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, Heraklion, Crete, Greece

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

Abstract

A method for human head pose estimation in multicamera environments is proposed. The method computes the textured visual hull of the subject and unfolds the texture of the head on a hypothetical sphere around it, whose parameterization is iteratively rotated so that the face eventually occurs on its equator. This gives rise to a spherical image, in which face detection is simplified, because exactly one frontal face is guaranteed to appear in it. In this image, the face center yields two components of pose (yaw, pitch), while the third (roll) is retrieved from the orientation of the major symmetry axis of the face. Face detection applied on the original images reduces the required iterations and anchors tracking drift. The method is demonstrated and evaluated in several data sets, including ones with known ground truth. Experimental results show that the proposed method is accurate and robust to distant imaging, despite the low-resolution appearance of subjects.

1 Introduction

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. In such situations, a human head is imaged in relatively low resolution, illumination artifacts are frequent, and occlusions are expected.

Existing multiview head pose estimation methods perform conventional single-view head pose estimation and then, fuse the results. Thus, they inherit the requirement for unoccluded face appearances and do not treat occlusions systematically. Due to lack of support, additional problems in face detection (FDn) itself are encountered. In this paper, we attempt to improve current state-of-the art results in two ways. First, information about surface structure is combined with FDn to assist head localization and its coarse orientation estimation. Second, visibility information is utilized to compensate for the large viewing distance and

BMVC 2009 doi:10.5244/C.23.118

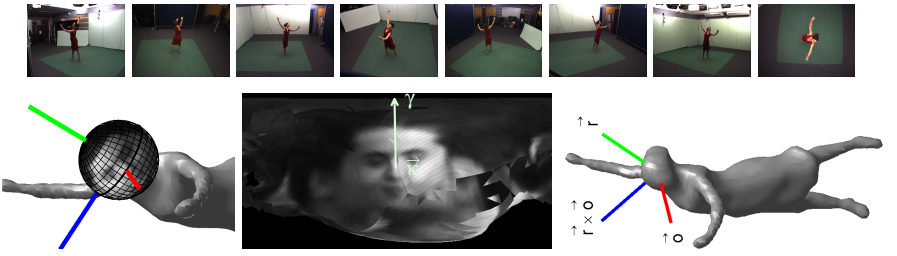


Figure 1: Method overview. Top row: input images. Bottom row: visual hull with face texture mapped on a localized hypothetical spherical head (left); spherical head image (middle); visual hull (right). 3D head pose results are superimposed, in 3D illustrations.

occlusions, by collecting all visible facial texture fragments in a single image, where exactly one frontal view of the face is guaranteed to appear at a known spatial scale. This not only provides substantially better support for FDn but also simplifies it, as only a frontal face appearance is sought.

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. 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 forms a spherical image I_s of the head, containing one frontal face appearance at a known spatial scale. 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. To reduce the spherical image distortions complicating FDn, the parameterization of S is continuously rotated so that the center of the face projects on its equator. In addition, FDn in the original images supports pose estimation by providing a coarse orientation estimate, which accelerates the method and improves robustness.

The remainder of this paper is organized as follows. In Sec. 2, related work is reviewed. In Sec. 3, the modules that support the proposed method, are presented. In Sec. 4, the proposed method for 3D head pose estimation is formulated and, in Sec. 5, is evaluated through several experiments. Sec. 6 summarizes this work.

2 Related work

This section reviews pertinent work in head localization and pose estimation. A recent overview of head pose estimation in computer vision is provided in [2].

Towards solving the head pose estimation problem, several methods assume that the face occupies most of the image. These methods encounter significant challenges when applied to distant views of a human subject. *Templates* [24] and *detector-arrays* [16], which coarsely pose-classify the observed face, require extensive training and exhibit a large rate of spurious detections; [10, 25] use 3D information to reduce this rate. *Nonlinear regression* [27] and *manifold embedding* [3, 12, 17] methods, vectorize the face image region or its features [8, 27] into a space where coordinates correspond to head poses. However, in wide range imaging, it is difficult to accurately segment the face and align it with the vectorized image

region. Similarly, methods that track rigid formations of landmarks [11, 13, 22, 30] or feature points [15, 23], are difficult to apply in poor resolution images due to inaccurate localization of such points. Methods that use *flexible models* instead [33] are more robust, but require that all or most of the formation is visible, which is not guaranteed for general subject motion.

