In-plane face orientation estimation in still images

Taner Danisman, Ioan Marius Bilasco

Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

In-plane face orientation estimation in still images

Taner Danisman · Marius Bilasco

Received: date / Accepted: date

Abstract This paper addresses a fine in-plane (roll) face orientation estimation from the perspective face analysis algorithm that requires normalized frontal faces. As most of the face analysers (e.g., gender, expression, and recognition) require frontal up-right faces, there is a clear need for the precise roll estimation, as precise face normalization has an important role in classification methods. The in-plane orientation estimation algorithm is constructed on top of regular Viola-Jones frontal face detector. When a face is detected for the first time, it is rotated with respect to the face origin to find the boundaries of the detection. Mean value of these angles is said to be the measurement of the in-plane rotation of the face. Since we only need a face detection algorithm, proposed method can work efficiently on very small sized faces where traditional landmark (eye, mouth) or planar detection based estimations fail. Experiments on controlled and unconstrained large-scale datasets (CMU Rotated, YouTube, Boston University Face Tracking, Caltech, FG-NET Aging, BioID and Manchester Talking-Face) showed that the proposed method is robust to various settings for in-plane face orientation estimation in terms of RMSE and MAE. We achieved less than ±3.5 ° mean absolute error for roll estimation which proves that the accuracy of the proposed method is comparable to that of the state-of-the-art tracking based approaches for the roll estimation.

Keywords in-plane rotation estimation · roll estimation · head-pose estimation

1 Introduction

Face orientation estimation has been ones of the most fundamental computer vision problems studied so far in the literature. The natural movement of head often generates in-plane (roll) and out-of-plane (pitch, yaw) rotations. In-plane rotation of the face is observed when the face is rotated along the axis from the face to the observer (rotation with respect to z axis) as show in Fig. 1. In the literature, majority of

T. Danisman · M. Bilasco
Lille 1 University, LIFL
IRCICA, Parc Scientifique de la Haute Borne
50 Avenue Halley 59655 Villeneuve d'Ascq - France
Tel.: +33-3-62531546
Fax: +33-3-28778537
T. Danisman
E-mail: taner.danisman@lifl.fr
M. Bilasco
E-mail: marius.bilasco@lifl.fr

the studies related to face analysis (face recognition, gender recognition, age estimation, facial expression recognition, etc.) assume only frontal upright faces and ignore in-plane face rotation. These assumptions limit the scope of face analysis methods in the wild application areas (e.g. indexing personal photos with gender and expression information) as the frontal upright position under these conditions is not guaranteed anymore. Robust face orientation estimation is a bridge filling the gap between these assumptions and the reality. As indicated by Demirkus et al [9], roll and pitch brings more challenge to face detection and tracking. Robust roll estimation provides more consistent facial images for the latter steps (e.g., machine learning, classification, template matching and appearance based methods). In our study, we mainly address robust estimation of in-plane face orientation for both frontal and in-plane rotated still images with the aim of using fine-grained roll estimation for coping with normalization requirements of frontal upright face analysers Danisman et al [8]. In the current version of this work, we did not consider a real-time solution as we aim at retrieving the entire set of faces contained within an image, but speed-up techniques can be envisioned under specific application assumptions (analyzing the expression of only one person over time, etc.).

The study of out-of-plane rotation is also an important and complex research topic in the field of facial analysis. However, with regard to our main objective, which is providing normalized up-right faces to classifiers, out-of-plane rotations cannot be easily corrected to frontal poses, without specific assumptions (generic 3D face model, face symmetry, etc.) that impact largely the underlying classification process. Besides, most of the out-of-plane orientation estimation approaches requires frontal faces for initializing. Our method can improve the correct selection of the initial frame as besides providing precise roll estimation, the solution proposed here reduces the false positive rate while keeping recall high. At last, the proposed solution cannot straight forward address the out-of-plane rotations, as it is constructed on the assumption that the face topology is conserved while exploring various roll orientation. In the following, we focus only on in-plane rotation problems.

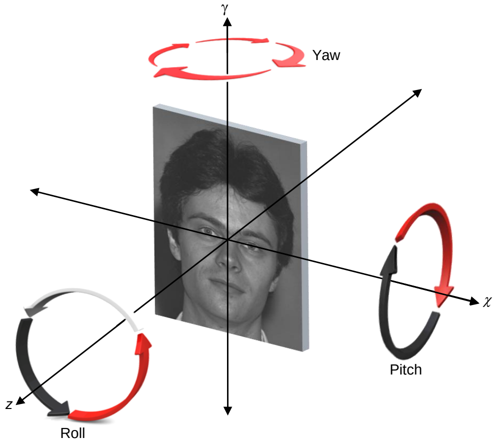


Fig. 1 In-plane (roll) and out-of-plane (pitch, yaw) rotations

Considering the huge amount of still images available in social media (facebook, flickr, twitter etc.), detecting the location and orientation of the faces is an important step towards automatic face analysis studies. In this user photos, majority of the people is looking at the camera (with subject's active participation) either in frontal upright position or with in-plane rotation. In-plane face orientation (roll) can be estimated prior to, or after the face detection. Therefore, many frameworks have at least two levels that corresponds generally to the detection and the estimation. Prior orientation estimation mainly targets rotation invariant and multi-view face detection e.g., [26, 30]. However, as indicated in Viola et al [30], one flaw of two level approaches is that final detection rate is dependent on the product of the correct classification rates of the two classifiers in each level. When the orientation is estimated after the face detection e.g., Wu et al [38], it generally needs facial feature detection e.g., eye detection [7, 43] or symmetry computation [6, 25, 39]. All of these approaches use complementary methods to better solve this problem. Since these methods are open computer vision problems, they also bring their flaws to the existing problem. For example, when roll of the face is estimated by eye detection, the flaws in eye detection impacts roll estimation. On the other hand, tracking based methods [15, 17, 21, 28] tends to provide substantially less error than systems that estimate the head-pose (in our case the roll) from individual frames and temporally filter the results as indicated by Murphy-Chutorian and Trivedi [20]. However, they usually require a sequence of images and strict initialization procedures (which most of the time they assumes the detection of a frontal face). For this reason, they can only be applied to temporal data. However, these works are interesting in the validation of our work as they generally provide fine-grain roll orientation estimation that is also required by the face normalization step for frontal upright classifiers. We estimate that providing similar performances in terms of roll estimation from single images is a challenging problem as no temporal information or any other assumption is used.

In our study, we addresses the estimation of the face orientation by exploiting in-plane rotation disadvantage of existing Viola-Jones (VJ) face detection algorithm [31]. According to Viola et al [30], the rotation disadvantage occurs when movement of face generates more than $\pm 15\,^\circ$ of in-plane rotation with respect to z axis. Another say, a true positive (TP) face is detectable approximately within $30\,^\circ$ out of full $360\,^\circ$ by frontal VJ face detector. This behavior is illustrated in Fig. 2 where all face detection windows at different scales are represented by red rectangles. The VJ face detector relays on these red windows for computing the matched face region. In this example, a frontal upright face is virtually rotated with respect to the image center. The number of potential rectangles describing faces increases when the rotated face is close to the frontal upright position. The in-plane rotation has negative effect on the face detection process for clockwise and counter clockwise direction as the number of candidate rectangles decrease in presence of high in-plane rotation.

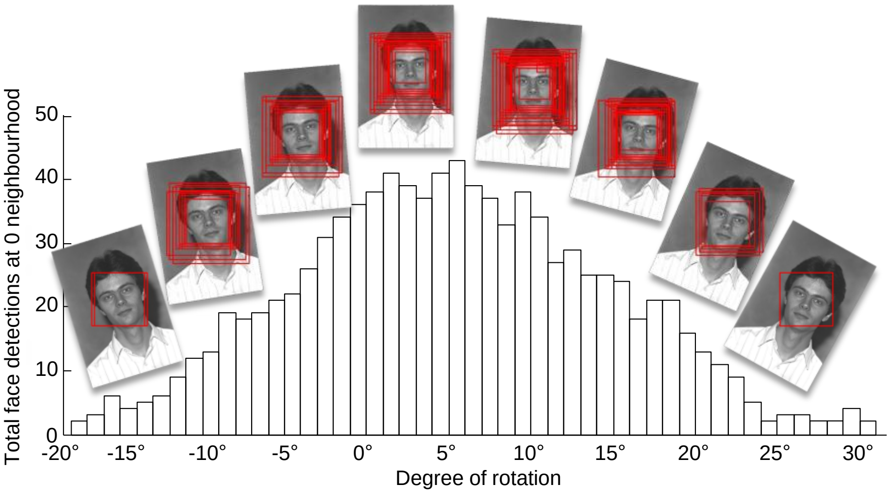


Fig. 2 Changes in total VJ face detections for different roll angles at zero neighborhood. Each bar shows the number of detections obtained from different scales of the VJ detector for specified degree of rotation

Our two-level estimation algorithm first performs a coarse frontal face detection by repeating frontal detector at different orientation angles. Then, the second level uses these faces as an input to find the minimum and maximum angles that the face is still detected. Mean value of the detections determine the roll angle of the face. Second level also provides confidence with respect to cumulative value of these multiple detections. According to the experiments, the proposed algorithm provides robust roll estimation by providing ±1.1 ° to ±3.5 ° of mean absolute error (MAE) for publicly available datasets (CMU Rotated, Boston University Face Tracking (BUFT), Caltech, FG-NET Aging, BioID and Manchester Talking-Face).

In comparison with previous studies, the main contribution of this study is to use the same frontal face detector for fine-grain roll estimation. Thus in our approach, in-plane rotation estimation is an immediate result of the traditional face detection process without additional steps. More complex tracking based models (e.g., 3D head tracking, planar, ellipsoid and cylindrical models) could get benefit from the temporal information. However, they require precise initialization, re-registration and tracking steps. Occlusions and disappearance of tracked points and accumulation of errors in temporal domain may also require re-initialization. In addition, these methods can only be applied to data with temporal dimension and their application to still images is limited. In science, it is a standard preference to have method as simple as possible. According to Occam's Razor theory, if two models otherwise equivalent and producing similar results, the simpler one is the better. We would like to recall that here we do not address out-of-plane rotations that are generally supported by more complex methods, but we address only the in-plane rotation problem with the aim of obtaining normalized frontal upright faces. Hence, our study argue that for in-plane face orientation estimation, VJ face detection algorithm can be effectively used for in-plane roll estimation. Our method can also be used for precise initialization (with regard to in-plane rotations) of face in tracking based methods.

The remainder of this paper is organized as follows. Section 2 presents related works for in-plane orientation estimation. Section 3 describes our methodology, face candidate selection, and face orientation estimation. Section 4 present datasets before discussing deeply; experimental setup, parameter selection, evaluation metrics, computational cost, results and comparison with the state-of-the-art methods. Effect of step factor and out-of-plane rotations on the roll estimation process is also discussed in section 4. Finally, section 5 concludes the paper. Appendix section provides examples from different datasets.

2 Related works

A common functional taxonomy covering the large variations in face orientation estimation studies can be found in Murphy-Chutorian and Trivedi [20]. In the current work, we selected representative approaches that report results on public datasets such as CMU rotated and BUFT dataset.

We distinguish between coarse grain approaches that estimates the pose if it corresponds few discrete poses and fine grain approaches that estimates continuous poses. A model can be generated to estimate discrete or continuous pose estimations for new data samples. Discrete pose represents the orientation at fixed intervals (e.g., ±5 degree) and they are only capable of estimating coarse pose space. On the other hand continuous estimations can handle fine (precise) poses. In each of this two categories we are considering only approaches that includes roll estimation, as the main objective of this work is to propose precise roll estimation which is adequate for upright face normalization.

