Fusion of Keypoint Tracking and Facial Landmark Detection for Real-Time Head Pose Estimation

Jilliam María Díaz Barros[⋆†‡*] Bruno Mirbach[‡] Frederic Garcia[‡] Kiran Varanasi[⋆] Didier Stricker[⋆†]
[⋆] German Research Center for Artificial Intelligence (DFKI), Germany
[†] Technische Universität Kaiserslautern, Germany
[‡]IEE S.A., Luxembourg
Jilliam_Maria.Diaz_Barros@dfki.de, {Bruno.Mirbach, Frederic.Garcia} @iee.lu
{Kiran.Varanasi, Didier.Stricker} @dfki.de

Abstract

In this paper, we address the problem of extreme head pose estimation from intensity images, in a monocular setup. We introduce a novel fusion pipeline to integrate into a dedicated Kalman Filter the pose estimated from a tracking scheme in the prediction stage and the pose estimated from a detection scheme in the correction stage. To that end, the measurement covariance of the Kalman Filter is updated in every frame. The tracking scheme is performed using a set of keypoints extracted in the area of the head along with a simple 3D geometric model. The detection scheme, on the other hand, relies on the alignment of facial landmarks in each frame combined with 3D features extracted on a head mesh. The head pose in each scheme is estimated by minimizing the reprojection error from the 3D-2D correspondences. By combining both frameworks, we extend the applicability of head pose estimation from facial landmarks to cases where these features are no longer visible. We compared the proposed method to other related approaches, showing that it can achieve state-of-the-art performance. We also demonstrate that our approach is suitable for cases with extreme head rotations and (self-) occlusions, besides being suitable for real time applications.

1. Introduction

Head pose estimation (HPE) refers to the problem of recovering the 6 degrees of freedom (D.o.F.) pose of the person's head, consisting of its location and orientation with respect to the camera coordinate system. It serves either as an intermediate step for other tasks as face alignment [46], face recognition [2], facial expression recognition [22] or gaze estimation [41, 42], or directly for a wide range of applications as human-computer interaction, driver monitoring [43, 5] and augmented reality[28]. Depending on the requirements, the estimation can be performed from 2D data, including RGB, infrared (IR), or intensity images, from depth images or from a combination of both.

We focus our investigation on HPE from intensity images, where the gaze could be extracted from the input data in a follow-up project. In that way, we extend the relevance of our approach to those cases where only RGB or monochrome cameras are available. We are also interested in a HPE approach that could be used in realistic scenarios, where a set of constraints are satisfied: the method should be able to perform in real time, preferably with no need of power demanding devices like graphic hardware; it should be able to work for different users, regardless of age, gender or ethnicity, with no need of any additional calibration; the method must be robust to handle extreme head poses, *i.e.*, when part of the head is (self-)occluded due to large rotations; it should be able to initiate the estimation task from different head poses, not only from frontal faces.

Although the use of consumer RGB-D cameras in HPE research has increased in the last years [12, 5, 38], we opt for intensity images captured in a monocular setup for reasons of cost and processing power restrictions in our targeted applications.

In this work, we introduce a method for robust HPE in real time. The approach is composed of two schemes, performed by separate: HPE from 2D keypoints tracking, using a geometric model and HPE from facial landmark alignment, using a synthetic head mesh. The final HPE results from the fusion of both schemes using a dedicated Kalman Filter, which combines the strengths of both pipelines: the precision of HPE from facial landmarks and the robustness to handle large head pose variations of HPE from keypoints tracking. We show that our method is suitable to work in

*This work was funded by the National Research Fund, Luxembourg. Project ID 9235599.

real time and is also robust to extreme head rotations.

The major contributions of our work are:

- A novel head pose estimation technique, where the pose computed from keypoints is fused with the pose estimated from facial landmarks, using a dedicated Kalman Filter. To the best of our knowledge, this is the first method that combines a local motion estimated by keypoints tracking, with a global head pose estimated from facial landmark alignment. The local motion is integrated at the prediction step of the Kalman Filter, while the global motion is included for the correction step.
- The proposed approach enables the combination of the pose estimated with a simple geometric model with the pose from a 3D mesh, resulting in a HPE method robust to extreme head rotations, without the need for an extensive training stage with manual annotation in large datasets.

We compared our method to other approaches of the state of the art, using a publicly available dataset for HPE. We also performed experiments for time-consumption analysis and to verify the robustness of our method to extreme head rotations.

2. Related work

Although methods based on 2D input data have been extensively studied [30], with the introduction of consumer RGB-D cameras in the last years, the number of depth-based approaches has increased recently [13, 12, 27, 31, 5, 8]. These approaches include combined pipelines using both RGB and depth data [4, 34], or IR and depth data [36].

Following the classification proposed in [5], we divide HPE approaches in three main categories: model-based, appearance-based and 3D head model registration approaches. It should be noted that some methods might fall in more than one category.

Model-based approaches. These HPE methods use rigid or non-rigid face models, facial landmark detection and/or any other prior information regarding the geometry of the head. HPE based on registration of texture map images with a cylindrical head model (CHM) was proposed by La Cascia *et al*. in [24]. Choi and Kim [7] used templates for HPE, combining a particle filter with an ellipsoidal head model (EHM). Sung *et al*. [37] combined active appearance model (AAM) with a CHM. An and Chung [2] used an EHM to formulate the HPE as a linear system, assuming a rigid body motion under perspective projection. Kumano *et al*. [22] used a face model given by a variable-intensity template with a particle filter, for simultaneous HPE and facial expression recognition. Jang and Kanade [18, 19] designed a user-specific CHM-based framework, by combining into a Kalman Filter the estimated motion and a pose retrieved from a dataset of SIFT feature points. In [42, 41], Valenti *et al*. used a CHM for simultaneous HPE and eye tracking, based upon a crossed feedback mechanism, which compensated the estimated values and allowed to re-initialize the head pose tracker. Asteriadis *et al*. [3] used a facial-feature tracker with Distance Vector Fields (DVFs) for HPE. In [32], Prasad and Aravind computed the pose using POSIT from the 3D-2D correspondences from a parametrized 3D face mask and SIFT feature points. Diaz *et al*. [9], used random feature points and a CHM to estimate the pose by minimizing the reprojection error of the 3D features and the 2D correspondences. On the other hand, Vicente *et al*. [43] used facial landmarks and a deformable head model, namely parameterized appearance models, to minimize the reprojection error for HPE. Yin and Yang [47] used a pixel intensity binary test for face detection, with pose regression along with local binary feature for face alignment. From a rigid head model, the pose was retrieved by solving the 2D-3D correspondences. Wu *et al*. in [45] presented a pipeline for simultaneous facial landmark detection, HPE and deformation estimation using a cascade iterative procedure augmented with model-based HPE. Similarly, Gou *et al*. [17] proposed a Coupled Cascade Regression (CCR) framework for simultaneous facial landmark detection and HPE.