Better performing in poor resolution images, some methods use a frontal view of the face and map it as texture on a hypothetical surface. Pose is estimated as the posture of this surface that exhibits the highest photoconsistency with the acquired image. In [18], a hypothetical cylinder was utilized to explain self-occlusions of this texture. In [29], an artificial 3D model of the human head was employed. In [6], the surface is pre-textured, thus not requiring a training view. Expressions and small occlusions are treated by periodic update of the reference texture [18, 29] but then, the estimate is subject to the error at the update frame.

In multicamera systems, pose has been estimated by single-camera estimation methods applied individually in each view, followed by a fusion of the individual estimates based on conditional probability [32], a Bayesian classifier [34], and a joint likelihood-estimator [28]. However, this strategy is reported to yield only coarse pose estimation, probably due to the overly coarse pose-classification obtained from each view. In [25], correspondence of FDns across cameras is required to increase robustness. In a different approach [4], skin-colored pixels are backprojected from multiple views onto a hypothetical ovaloid; the centroid of the skin-colored blob on it yields a coarse orientation estimate. In [26], a textured visual hull is employed to detect the face of a person, but it is assumed that the person is facing forward and head pose is determined by his/her motion trajectory.

The *proposed method* is novel in the following ways. It combines a face detector (FDr) with the 3D structure information of the visual hull, to locate the head in the acquired images and assist FDn. A hypothetical surface approximating the head's surface is employed in the composition of a "multiview image" (I_s), whose formation collects visible texture fragments of the head from multiple views. This compensates for the reduction in visible facial image area imposed by occlusions. In this way, not only occlusions are treated, but the constraint that a significant portion of the face should be visible in the same view, is relaxed. Finally, FDn in I_s is simplified, as exactly one frontal face appears, at a known spatial scale.

3 Building blocks

A synchronized multicamera system is assumed. Each camera i is located at 3D point K_i, has a projection matrix P_i and provides image I_i. Besides background subtraction in the computation of the visual hull, images I_i are treated as monochromatic. The term *frame* refers to all images I_i, synchronously acquired at a certain moment in time.

Visual hull. Images I_i are background subtracted [35] yielding binary images B_i. The volume occupied by humans is approximated by their visual hulls and efficiently computed from B_is as a volumetric occupancy grid, as in [20]. In our case, though, a voxel takes the value of 1 if occupied and -1 if not. This space is then finely smoothed and the visual hull is extracted as its 0-isosurface, using [19]. For each frame, the visual hull is encoded as a mesh $\mathscr{M}$ of triangles. In all experiments, voxel size was set to $1\,cm^3$.

Head localization. Due to the shape of the human head, it is assumed that its 3D position can be approximated with the center $\vec{c}$ of a hypothetical sphere S of radius ρ ($16\,cm$) registered with the head. Let $\vec{s}_0$ be a coarse estimate of the head's center. At each frame, S is centered at $\vec{s}_0$, which is the head center position as it has been estimated in the previous

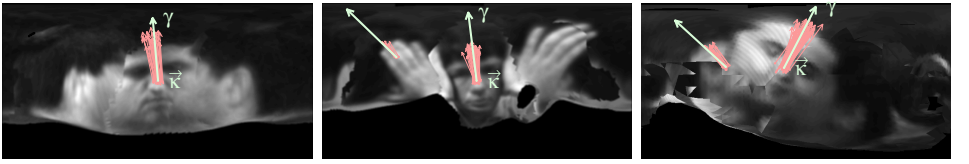


Figure 2: Detection of at most one face in three spherical images.

frame. In the first frame, $\vec{s}_0$ is provided by triangulation of FDns in I_i.

The estimate $\vec{s}_0$ is iteratively refined with the following variant of the Mean-Shift algorithm [9]. In each iteration j, the points $\vec{p}_k \in \mathscr{M}$ within S are retrieved. Their distances from S are $d_k = ||\vec{p}_k - \vec{s}_{j-1}| - \rho|$. The center of S is translated to $\vec{s}_j = \sum(\vec{p}_k \cdot w_k)/\sum w_k$, where $w_k = \rho - d_k$. The process terminates when distance $|\vec{s}_j - \vec{s}_{j-1}|$ falls below $1\,mm$ (92% of the times, in less than 6 iterations) or a maximum number of iterations (10) has been reached. The final $\vec{s}_j$ constitutes the estimate of the head's mass center $\vec{c}$. The areas A_i where S projects in images I_i and corresponding radii ρ_i are calculated for later use (see Sec. 4.1).