Generally, coarse grain approaches can be associated to classification problems, where per discrete pose specific training and analysis is done. Du, Zheng, You, Wu, Yuan, and Wu [10] introduced new set of rotated haar-like features at 26.5 ° for in-plane rotated face detection. They build 12 different rotated detectors by rotating an upright face detector directly to achieve in-plane rotated face detection and perform coarse roll estimation. Viola et al [30] extended VJ detector in the context of multi-view settings for non-upright faces using two-stage approach. First stage estimates the pose using a decision tree and the second stage uses N pose VJ detectors for the classification. Since the frontal VJ face detector handles ±15 ° of in-plane rotation, they developed 12 detectors locally covering 30 ° and globally covering full 360 ° of in-plane rotation. They also showed that trying all pose orientations leads to superior ROC curves, but still generates considerable false positives. Wu, Ai, Huang, and Lao [37] proposed a rotation invariant multi-view face detection method based on Real Adaboost algorithm. In order to handle the rotated faces, they divide the whole 360° range into 12 sub-ranges where a special view-based detector is applied. Their multi-view face detector subsystem retrieves 89.8% of the faces for CMU profile face test set. Pan, Zhu, and Xia [24] proposed a face detection system based on heterogeneous feature representation and feature selection using Particle Swarm Optimization and Adaboost. They used SVM to remove non-face patterns in the last stage of their three-stage hierarchical classifier. A test image is scanned by different view detectors using PSO-Adaboost algorithm. In order to improve the detection speed, they employed Width-First-Search (WFS) tree structure. They focused on in-plane and out-of-plane face detection in coarse pose space. According to experimental results, they obtained 92% detection accuracy on CMU profile face dataset up to ±22.5 ° out-of-plane rotations. However the detection rate decrease to 24% when yaw rotation angle reaches 90 °.

The above approaches do not conform to the requirements of upright face normalization where the required precision is largely inferior to the common 15° span of discrete poses. In the following we focus on approaches that consider continuous or fine grain rotation estimation. Although, our main objective is providing roll estimation from static images, we also discuss works that were proposed in a video context as they are predominant in the field of fine-grain estimation and they generally yield better results as they can take advantage of the temporal information. We pay special attention to their initialization procedure as they usually requires frontal faces in order to study if our method conforms their requirements.

As underlined in Murphy-Chutorian and Trivedi [20] it is a challenging task to provide a single taxonomy for the multitude of head pose approaches. Most of the time solutions are implemented in several layers supporting hybrid methods. Other than still image based approaches, there are also tracking and motion based approaches that uses information from the temporal domain. In this study, we decide to organize the related works in the following categories: head model based tracking, point of interest (landmarks, corners) detection and tracking, holistic, taking into account their predominant characteristics. We start with the later as they appear closely related to our approach in term of constraints and initial assumptions.

2.1 Holistic approaches for in-plane rotations

Rowley and Kanade [26] proposed a two-step rotation invariant face detection using two neural network components: a router and a detector network. The router network assumes that a sample window contains a face and then tries to estimate its orientation. After the de-rotation process, detector network performs conventional frontal detection. Their router network generates ±10 ° of angular error for in-plane face orientation estimation for 92.0% of the faces from CMU dataset. Zhou, Lu, Zhang, and Wu [44] studied orientation analysis of human head by using orientation histograms. They compute the orientation histogram from gradient images where the magnitude value of the pixel exceeds a defined threshold. Since this approach employs statistical measurement of the orientation, it is less sensitive to the noise (e.g., gaussian, black-white or speckle). They detect orientation of the principal axis up to 98% of the faces at ±5 ° of error. Osadchy [23] propose a method for simultaneously detecting faces and estimating their pose in real time. A convolutional network is used for map faces on pose-dependent manifolds and non-faces outside existing manifolds. In the reported experience, the authors annotate and crop training and tests images so that the midpoint between the eyes and the center of the mouth are positioned in the center of the image, 10 pixels apart in 32×32 pixel image with some moderate variation of location and scale. We can argue that in this work, there is a strong assumption on the fact that the image presented either a canonical view of a face or an unrelated patch. Our proposition is more generic with regard to face detection and roll detection, as it does not use any prior knowledge and do not require specific training other than the generic frontal training. Guo, Kotsia, and Patras [12] used Supervised Multilinear Learning Model for head-pose estimation using Support Tensor Regression (STR). They showed that STR gives better and clearer spatial patterns for head-pose estimation than Support Vector Regression (SVR). They obtain 5.8 ° and 5.6 ° of MAE for SVR and STR on BUFT dataset respectively. [3] uses a two layer approach : the first layer uses a particle filter for keeping track of the face region and the second layer classify the pose of the tracked region. The authors used 140 persons from PIE and FERET for training the model. They report on the importance of face tracking, as errors in the tracking phase (larger bounding boxes) influence negatively on the classification process. Besides, their study require an initialization phase for the face tracking. My and Zell [21] proposed a real-time method to combine an adaptive correlation filter with Viola-Jones object detector for unconstrained face tracking. They also used depth sensor to estimate the face size for improving the tracking performance. They estimate the roll and yaw angles on BUFT dataset and obtained 3.53 ° and 5.67 ° of MAE respectively. Yan et al [41] proposes a person-independent pose estimation solution that uses person-dependent manifolds but all the data used for training and testing is fully annotated with head bounding boxes. They reported a MAE of 4.03 ° in the CHIL multi-view dataset from CLEAR07 evaluation [32]. The work could benefit from our approach as all poses involving limited yaw and pitch could be detected precisely and hence serve to better tune the proposed system.

Although the above approaches consider the whole face, except for Rowley and Kanade [26], which is orientation-independent, the other requires either close to frontal faces or assumptions about the location of the face.

2.2 Point-of-interest based approaches for in-plane rotations

Face orientation can also be estimated through facial feature analysis (e.g., eye detection [7, 43]). Slop θ of the line passing through left and right eye has a direct correlation with the orientation of the face. However, robust eye detection is also a difficult computer vision problem and can easily be affected by lighting, occlusions (e.g., hair, sunglass, and face rotation), eye closure and eye-glass reflections. In order to estimate θ value, both of the eyes must be correctly detected. More complex solutions based on specific or generic landmarks and landmark tracking systems are available in the literature. Asteriadis et al [2] address the problem of head pose estimation using one camera in uncontrolled environments. They uses Distance Vector Fields (distance to the nearest corners) for facial feature tracker and relate features' location to face boundaries. The roll angle is computed from eyes coordinates. The tracker need to be initialized by a frontal face with limited in-plane and out-of-plane rotations. Subsequent faces are tracked after skin filtering. Oka et al [22] uses a stereo vision systems and relays on the usage of OKAO vision library that provide information about face and 6 face landmarks (eyes and mouth corner). Four more points (nostrils and inner brows) are detected using dedicated algorithm. Tracking of landmarks is then performed. In this work, the initialization phase is subject to local noise occlusions, cluttering, etc., that might impact the feature selection process, whereas, this noise can be better supported by face frontal detector as in our approach. The authors report 0.7 ° of MAE for roll, but the results were reported from analyzing only 2 videos (20 seconds long). The videos do not seem available anymore, but the captions included in the article show images containing one big face with limited in-plane rotations. Their initial face detection methods seems in a sense similar to ours, but they do not compute confidence levels and neither extract in-plane orientation angles.

As landmark detection is subject to local noise, some authors exploit 3D information available in multi-view settings to make consistent landmark tracking. Wang and Sung [33] uses a stereo-vision system for reconstructing 3D model of the face, but also consider 2D facial landmarks (eyes, mouth) for estimating the orientation considering the vanishing point of the lines passing by eye corners and moth corners respectively. They use jointly the 2D and 3D model in order to better support changes in expressions with regard to the pose. The initial assumption about the face position seems to be quite important, and landmarks are subject to much noisier detections than the global face detector used in our approach. Authors report 3.54 ° error for a restricted collection of videos where the head orientation ground truth was recorded using an inertial sensor. Zhao, Chen, Song, and Chen [42] uses a stereo vision system together with SIFT points tracking in 2D consecutive frames. The tracked points are reconstructed in 3D and then a 3D registration method is used for detecting the head pose. In this approach the initialization step is subject to frontal face detection with limited roll. The experiments seems quite limited as only 3 sequences were tested yielding 2.86 ° MAE for roll. A similar approach is proposed in Tran, Ababsa, Charbit, Feldmar, Petrovska-Delacrtaz, and Chollet [28] where a Bayesian method deals with the problem of face tracking using a 3D adaptive model through eigen decomposition. They use SIFT features to track the landmarks constrained by the 3D shape and Singular Value Decomposition (SVD) to update the tracking model. They obtained 2.4 ° of MAE on BUFT dataset. Jung and Nixon [15] mainly addressed the face recognition problem by reconstructing high-quality frontal face images from a 3D ellipsoidal head-pose model and gait information. They used region and distance based refinement of head-pose estimation by SIFT features where they obtained 2.1 ° of MAE on BUFT dataset.

Although the joint use of 2D and 3D methods make the orientation more precise, in the initialization phase, these approaches are subject to problems that might arise in the landmarks detection process. The local noise on the face (occlusions, cluttering, etc.) is generally better supported by the face detector than landmarks detectors in a static context corresponding to the initialization step.

2.3 Head Models approaches for in-plane rotations

As stated earlier, we have included planar/cylindrical tracking approaches as they provide a challenging context for validating our method of roll estimation.

Sung, Kanade, and Kim [27] used video frames to find global head motion parameters for their Cylinder Head Model (CHM). They achieved 3.1 ° of RMSE (Root Mean Square Error) error on BUFT dataset. [16] uses a cylindrical tracker and image registration techniques for estimating the head orientation in inner and outer plan rotations. From plots provided in the paper on BUFT, Murphy-Chutorian and Trivedi [20] estimated the roll MAE at 9.8 °. Valenti, Yücel, and Gevers [29] used both CHM pose tracker and isophote based eye center locator for visual gaze estimation. They achieved, at best, 3.0 ° of RMSE error with eye detection support on BUFT dataset. For their fixed initial template experiments with and without eye cues they obtained 3.0 ° and 3.85 ° of RMSE error respectively. Similarly, they obtained 3.93 ° and 4.15 ° of RMSE error for updated template experiments. An and Chung [1] replaced the traditional CHM by a 3D ellipsoidal model and achieved 2.83 ° of MAE error on BUFT dataset. However, the initialization phase of the above methods requires frontal face detection. Hence, leaving out the tracking phase where the head orientation can be tracked outside de limits of a frontal tracker, the method presents similar faults to our first layer as they both require frontal detection.

Morency, Whitehill, and Movellan [18] present a stereo camera system that jointly perform head tracking and head orientation using Generalized Adaptive View-based Appearance Model for head-pose estimation. Their probabilistic framework provides frame-by-frame head-pose tracking using keyframe updates with Kalman filter. The pose is estimated using linear regression. They obtain 2.91 ° of MAE on BUFT dataset. The tracking system is initialized with the assumption that the object of interest is the front-most object in the scene, as determined by the range images. This kind of assumptions could be simulated by metrics related to contours or other dominant 2D features (skin colorimetry), however, in our case, we are not considering only one face, but all the faces presented in an image. Murphy-Chutorian and Trivedi [19] uses an appearance-based particle filter for tracking the head combined with a static estimator for checking consistency. The particle samples generated corresponds to projection and translation of a generic head model considering slight variations in terms of roll, yaw, tilt and offset. They achieve an MAE of 2.38 ° for roll angles in a multi-view system deployed in a car for surveilling driver activities. As other tracking techniques, the initialization part requires frontal poses and it impacts the global process. The initialization part is also subject to frontal detection flaws. Xiao et al [40] uses a planar/cylindrical based approach constructed of three levels dealing each one with specific motions (translations and roll for the first level, yaw and pitch on the following). The system requires a frontal pose or the annotation of the head region in a given frame as an initialization step. On average, this methods provides on average 1.9 ° of errors on BUFT dataset. Lefevre and Odobez [17] proposed a deformable 3D face model and an array of view-based templates to model the head appearance online. In addition to the frontal faces, they consider head side to robustly track challenging head movements. In their approach, different appearances and poses addressed by different templates. Their head tracking method obtained 1.9 ° of MAE on BUFT dataset.

Globally, the drawbacks reproached to our approach with regard to out of plane detection are also common to other head orientation techniques that require the initialization from a frontal face. Our solution is not to be seen as a replacement of existing head tracking techniques as it only concerns the roll detection, but more over as a more robust face detector (to in-plane rotation) for the initialization phase. The exhaustive search of faces, the confidence levels and the low false-positive rates yielded by our approach may support in loosely constrained settings existing head tracking solutions (without presumption of a face present in the scene). Summary of the results in the literature can be found in Section 4.4.1, Table 4.