Appearance-based approaches. They use machine learning techniques for HPE, based on visual features of the face appearance. Even though these methods are robust to extreme head poses, usually the output corresponds to discrete head poses, thus assigning the pose to specific ranges instead of continuous estimation. These approaches usually have a higher performance for low-resolution face images [1, 11]. In [13], Fanelli *et al*. used random regression forests for HPE and facial feature detection, from depth data. Patches from different parts of the face were used to recover the pose through a voting scheme. For the training, it was necessary a large dataset with annotated data. Wang *et al*. presented in [44] a head tracking approach from invariant keypoints. Simulation techniques and normalization were combined to create a learning scheme. Ahn *et al*. [1] introduced a deep-learning-based approach for RGB images, with a particle filter to refine and increase the stability of the estimated pose. In [26], Liu *et al*. used convolutional neural networks, where HPE was formulated as a regression problem. The network was trained using a large synthetic dataset obtained from rendered 3D head models. [1] and [26] used a GPU to reach real-time capabilities. Tulyakov *et al*. introduced in [40] a person-specific template scheme using a depth camera, which combined template-matching-based tracking with a frame-by-frame decision-tree-based estimator. Borghi *et al*. [5] presented a real time deep-learning-based approach for HPE from depth images,

using a regression neural network, POSEidon, which integrated depth with motion features and appearance. In [36], Schwarz presented a deep learning method for HPE which fused IR and depth data with cross-stitch units. Derkach *et al*. [8] proposed a system intended for depth input data, which integrated three different approaches for HPE, two based on landmark detection and one on a dictionary-based method for extreme head poses.

3D head model registration approaches. These methods register the measured data to reference 3D head models. Meyer *et al*. [27] combined particle swarm optimization and the iterative closest point (ICP) algorithm to register a 3D morphable model (3DMM) to a measured depth face. Yu *et al*. [48] extended this with an online 3D reconstruction of the full head, to handle extreme head rotations. Ghiass *et al*. [15] estimated the pose through a fitting process with a 3D morphable model and RGB-D data. Papazov *et al*. [31] introduced triangular surface patch descriptors for HPE from depth data. The pose was computed from a voting scheme resulting from matching the descriptors to patches from synthetic head models. Jeni *et al*. [20] presented an approach for 3D registration of a dense face mesh from 2D images, through a cascade regression framework trained using a large database of high-resolution 3D face scans. Tan *et al*. [38] used RGB-D data to regress the 3D head pose using random forest in a temporal tracking scheme.

Other methods define HPE as an optimization problem. That is the case of [29], where Morency *et al*. presented a probabilistic scheme, namely Generalized Adaptive View-based Appearance Model (GAVAM), using an EHM. The pose was estimated by solving a linear system with normal flow constraint (NFC). Baltrusaitis *et al*. presented in [4] an extension, which combined head pose tracking with a 3D constrained local model, using both depth data and intensity information. Saragih *et al*. introduced in [35] a HPE approach which fits a deformable model using an optimization strategy through a non-parametric representation of the likelihood maps of landmarks locations. Drouard *et al*. [11] used a Gaussian mixture of locally-linear mapping model to map HOG features extracted on a face region to 3D head poses.

One of the issues of most tracking-based methods is that their robustness to initial HPE when the head is not frontal is not clear [8]. For facial-landmarks-based HPE methods, the accuracy of the head pose relies on the precision of the estimated facial landmarks. Since they strongly depend on the detection of facial landmarks, the misalignment of the landmarks in a frame might lead to erroneous estimations. Hence, these methods might be sensitive to extreme head poses, partial occlusions, facial expressions and low resolution images.

In this work, we introduce a model-based framework for HPE using only intensity images. We combine two independent pipelines for pose estimation, extending the HPE to extreme head poses. The proposed method does not have any constraint for initialization, as facing the camera for the first frame, and is suitable for real time applications, making it useful for HPE in realistic scenarios.

3. Proposed HPE pipeline

Several methods of the state of the art rely on facial landmarks for HPE. Even though they might be a reliable source for HPE for frontal and near-frontal faces, facial landmarks are sensitive to extreme head rotations and (self-) occlusions, where important reference regions of the face such as the eyes or nose are partially or totally occluded. In order to tackle this problem, we propose to integrate the head pose computed from a set of keypoints that can be tracked continuously, even when the facial landmarks are not visible. Although a keypoint-based HPE approach could be used alone, it might suffer from drifting in long sequences [18, 19, 9]. Accordingly, a mechanism to reinforce and correct the head pose from keypoint tracking, using the facial-landmark HPE scheme must be included.

