3D head pose estimation from multiple distant views

Xenophon Zabulis[1]
zabulis@ics.forth.gr
Thomas Sarmis[1]
sarmis@ics.forth.gr
Antonis A. Argyros[12]
argyros@ics.forth.gr

[1] Institute of Computer Science, Foundation for Research and Technology - Hellas, N. Plastira 100, Vassilika Vouton, 700 13 Heraklion, Crete, Greece

[2] Department of Computer Science University of Crete, 714 09 Heraklion, Crete, Greece

3D head pose estimation constitutes a special problem of human motion modeling. An accurate and robust solution to this problem is of particular interest, because the 3D head pose of a human conveys important information on his/her behavior. Significant advances have been achieved in human head pose estimation for relatively close-range images, but the related available methods are not directly applicable in wider-range imaging conditions.

The proposed method is overviewed in Fig. 1. The visual hull of a person is obtained from images acquired synchronously from multiple viewpoints. While moving, the person's head is tracked in 3D employing a variant of the Mean-Shift algorithm and a spherical kernel. The texture on the surface of the hull is collected from multiple views and projected on a hypothetical sphere S that is concentric to the person's head. This gives rise to spherical image I_s within which face detection is simplified, because exactly one frontal face is guaranteed to appear in it at a known spatial scale.

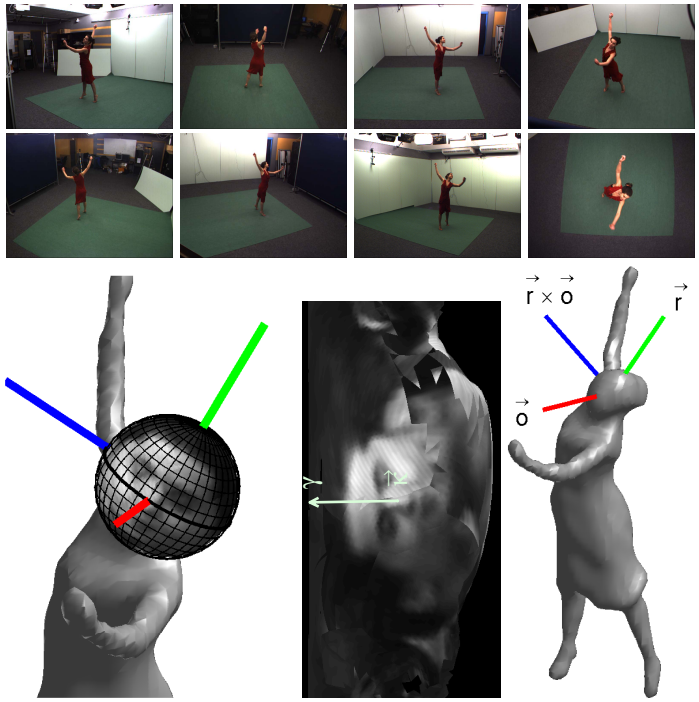

Figure 1: Overview of the proposed method. Top row: input images. Bottom row: visual hull with face texture mapped on a localized hypothetical spherical head (left); spherical head image (middle); visual hull (right). 3D head pose results are superimposed, in 3D illustrations.

Detecting the face center $\vec{\kappa}$ in I_s, yields an estimate of the head's 3D *orientation* $\vec{o}$, whose spherical coordinates are the *pitch* and *yaw* components of an absolute pose estimate. The 2D orientation γ of the face in the spherical image yields vector $\vec{r}$, which determines the *roll* component of this estimate. In each frame, face detection in I_s is iterative, in order to reduce the spherical image distortions that complicate this process. In particular, the parameterization of points on S, is iteratively rotated, so that the center of the face eventually projects on the equator of S. In each iteration, points on S are re-parameterized, so that the face detection of the previous iteration occurs on the equator of S. Then, a new I_s is formed and the face is redetected in the new I_s. Correspondingly, estimates $\vec{\kappa}$ and γ are updated and iterations terminate upon convergence, of the face detection to a point on S.

To accelerate the method and improve its robustness against tracking drift, a coarse orientation estimate of the head is computed prior to the application of the above method. This estimate $\vec{o}_c$ is obtained by performing face detection in the image regions where the head projects and finding the intersections of the optical rays passing through the image face centers, with S (Fig. 2). Estimate $\vec{o}_c$ is then computed as a representative of the above intersections, which is robust to outliers. In the proposed method, the coarse orientation estimate $\vec{o}_c$ plays a dual role. First, it conserves computational time as it places the sphere's equator near the face center and, thus, fewer iterations are required for convergence. It also serves as an anti-drift mechanism, so that even if pose estimation was erroneous in the previous frame it has no consequence in the current frame. Whatsoever, $\vec{o}_c$ is not prerequisite for the method and, thus, the system can cope with failures of face detection in areas A_i. If such face detections occur, they are utilized. If not, the process is based on the previous-frame estimate of $\vec{o}$, until the next face detection occurs.

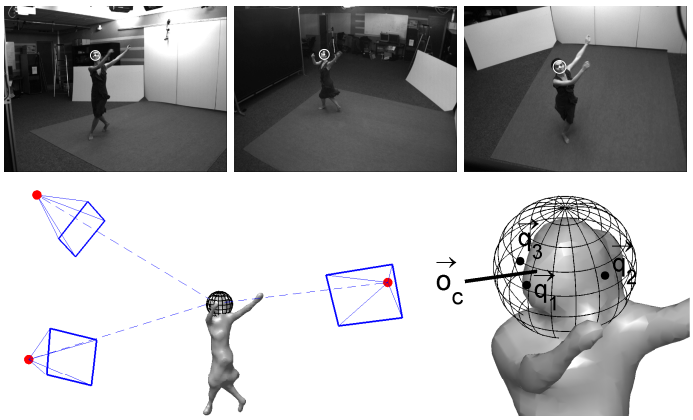

Figure 2: Coarse orientation estimation of the head. FDn in regions A_i of the original images (top) determines the optical rays through the face centers that intersect $\mathscr{S}$ (bottom, left); circle radii indicate the returned size of the face detection, which is required to be compatible with the projection size of $\mathscr{S}$ in each image. This yields intersection points $\vec{q}_{i'}$, through which the coarse orientation estimate $\vec{o}_c$ is robustly estimated (bottom, right).

The proposed method has been evaluated based on a series of experiments, which are all presented in the accompanying video. The specific camera setup was not optimized for the particular task of 3D head pose estimation, but has been decided to generically serve the purposes of human activity interpretation. It, therefore, features wide FOV cameras that are at more than $3.5\,m$ from the subject(s). Extensive experimental evaluation of the proposed method demonstrates that it is accurate, robust and can cope with several challenging situations including distant head views, frontal face occlusions, etc. In addition, the method was quantitatively evaluated utilizing a distant-viewpoint and multiview dataset annotated with ground-truth. The results of this evaluation are compared against state-of-the-art methods in the context of distant viewing, indicating that the proposed method is by $\approx 10^{\circ}$ more accurate than other distant-viewing methods. Tracking was not included in the experiments, so that the accuracy of pose estimation is assessed without its improvement. Furthermore, the proposed method estimates all 3 components of head pose, whereas similar methods in the context of distant-viewing compute only pitch and yaw. Finally, the employed ground truth annotated dataset includes a model of the background and is publicly availed.