3 Methodology

The first step towards robust in-plane face orientation estimation is rotation invariant face detection. The simplest method is to repeat frontal face detectors at different orientation angles e.g., [30]. However, face detection at different orientation angles is a hard asymmetric classification problem as there is no balance between positive and negative classes which increase the possibility of false positive (FP) detections more than traditional frontal detection at a single angle. In order to handle both in-plane rotations and false positive detections, we used a two-level estimation algorithm based on VJ face detector. Other face detection algorithms having the rotation disadvantage can also be used as a base detector. Figure 3 shows flow diagram of proposed algorithm. First level takes into account coarse face detection at generic orientation angles (0 °, 90 °, 135 °, 180 °, -135 ° and -90 °) as well as common upright head-pose angles (from -50 ° to $+50$ ° by 15 ° of increments). The second level detects in-plane face orientation by multiple face detections at orientation angles from -40 ° to $+40$ ° by 3 ° of increments.

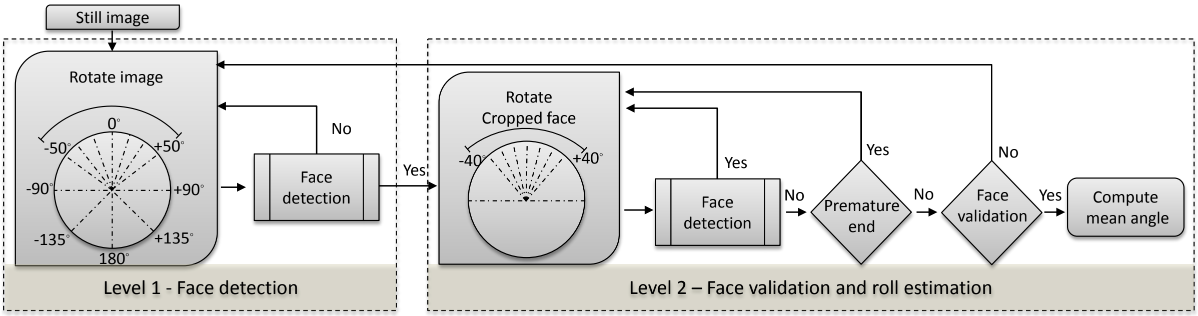


Fig. 3 Flow diagram of proposed algorithm

3.1 Face candidate selection

We used the frontal face detection model available in OpenCV. Considering Fig. 3, face detector at Level 1 and Level 2 has different parameter settings (scaling factor, neighborhood, min/max face size).

Scaling factor parameter specifies how much the image size is reduced at each image scale. Therefore, it is used to create user defined scale pyramid. Using a small scaling factor such as 1.1 means that the size is reduced by 10% in the next step which increases the chance of a matching size with the trained model for detection is found. Smaller scaling factor provides more search windows thus increase both the response probability and computational time. For this reason, the first level focuses more on permissive strategies to detect the face while the second level focus on face validation and orientation estimation.

Neighborhood parameter specifies how many neighbors each candidate rectangle should have to retain it. It affects overall quality of the detected faces. Higher neighborhood parameter value results in less detections but with higher quality. Min/Max face size specifies the minimum and maximum possible object size. Objects that are smaller than the min face size and objects that are larger than the maximum face size are ignored. We choose the scaling factor as 1.15, neighborhood parameter as 3 and min/max face size factor is set to 2^{-5} of the input image *width* and *height* on Level 1 respectively.

First level detections are performed for each predetermined orientation as shown in Fig. 3. Every face detection action in Level 1 is overlaid by black rectangle to ignore these face candidates for the rest of Level 1 detections at different orientations. However, if Level 2 does not validate the Level 1 detection then the overlaid face is restored back to its original intensity values. Validated faces in Level 2 are replaced by gray level filled elliptical mask as shown in Fig. 4 (b).

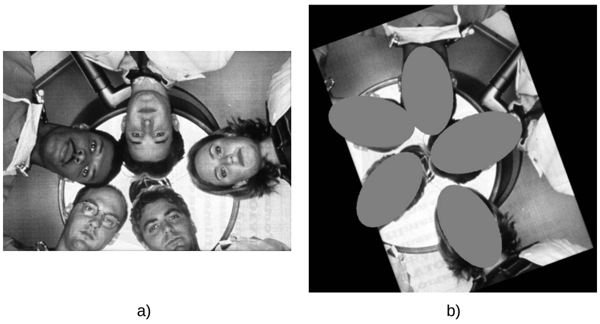


Fig. 4 a) Original image from CMU rotated dataset. b) Validated faces in Level 1 (intermediate state)

3.2 Face orientation estimation

Level 2 uses 25% enlarged crop of the candidate faces detected in Level 1 with respect to the face center. These faces are scaled up or down to 100×100 pixels resolution for Level 2 processing. Cropped faces are in candidate status unless they are validated by the second level. In Level 2, it is expected that there is a face inside the cropped region. Therefore, more relaxed scaling parameter (1.2) and bigger min and max face size factor (2^{-3}) is used in Level 2. It allows faster processing as well as reduced false positive detections. Note that Level 2 both validates the face and estimates the precise orientation.

Level 2 starts by rotating the face -40° in counter clockwise direction while assuming that the image contains a frontal upright image. Level 2 estimates the in-plane face orientation by multiple face detections at different orientation angles from -40° to $+40^\circ$ by 3° of increments. In each iteration i, the original cropped face from the first level is rotated by $-40 + (i * step\_factor)^\circ$. It provides detection of a frontal face within the 30° in-plane rotation in theory. The 30° comes from the $\pm15°$ of in-plane rotation capability of frontal VJ detector. We added $\pm10°$ of extension to the $\pm30°$ degree theoretical detection boundary of the VJ detector. Experimental observations on CMU rotated dataset showed that $\pm10°$ of extension is sufficient enough to obtain high face detection accuracy (Precision: 97.2%, Recall: 93.2%, F1: 95.2% for $\pm9.0°$ of MAE). Note that this is an arbitrarily chosen extension to the theoretical detection boundary and it can be further increased or decreased depending on the problem space.

In order to validate a face we defined a confidence value C. The second level validates and estimates the orientation considering the confidence and step factor parameters. Confidence is the total length of sequential face detections in Level 2 as shown in Fig. 5. Higher confidence value of a face leads to better accuracy for the orientation estimation, since the face can be detected at many different orientation angles. This is similar to the frontal face detector of Viola and Jones, where a single face is detected by multiple search windows at different scales and positions. In VJ detector, among others only one scale is selected to show the detected face. In our method, only one angle (mean angle) is selected among the sequential detections represented by the confidence value.

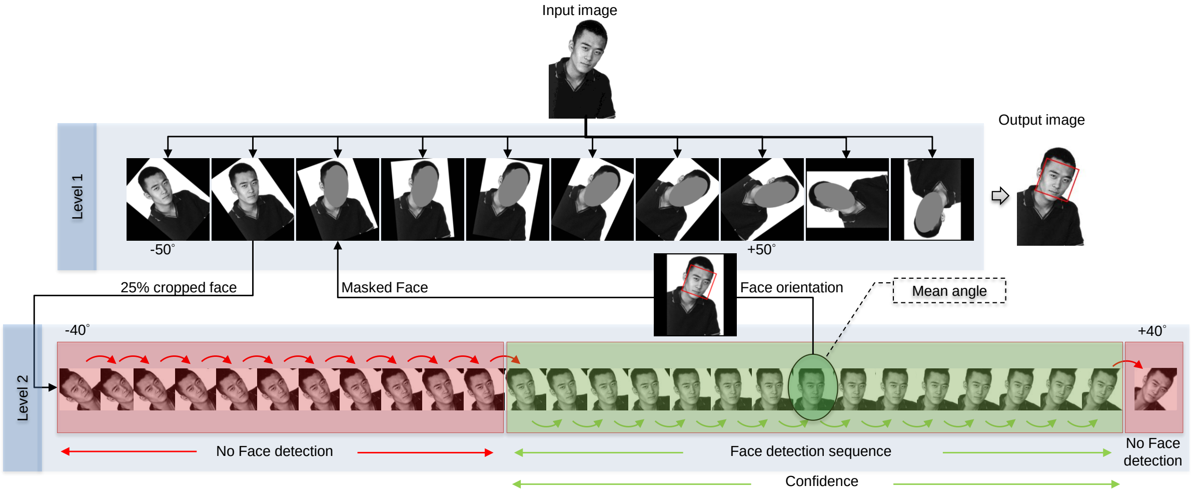


Fig. 5 Sample processing diagram showing Level 1 and Level 2 operations for a still input image

In order to determine the optimal value for the confidence C, we defined a threshold value T that maximizes F_1 value (harmonic mean of *precision* and *recall*) in CMU rotated dataset. Main reason to select CMU dataset is its large variety of in-plane rotation samples. According to experimental studies illustrated in Fig. 6, we set the threshold value $T = 21^\circ$ for face validation in Level 2 where F_1 value is maximized (95.25%). As a result, a face is validated in Level 2 only if there exists at least an angular confidence of $C = 21^\circ$ of successive detections. When a face is validated in Level 2 ($C \geq T$), the face is no more visible for Level 1 as shown in Fig. 4 (b). This behavior further eliminates the false positives.

As a simple example, a perfect frontal upright face without any out-of-plane rotation has a multiple detection sequence starting from min$=-15^\circ$ to max$=+15^\circ$ having the mean angle $\theta = 0^\circ$ (-15, -12, -9, -6,..., +9, +12, +15) and confidence value as $C = 30^\circ$ as the face was continuously detected over a range of 30°. In order to estimate the correct θ_D angle (detected orientation angle) of the face ($\theta_D \approx \theta_A$), with regard to the θ_A (ground-truth orientation angle), we keep track of the multiple detections by 3° of increments. Algorithm 1 shows the details of Level2 operations.

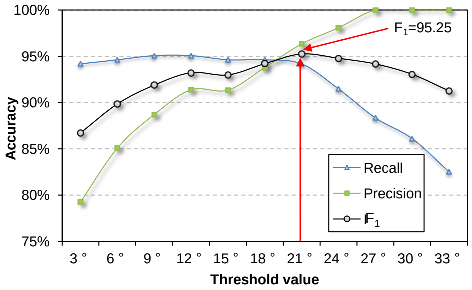


Fig. 6 $T = 21\,°$ provides the maximum $F_1 = 95.25$ on CMU rotated dataset. Extracted parameter $T = 21\,°$ is used statically in the rest of the experiments for all datasets

Algorithm 1 *In-plane face orientation estimation for the second level*

```
procedure LEVEL2($croppedFaceImage$,$angularThreshold$,$stepFactor$)
    $sequenceStarted \leftarrow false$
    $angleIteratorCount \leftarrow 0$
    $cummulativeAngle \leftarrow 0$
    for all $angle$ such that $-40 \leq angle \leq 40$ do
        $rotatedImage \leftarrow$ rotate($croppedFaceImage, angle$)
        if VJDetectFace($rotatedImage$) then
            if Not $sequenceStarted$ then
                $cummulativeAngle \leftarrow angle$
                $angleIteratorCount \leftarrow 1$
                $sequenceStarted \leftarrow true$
                $prevAngle \leftarrow angle$
            else
                if $angle - prevAngle = stepFactor$ then
                    $cummulativeAngle \leftarrow cummulativeAngle + angle$
                    $angleIteratorCount \leftarrow angleIteratorCount + 1$
                    $sequenceStarted \leftarrow true$
                    $prevAngle \leftarrow angle$
                else
                    if $angleIteratorCount \times stepFactor \leq 15$ then
                        $sequenceStarted \leftarrow false$
                    end if
                end if
            end if
        end if
        $angle \leftarrow angle + stepFactor$
    end for
    if $angleIteratorCount \times stepFactor \geqq angularThreshold)$ then
        return ($cummulativeAngle/angleIteratorCount$)
    end if
    return $NoFaceDetected$
end procedure
```

end

croppedFaceImage parameter is 25% enlarged crop of the face detected in Level 1. *angularTreshold* parameter is the face validation threshold value (T=21 for the experiments). *stepFactor* parameter determines the amount of rotation for each iteration in Level 2. *sequenceStarted* variable controls the start of a detection sequence. It is true when the first face is detected in Level 2 and it is false when a detection sequence is interrupted. *angle* variable represents the amount of rotation applied to the *croppedFaceImage*. *rotatedImage* variable is the rotated *croppedFaceImage* variable. *cumulativeAngle* variable holds the amount of the rotation applied to the *croppedFaceImage* in a valid sequence. *angleIteratorCount* variable holds the number of iterations for the current sequence. *prevAngle* variable is used to detect interrupts in the detection sequence. Line 20 in Algorithm 1 controls the premature ending of the detection sequence. If the sequence is longer than 15 ° then the process continues with the next iteration, otherwise it stops. This is different than T=21 which validates the complete sequence.