Texture mapping. Texture is mapped on $\mathscr{M}$ from I_is, utilizing a Z-buffer [5] per view to account for visibility. The Z-buffer of each view determines which triangles are visible to it, thus facilitating the association of each triangle in $\mathscr{M}$ to the views from which it is visible. The computation of texture intensity ϕ at a point $\vec{x}$ on a triangle $T \in \mathscr{M}$ proceeds as follows. Point $\vec{x}$ is projected on the images that T is visible. Let i' index these images. The blending ϕ of the appearances of $\vec{x}$ in $I_{i'}$ is $\phi = \sum \beta_{i'} \phi_{i'} / \sum \beta_{i'}$, where $\phi_{i'} = I_{i'}(P_{i'}\vec{x})$ and β_i is the projection area of T in $I_{i'}$, so that distal and oblique appearances of T are weighted less.

Face detection. We approach 2D orientation-invariant FDn in a generic way, utilizing a publicly available frontal face detector [31] much like a template. A FDr though is advantageous to a template (e.g. a reference view) as it is more robust to individual differences of human subjects, expressions, and illumination effects. The images on which FDn is applied are, by construction, guaranteed to contain exactly one frontal face view at a known spatial scale, but in an arbitrary 2D orientation (see Sec. 4.2).

The employed FDr detects a frontal face in an image ι, but assumes that the face is vertically oriented. To cast FDn invariant to 2D orientation, ι is center-rotated for a range of orientations $\Gamma = [0, 2\pi)$ (quantized by 1°). Input parameters to this operation is Γ and the range of acceptable face sizes M. This returns multiple FDns, indexed by j, each one associated with position $\vec{\kappa}_j \in \iota$ and orientation $\gamma_j \in \Gamma$. Orientation γ is computed assuming that a frontal face appearance has a vertical symmetry. Let the mean angle μ of the orientations in g and the angular interval $\alpha = [\mu - \omega, \mu + \omega]$ ($\omega = 10^\circ$). For every orientation $\gamma_j \in \alpha$, the region around $\vec{\kappa}$ is rotated by γ_j and bisected at its middle column. One of the counterparts is horizontally reflected and the Normalized Cross Correlation (NCC) of both is computed, as a measure of symmetry between them. Orientation γ is set equal to that minimizing NCC, thus selecting the candidate FDn exhibiting the greatest symmetry. For a spherical ι, the image is "wrapped-around" on its edges and M is defined in units of solid angle. FDns are grouped by proximity and orientation, to be pairwisely closer than $0.1 \cdot \max(M)$ and differ less than 10°. The group g with the greatest cardinality is selected. Let the centroid of its members be $\vec{\kappa}$. If not null, the result of FDn is locus $\vec{\kappa} \in \iota$ and orientation γ of the FDn.

The technique is illustrated in Fig. 2. In the figures, large arrows indicate $\vec{\kappa}$ and γ and small ones $\vec{\kappa}_j$ and γ_j in g. The two rightmost images feature spurious detections that are

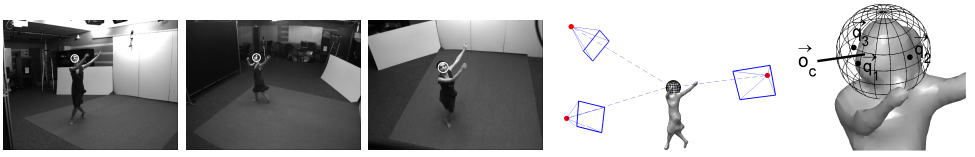


Figure 3: Coarse orientation estimation. FDn in A_i of I_i (three rightmost images) determines the optical rays through face centers intersecting S (2^{nd} from right); circle radii indicate FDn sizes, required to be compatible with the projection size of S in each I_i. This yields intersection points $\vec{q}_{i'}$, through which coarse orientation estimate $\vec{o}_c$ is robustly estimated (right).

rejected, as their groups have less members than the corresponding gs.

4 Head pose estimation

4.1 Coarse orientation estimation

A coarse estimate $\vec{o}_c$ of the head's 3D orientation is obtained by performing FDn (as in Sec. 3) in regions A_i and finding the intersection of the optical ray(s) passing through the image face center(s), with S (Fig. 3). The process is triggered by a FDn occurring in some A_i.

When executing FDn in A_i, its input size range M is individually modulated per view. As in [25], this casts M to be equivalent to the projection area A_i, pruning spurious FDns that occur centered at A_i but at incompatible sizes to the true face appearance.

