Real-time 3D Head Pose and Facial Landmark Estimation from Depth Images Using Triangular Surface Patch Features

Chavdar Papazov[1], Tim K. Marks[1], Michael Jones[1],
[1]Mitsubishi Electric Research Labs (MERL), Cambridge, MA.

We address the problem of 3D head pose estimation and facial landmark localization using a commodity depth sensor such as Microsoft's Kinect. Our method is robust to noise, is rotation and translation invariant, and runs on a frame-by-frame basis without the need for initialization. It can process 10 to 25 frames per second, depending on the desired accuracy, on a single CPU core.

Our method consists of an offline training and an online testing phase. Both phases rely on a novel triangular surface patch (TSP) descriptor. A TSP is specified by any equilateral base triangle of a fixed size such that all three vertices are on or very close to the 3D surface of the face. Given such a triangle, the TSP consists of those points of the 3D surface of the face that are above or below the triangle (see Figure 1(a,b)). Intuitively, a TSP is a triangular patch of the 3D surface of a face. We create an efficient descriptor of a TSP by discretizing it. The base triangle defining a TSP is subdivided into smaller sub-triangles, and the descriptor consists of the average height of the surface points above or below each sub-triangle. A TSP descriptor thus consists of a small number (we use 25 in our experiments) of floats representing the mean height for each sub-triangle. See Figure 2 for an illustration of a TSP descriptor. TSPs are viewpoint-independent and robust to changes in pose, scale (distance to camera), and resolution.

In the training phase, a large number of TSPs are sampled from synthetic 3D heads, and a descriptor for each TSP is stored in a library along with 3D displacement vectors from the centroid of the base triangle to the centroid of the head and to facial landmarks (see Figure 1(c)). In the testing phase, TSPs are sampled from a 3D point cloud computed from an acquired depth image. The nearest neighbors to each TSP's descriptor are found in the training library using a fast approximate nearest neighbor method [3]. The nearest TSPs in the training library are used to estimate the 3D head pose and facial landmark positions.

Each sampled TSP gives an estimate of the 3D head pose as well as the 3D locations of the head centroid and facial landmark points. To get the 3D head pose estimate, we compute the rotation matrix R that rotates the nearest-neighbor TSP from the training library to the test TSP. This rotation matrix R is the estimate of head orientation. The estimated landmark positions are computed by transforming the landmark position displacements from the nearest-neighbor TSP according to the rotation R and adding the centroid of the sampled test triangle. The head centroid location is estimated similarly to the landmark locations.

To get a final estimate of head pose and landmark positions given all of the estimates from sampled TSPs, we use an algorithm that jointly clusters in both head orientation space, SO(3), and head centroid space, $\mathbb{R}^3$. The basic idea is to only use estimates from TSPs for which the estimated pose angles and the estimated head centroid locations form a cluster in *both* of their respective spaces.

We tested our algorithm on the Biwi Kinect Head Pose Database [2] and compared against the results of Fanelli et al. [2] (see Figure 3) and Baltrusaitis et al. [1]. Our accuracy in terms of both 3D head pose angles and 3D facial landmark positions show significant improvement over these previous algorithms while maintaining similar (real-time) speed.

[1] T. Baltrusaitis, P. Robinson, and L-P. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *CVPR*, 2012.

[2] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *Int. J. of Computer Vision*, 101: 437–458, 2013.

[3] Marius Muja and David G. Lowe. Scalable nearest neighbor algorithms for high dimensional data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36, 2014.

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

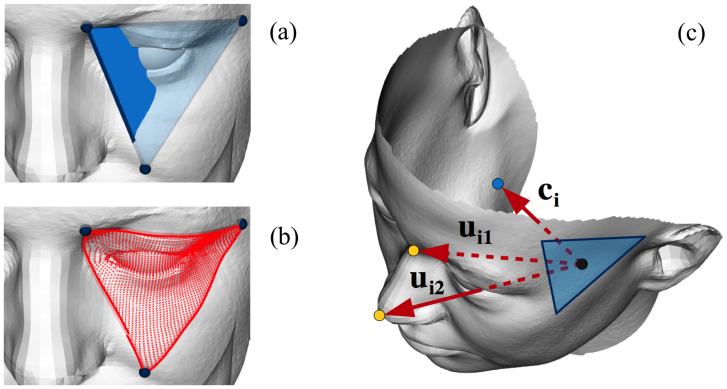

Figure 1: (a) An equilateral base triangle sampled from the vertices, **S**, of a 3D head model. The light blue part of the base triangle is occluded by the face's surface, while the dark blue part occludes the face. (b) The corresponding triangular surface patch (TSP), shown in red, consists of the points in **S** that lie above or below the base triangle. (c) A base triangle (shown in blue) sampled from a training head model, along with the vectors $\mathbf{c}_i, \mathbf{u}_{i1}$, and $\mathbf{u}_{i2}$, which originate at the centroid of the base triangle and point to the head model's centroid (blue dot) and two facial landmarks (yellow dots), respectively.

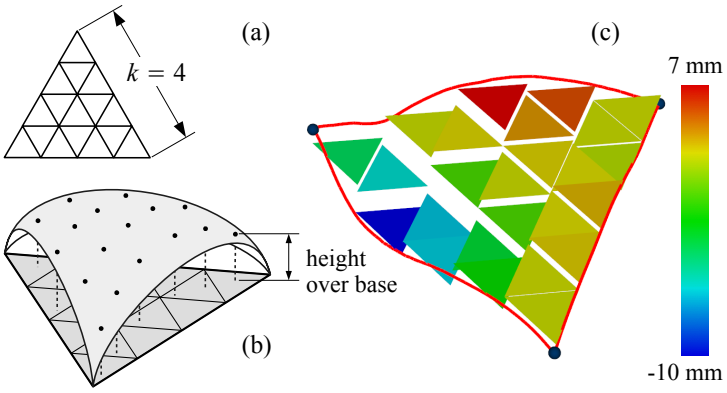

Figure 2: (a) Subdivision of the base triangle into $k = 4$ sub-triangles per side results in a total of k^2 sub-triangles. (b) Each TSP point (black dot) belongs to the sub-triangle in which it would lie if it were projected perpendicularly onto the base triangle. (c) Visualization of the descriptor for the TSP shown in Figure 1(b), using $k = 5$ sub-triangles per side. Each sub-triangle is displaced above or below the base triangle and colored according to the mean height of the points it contains.

Method	Nose tip (mm)	Direction (°)	Yaw (°)	Pitch (°)	Roll (°)	Time (ms)
Ours ($\triangle = 200$)	6.8	3.2	2.5	1.8	2.9	75.1
Ours ($\triangle = 100$)	8.6	4.4	3.5	2.5	4.2	38.9
[2] Trained on Biwi	12.2	5.9	3.8	3.5	5.4	44.7
[2] Synthetic Training	19.7	8.5	6.0	4.8	5.8	44.0

Figure 3: Position and orientation errors on Biwi Kinect Head Pose Database.