Therefore, we develop two different strategies for HPE that run independently and which are later fused using a Kalman Filter. The proposed framework is illustrated in Figure 1. The first strategy (blue area in Figure 1) includes a temporal tracking scheme, which uses optical flow to compute the correspondences of a set of keypoints in every pair of frames (step 1). The keypoints are projected onto a geometric head model, to estimate the 3D keypoints (step 2). Both the 2D and 3D keypoints are used to compute the head pose from frame $k-1$ to frame k (step 3). The second strategy (red area in Figure 1) includes a tracking-by-detection algorithm, which estimates the pose independently in each frame by aligning 2D facial landmarks to every input image (step 4). Along with a set of the corresponding 3D facial landmarks extracted from a head mesh (step 5), the head pose from facial landmarks is estimated in step 6. We use the HPE from the keypoints for the prediction step of the Kalman Filter (step 7) and the HPE from the facial landmarks in the correction step (step 8). The final head pose corresponds to the output of the Kalman Filter (step 9). As the two different strategies for HPE run in parallel independently of each other, time consumption of the algorithm can be reduced considerably.

The head pose is represented with a transformation, composed of a rotation $\mathbf{R}$ and a translation $\mathbf{t}$. The pose of every 3D point $\mathbf{P}$ in the head is updated following a rigid transformation. $\mathbf{R}$ can also be denoted by the rotation angles $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]$ with respect to the X, Y and Z axes of a known coordinate system. ω_x, ω_y, and ω_z are usually termed as pitch, yaw and roll angles. For our framework, the calibration of the camera is required in advance.

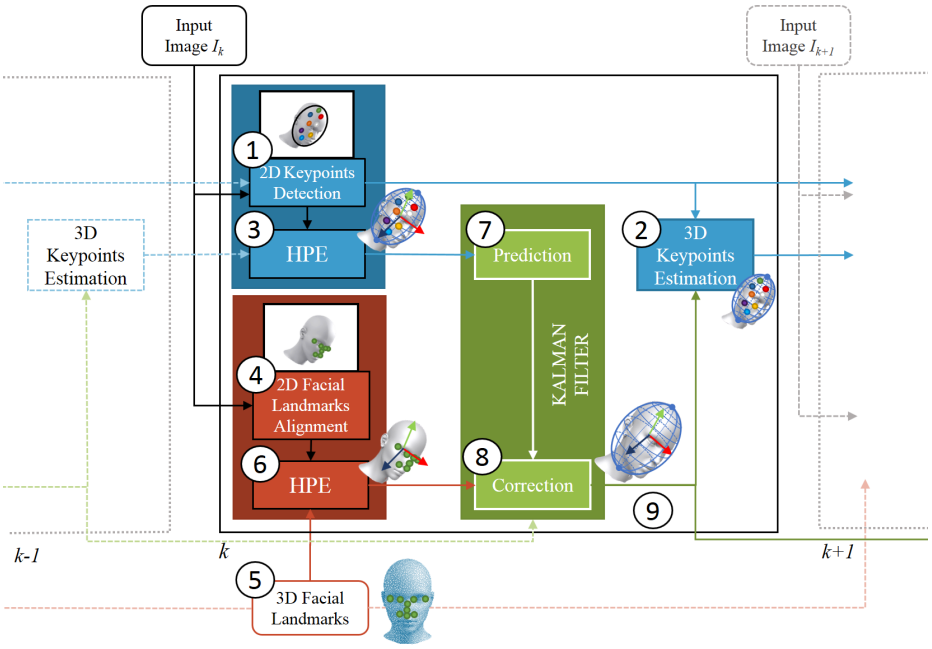

Figure 1. Proposed HPE pipeline, from Keypoints (blue) and Facial Landmarks (red).

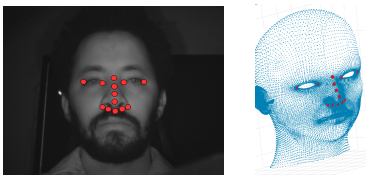

Figure 2. Facial Landmarks in 2D (left) and 3D (right).

3.1. Facial landmarks

In the following section, we describe the procedure to compute the set of facial landmarks, prior to the HPE. We denote the set of n 2D facial landmarks as $\{\mathbf{p}_F\}_{i=1}^{n}$ and the 3D facial features as $\{\mathbf{P}_F\}_{i=1}^{n}$.

2D facial landmarks. A set of facial landmarks are detected in each input image using the approach proposed in [21]. This method uses an ensemble of regression trees to align the facial landmarks, from a sparse subset of intensity values indexed to an initial estimate of the shape. As we are modeling the head as a rigid body, we select a set of fiducial features which are robust to facial expressions, blinking and other non-rigid motions from the face. The set includes the corners of the eyes and the points around the nose, as shown in Figure 2 (right), for a total of 13 features.

3D facial landmarks. Given the robust 2D facial landmarks described earlier and illustrated on Figure 2 (left), the corresponding 3D facial landmarks are extracted offline on a reference head mesh (see Figure 2 (right)). These pre-defined 3D features were manually annotated from an open-source 3D face model. With the proposed method, we avoid the time-consuming process of manual [20] or semi-automatic facial landmarks annotation [4] on large datasets of 3D face scans, yet providing a robust estimated head pose as long as the facial features are visible in the image.

3.2. Keypoints extraction

The sets of 2D and 3D keypoints are denoted as $\{\mathbf{p}_K\}_{i=1}^{m}$ and $\{\mathbf{P}_K\}_{i=1}^{m}$, respectively. These sets are updated every frame, therefore the number of keypoints m varies through the entire sequence.

2D keypoints. Keypoints are extracted on the area of the head using the Features from Accelerated Segment Test algorithm, better known as FAST [33]. This corner detection method is suitable for real-time applications and extracts robust feature points for tracking.

From a set of keypoints extracted at time $k-1$, the correspondences at frame k are estimated with optical flow, specifically using the pyramidal Lucas-Kanade feature tracker proposed in [6].