Due to occlusions and FDn failures, FDns are less than views. Let FDns be indexed by t. For each FDn, the line L_t from camera center K_t through the center point of the FDn $\vec{\kappa}_t$, in A_t, intersects S at $\vec{q}_t$. Then, $\vec{v}_t = \vec{q}_t - \vec{c}$ approximates the head's orientation. Point $\vec{q}_t$ is the intersection of S with L_t that is closest to K_t. Two points on L_t are K_t and $(P_t^T(P_tP_t^T)^{-1})[\vec{\kappa}_t^T\,1]^T$ [14, p. 148]. If FDn returns a null result, $\vec{o}_c$ is assigned with the previous orientation estimate. A more accurate approach would be to find the intersections of all L_t with $\mathscr{M}$ instead of S, but it is computationally more complex and of no impact on the outcome of pose estimation.

A representative $\vec{o}_c$ of all $\vec{v}_t$ that is robust to outliers, such as q_2 in Fig. 3, is computed as follows. For each $\vec{v}_t$, its relative angles to all other $\vec{v}_t$s are computed; let h_t be the median of these angles for each $\vec{v}_t$. Let also, $\delta = \min_t(h_t)$ and $t = \arg\min_t(h_t)$. Then, the mean of the input vectors with relative angles less than 2.5δ to $\vec{v}_t$ is assigned to $\vec{o}_c$.

4.2 Pose estimation

Given $\vec{o}_c$, the following process estimates the head pose. The texture of the portion of $\mathscr{M}$ inside S is projected on S, with $\vec{c}$ as the projection center, yielding spherical image I_s on S. Pose estimation is based on FDn in I_s and is computed from the $\vec{\kappa}$ and γ of this FDn. Note that the structure of I_s (spherical) is independent to the shape of S, but determined by the parameterization by which points are sampled on S to form an image.

A convenient way to create I_s is to sample the projected texture on S, assuming points on S in a spherical coordinate parameterization $\mathscr{P}$, its coordinates quantized by angle ψ.

Unfortunately, if the face is projected near one of the poles of the sphere, it becomes highly distorted in I_s and FDn becomes problematic. Thus, points of $\mathscr{P}$ are rotated about $\vec{c}$, so that the middle of the face in I_s ($\vec{\kappa}$) is sampled by a point on the equator of $\mathscr{P}$.

The whole process proceeds iteratively. This is because a face projected further from the equator appears distorted in I_s and, consequently, can be inaccurately localized. In the worst case, the face could be projected near a pole causing a failure in FDn. To avoid this in the first iteration and, also, to avoid potential drift from the previous frame, $\mathscr{P}$ is initially rotated so that its equator occurs in the direction pointed by $\vec{o}_c$.

The result of pose estimation is vectors $\vec{o}$ and $\vec{r}$, from which the resulting, absolute yaw Θ_τ, pitch Φ_τ, and roll Ω_τ estimates are derived. In each iteration n, the components of the current pose estimate are Θ_n (yaw), Φ_n (pitch) and Ω_n (roll). Initially, $\vec{o}_0$ is set equal to $\vec{o}_c$ and Θ_0, Φ_0 are set equal to the spherical coordinates of $\vec{o}_0$. In addition, Ω_0 is set equal to 0°. In each iteration n, the following three operations are performed:

(1) Parameterization: $\mathscr{P}$ is rotated about $\vec{c}$ so that $\vec{o}_{n-1}$ points its equator. The rotation matrix is $R = R_1 R_x(\Omega_n)$, where $R_1[1\,0\,0]^T = \vec{o}_{n-1}$, and $R_x(\Omega_n)$ a rotation of Ω_n about the xx' axis.

(2) Spherical image formation: Each $\vec{p} \in \mathscr{P}$ corresponds to a pixel ε in I_s. The line segment from $\vec{p}$ to $\vec{c}$ intersects a triangle of $\mathscr{M}$ at $\vec{x}$, computed as in [21]. The texture intensity at $\vec{x}$ is assigned to $I_s(\varepsilon)$. If multiple intersections are found, the one closest to S is selected; this occurs in concave parts of the head (e.g. a ponytail). In points where S occurs within $\mathscr{M}$ (the neck), the intersection is null and the intensity is set to 0.

(3) Face detection: Performed on I_s, as in Sec 3, yielding $\vec{\kappa}_n$ and γ_n, or null. The point $\vec{p}_n \in \mathscr{P}$, corresponding to $\vec{\kappa}_n$, is looked-up and the current orientation estimate, $\vec{o}_n$, is set as $\vec{o}_n = \vec{p}_n - \vec{c}$. Input size range M was restricted to $[60^\circ, 160^\circ]$ of solid angle.