4 Experiments

In this section we provide datasets, parameter selection, evaluation methodology and experimental results obtained from in-plane, out-of-plane, and low resolution tests.

4.1 Datasets

In order to demonstrate the effectiveness of the proposed method, we performed quantitative experiments on variety of controlled and unconstrained datasets including CMU Rotated, BUFT, YouTube-Faces, Caltech, FG-NET Aging, BioID and Manchester Talking-Face datasets. Experimentation datasets are selected from publicly available datasets providing the left and right eye position in order to estimate roughly the ground-truth for the in-plane rotation of the head. Main reason to include frontal datasets in the experiments is to show that proposed method is robust to both in-plane rotated and near frontal upright faces for in-plane face orientation estimation.

General in-plane rotation information of these datasets is characterized by Mean Absolute Rotation (MAR) as shown in Eq. 1:

$$MAR = \frac{1}{n}\sum_{k=1}^{n} abs(\theta_{A_k}) \tag{1}$$

Where n is the number of detected faces in the dataset and θ_A is the ground-truth roll angle. The near frontal datasets have an MAR close to 0 and more challenging datasets with a large palette of in-plane rotations have larger MAR value. Table 1 summaries the characteristics of the datasets used in the experiments. The variety of the features of the selected datasets guarantees a thought validation of our method in a wide collection of settings.

BioID dataset [13] consists of 1521 grayscale images of 23 subjects having variety of illumination background and face size. Left and right eye locations are provided as ground-truth data.

Boston University Face Tracking (BUFT) dataset [16] is a low resolution (320 × 240 pixels) video dataset having 45 video sequences of 5 subjects performing 9 different head motions in a controlled office setting. The dataset has annotations for the left and right eye as ground-truth information. However, there are two different ground-truth definitions available for BUFT dataset. The first ground-truth is captured by magnetic sensors "Flock and Birds" while the latter is eye-based manual annotation. Therefore first we evaluate the null hypothesis regarding the probabilities of two different measurements (head-position orientation and eye-based orientation). Using the Pearson's method, we obtained ($r = 0.9843$, $p = 1.19e - 151$). Since there is a very high positive correlation and the p-value is a lot less than 0.01 (approximating to

Table 1 Summary of the datasets. C=Controlled, F=Frontal, GT=Ground-truth, MF=Multiple faces, R=Rotated, U=Unconstrained, MAR= Mean absolute rotation from upright

Dataset	Type	Images	Faces	MAR	Roll Range	Yaw Range	Pitch Range
BioID [13]	C, F,GT	1521	1521	2.20 °	[−32.6 ° + 14.8 °]	–	–
BUFT [16]	C, R,GT	8955	8955	7.03 °	[−38.7 ° + 39.8 °]	[−46.3 ° + 39.7 °]	[−29.7 ° + 35.6 °]
Caltech [35]	C, F,GT	450	450	2.03 °	[−11.6 ° + 13.2 °]	–	–
CMU [26]	U, R,MF,GT	50	223	35.60 °	[−101.4 ° + 180 °]	–	–
FG-NET [11]	C, F,GT	1002	1002	5.07 °	[−32.0 ° + 29.4 °]	–	–
Manchester [5]	C, F,GT	5000	5000	5.03 °	[−23.6 ° + 11.5 °]	–	–
YouTube C [36]	U, R,MF,GT	4000	> 4000	8.86 °	[−32.4 ° + 29.0 °]	[−92.6 ° + 95.2 °]	[−46.8 ° + 44.0 °]

zero) we can say that the data is statistically significant and the relation really exists. There is an average of 2.1 ° of shift between the two ground-truth values. Since the orientations obtained from the manually annotated eye positions are more robust than measurement by an electronic device, we used eye based annotations in the experiments. Figure 7 shows the two different ground-truth value and the correlation.

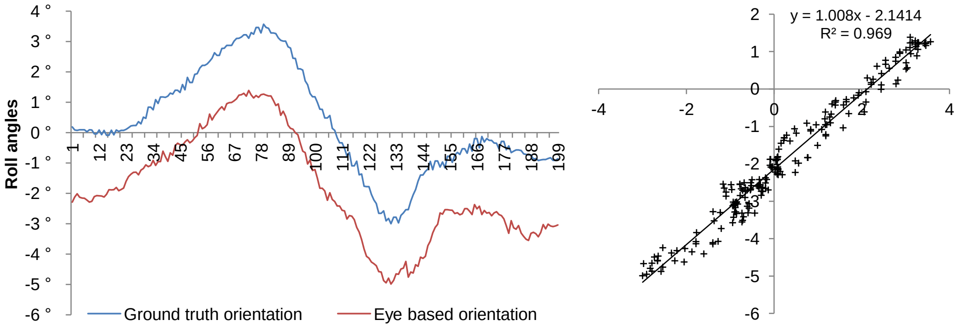


Fig. 7 Head-position based and eye-based ground-truth for BUFT dataset and corresponding correlation obtained from averaging 45 video sequences

Caltech frontal faces dataset [35] consists of 450 frontal face images about 27 unique people under different lighting, expression and backgrounds. In order to estimate the face orientation, we manually annotated the left and the right eye locations.

CMU Rotated dataset [26] consists of 50 images containing 223 frontal faces with in-plane rotations. The dataset is publicly available and provides ground-truth position information for left-eye, right-eye, nose, left-corner-mouth, center-mouth and right-corner-mouth.

FG-NET Aging dataset [11] is an image dataset containing 1002 face images from 82 subjects at different ages. The dataset is annotated with 68 point facial feature points including the left and right eye locations.

Manchester Talking-Face video dataset [5] consists of 5000 frames taken from a video of a single person during a conversation. It presents the behavior of the face in natural conversations. Ground-truth data is provided using an AAM based tracker giving 68 facial feature points including the eye positions. Ground-truth auto-annotations were visually validated by human observers.

YouTube-Faces [36] dataset is an unconstrained large-scale collection of videos along with labels indicating identities and pose (pitch, roll and yaw angles) of the faces. There are 3425 videos of 1595 subjects providing a grand total of 614895 auto-annotated frames. Three rotation angles of the head are auto-annotated by the Web-API of face.com. Since provided ground-truth does not refer to manual pose an-

Table 2 Description of the YouTube datasets

Dataset	Description	Ground-truth	MAR
YouTube B	4000 frames having the worst RMSE	auto-annotated	16.90 °
YouTube C	The same 4000 frames	manually-annotated	8.86 °

notation, we manually annotated the eye positions for a special subset of 4000 images. This is because that, auto-annotated pose information is not stable for sequential frames which affect overall evaluation performance of our algorithm. The subset is selected from the images having the worst 4000 estimations for which we obtain the highest RMSE. Table 2 presents short comparison of provided ground-truth and manually annotated ground-truth.

4.2 Parameter selection

Selection of near optimal parameters has a crucial impact on results. Considering the VJ algorithm, we set the search flags as: CV HAAR FIND BIGGEST OBJECT and CV HAAR DO ROUGH SEARCH in both of the levels. The search flag CV HAAR FIND BIGGEST OBJECT is set to terminate after the first window was declared as face. It affects the search time since the rest of the user pyramid is not tested after the first positive match which is the largest face. CV HAAR DO ROUGH SEARCH skips some steps that would shrink the region of this match.

4.3 Evaluation

For each dataset, we compute annotated in-plane face orientation angle θ_A with respect to the vertical axis using the slope m of the line passing through the left and right eye. Figure 8 shows sample images from CMU rotated dataset with annotated θ_A angles. Difference between θ_A and θ_D is the error value $\theta_{Err} = \theta_D - \theta_A$ with respect to the vertical axis.

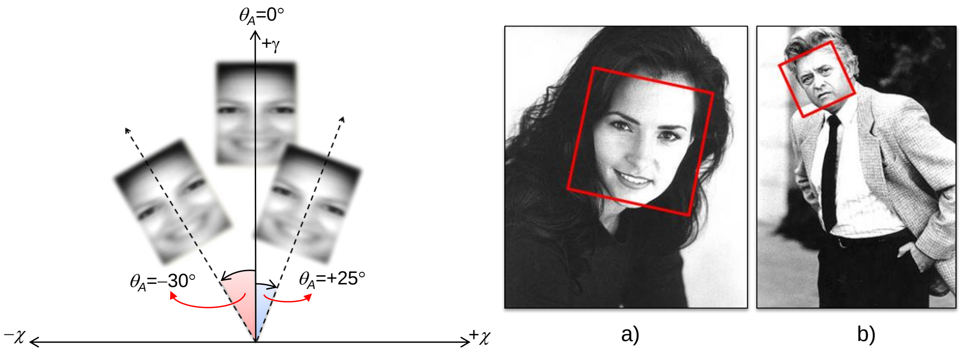


Fig. 8 Sample ground-truth values from CMU rotated dataset a) $\theta_A = +12.63\,°$ b) $\theta_A = -26.56\,°$

In order to provide quantitative evaluation of proposed face orientation estimation algorithm, we presented our results in terms of $RMSE(\hat{\theta})$ (Root mean square error), $MAE(\hat{\theta})$ (Mean absolute error) and STD (Standard deviation).

Since CMU rotated and YouTube face datasets have multiple faces for the majority of the images, we employed face matching based on Euclidean distance between the annotated and detected faces. First, we find the center point between the left and right eye in annotated faces represented by Ct_A. Then we find the center of gravity of the detected face Ct_D. ΔCt is the Euclidean distance between Ct_A and Ct_D. We used Inter Pupillary Distance (IPD) based threshold to match the faces. IPD is the Euclidean distance between the two ground-truth eye centers. Finally, a face is said to be matched if $\Delta Ct \leq 2.0 \times IPD$ as shown in Eq. 2:

$$Face_i = \left\{ \begin{matrix} Matched & \Delta Ct_i \leq 2.0 \times IPD_i \\ Nomatch & otherwise \end{matrix} \right\} \tag{2}$$

Figure 9 presents the illustration of Eq. 2 on a sample image from CMU dataset. According to Eq. 2, center of gravity of the face Ct_D must be inside of the yellow circle. Matched face is represented by the white rectangle. Considering all of the matched faces, we compute the MAE, RMSE and STD. When a face is detected as a FP, logically there is no ground-truth available (no Ct_A) to compute the MAE since there is no actual face. Therefore we did not use FP detections in the MAE computation.

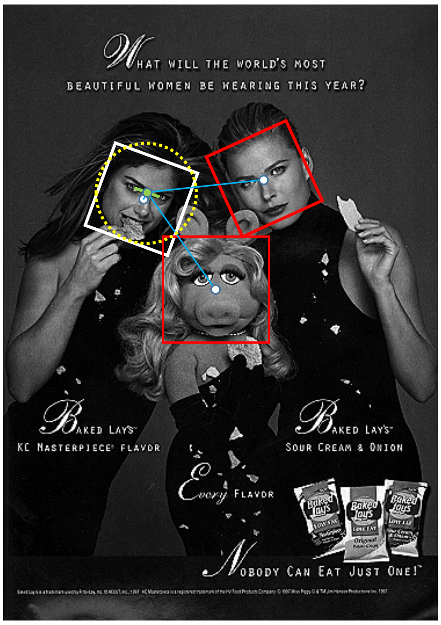

What will the world's most beautiful women be wearing this year?
Baked Lay's
KC Masterpiece® flavor
Baked Lay's
Sour Cream & Onion
Every Flavor
Nobody Can Eat Just One!™

Fig. 9 Face matching example from CMU database. Ct_D (Center of gravity of the face) is represented by white filled circles. Green filled circle presents Ct_A (center point between the left and right eye) of the current face. IPD distance is presented by the green line. Yellow circle shows the boundaries of the maximum ΔCt for the given face (radius= 2 × IPD). Blue lines present Euclidean distance between Ct_A and Ct_D.