3D keypoints. Contrary to the 3D facial landmarks, where points are extracted offline from a 3D head mesh, we compute $\{\mathbf{P}_K\}_{i=1}^{m}$ using a simple geometrical model, namely an ellipsoidal shape, which gives an approximate location of the keypoints on the 3D space. These 3D points result

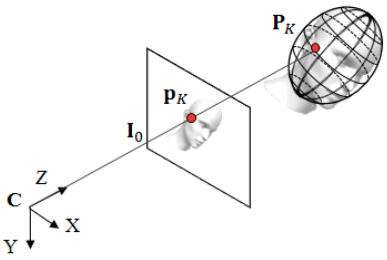

Figure 3. Computation of 3D keypoints.

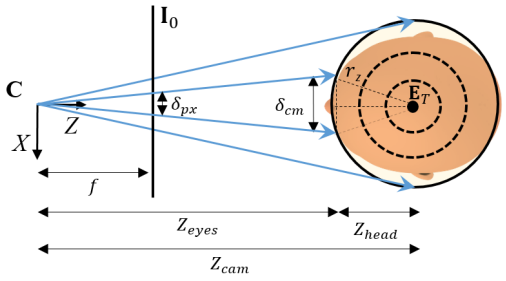

Figure 4. Estimation of the EHM depth.

from the intersections on the ellipsoidal head model (EHM) of the projection lines $\mathbf{l}$ which pass through the optical center of the camera $\mathbf{C}$ and the corresponding 2D keypoints $\{\mathbf{p}_K\}_{i=1}^{m}$ in the image plane $\mathbf{I_0}$, as shown in Figure 3. The dimension and pose of the ellipsoid on the 3D space with respect to the camera coordinate system are computed in advance, as detailed in Section 3.3.

The projection line can be written as $\mathbf{l} = \mathbf{C} + \lambda\mathbf{d}$, where $\mathbf{d}$ is a parallel line and λ is a scalar computed from the quadratic equation of the ellipsoid given by:

$$|\mathbf{a}|^2 \lambda^2 + 2(\mathbf{a} \cdot \mathbf{b})\lambda + |\mathbf{b}|^2 - 1 = 0 \tag{1}$$

$\mathbf{a} = \mathbf{G}\mathbf{R}^T\mathbf{d}$ and $\mathbf{b} = \mathbf{G}\mathbf{R}^T(\mathbf{C} - \mathbf{E_0})$, with $\mathbf{G}$ being a 3×3 diagonal matrix of the inverses of the ellipsoid radii $\{\frac{1}{r_x}, \frac{1}{r_y}, \frac{1}{r_z}\}$, $\mathbf{R}$ the rotation matrix of the ellipsoid and $\mathbf{E_0}$ its center.

Head ROI computation. As mentioned earlier, keypoints are detected only in the area of the head in the 2D image. This area could be defined by a face detection algorithm as in [23] or by the aligned facial landmarks. However, both methods would fail in every frame where the face is not detected, particularly for extreme head poses where the face might not be visible in several consecutive frames. In order to address this situation, we define this area from the projection of the 3D geometric model onto the image plane. It is important to note that the projected area correspond to the visible part of the head of the EHM and not only the face.

The head ROI is computed by estimating the plane parallel to the horizontal axis of the image plane and to the axis of the geometric model, which cuts the ellipsoid in two parts. The elliptical surface that results from the intersection of the plane and the geometric model is then projected in the image, assuming a perspective camera model.

3.3. Initialization

Contrary to other methods of the state of the art which use 3D morphable models for HPE [4, 27, 48], we propose to use a simple EHM for head tracking. Complex head models can be computationally expensive, requiring the use of graphics hardware. The EHM, although not precise, provides a reliable approximation of the head's shape for HPE. In Section 4, we demonstrate that the use of this simple geometric model yields good results for the tracking task.

Similarly to [9], the EHM is adapted to the dimension of the head, based on the 2D facial landmarks detected in the first frame. As the calibration of the camera is available, the initial depth of the EHM with respect to the camera's optical center $\mathbf{C}$ can be computed from the relation between the interpupillary distance in pixels of the eyes extracted in the image δ_{px} and an approximate distance between a person's eyes in cm. According to the measurements reported in [10, 16], the averaged distances for men is 6.47 cm and for women 6.23 cm. For our method, we assume this distance δ_{cm} to be of 6 cm.

As can be observed in Figure 4, the distance Z_{cam} between $\mathbf{C}$ and the center of the head is given by the sum of Z_{eyes} and Z_{head}. Z_{eyes}, the distance from $\mathbf{C}$ to the eyes' baseline can be calculated from Eq. (2), while Z_{head}, the distance from the eyes' baseline to the center of the head is given by Eq. (3). f is the focal length of the camera and r_z is the radius of the ellipsoid in the Z axis.

$$Z_{eyes} = f \cdot \frac{\delta_{cm}}{\delta_{px}} \tag{2}$$

$$Z_{head} = \sqrt{{r_z}^2 - (\delta_{cm}/2)^2} \tag{3}$$

The dimension of the ellipsoid is defined by the 2D bounding box of the detected head ROI, given by points $\{\mathbf{p}_{TL}, \mathbf{p}_{TR}, \mathbf{p}_{BL}, \mathbf{p}_{BR}\}$, (top-left, top-right, bottom-left, bottom-right). The radii r_x and r_z of the EHM correspond to half of the width of the detected head ROI and is computed from Eq. (4), while the radius r_y correspond to half of the height of the ROI, calculated from Eq. (5). Consequently, the EHM is a prolate ellipsoid, or spheroid.

$$r_x = r_z = \frac{1}{2}|\mathbf{p}_{TR} - \mathbf{p}_{TL}| \cdot \frac{\delta_{cm}}{\delta_{px}} \tag{4}$$

$$r_y = \frac{1}{2}|\mathbf{p}_{TR} - \mathbf{p}_{BR}| \cdot \frac{\delta_{cm}}{\delta_{px}} \tag{5}$$

3.4. Head Pose Estimation

In this section, we refer to global transformation as the function that maps the head pose from the first given frame at time k_0, $\mathbf{R}_0$ and $\mathbf{t}_0$, to pose at time k. This transformation is denoted by $\mathbf{R}_0^k$ and $\mathbf{t}_0^k$. Additionally, a frame by frame transformation from frame at time $k-1$ to k is referred to as a local transformation, denoted by $\mathbf{R}_{k-1}^k$ and $\mathbf{t}_{k-1}^k$. These definitions are of great importance in our pipeline, as we introduce a method to combine a local with a global transformation using a dedicated Kalman Filter.