In iteration n, $\vec{o}_n$ is the current orientation estimate. The spherical coordinates of $\vec{o}_n$, Θ_n and Φ_n, are absolute yaw and pitch estimates, respectively. The absolute roll estimate, Ω_n, can be computed as follows. Let $\vec{r}$ be a vector tangent to S at $\vec{p}_n$ and oriented as γ_n. Let, also, R_o be the rotation matrix that maps $\vec{o}_n$ to zz' ($R_o[0\,0\,1]^T = \vec{o}_n$). The angle of $\vec{r}_o$ with the yy' axis is Ω_n, where $\vec{r}_o$ is the projection of a unit vector $R_o \cdot \vec{r}$ on the XY plane. Thus, in the "parameterization" step of the iterative process, R_1 brings the detected face to the center of I_s and $R_x(\Omega_n)$ aligns its vertical axis with the image columns.

The process terminates at iteration τ, in which any of the following conditions is met: (a) The angle between $\vec{o}_n$ and $\vec{o}_{n-1}$ becomes below 1°, (b) A maximum number (5) of iterations is reached, (c) No face is detected; in this case, $\vec{o}_c$ is returned. The pose estimate is $\vec{o} = \vec{o}_\tau$ and $\vec{r} = \vec{r}_\tau$, from which $(\Theta_\tau, \Phi_\tau, \Omega_\tau)$ are calculated. The first iteration is the most important, as convergence occurs in 2-3 iterations with the last ones refining $\vec{o}_n$ by less than 3°.

In the proposed method, the coarse orientation estimate $\vec{o}_c$ plays a dual role. First, it conserves computational time as it places $\mathscr{P}$'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 a prerequisite for the method and, thus, the system can cope with failures of FDn in areas A_i. If such FDns occur, they are utilized. If not, the process is based on the previous-frame estimate of $\vec{o}$, until the next FDn occurs.

From a computational viewpoint, I_s's formation is accelerated by an index $\mathscr{E}$, which is a 2D buffer equal in size to I_s. Triangles within S are projected from $\vec{c}$ to S and corresponding elements in $\mathscr{E}$ are marked with their labels. When forming $I_s(\varepsilon)$, the intersection triangle is looked-up in $\mathscr{E}(\varepsilon)$. In the experiments of Sec. 5.2, processing a frame takes $90\,sec$ on average by a MATLAB implementation, the most time-consuming part being the synthesis

of I_s. A GPU implementation may accelerate this process, treating it as texture-mapping.

5 Experiments

The proposed method has been evaluated based on a series of experiments[1]. Most experiments took place in a $5 \times 5 m^2$ room where cameras are used to visually interpret human activity. Eight cameras are mounted at the corners and at the in-between mid-wall points of the room viewing it in yaw-steps of $\simeq 45°$. The cameras are pointing at the floor center in a relative pitch of $\simeq -43°$. Their height is $\simeq 2.6m$ from the floor. At the same height, a ninth camera is mounted on the ceiling, overlooking the floor. All cameras have $66° \times 51°$ FOV and 960×1280 resolution. It is worth noting that the specific camera setup has been decided to generically serve the purposes of human activity interpretation and was not optimized for the particular task of 3D head pose estimation. A public dataset [2] was also utilized in Sec. 5.2, as well as in Figs. 1 and 3. Experiments in Sec. 5.1 compare results quantitatively against state-of-the-art in the context of distant viewing, indicating that the proposed method is more accurate than other distant-viewing methods [28, 32, 34].

In Sec. 5.2, the proposed method is demonstrated in challenging situations not handled by existing methods (broad range of poses and scene arrangements, severe occlusions, etc). Throughout the experiments, head shapes and sizes of the subjects exhibit significant variability (different adult subjects, small female mannequin). Therefore, we find no reason to expect individual differences to dramatically affect performance as long as faces are detectable by the FDr. It is noted that tracking was not included in the experiments, so that accuracy of pose estimation is assessed without its improvement; the incorporation of the proposed method within a robust tracking framework is left for future work.