4.4 Experimental results

We provide results for continuous roll estimation in Table 3. As we emphasized before, the YouTube (original) dataset does not provide manual annotations, therefore we present results from YouTube C set having manual annotations for 4000 images.

4.4.1 Face orientation experiments

Table 3 shows RMSE, MAE and STD obtained from the test datasets. Mean value of the MAE for the seven different datasets is 2.16 °. The first two minimum MAE (1.13 ° and 1.69 °) and STD (0.91 ° and 1.18 °) are obtained from Caltech and Manchester datasets respectively. Compared to the other datasets, face size to frame size ratio is relatively bigger and there are small changes in yaw and pitch rotations in these two datasets.

Table 3 Experimental results for face orientation estimation (roll) in terms of RMSE, MAE and STD for different datasets.

Dataset	RMSE	MAE	STD
BioID	2.73 °	2.20 °	1.61 °
BUFT	2.68 °	1.93 °	1.83 °
Caltech	1.46 °	1.13 °	0.91 °
CMU	3.75 °	2.87 °	2.42 °
FG-NET	2.63 °	1.86 °	1.86 °
Manchester	2.07 °	1.69 °	1.18 °
YouTube C	5.59 °	3.50 °	4.36 °

The maximum MAE (3.50 °) is obtained from the most difficult YouTube C set. This is because that, YouTube C has large variety of illumination and mixed pose changes (more than 30 ° of yaw and pitch) in very low resolution images which increases MAE. Resolution is an important feature to consider since a few pixels of human annotation error in a low resolution image may generate considerable annotation errors. For example, a single pixel error in vertical position (y value) of either left or right eye in a 10 pixels wide face generates approximately 5.71 ° error. Therefore face width with respect to the frame resolution is an important factor for manually annotated datasets. Figure 10 shows sample detections and estimation errors from YouTube C dataset where the white line presents the ground-truth. Main reason for the estimation errors come from false positive detections from Level 1. In general, partially wrong face detections in Level 1 corresponds to false positive faces but technically they include faces inside the window as shown in the wrong detections in Fig. 10. Another say, they include much more background pixels than face pixels. When Level 2 operates with 25% enlarged crop of this window, face size with respect to the cropped region more decreases. Therefore, precise estimation is not possible in such cases.

According to the experiments, our method reduced the MAE to less than 2.0 ° for frontal datasets (e.g., BioId, Caltech, Manchester) having 2.03 ° to 5.03 ° MAR from the vertical axis (upright position). In case of rotated datasets, all RMSE values are less than 3.0 ° except CMU. However, compared to the other datasets, CMU has the highest MAR (35.60 °) while having the smallest number of images and faces (see Table 1). Figure 11 (a) shows that our method works well regardless of original face orientation angle. The MAE is stable (approximates to horizontal line) over the all the test datasets. Figure 11 (b) presents θ_{Err} coverage for each test dataset. In average, our method covers 91.69% and 95.14% of the test datasets at $\theta_{Err} = \pm 5.0^\circ$ and $\theta_{Err} = \pm 6.0^\circ$ of error respectively.

Table 4 compares our face orientation estimation method on BUFT dataset with other state-of-the-art methods in terms of RMSE, MAE and STD. The performance of our method in terms of MAE is comparable to the studies provided in Xiao et al [40] and Lefevre and Odobez [17]. Note that we did not use any temporal data during the experiments. Figure 12 shows images having the highest (first row) and lowest (second row) θ_{Err} for face orientation estimation. The worst orientation estimation occurs when the face has more than 30 ° yaw rotation (Fig. 12 (e)(f)(g)(h)). More examples from the BUFT and CMU datasets can be found in Appendix A, Fig. 19 and Fig. 20.

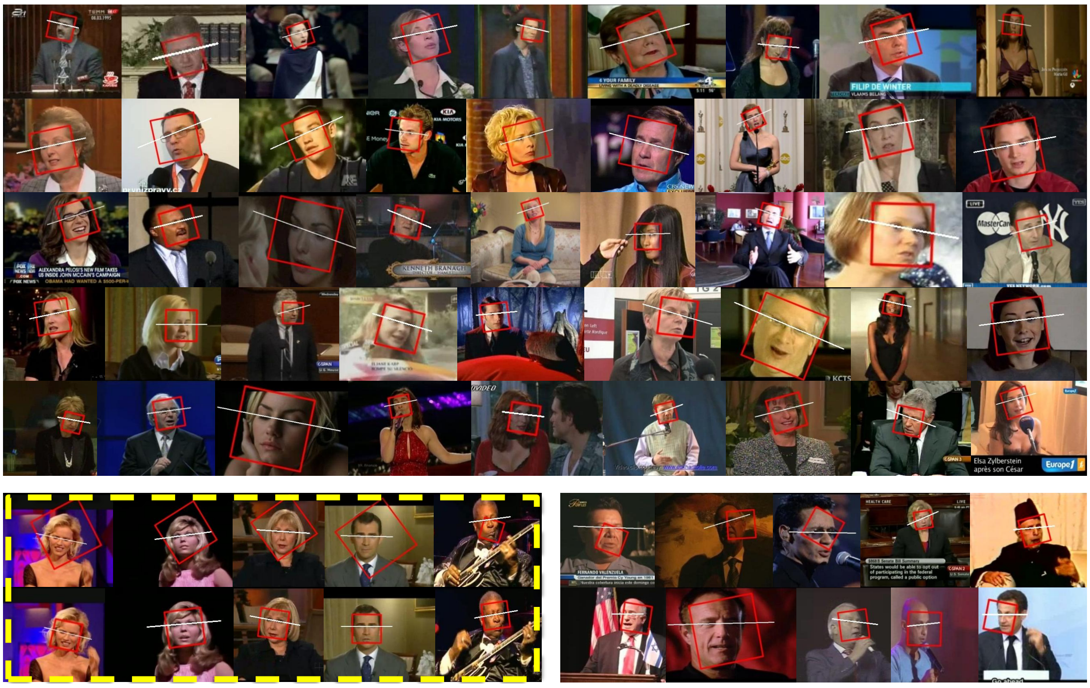


Fig. 10 Good (first five row) and problematic (last two row) estimations from YouTube C dataset where the white line presents the ground-truth orientation. Some of the problematic examples (e.g., first five column of the last two row highlighted by the yellow rectangle) are singular incidents and they are corrected in the following frames. The problem occurs when the frontal face detector in Level 1 detects a large false positive face which actually includes a real face. The errors from these examples also used in MAE computation

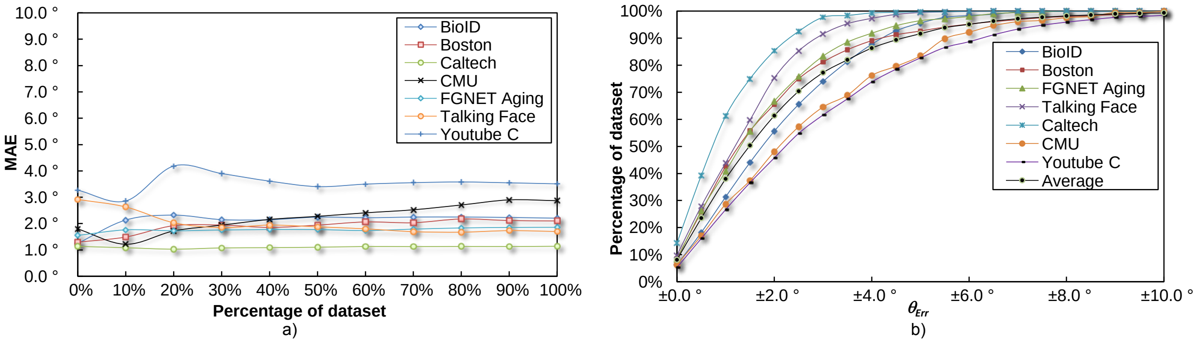


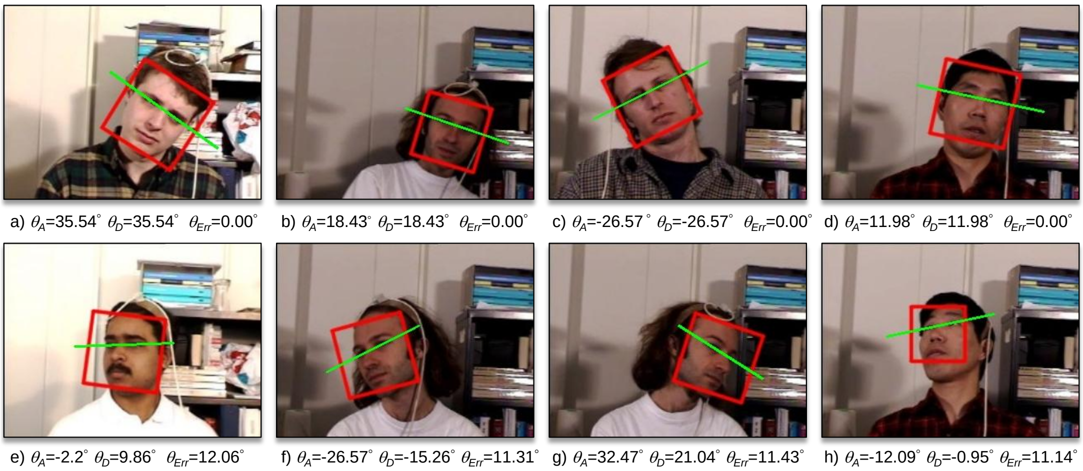
Fig. 11 a) Change of $MAE(\hat{\theta})$ in test datasets b) Angular error θ_{Err} coverage for test datasets

4.4.2 Effect of the step factor

To illustrate the effect of the step factor parameter in roll estimation, we vary the step factor over the range [1.0, 20.0] while keeping other parameters constant. Change of angular granularity in the first level only affects the amount of face candidates. However, the angular granularity in the second level affects both of the validation of the faces and the roll estimation process. Figure 13 shows the result of the step factor experiment for BUFT dataset. The MAE increases with the increasing size of the step factor. This is an

Table 4 Comparison of face orientation studies (roll) for BUFT dataset in terms of RMSE, MAE and STD results

Study	RMSE	MAE	STD
AAM + cylinder head model, Sung et al [27]	-	3.10 °	-
3D ellipsoidal head model, An and Chung [1]	-	2.83 °	-
3D cylinder head model, An and Chung [1]	-	3.22 °	-
2D planar model, An and Chung [1]	-	2.99 °	-
Fixed templates and eye cue, Valenti et al [29]	3.00 °	-	2.82 °
Updated templates and eye cue, Valenti et al [29]	3.93 °	-	3.57 °
Distance vector fields, Asteriadis et al [2]	3.56 °	-	-
Generalized Adaptive View-based Appearance Model, Morency et al [18]	-	2.91 °	-
Support Tensor Regression, Guo et al [12]	-	5.60 °	-
Support Vector Regression, Guo et al [12]	-	5.80 °	-
3D face model + Eigen-Decomposition Based Bayesian Approach, Tran et al [28]	-	2.40 °	1.40 °
3D deformable face tracking, Lefevre and Odobez [17]	-	1.90 °	-
3D ellipsoidal head pose model + gait, Jung and Nixon [15]	-	2.10 °	-
Adaptive correlation filter + Viola-Jones detector, My and Zell [21]	-	3.53 °	-
Dynamic templates and re-registration, Xiao et al [40]	-	1.40 °	-
Our method (step factor =3)	**2.89 °**	**2.10 °**	**1.99 °**
Our method (step factor =1)	**2.68 °**	**1.93 °**	**1.83 °**



a) θ_A=35.54° θ_D=35.54° θ_{Err}=0.00° b) θ_A=18.43° θ_D=18.43° θ_{Err}=0.00° c) θ_A=-26.57 ° θ_D=-26.57° θ_{Err}=0.00° d) θ_A=11.98° θ_D=11.98° θ_{Err}=0.00°

e) θ_A=-2.2° θ_D=9.86° θ_{Err}=12.06° f) θ_A=-26.57° θ_D=-15.26° θ_{Err}=11.31° g) θ_A=32.47° θ_D=21.04° θ_{Err}=11.43° h) θ_A=-12.09° θ_D=-0.95° θ_{Err}=11.14°

Fig. 12 Examples of the best (first row) and worst (second row) results obtained from BUFT dataset. Oriented rectangles in red color presents our face detection result with the corresponding orientation while the green lines shows the ground-truth orientation

expected result since increasing value of the step factor provides faster but coarser estimation of the roll angle. In the best case, we obtained 1.93 ° of MAE when the step factor is 1. However, the computational complexity also increases with lower step factors. Higher step factors provides more coarse estimation of the orientation. An example of the high and low precision obtained from different step factors are given in Fig. 14.

