unconstrained Realtime Facial Performance Capture

Pei-Lun Hsieh[1], Chongyang Ma[1], Jihun Yu[2], Hao Li[1]
[1] University of Southern California, [2] Industrial Light & Magic.

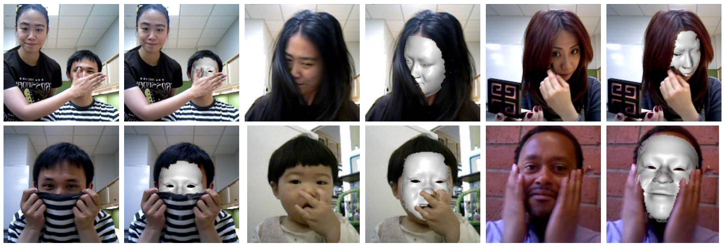

Figure 1: Calibration-free realtime facial performance capture on highly occluded subjects using an RGB-D sensor.

Introduction. Facial performance capture is well-established in the film and game industries for efficient and realistic animation production. While professional studios tend to rely on sophisticated solutions, *realtime* and *markerless* tracking technologies using *lightweight* monocular sensors (video or depth cameras) are becoming increasingly popular, due to their ease of adoption, cost, and deployability. In production, the capture process is typically *constrained* for optimal performance: face visibility is maximized; the environment is well lit; and an optimal facial tracking model is built before tracking. Unconstrained facial performance capture, on the other hand, has the potential to impact surveillance, recognition, and numerous applications in the consumer space, such as personalized games, make-up apps, and video chats with virtual avatars. In these unconstrained scenarios, new challenges arise: (1) occlusions caused by accessories, hair, and involuntary hand-to-face gesticulations challenge the face segmentation problem; (2) a facial tracking model needs to be constructed on-the-fly to enable instantaneous tracking and user switching for unobtrusive performance capture.

As demonstrated by Li and colleagues [4], the combination of sparse 2D facial features (e.g., eyes, eyebrows, and mouth) with dense depth maps are particularly effective in improving tracking fidelity. While recent advances have shown promising results in facilitating unconstrained facial tracking with data-driven methods, they do not ensure uninterrupted tracking in the presence of large and unexpected occlusions. Driven by the growing availability of consumer-level realtime depth sensors, we leverage the combination of reliable depth data and RGB video and present a realtime facial capture system that maximizes uninterrupted performance capture in the wild. It is designed to handle large occlusion and smoothly varying but uncontrolled illumination changes. Our system also allows instant user switching without any facial calibration.

The input data for our system is obtained from a PrimeSense Carmine 1.09 sensor. There are three main components in our system: realtime facial tracking, face segmentation with occlusion completion, and tracking model personalization which runs concurrently to the tracking and segmentation thread.

Tracking. Facial tracking is achieved by fitting a textured 3D tracking model to every captured RGB-D frame. First the rigid motion is estimated between the input data and the tracking model obtained in the previous frame. A user-adaptive tracking model is then used to solve for the linear blendshape expressions. Next a Laplacian deformation is applied to the tracked blendshape for accurate per-vertex displacements in the final output. Similar to the tracking method of [4] and [2], we use sparse facial features detected in the RGB channels to improve the facial tracking fidelity and to better handle fast motions in *xy*-directions. We use 36 out of 49 landmarks (eyebrows, eye and mouth contours) obtained from the supervised descent method of [5].

Segmentation. The facial expressions should be solved with constraints that are defined in unoccluded regions and visible to the sensor. We therefore compute a binary segmentation map for every frame by labeling each pixel as face region or occlusion in the UV map of the tracking model. We model occlusions as outliers in the input data using the vertex positions and texture of the exponentially smoothed tracking model as reference. Naïve per-pixel thresholding is prone to errors because of noise in the input video, we therefore enforce smooth spatial and color coherence in the segmentation result using an outlier voting scheme in the superpixel space. While only unoccluded regions are used for tracking, we fill the occluded ones with textures that are aggregated on the tracked face model from the previous frames. By synthesizing the facial features behind occlusions, landmark detection becomes significantly more reliable.

Personalization. Our tracking model is initialized with a generic blendshape model (statistical mean and 28 generic FACS-based expressions) which is adapted to the user during tracking. Every time the template personalization updates its shape and appearance, the latest one is retrieved and used for tracking. For every input frame, the blendshape coefficients computed by the facial tracking are used to solve for the shape of the user's identity using a linear PCA model. Next, the mesh vertices of the expression shapes are refined to match the captured subject. To account for the entire history of personalized tracking models, we recursively aggregate the new shapes and recorded texture to the previous ones via exponentially weighted moving average. Only those expressions are solved, if the currently tracked model is closer to the corresponding expression than all previous observations. The tracking model effectively improves over time and uncontrolled shape variations of the tracking model can be significantly reduced.

Conclusion. We provide the implementation details of our system and compare our approach with several state-of-the-art realtime facial tracking techniques in the paper. Our system demonstrates the ability to handle extremely challenging occlusions via an explicit face segmentation approach during tracking using both depth and RGB channels. By simply voting inliers in superpixel space using an appearance adaptive tracking model, our system produces clean segmentations even when the illumination changes in the environment. Unlike existing data-driven methods, our approach does not require a dedicated appearance modeling, since its construction would require a prohibitively large amount of training data to capture all the possible variations. We have also demonstrated that synthesizing face textures in the occluded regions is a crucial step to enable reliable use of landmark detection and provide accurate and continuous tracking when the face is occluded. Even though there is no solution yet for an accurate prediction of identity and expressions shapes for arbitrary users, our on-the-fly blendshape modeling solution prevents uncontrolled shape variations using localized expression optimization and blendshape coefficient monitoring. While our online generated models are close to pre-calibrated ones [3], our depth-based personalization algorithm significantly outperforms pure RGB systems [1] in terms of accuracy.

[1] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4):43:1–43:10, 2014.

[2] Yen-Lin Chen, Hsiang-Tao Wu, Fuhao Shi, Xin Tong, and Jinxiang Chai. Accurate and robust 3D facial capture using a single rgbd camera. In *ICCV*, pages 3615–3622, 2013.

[3] Faceshift, 2014. http://www.faceshift.com/.

[4] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42:1–42:10, 2013.

[5] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its application to face alignment. In *CVPR*, pages 532–539, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.