The pose estimated from keypoints (Section 3.2) results from a frame by frame tracking scheme. In our case, that transformation is local ($\mathbf{R}_{k-1}^k$ and $\mathbf{t}_{k-1}^k$), as it is computed from two consecutive frames. Moreover, the pose estimated from facial landmarks (Section 3.1) is computed from an initial reference set of 3D landmarks aligned to the initial pose $\mathbf{R}_0$ and $\mathbf{t}_0$, which means that the pose is given by a global transformation, $\mathbf{R}_0^k$ and $\mathbf{t}_0^k$.

To estimate the head pose from each set of features, we minimize the error between the reprojected 3D features points $\{\mathbf{P}\}_{i=1}^{\eta}$ on the image plane and their 2D correspondences on the image $\{\mathbf{p}\}_{i=1}^{\eta}$ at time k. These features correspond to the keypoints $\{\mathbf{P}_K, \mathbf{p}_K\}$, with $\eta = m$, or to the facial landmarks $\{\mathbf{P}_F, \mathbf{p}_F\}$, with $\eta = n$. For the keypoints, the reprojected 3D features $\{\mathbf{P}_K\}_{i=1}^{m}$ are given at time $k-1$, while for the facial landmarks, the 3D features $\{\mathbf{P}_F\}_{i=1}^{n}$ are given at the initial frame. The minimization is then expressed as

$$\arg\min \sum_{i=1}^{\eta} \|\pi(\mathbf{R}\mathbf{P}_i + \mathbf{t}) - \mathbf{p}_i\|_2^2 \tag{6}$$

$\pi(\mathbf{P})$ denotes the perspective projection operator, where $\pi : \mathbb{R}^3 \mapsto \mathbb{R}^2$, and i is the index of the i-th feature point. As the calibration of the camera is known, Eq. (6) is minimized in the least squared sense with respect to the pose parameters $\mathbf{R}$ and $\mathbf{t}$, using Levenberg-Marquardt iteration.

For the Kalman Filter, the head rotation is denoted using a quaternion $\mathbf{q} = [q_x, q_y, q_z, q_w]^T$, where q_w is the scalar part and $\{q_x, q_y, q_z\}$ the vector part. The translation is denoted in homogeneous coordinates as $\tilde{\mathbf{t}} = [t_x, t_y, t_z, 1]^T$. As we are interested in HPE providing 6 D.o.F., we concatenate the rotation and translation vectors to have a 8×1 state vector $\mathbf{x}$ in the Kalman Filter, $\mathbf{x} = \left[\mathbf{q}^T, \tilde{\mathbf{t}}^T\right]^T$. This vector comprises the overall estimated head pose from the first given frame, *i.e.*, the global head pose.

Initial HPE. The head pose in the first frame, $\mathbf{R}_0$ and $\mathbf{t}_0$, is obtained exclusively from facial landmarks. We do so using Eq. (6) with $\{\mathbf{P}_F\}_{i=1}^{n}$ from the reference head model and the aligned facial landmarks $\{\mathbf{p}_F\}_{i=1}^{n}$ obtained at the first frame. The Kalman Filter is then initialized with the pose estimated from the facial landmarks.

HPE for the other frames. Since the rotation is represented with quaternions and the translation with homogeneous coordinates, we can define a linear process model and thus, a linear Kalman Filter. The predicted state estimate is computed from Eq. (7), where $\hat{\mathbf{x}}_k^-$ represents the *a priori* estimate at time k and $\mathbf{A}$ is the state transition matrix.

$$\hat{\mathbf{x}}_k^- = \mathbf{A}_k \hat{\mathbf{x}}_{k-1} \tag{7}$$

In the proposed approach, $\mathbf{A}$ is a 8×8 matrix given by Eq. (8) with a normal distributed process noise with covariance $\mathbf{Q}$. This matrix needs to be updated at each iteration.

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_\rho & 0 \\ 0 & \mathbf{A}_\mathbf{t} \end{bmatrix} \tag{8}$$

$\mathbf{A}_\rho$ is the state transition sub-matrix to project the rotation ahead and is defined by the local rotation $\mathbf{R}_{k-1}^k$. This rotation is computed from the keypoint-based tracking scheme and is denoted by $\boldsymbol{\rho} = [\rho_x, \rho_y, \rho_z, \rho_w]^T$. $\mathbf{A}_\rho$ is given by

$$\mathbf{A}_\rho = \begin{bmatrix} \rho_w & -\rho_z & \rho_y & \rho_x \\ \rho_z & \rho_w & -\rho_x & \rho_y \\ -\rho_y & \rho_x & \rho_w & \rho_z \\ -\rho_x & -\rho_y & -\rho_z & \rho_w \end{bmatrix} \tag{9}$$

Meanwhile, the state transition sub-matrix $\mathbf{A}_\mathbf{t}$ to update the translation is given by

$$\mathbf{A}_\mathbf{t} = \begin{bmatrix} \mathbf{R}_{k-1}^k & \mathbf{t}_{k-1}^k \\ 0 & 1 \end{bmatrix}, \tag{10}$$

where the new translation estimate $\mathbf{t}_0^{k-}$ is computed from

$$\mathbf{t}_0^{k-} = \mathbf{R}_{k-1}^k \mathbf{t}_0^k + \mathbf{t}_{k-1}^k. \tag{11}$$

The measurement model is given by Eq. (12). $\mathbf{z}_k$ is the new measurement at time k, $\mathbf{H}$ is a 7×8 matrix that relates the current state $\mathbf{x}_k$ to the measurement and $\mathbf{v}_k$ denotes the measurement noise in the observation. $\mathbf{H}$ is given by $\mathbf{H} = [\mathbf{I}_7\ 0]$, where $\mathbf{I}_7$ is a 7×7 identity matrix.

$$\mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k \tag{12}$$

Subsequently, the state estimate is updated at the correction step using Eq. (13). $\mathbf{K}_k$ denotes the Kalman gain and $\hat{\mathbf{x}}_k$ the *a posteriori* estimate.