We obtained 2.49 ° and 4.06 ° of MAE from BUFT dataset for the step factors 10 and 20 respectively. Considering the step factor change from 3 to 6, the MAE increases from 2.07 ° to 2.21 °. Since there is a small increase in the MAE while doubling the step factor, the step factor parameter can be selected in the range [1, 6] in practical applications. We used the value of 3 in our examples.

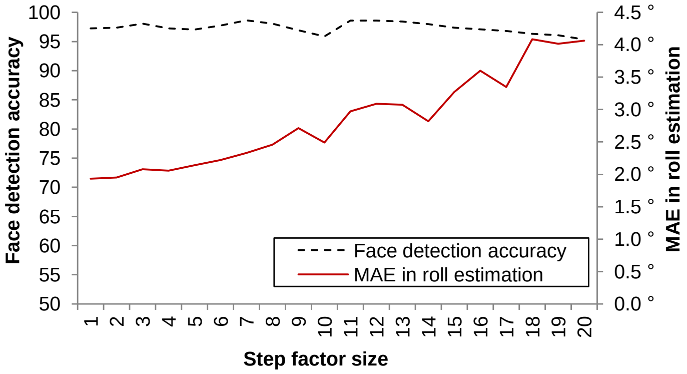


Fig. 13 Effect of angular granularity on MAE and face detection accuracy in the Level 2 for the BUFT dataset

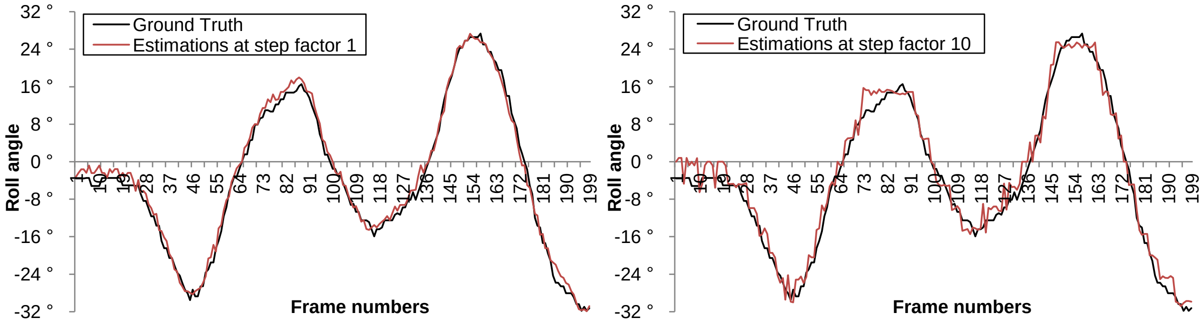


Fig. 14 An example in-plane rotation estimation from BUFT dataset at step factor 1 and 10 (Jam 1 sequence)

4.4.3 Effect of out-of-plane rotations

Considering the out-of-plane rotations, the roll estimation process depends on the capability of VJ detector which is known to be limited. Main disadvantage of our method occurs in case of out-of-plane rotations. This is mainly the same problem of underlying VJ face detector which is trained on frontal upright faces only. The roll estimation can be effectively done when an in-plane rigid rotation occurs. But in reality, it is unlikely that the center of rotation is in the center of the face. Out-of-plane rotations change the shape of the face which cannot be handled by the frontal VJ detector.

We divide the ground-truth yaw and pitch angles from the BUFT dataset into 1 ° segments (from −50 ° to +50 °). Then we created a histogram representing the amount of faces in each of the 1 ° segment as shown in Fig. 15.

Yaw angles in BUFT dataset vary from −47 ° to +40 ° and pitch angles vary from −30 ° to +36 °. However majority of the faces are in the range −15 ° to +15 ° degrees. In order to present out-of-plane rotation capability of the VJ detector using our methodology, we performed three tests.

In the first test, we measured the face detection accuracy of our method on each of these segments to see the out-of-plane detection capability of our method. We obtained 98.05% detection rate on BUFT dataset (8781/8955) and missed only 174 faces out of 8955. Then, we create corresponding histogram presenting the distribution of missed faces with respect to the out-of-plane yaw and pitch angles as shown in Fig.

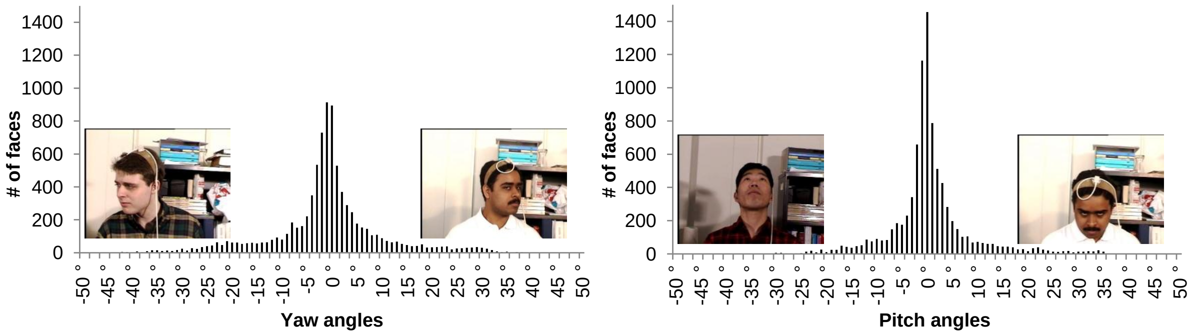


Fig. 15 Distribution of faces with respect to ground-truth yaw and pitch angles for BUFT dataset

16 (a). Figure 16 (b) shows the distribution of missed faces (False Negatives) with respect to the yaw and pitch angles. According to the experimental results, our method is less sensitive to pitch changes than yaw changes on BUFT dataset.

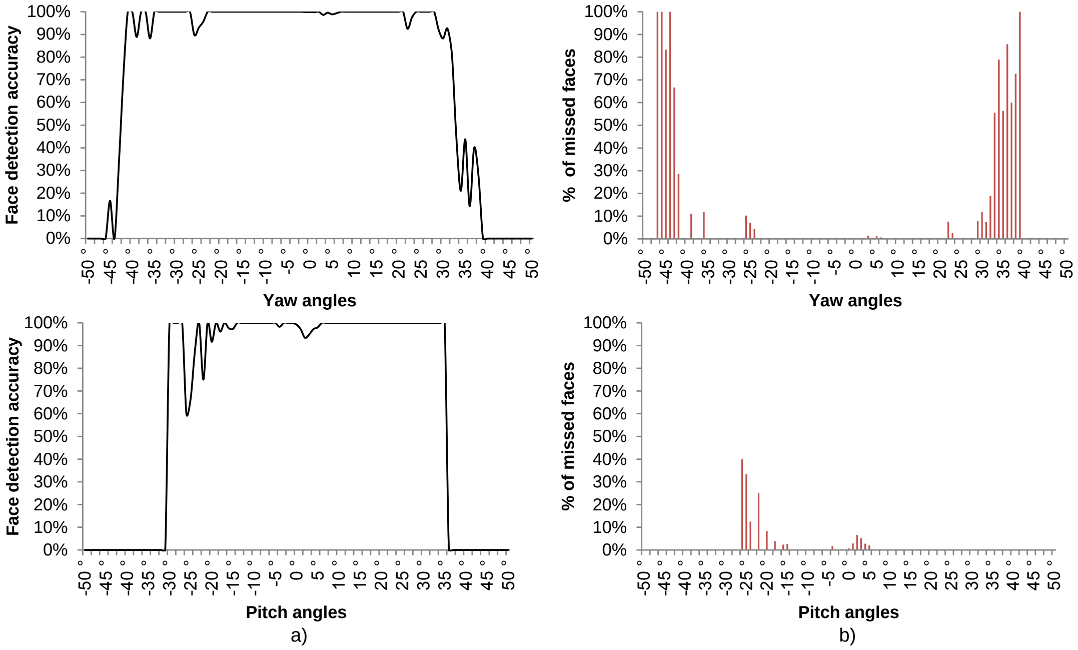


Fig. 16 a) Face detection accuracy for individual yaw and pitch angles b) Distribution of missed faces with respect to the yaw and pitch angles for BUFT dataset

In the second and third test, we measure the effect of changing the out-of-plane yaw and pitch rotation on the roll estimation accuracy. Figure 17 shows the histogram of MAE in the roll estimation with respect to the ground-truth yaw and pitch angles. Since the BUFT dataset has limited yaw and pitch angles (from -47° to $+40^{\circ}$ for yaw and from -30° to $+36^{\circ}$ for pitch), we test our method within these ranges. According to Fig. 17 (a), it is clear that, the change in yaw angle has a negative effect on the roll estimation process. MAE starts to increase after ±10 degree of yaw rotation. Here we should note that not all angles have the same amount of faces. However, the effect of yaw rotation is clearly visible. According to the last experiment, pitch changes do not affect the roll estimation from -30° to $+30^\circ$ as shown in Fig. 17 (b). For this range, MAE is stable between 1.0° to 3.5° and it does not present a linear relationship with the increasing or decreasing value of the pitch angle. Because, the amount of visual change on the face in a pitch rotation is much less than that of in a yaw rotation where discriminative facial features are still partially visible. In case of yaw rotation the most discriminative features of the face starts to disappear which affects the amount of sequential face detections in Level 2. Haar-like features provides better performance in case of visible eye pair and noise. Therefore, the changes in the yaw angle affect the roll estimation much more than the pitch angle as shown in Fig. 17.

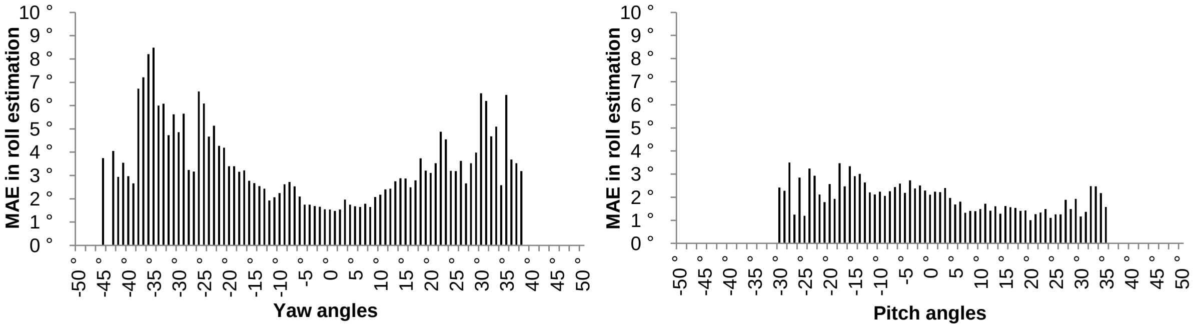


Fig. 17 Histogram of MAE in the roll estimation for ground-truth yaw and pitch angles for step factor 3

4.4.4 Low resolution tests

Main advantage of the proposed method is its ability to use a face detection algorithm for the roll estimation. In order to present the advantage of the proposed method, we performed experiments in low resolution images. First, we reduced the image resolution in BUFT (320 × 240), Manchester (720 × 576) and Caltech (896 × 592) datasets to (80 × 60), (80 × 64) and (80 × 52) pixels respectively. After that we interpolated the images upwards to get the (240 × 180), (240 × 192) and (240 × 156) resolution as shown in Fig. 18. This process makes the images distorted enough to have JPEG compression defects. We compared the new results with the initial results obtained from the original datasets. For the low resolution test, we obtained 2.24°, 1.39° and 1.33° of MAE for BUFT, Manchester and Caltech datasets as shown in Fig. 18 (a)(b)(c). Even very low resolution, the proposed method provides similar MAE obtained from higher resolution images.