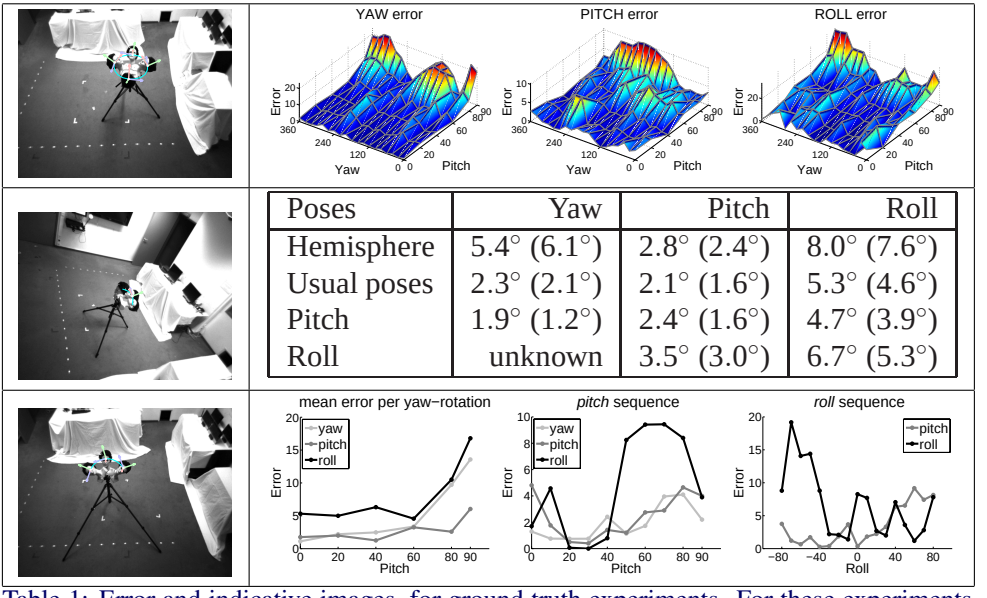
5.1 Ground truth dataset

To the best of our knowledge, there is currently no publicly available multiview dataset which (a) includes high-precision head pose ground truth data, (b) provides images for building a background model of the scene or the model itself, and (c) observes human subjects from distant viewpoints. Thus, such a dataset was created [1]. The dataset was collected using a mannequin's head, mounted on a tripod with 2 degrees of freedom (pitch, yaw) and marked rotation gratings. The head's was $\simeq 1.3m$ from the floor, emulating the head locus of a sitting person. To modulate roll, the head was unmounted and rotated by 90°; thus, during this modulation, ground truth for yaw was unavailable.

The main part of the dataset sampled a hemisphere of poses, consisting of six 360° yaw-rotations, in steps of 20°. In each, the first and last frame imaged the same pose. The pitch angles of the yaw-rotations were $\{0°, 20°, 40°, 60°, 80°, 90°\}$. In two additional sequences, *pitch* and *roll* were individually modulated within the rotation limits of the tripod, that is, $[0°, 90°]$ and $[-80°, 80°]$, respectively, in steps of 10°. During the *roll* modulation the head had a 40° pitch. Tripod and world coordinate frames were aligned. Resolution of I_s was set to 360×720 ($\psi = .5°$). Except for the *roll* sequence, the tripod was still and head centers occurred on a hemisphere. As an indication of head localization error, the average distance of the estimated centers from their least-squares fitting sphere was $2.7mm$.

In Table 1, head pose results for the ground truth dataset are presented. In its left column, key poses of the 60°-pitch yaw-rotation (top), the *pitch* (middle) and the *roll* (bottom)

[1] Results from all experiments and all views are presented in the accompanying video.



Poses	Yaw	Pitch	Roll
Hemisphere	5.4° (6.1°)	2.8° (2.4°)	8.0° (7.6°)
Usual poses	2.3° (2.1°)	2.1° (1.6°)	5.3° (4.6°)
Pitch	1.9° (1.2°)	2.4° (1.6°)	4.7° (3.9°)
Roll	unknown	3.5° (3.0°)	6.7° (5.3°)

Table 1: Error and indicative images, for ground truth experiments. For these experiments, the minor table in the middle row reads mean and standard deviation of estimation error.

sequences are shown; the cyan line plots $\vec{c}$'s trajectory. In the right column, the following data are reported. In the top row, errors for the hemisphere of poses are plotted: mean errors per yaw-rotation are plot on leftmost graph; middle and right graphs plot the errors for the additional *pitch* and *roll* sequences, respectively. The middle row, contains a table within which errors are averaged. Labels of this table correspond to the aforementioned rotations of the mannequin head: in *hemisphere* the error for the 6 yaw rotations is averaged and in *usual poses* head pitch was in $[0°, 60°]$ emulating typical poses of a human subject in the room. In the bottom row of Table 1, average error in the *hemisphere*, *pitch* and *roll* sequences is plot as a function of pitch or roll.