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}\hat{\mathbf{x}}_k^-) \tag{13}$$

In this method, $\mathbf{z}_k$ represents the global head pose extracted only from the facial landmarks. In general, the Kalman gain $\mathbf{K}_k$ depends on the ratio between the covariances of the predicted and the measured state. Herein, we adapt the measurement covariance in every frame according to the expected accuracy of the extracted global head pose.

Method	Year	RMSE ± STD			MAE			Average	Time (FPS)
		Roll	Pitch	Yaw	Roll	Pitch	Yaw		
La Cascia *et al*. [24]	2000	-	-	-	9.8	6.1	3.3	6.4	-
Sung *et al*. [37]	2008	-	-	-	3.1	5.6	5.4	4.7	-
Morency *et al*. [29]	2008	-	-	-	2.91	3.67	4.97	3.85	6
Jang and Kanade [18]	2008	-	-	-	2.1	3.7	4.6	3.46	-
An and Chung [2]	2008	-	-	-	2.83	3.95	3.94	3.57	-
Choi and Kim [7]	2008	-	-	-	2.82	3.92	4.04	3.59	14
Kumano *et al*. [22]	2009	-	-	-	2.9	4.2	7.1	4.73	-
Lefevre and Odobez [25]	2009	-	-	-	**2.0**	3.3	4.4	3.23	3
Asteriadis *et al*. [3]	2010	3.56	4.89	5.72	-	-	-	-	-
Prasad and Aravind [32]	2010	-	-	-	3.6	**2.5**	3.8	3.3	-
Jang and Kanade [19]	2010	-	-	-	2.07	3.44	4.22	3.24	-
Saragih *et al*. [35]	2011	-	-	-	2.55	4.46	5.23	4.08	8
Valenti *et al*. [41]	2012	3.00± 2.82	5.26± 4.67	6.10± 5.79	-	-	-	-	-
Wang *et al*. [44]	2012	-	-	-	**1.86**	2.69	**3.75**	2.76	15
Baltrusaitis *et al*. [4]	2012	-	-	-	2.08	3.81	**3.00**	2.96	-
Tran *et al*. [39]	2013	-	-	-	2.4	3.9	5.4	3.90	5
Vicente *et al*. [43]	2015	-	-	-	3.2	6.2	4.3	4.56	25
Jeni *et al*. [20]	2017	-	-	-	2.41	**2.66**	3.93	3.0	50
Wu *et al*. [45]	2017	-	-	-	3.1	5.3	4.9	4.43	-
Diaz Barros *et al*. [9]	2017	3.36±2.98	4.46±3.84	5.09±4.56	2.56	3.39	3.99	3.31	56
Gou *et al*. [17]	2017	-	-	-	3.3	4.8	5.1	4.4	-
HPE from keypoints	2018	3.42±3.05	4.61±3.83	5.49±4.84	2.61	3.51	4.29	3.47	-
HPE from facial landmarks	2018	**2.58±2.42**	5.68±4.31	7.18±6.71	**2.01**	4.45	5.87	4.11	-
HPE in fused framework	2018	3.06±2.78	**4.38±3.76**	**4.93±4.56**	2.32	3.41	3.90	3.21	40

Table 1. Comparison with other methods of the state of the art on BU dataset with different metrics for head rotation.

Handling occlusion and pose recovery. We are interested in a method which is robust to extreme head rotations, *i.e*., that is able to estimate the pose, even when the face is not detected or is occluded. In this situation, the updated head pose relies only on the keypoint tracking by updating only the prediction step at the Kalman filter with no correction step. This means that the approach is able to estimate the pose even when the face is not detected, though with a state covariance increasing in every frame.

As soon as the face is detected again, the new measurement is incorporated to correct the estimated pose. This step is crucial for pose recovery, especially if the head had not been detected for several frames. Moreover, as in this case the predicted state covariance increases with every frame, it's weight will be relatively small in the correction step, once a reliable global head pose is detected. Thus, the recovery will take place rapidly. We update the covariances of the process noise and the measurement noise each frame, depending on the current pose, to give more weight to the detection scheme for (near) frontal head poses and to the tracking scheme for extreme head poses.

4. Experiments and results

The performance of the proposed scheme has been assessed using a publicly available database, to compare to related works and with our own dataset, for testing extreme head poses. We also investigated the contribution of each pipeline in our HPE framework, by evaluating the performance of each one by separate. The pipelines were implemented in C++ and were tested using an Intel Core(TM) i5-4210U processor.

4.1. Comparison with related methods

We evaluated our method using the Boston University (BU) head tracking database [24]. This database is composed of 45 video sequences under uniform illumination, with 5 different subjects performing several head movements. Ground truth was acquired using a magnetic tracker attached to the users' heads, with a nominal accuracy of 1.8 mm in translation and 0.5 degrees in rotation.

A comparison with relevant works in 2D input data is reported in Table 1. We evaluated the accuracy of the estimated pose using three metrics: root mean square error (RMSE), mean absolute error (MAE) and the standard deviation (STD). We reported also the average of the MAE for each method. From the results, it can be noted that the accuracy of the proposed approach is similar to other methods of the state of the art. Only [20] presents a smaller average error with a higher estimation rate.