4.4.5 Computational cost

In their work, Wang [34] presents algorithm level analysis of the VJ algorithm. Feature level, cascade level and image level parallelism analysis of VJ algorithm is presented in Jia et al [14]. Our method can be applied using image level parallelism where the input image is rotated iteratively with respect to the step factor considering the defined boundaries. Each rotation of the input image can be processed in parallel without any synchronization in both of the levels. On the other hand, as indicated by Viola and Jones [31], the speed of a cascaded detector is directly related to the number of features evaluated per scanned

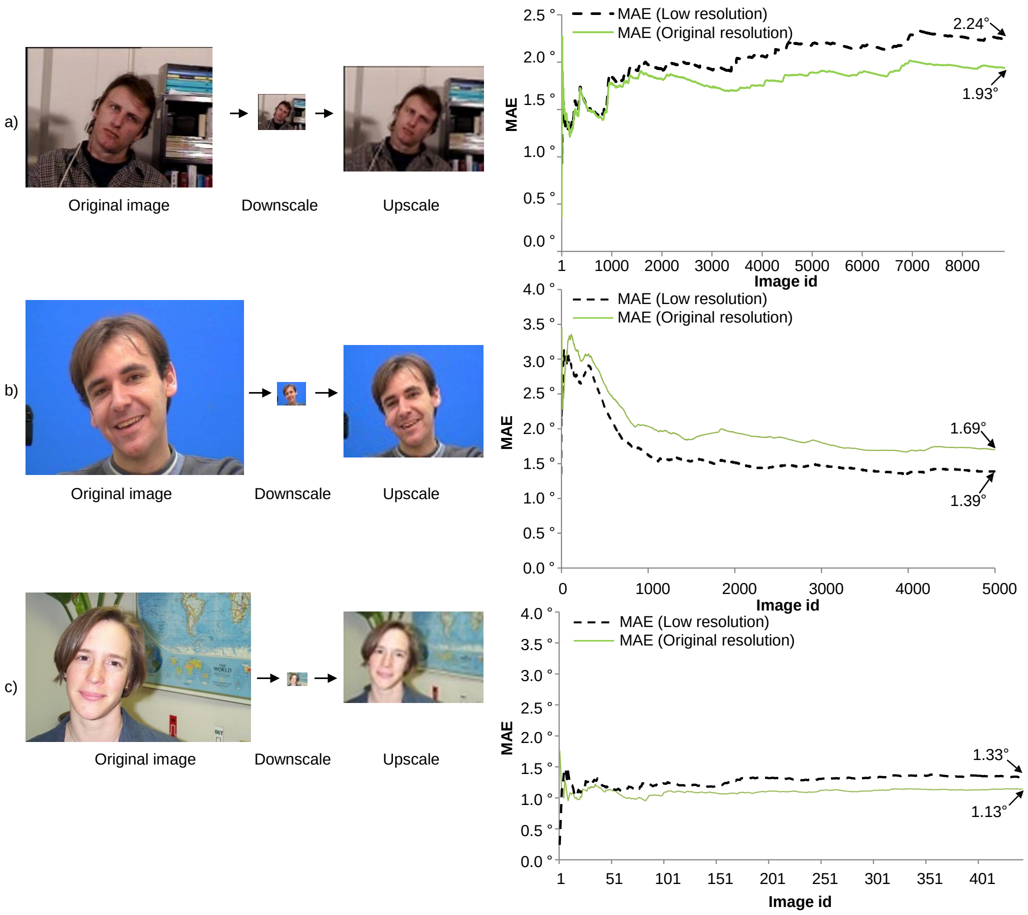


Fig. 18 Effect of low resolution processing on BUFT, Manchester and Caltech datasets. Our method provides consistent MAE in low resolution images for the test datasets. a) MAE for the BUFT dataset increases from 1.93 ° to 2.24 °. Original dataset MAR value is 7.03 °. b) MAE drops from 1.69 ° to 1.39 ° for the Talking-face dataset. Original dataset MAR value is 5.03 °. c) MAE for Caltech dataset increases from 1.13 ° to 1.33 °. Original dataset MAR value is 2.03 °

sub-window. Since the number of features evaluated depends on the content of the images being scanned, the worst case processing cost per image can be represented by Eq. 3:

$$\begin{aligned}
VJ\_cost &= O(Single\_VJ\_matching\_cost \times width \times height \times nscales) \text{ Castrillon et al [4]} \\
Rotation\_cost &= O(n^2) \text{ where } n = argmax(width, height) \\
Level1\_cost &= O((VJ\_cost + Rotation\_cost) \times n\_orientations) \\
Level2\_cost &= O((VJ\_cost + Rotation\_cost) \times n\_step\_factor \times nfaces)
\end{aligned} \tag{3}$$

where width × height × nscales present the cost of VJ for a single image. Width and height is the corresponding width and height of the input image. *nscales* parameter represents the number of scales in the user pyramid which depends on the scaling parameter. Rotation cost is defined by selecting the maximum of the width and height of the input image since in each iteration the whole image is rotated. *n_step_factor* is the number of steps executed for a given step factor. *nfaces* parameter represent the total number of faces detected in first level. For each face initially detected in Level 1 , the Level 2 process is executed. Since the size of the image in Level 2 (a cropped face) is much smaller than original input image in Level 1, cost of the Level 2 is smaller than that of the Level 1 in general. In our experiments, we also measured the computational time required to complete the roll estimation process. Computational cost of the proposed method mainly depends on the resolution and number of faces in the image. Resolution is a common factor that affects all methods in the domain. Since Level 2 operations performed on each face candidate from Level 1, the computational time is linearly dependent to the number of faces in the image. According to the experiments, datasets having single face per image (e.g., BioID, BUFT, Caltech, Manchester Talking Face) we obtained $5.1fps$ to $12.2fps$ on Xenon 3.0 Ghz processor with parallel execution setting. For other datasets, we measured an average of $2.5fps$ for the same configuration.

5 Conclusion

We proposed a robust in-plane face orientation estimation algorithm in still images using an existing frontal face detection algorithm. We used the same frontal detection algorithm for coarse and fine estimation of the roll angle. Considering the complete head-pose estimation problem, we showed that roll angle can be estimated effectively using a frontal face detector. Experiments on variety of well-known, controlled and unconstrained datasets showed that proposed method provides practical implementation and state-of-the-art performance for in-plane face orientation estimation. As a future work, we planned to extend our work for using profile detectors in combination with frontal detectors to improve out-of-plane capability of our method. Considering the guidelines in Murphy-Chutorian and Trivedi [20], our method can be classified as an accurate, autonomous, monocular, multi-person, identity and resolution independent roll estimation algorithm.

6 Acknowledgements

This study is supported by TWIRL (ITEA2 10029 - Twinning Virtual World Online Information with Real-World Data Sources) project.