Comparison of results with the overview tables in [7] indicates approximately 10° accuracy improvement of currently published results concerning distant head viewing. On average, [32] yields more than 12° and [34] 33.56° of error. In [28], precision is up to 30° as 12 yaw-poses of a person's head are distinguished. It is also pointed out that the proposed method estimates all 3 angles of pose, while [32, 34] only 2 (pitch and yaw). The method in Sec. 4.2 refines the coarse estimate of Sec. 4.1, by 14.2° on average. We also note that published results are evaluated only for smaller ranges of poses, which are referred to as *usual poses*, in the middle row table of Table 1.

In the hemisphere measurements, errors tend to grow with pitch (Table 1, bottom row). This is due to the reduced multiview coverage of the face area in high-pitch poses (80°, 90°). This is also suggested by the error plot of the *pitch* sequence (bottom row, middle graph), where error is less in the vicinity of 30°. In this angle, the face is approximately frontoparallel to the peripheral cameras. Roll is the weakest estimate, as its accuracy depends on the quality of the unfolded texture and is more sensitive to image noise and poor registration of textures from different views. As a general remark, the reported results demonstrate that head pose estimation accuracy degrades gracefully to the loss of visual coverage.

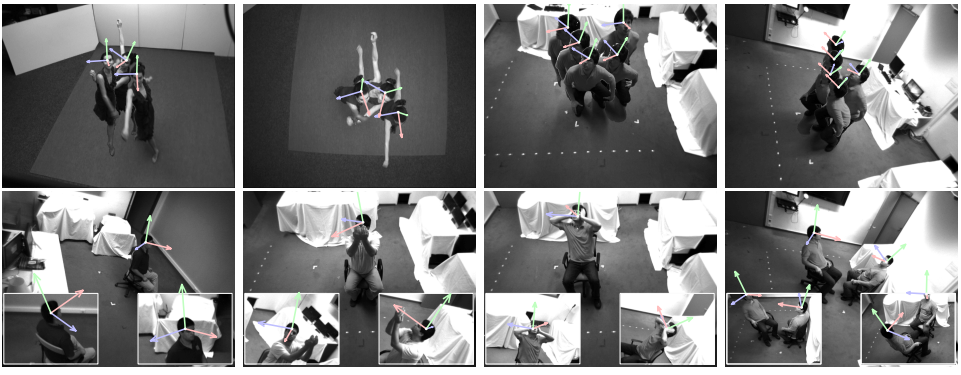


Figure 4: Top: Results for key frames of the *dance* (left 2) and *standing* sequences (right 2). Bottom, left to right: results from experiments *chair360*, *occlusions*, *hands* and *2persons*.

5.2 Estimation of head pose in human subjects

Datasets with human subjects capture indicative cases where head pose estimation can be challenging (see Fig. 4). In these experiments, $\psi = 1^\circ$ (I_s was 180×360) and the ceiling camera was excluded, due to hardware limitations, The *dance* sequence was acquired from a different group and employed 8 cameras in similar arrangement, but including a ceiling view (FOV $= 53^\circ \times 42^\circ$, resolution 780×582).

In the *dance* sequence, a person is dancing with head motion that includes significant yaw and roll rotations. Brief face occlusions occur in frontal views. In the *standing* sequence, the subject moves in space and the height of his head is varied about $0.8\,m$, while undergoing yaw, pitch, and roll rotation. In the *chair360* sequence, the subject is sitting on a rotating chair during four 360° yaw-rotations at 2 different pitch angles. Pose estimation results evolve smoothly as the subject's head faces towards all yaw-directions. In the *occlusions* sequence, the subject rotates his head while severely occluding it from its facing camera with his hands. The proposed method is capable of handling this challenging situation, as facial texture mapping is still possible from lateral cameras. In the *hands* sequence, the subject rotates his head while holding it with one and two hands. The goal of this experiment is to test head localization in the challenging situation encountered when the visual hull of the head is more complex than a simple protrusion. In the *2persons* sequence, pose estimation is performed individually for two subjects, demonstrating the capability of the method in handling multi-head pose estimation. In the details, notice that the subject's heads appear partially for frontal views, further demonstrating the collection of facial texture from partial face appearances.

6 Summary

A method that estimates all 3 components (yaw, pitch, roll) of 3D head pose from distant views has been described and evaluated. The main contributions of this work are the following. First, the visual hull is employed to assist FDn in multiview environments. Second, multiple views and visibility information are used to create an image that contains a frontal view of the face, despite occlusions and partial face appearances. In distant-viewpoint imag-

ing conditions and the presence of occlusions, this casts the proposed method more reliable and accurate than methods that perform individual-view pose estimation and fuse the results. 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. Additionally, as a byproduct of this work, a distant viewpoint and multicamera dataset annotated with ground truth was compiled and became available to the computer vision community, to assist the process of evaluating 3D head pose estimation methods.