In the last three rows of Table 1, we have compared the angular accuracy of our approach with the individual HPE

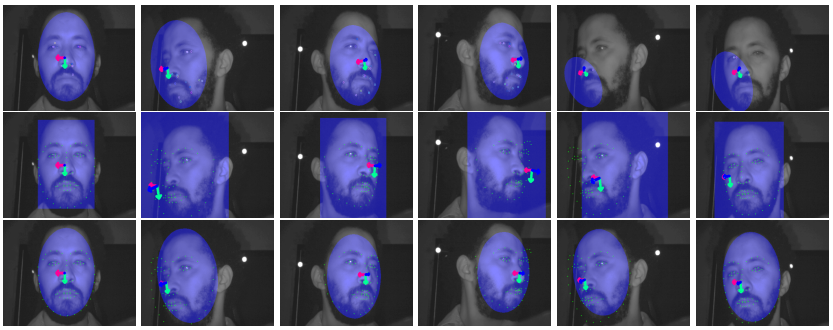

Figure 5. HPE for extreme yaw rotation with keypoints (top), facial landmarks (center) and the fusion scheme (bottom).

PROCESS	Time (ms)
Initial face detection	29.6
Initial HPE	1.9
Total	31.5
2D feature detection and matching	23.81
Estimation of 3D keypoints	0.13
HPE	0.93
Total	24.87

Table 2. Averaged runtime for 100 launches.

methods using keypoints or facial landmarks. The fused method performs better in yaw and pitch angle estimation than both individual methods. In the case of the roll angle, the method is better than the keypoint-based approach, but does not achieve the performance of the facial-landmark-based approach. This result can be addressed to the fact that in the case of pure "roll" rotation facial landmarks are visible constantly, leading to a high accuracy. But in other motions with out-of-plane rotation, facial landmarks are not stable and our fused approach outperforms it.

Most works do no report translation errors, since no calibration data is provided for the dataset. In our case, the MAE in the translation for the X, Y and Z axes correspond to 3.16, 1.64 and 1.17 respectively. For the HPE scheme from facial landmarks, these errors are 4.42, 2.39 and 1.24, while for the HPE scheme from keypoints, the errors are 3.25, 1.65 and 1.23, respectively, *i.e.*, the results of the proposed fused method are better than both individual methods.

Time consumption analysis. To account for real-time capabilities of our method, we evaluated the time consumption for each step on the BU dataset (Table 2). The consumption of others methods are shown on Table 1. During the experiments, the initialization step took around 31 FPS, while for the rest of the sequences the average was of 40 FPS. Comparing with related works, only [20] and [9] could reach >40FPS, with estimation errors similar to the proposed approach. In contrast to [20], no training with a large dataset of high-resolution 3D face scans was needed.

4.2. Experiments with our own dataset

We evaluated the proposed approach using our own dataset [14], which includes sequences with extreme head rotations. This means that the rotation around the X or Y axes (pitch and yaw) are higher than a threshold value (*e.g.*, $>$ 45 degrees).

Figure 5 shows some frames for a sequence with extreme yaw rotations. Results from the keypoint-based, facial-landmark-based and fusion approaches are shown on the top, second and last row, respectively. The blue area corresponds to the projection of the head model onto the 2D image, while the coordinate system depicting the estimated pose is shown with the RGB arrows. For the keypoint-based and fusion approaches, the projection of the EHM corresponds to an elliptical surface, as detailed in Section 3.2.

It can be observed that the estimated pose from facial landmarks is not consistent for extreme head rotations, although it is precise for frontal faces. Furthermore, the pose estimated from keypoints was continuous, but started drifting after several frames. On the other hand, the pose estimated from the fusion scheme enables the tracking of the head correctly through the entire sequence, even when the facial landmarks were self-occluded (2nd and 5th columns).

5. Conclusions

In this work, we presented a real-time approach for head pose estimation, based on intensity images. The method relies on facial landmarks detected in each frame and tracked keypoints, to compute two independent pose estimations that are later fused using a dedicated Kalman Filter. This scheme contributes to improve the estimation step, by making the method robust to extreme head rotation.

The experiments showed that our method has similar results to the state of the art, with an estimation rate of 40 FPS. Future work includes the refinement of the 3D keypoints on the geometric head model, to increase the accuracy of the HPE for the tracking scheme. We are also interested in gaze estimation from the detected head pose.

References

[1] B. Ahn, J. Park, and I. S. Kweon. Real-time head orientation from a monocular camera using deep neural network. In *Asian Conference on Computer Vision (ACCV)*, pages 82–96. Springer, 2014.

[2] K. H. An and M. J. Chung. 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 (IROS)*, pages 307–312. IEEE, 2008.

[3] S. Asteriadis, K. Karpouzis, and S. Kollias. Head pose estimation with one camera, in uncalibrated environments. In *Workshop on Eye Gaze in Intelligent Human Machine Interaction*, pages 55–62. ACM, 2010.

[4] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[5] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[6] J. Y. Bouguet. Pyramidal implementation of the affine Lucas-Kanade feature tracker description of the algorithm. *Intel Corporation*, 5:1–10, 2001.

[7] S. Choi and D. Kim. Robust head tracking using 3D ellipsoidal head model in particle filter. *Pattern Recognition*, 41(9):2901–2915, 2008.

[8] D. Derkach, A. Ruiz, and F. M. Sukno. Head pose estimation based on 3-D facial landmarks localization and regression. In *12th International Conference on Automatic Face & Gesture Recognition (FG'17)*, pages 820–827. IEEE, May 2017.

[9] J. M. Diaz Barros, F. Garcia, B. Mirbach, and D. Stricker. Real-time monocular 6-DoF head pose estimation from salient 2D points. In *International Conference on Image Processing (ICIP)*. IEEE, 2017.

[10] N. A. Dodgson. Variation and extrema of human interpupillary distance. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 36–46. SPIE, 2004.

[11] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, and R. Horaud. Head pose estimation via probabilistic high-dimensional regression. In *International Conference on Image Processing (ICIP)*, pages 4624–4628. IEEE, 2015.

[12] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3D face analysis. *International Journal of Computer Vision*, 101(3):437–458, February 2013.

[13] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 617–624. IEEE, 2011.