A Appendix: Example detections from test databases

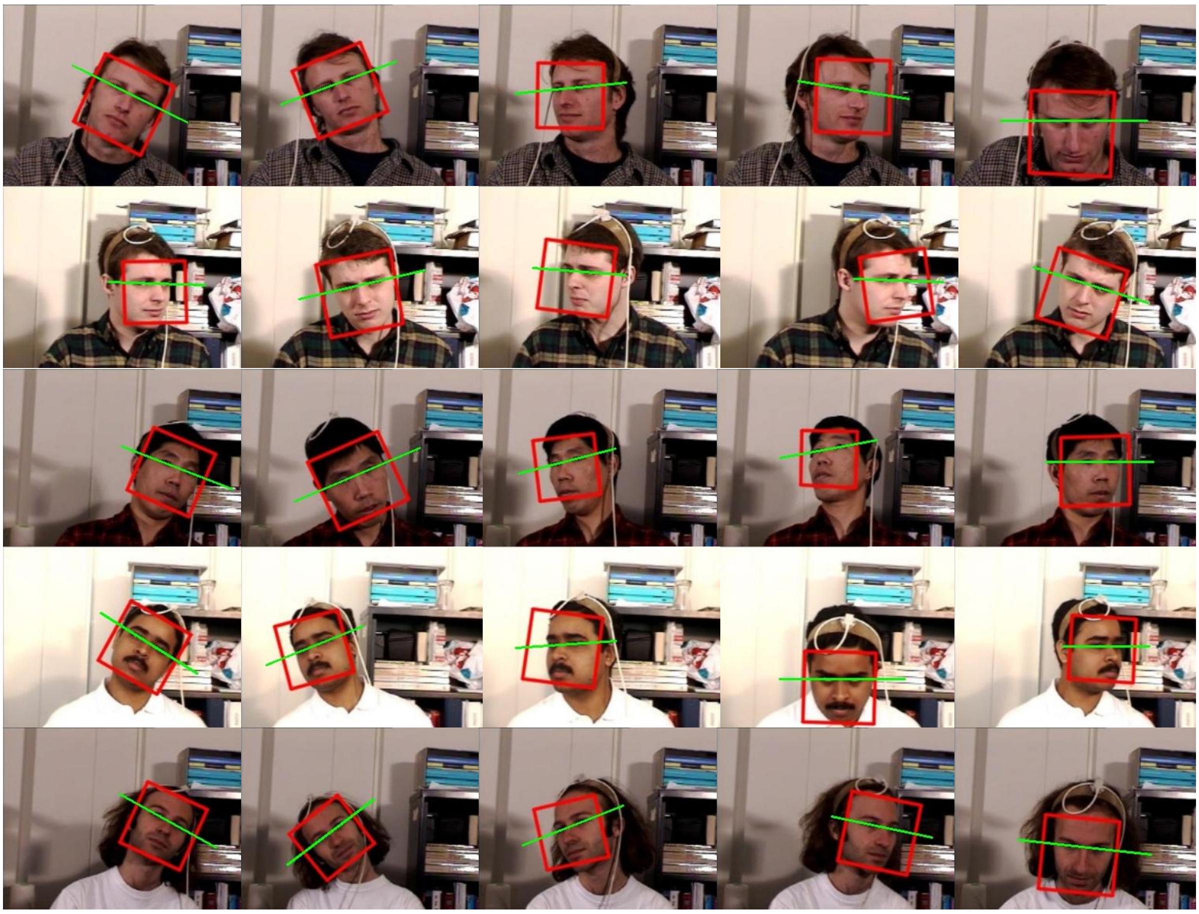


Fig. 19 Additional samples from BUFT rotated dataset

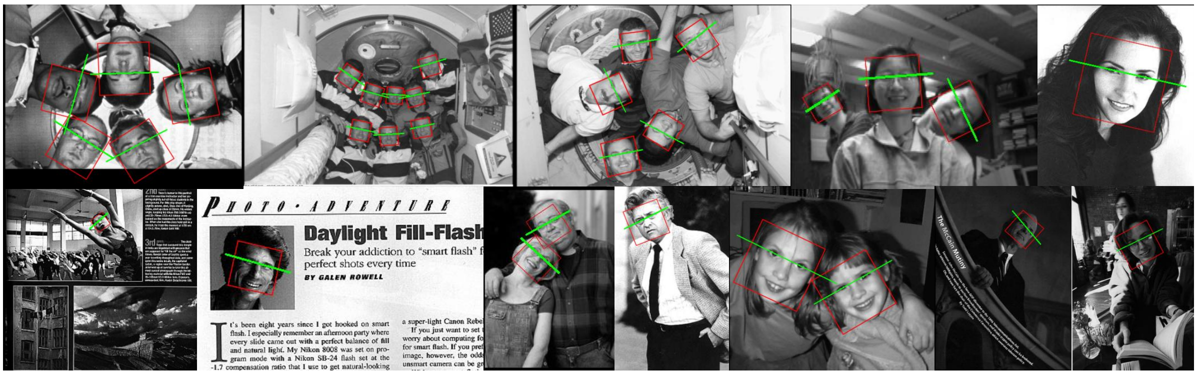


Fig. 20 Sample detections from CMU rotated dataset

References

1. An KH, Chung MJ (2008) 3D head tracking and pose-robust 2d texture map-based face recognition using a simple ellipsoid model. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp 307–312
2. Asteriadis S, Karpouzis K, Kollias S (2010) Head pose estimation with one camera, in uncalibrated environments. In: Proceedings of the 2010 workshop on eye gaze in intelligent human machine interaction, ACM, New York, NY, USA, EGIHMI '10, pp 55–62, DOI 10.1145/2002333.2002343, URL `http://doi.acm.org/10.1145/2002333.2002343`
3. Ba S, Odobez J (2004) A probabilistic framework for joint head tracking and pose estimation. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol 4, pp 264–267 Vol.4, DOI 10.1109/ICPR.2004.1333754
4. Castrillon M, Deniz O, Guerra C, Hernandez M (2007) Encara2: Real-time detection of multiple faces at different resolutions in video streams. J Vis Commun Image R 18(2):130 – 140, DOI http://dx.doi.org/10.1016/j.jvcir.2006.11.004, URL `http://www.sciencedirect.com/science/article/pii/S1047320306000782`
5. Cootes TF (2004) Manchester talking face video dataset. URL `http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html`, date last accessed: 02.02.2013
6. Dahmane A, Larabi S, Djeraba C (2010) Detection and analysis of symmetrical parts on face for head pose estimation. In: 17th IEEE International Conference on Image Processing (ICIP), pp 3249 –3252, DOI 10.1109/ICIP.2010.5651202
7. Danisman T, Bilasco IM, Ihaddadene N, Djeraba C (2010) Automatic facial feature detection for facial expression recognition. In: Richard P, Braz J (eds) VISAPP 2010 - Proceedings of the Fifth International Conference on Computer Vision Theory and Applications, Angers, France, May 17-21, 2010 - Volume 2, INSTICC Press, pp 407–412
8. Danisman T, Bilasco I, Djeraba C (2014) Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings. In: 22nd IEEE International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden
9. Demirkus M, Clark J, Arbel T (2013) Robust semi-automatic head pose labeling for real-world face video sequences. Multimed Tools Appl pp 1–29, DOI 10.1007/s11042-012-1352-1, URL `http://dx.doi.org/10.1007/s11042-012-1352-1`
10. Du S, Zheng N, You Q, Wu Y, Yuan M, Wu J (2006) Rotated haar-like features for face detection with in-plane rotation. In: Zha H, Pan Z, Thwaites H, Addison A, Forte M (eds) Interactive Technologies and Sociotechnical Systems, LNCS, vol 4270, Springer Berlin Heidelberg, pp 128–137, DOI 10.1007/11890881_15, URL `http://dx.doi.org/10.1007/11890881_15`
11. Face and gesture recognition working group (2000) FGNET Aging dataset. http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html, URL `http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html`, date last accessed: 22.05.2009
12. Guo W, Kotsia I, Patras I (2011) Higher order support tensor regression for head pose estimation. In: 12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011)
13. Jesorsky O, Kirchberg K, Frischholz R (2001) Robust face detection using the hausdorff distance. In: Bigun J, Smeraldi F (eds) Audio- and Video-Based Biometric Person Authentication, LNCS, vol 2091, Springer Berlin Heidelberg, pp 90–95, DOI 10.1007/3-540-45344-X_14, URL `http://dx.doi.org/10.1007/3-540-45344-X_14`
14. Jia H, Zhang Y, Wang W, Xu J (2012) Accelerating Viola-Jones face detection algorithm on gpus. In: High Performance Computing and Communication 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012 IEEE 14th International Conference on, pp 396–403, DOI 10.1109/HPCC.2012.60
15. Jung S, Nixon MS (2012) On using gait to enhance frontal face extraction. IEEE Trans Inf Forensics Security 7(6):1802–1811, DOI 10.1109/TIFS.2012.2218598, URL `http://dx.doi.org/10.1109/TIFS.2012.2218598`
16. La Cascia M, Sclaroff S, Athitsos V (2000) Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. IEEE Trans Pattern Anal Mach Intell 22(4):322–336, DOI 10.1109/34.845375, URL `http://dx.doi.org/10.1109/34.845375`
17. Lefevre S, Odobez J (2010) View-based appearance model online learning for 3D deformable face tracking. In: Richard P, Braz J (eds) VISAPP 2010 - Proceedings of the Fifth International Conference on Computer Vision Theory and Applications, Angers, France, May 17-21, 2010 - Volume 1, INSTICC Press, pp 223–230
18. Morency LP, Whitehill J, Movellan J (2010) Monocular head pose estimation using generalized adaptive view-based appearance model. Image Vis Comput 28(5):754 – 761, DOI 10.1016/j.imavis.2009.08.004, URL `http://www.sciencedirect.com/science/article/pii/S0262885609001735`, best of Automatic Face and Gesture Recognition 2008
19. Murphy-Chutorian E, Trivedi M (2008) Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking. In: Intelligent Vehicles Symposium, 2008 IEEE, pp 512–517, DOI 10.1109/IVS.2008.4621320
20. Murphy-Chutorian E, Trivedi M (2009) Head pose estimation in computer vision: A survey. IEEE Trans Pattern Anal Mach Intell 31(4):607 –626, DOI 10.1109/TPAMI.2008.106
21. My VD, Zell A (2013) Real time face tracking and pose estimation using an adaptive correlation filter for human-robot interaction. In: ECMR, pp 119–124
22. Oka K, Sato Y, Nakanishi Y, Koike H (2005) Head pose estimation system based on particle filtering with adaptive diffusion control. In: Proceedings of the IAPR Conference on Machine Vision Applications (IAPR MVA 2005), May

16-18, 2005, Tsukuba Science City, Japan, pp 586–589, URL http://b2.cvl.iis.u-tokyo.ac.jp/mva/proceedings/CommemorativeDVD/2005/papers/2005586.pdf

23. Osadchy M, Cun YL, Miller ML (2007) Synergistic face detection and pose estimation with energy-based models. J Mach Learn Res 8:1197–1215, URL http://dl.acm.org/citation.cfm?id=1248659.1248700
24. Pan H, Zhu Y, Xia L (2013) Efficient and accurate face detection using heterogeneous feature descriptors and feature selection. Comput Vis Image Understand 117(1):12 – 28, DOI http://dx.doi.org/10.1016/j.cviu.2012.09.003, URL http://www.sciencedirect.com/science/article/pii/S1077314212001294
25. Pathangay V, Das S, Greiner T (2008) Symmetry-based face pose estimation from a single uncalibrated view. In: 8th IEEE International Conference on Automatic Face Gesture Recognition, FG '08., pp 1 –8, DOI 10.1109/AFGR.2008.4813312
26. Rowley HA, Baluja S, Kanade T (1998) Rotation invariant neural network-based face detection. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 38–44
27. Sung J, Kanade T, Kim D (2008) Pose robust face tracking by combining active appearance models and Cylinder Head Models. Int J Comput Vis 80:260–274, DOI 10.1007/s11263-007-0125-1, URL http://dx.doi.org/10.1007/s11263-007-0125-1
28. Tran NT, Ababsa FE, Charbit M, Feldmar J, Petrovska-Delacrtaz D, Chollet G (2013) 3D face pose and animation tracking via eigen-decomposition based bayesian approach. In: Bebis G, Boyle R, Parvin B, Koracin D, Li B, Porikli F, Zordan V, Klosowski J, Coquillart S, Luo X, Chen M, Gotz D (eds) Advances in Visual Computing, LNCS, vol 8033, Springer Berlin Heidelberg, pp 562–571, DOI 10.1007/978-3-642-41914-0_55, URL http://dx.doi.org/10.1007/978-3-642-41914-0_55
29. Valenti R, Yücel Z, Gevers T (2009) Robustifying eye center localization by head pose cues. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 612–618
30. Viola M, Jones MJ, Viola P (2003) Fast multi-view face detection. Tech. Rep. TR2003-96, Mitsubishi Electric Research Laboratories, 201 Broadway Cambridge, MA 02139
31. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57:137–154, URL http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb, 10.1023/B:VISI.0000013087.49260.fb
32. Voit M (2007) Clear 2007 evaluation plan: Head pose estimation. URL http://isl.ira.uka.de/mvoit/clear07/CLEAR07HEADPOSE2007-03-26.doc
33. Wang JG, Sung E (2007) EM enhancement of 3D head pose estimated by point at infinity. Image Vis Comput 25(12):1864 – 1874, DOI http://dx.doi.org/10.1016/j.imavis.2005.12.017, URL http://www.sciencedirect.com/science/article/pii/S0262885606002848, the age of human computer interaction
34. Wang YQ (2014) An Analysis of the Viola-Jones Face Detection Algorithm. Image Processing On Line 4:128–148, DOI 10.5201/ipol.2014.104
35. Weber M (1999) Caltech frontal face dataset. http://www.vision.caltech.edu/html-files/archive.html, URL http://www.vision.caltech.edu/html-files/archive.html, date last accessed: 02.02.2013
36. Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20-25 June 2011, IEEE, pp 529–534, DOI 10.1109/CVPR.2011.5995566, URL http://dx.doi.org/10.1109/CVPR.2011.5995566
37. Wu B, Ai H, Huang C, Lao S (2004) Fast rotation invariant multi-view face detection based on real adaboost. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp 79–84, DOI 10.1109/AFGR.2004.1301512
38. Wu S, Jiang L, Xie S, Yeo AC (2006) A robust method for detecting facial orientation in infrared images. Pattern Recogn 39(2):303 – 309, DOI 10.1016/j.patcog.2005.06.003, URL http://www.sciencedirect.com/science/article/pii/S003132030500227X, part Special Issue: Complexity Reduction
39. Wu S, Lin W, Xie S (2008) Skin heat transfer model of facial thermograms and its application in face recognition. Pattern Recogn 41(8):2718 – 2729, DOI 10.1016/j.patcog.2008.01.003, URL http://www.sciencedirect.com/science/article/pii/S0031320308000113
40. Xiao J, Kanade T, Cohn JF (2002) Robust full-motion recovery of head by dynamic templates and re-registration techniques. In: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, Washington, DC, USA, FGR '02, pp 163–, URL http://dl.acm.org/citation.cfm?id=874061.875442
41. Yan S, Zhang Z, Fu Y, Hu Y, Tu J, Huang T (2008) Learning a person-independent representation for precise 3d pose estimation. In: Stiefelhagen R, Bowers R, Fiscus J (eds) Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science, vol 4625, Springer Berlin Heidelberg, pp 297–306, DOI 10.1007/978-3-540-68585-2_28, URL http://dx.doi.org/10.1007/978-3-540-68585-2_28
42. Zhao G, Chen L, Song J, Chen G (2007) Large head movement tracking using sift-based registration. In: Proceedings of the 15th International Conference on Multimedia, ACM, New York, NY, USA, MULTIMEDIA '07, pp 807–810, DOI 10.1145/1291233.1291416, URL http://doi.acm.org/10.1145/1291233.1291416
43. Zhao S, Yao H, Sun X (2013) Video classification and recommendation based on affective analysis of viewers. Neurocomputing 119(0):101 – 110, DOI http://dx.doi.org/10.1016/j.neucom.2012.04.042, URL http://www.sciencedirect.com/science/article/pii/S0925231212009149, Intelligent Processing Techniques for Semantic-based Image and Video Retrieval

44. Zhou J, Lu XG, Zhang D, Wu CY (2002) Orientation analysis for rotated human face detection. Image Vis Comput 20(4):257 – 264, DOI 10.1016/S0262-8856(02)00018-5