References

[1] http://www.ics.forth.gr/cvrl/headpose/.

[2] https://charibdis.inrialpes.fr.

[3] N. Balasubramanian, J. Ye, and S. Panchanathan. Biased manifold embedding: A framework for person-independent head pose estimation. In *CVPR*, pages 1–7, 2007.

[4] C. Canton-Ferrer, J. Ramon, and M. Pardas. Head pose detection based on fusion of multiple viewpoint information. In *CLEAR*, pages 305–310, 2007.

[5] E. Catmull. *A Subdivision Algorithm for Computer Display of Curved Surfaces*. PhD thesis, U. of Utah, 1974.

[6] E. Chutorian and M. Trivedi. Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking. In *Intelligent Vehicles*, pages 512–517, 2008.

[7] E. Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *PAMI*, 31(4):607–626, 2009.

[8] E. Chutorian, A. Doshi, and M. Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *ITSC*, pages 709–714, 2007.

[9] D. Comaniciu and P. Meer. Mean shift : A robust approach toward feature space analysis. *PAMI*, 24:603–619, 2002.

[10] M. Dixon, F. Heckel, R. Pless, and W. Smart. Faster and more accurate face detection on mobile robots using geometric constraints. In *IROS*, pages 1041–1046, 2007.

[11] F. Fleuret and D. Geman. Fast face detection with precise pose estimation. In *ICPR*, page 10235, 2002.

[12] Y. Fu and T. Huang. Graph embedded analysis for head pose estimation. In *FGR*, pages 3 – 8, 2006.

[13] A. Gee. and R. Cipolla. Estimating gaze from a single view of a face. In *ICPR*, 1994.

[14] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[15] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3D head orientation from a monocular image sequence. In *FGR*, page 242, 1996.

[16] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using SVMs. In *ICPR*, pages 154–156, 1998.

[17] M. Ju and H. Kang. A new partially occluded face pose recognition. In *Advanced Concepts for Intelligent Vision Systems*, volume 4678/2007, pages 322 – 330, 2007.

[18] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3D models. *PAMI*, 22:322 – 336, 2000.

[19] W. Lorensen and H. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *SIGGRAPH*, pages 163–169, 1987.

[20] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara. Real-time 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video. *CVIU*, 96 (3):1077–3142, 2004.

[21] T. Moller and B. Trumbore. Fast, minimum storage ray-triangle intersection. *Graphics Tools*, 2(1):21–28, 1997.

[22] L. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *In CVPR*, pages 803 – 810, 2003.

[23] R. Niese, A. Al-Hamadi, and B. Michaelis. A stereo and color-based method for face pose estimation and facial feature extraction. In *ICPR*, pages 299–302, 2006.

[24] S. Niyogi and W. Freeman. Example-based head tracking. In *FGR*, page 374, 1996.

[25] G. Potamianos and Z. Zhang. A joint system for single-person 2D-face and 3D-head tracking in CHIL seminars. In *CLEAR*, pages 105–118, 2007.

[26] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *CVPR*, pages 439–46, 2001.

[27] J. Sherrah, S. Gong, and E. Ong. Face distributions in similarity space under varying head pose. In *Image and Vision Computing*, volume 19, pages 807 – 819, 2001.

[28] Y. Tian, L. Brown, J. Connell, S. Pankanti, A. Hampapur, A. Senior, and R. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *FGR*, page 92, 2003.

[29] J. Tu, T. Huang, and H. Tao. Accurate head pose tracking in low resolution video. In *FGR*, pages 573–578, 2006.

[30] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based head pose estimation from images. In *Humanoids*, 2007.

[31] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–8, 2001.

[32] M. Voit, K. Nickel, and R. Stiefelhagen. Neural network-based head pose estimation and multi-view fusion. In *CLEAR*, pages 291–298, 2007.

[33] J. Wu and M. Trivedi. A two-stage head pose estimation framework and evaluation. *PR*, 41(3):1138 – 1158, 2008.

[34] Z. Zhang, Y. Hu, M. Liu, and T. Huang. Head pose estimation in seminar room using multi view face detectors. In *CLEAR*, pages 299–304, 2007.

[35] Z. Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In *ICPR*, pages 28–31, 2004.