[14] German Research Center for Artificial Intelligence (DFKI). Head pose estimation dataset. `http://av.dfki.de/publications/fusion-of-keypoint-tracking-and-facial-landmark-detection-for-real-time-head-pose-estimation/`, 2017.

[15] R. S. Ghiass, O. Arandjelović, and D. Laurendeau. Highly accurate and fully automatic head pose estimation from a low quality consumer-level RGB-D sensor. In *2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, pages 25–34. ACM, 2015.

[16] C. C. Gordon, B. Bradtmiller, C. E. Clauser, T. Churchill, J. T. McConville, I. Tebbetts, and R. A. Walker. Anthropometric survey of u.s. army personnel: Methods and summary statistics. In *Technical report 89-044. Natick MA: U.S. Army Natick Research, Development and Engineering Center.*, 1989.

[17] C. Gou, Y. Wu, F.-Y. Wang, and Q. Ji. Coupled cascade regression for simultaneous facial landmark detection and head pose estimation. In *International Conference on Image Processing (ICIP)*. IEEE, 2017.

[18] J. S. Jang and T. Kanade. Robust 3D head tracking by online feature registration. In *8th International Conference on Automatic Face & Gesture Recognition (FG'08)*. IEEE, 2008.

[19] J. S. Jang and T. Kanade. Robust 3D head tracking by view-based feature point registration. Technical report, People Image Analysis (PIA) Consortium, Carnegie Mellon University, 2010.

[20] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D face alignment from 2D video for real-time use. *Image and Vision Computing*, 58:13–24, 2017.

[21] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874. IEEE, 2014.

[22] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International Journal of Computer Vision*, 83(2):178–194, Jun 2009.

[23] J. Kun, S. Bok-Suk, and K. Reinhard. Novel backprojection method for monocular head pose estimation. *International Journal of Fuzzy Logic and Intelligent Systems*, 13(1):50–58, 2013.

[24] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. *Transactions on Pattern Analysis and Machine Intelligence*, 22(4):322–336, 2000.

[25] S. Lefevre and J. M. Odobez. Structure and appearance features for robust 3D facial actions tracking. In *International Conference on Multimedia and Expo*, pages 298–301. IEEE, June 2009.

[26] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei. 3D head pose estimation with convolutional neural network trained on synthetic images. In *International Conference on Image Processing (ICIP)*, pages 1289–1293. IEEE, 2016.

[27] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz. Robust model-based 3D head pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 3649–3657. IEEE, 2015.

[28] P. Mohr, M. Tatzgern, J. Grubert, D. Schmalstieg, and D. Kalkofen. Adaptive user perspective rendering for hand-

held augmented reality. In *Symposium on 3D User Interfaces (3DUI)*, pages 176–181. IEEE, 2017.

[29] L. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *8th International Conference on Automatic Face & Gesture Recognition (FG'08)*, pages 1–8. IEEE, 2008.

[30] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[31] C. Papazov, T. K. Marks, and M. Jones. Real-time 3D head pose and facial landmark estimation from depth images using triangular surface patch features. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[32] B. H. Prasad and R. Aravind. A robust head pose estimation system for uncalibrated monocular videos. In *7th Indian Conference on Computer Vision, Graphics and Image Processing*, pages 162–169. ACM, 2010.

[33] E. Rosten, R. Porter, and T. Drummond. FASTER and better: A machine learning approach to corner detection. *Transactions on Pattern Analysis and Machine Intelligence*, 32:105–119, 2010.

[34] A. Saeed and A. Al-Hamadi. Boosted human head pose estimation using kinect camera. In *International Conference on Image Processing (ICIP)*, pages 1752–1756. IEEE, 2015.

[35] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[36] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen. Driveahead - a large-scale driver head pose dataset. In *International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017.

[37] J. Sung, T. Kanade, and D. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 80(2):260–274, 2008.

[38] D. J. Tan, F. Tombari, and N. Navab. Real-time accurate 3D head tracking and pose estimation with consumer rgb-d cameras. *International Journal of Computer Vision*, pages 1–26, 2017.

[39] N.-T. Tran, F.-E. Ababsa, M. Charbit, J. Feldmar, D. Petrovska-Delacrétaz, and G. Chollet. 3D face pose and animation tracking via eigen-decomposition based bayesian approach. In *International Symposium on Visual Computing*, pages 562–571. Springer, 2013.

[40] S. Tulyakov, R.-L. Vieriu, S. Semeniuta, and N. Sebe. Robust real-time extreme head pose estimation. In *22nd International Conference on Pattern Recognition (ICPR)*, pages 2263–2268. IEEE, 2014.

[41] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. *Transactions on Image Processing*, 21(2):802–815, 2012.

[42] R. Valenti, Z. Yucel, and T. Gevers. Robustifying eye center localization by head pose cues. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 612–618. IEEE, 2009.

[43] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi. Driver gaze tracking and eyes off the road detection system. *Transactions on Intelligent Transportation Systems*, 16(4):2014–2027, 2015.

[44] H. Wang, F. Davoine, V. Lepetit, C. Chaillou, and C. Pan. 3D head tracking via invariant keypoint learning. *Transactions on Circuits and Systems for Video Technology*, 22(8):1113–1126, 2012.

[45] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] X. Xu and I. A. Kakadiaris. Joint head pose estimation and face alignment framework using global and local cnn features. In *12th International Conference on Automatic Face & Gesture Recognition (FG'17)*, volume 2, pages 642–649. IEEE, May 2017.

[47] C. Yin and X. Yang. Real-time head pose estimation for driver assistance system using low-cost on-board computer. In *15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry*, volume 1, pages 43–46. ACM, 2016.

[48] Y. Yu, K. A. Funes Mora, and J.-M. Odobez. Robust and accurate 3D head pose estimation through 3DMM and online head model reconstruction. In *12th International Conference on Automatic Face & Gesture Recognition (FG'17)*, pages 711–718. IEEE, May 2017.