Head Pose Estimation via Probabilistic High-Dimensional Regression

Vincent Drouard, Silèye Ba, Georgios Evangelidis, Antoine Deleforge, Radu Horaud

HEAD POSE ESTIMATION VIA PROBABILISTIC HIGH-DIMENSIONAL REGRESSION

Vincent Drouard [1], *Silèye Ba* [1], *Georgios Evangelidis* [1], *Antoine Deleforge* [2], *and Radu Horaud* [1]

[1] INRIA Grenoble Rhône-Alpes, France [2] Friedrich-Alexander-Universität, Erlangen, Germany

ABSTRACT

This paper addresses the problem of head pose estimation with three degrees of freedom (pitch, yaw, roll) from a single image. Pose estimation is formulated as a high-dimensional to low-dimensional mixture of linear regression problem. We propose a method that maps HOG-based descriptors, extracted from face bounding boxes, to corresponding head poses. To account for errors in the observed bounding-box position, we learn regression parameters such that a HOG descriptor is mapped onto the union of a head pose and an offset, such that the latter optimally shifts the bounding box towards the actual position of the face in the image. The performance of the proposed method is assessed on publicly available datasets. The experiments that we carried out show that a relatively small number of locally-linear regression functions is sufficient to deal with the non-linear mapping problem at hand. Comparisons with state-of-the-art methods show that our method outperforms several other techniques.

1. INTRODUCTION

Head pose is an important visual cue in many scenarios such as social event analysis [1], human-robot interaction (HRI) [2] and driver-assistance systems [3] to name a few. For example, in social event analysis, 3D head-pose information drastically helps to determine the interaction between people and to extract the visual focus of attention [4]. The pose is typically expressed by three angles (pitch, yaw, roll) that describe the egocentric orientation of the head. Its estimation becomes challenging when several people are present in an image, so that their faces have a small support area, typically lower than 100×100 pixels (Fig. 1). Even if the face position in the image is known, one has to extract the pose angles from low-resolution data. The detection of local features, e.g., facial landmarks, is problematic in this case and one can only use global visual information, e.g, HOG [5].

Such a global face descriptor is used as input in this paper to estimate the 3D pose. Therefore, we must solve a high-dimensional to low-dimensional (high-to-low) mapping problem. It is well known that high-to-low regression is very challenging because of the large number of parameters that need to be estimated. This is often solved using kernel methods, such as Gaussian process regression. However, this implies an ad-hoc choice of a kernel function as well as the estimation of hyper parameters which leads to a non-linear/non-convex optimization problem. Recently, we proposed a high-to-low regression that learns a *low-to-high* regression, from which a *high-to-low* prediction is then inferred based on Bayes inversion, namely the Gaussian mixture of locally-linear mapping (GLLiM) model [6]. The advantage of this approach over existing mixture of linear regression techniques is that it avoids the estimation of the huge number of parameters associated with high-to-low learning.

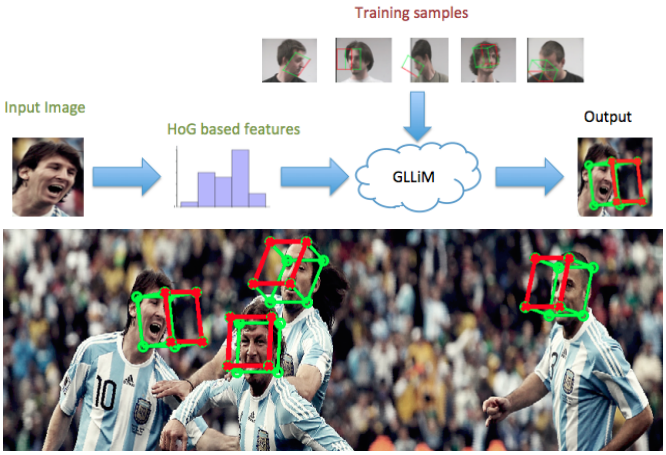

Fig. 1: ***Top*****: Method pipeline; a Gaussian locally-linear mapping (GLLiM) model, learnt using training data, maps a HOG face descriptor onto the space of head poses.** ***Bottom*****: Pose estimation on low-resolution faces, e.g.,** 100×100 **pixels.**

This paper proposes head pose estimation based on GLLiM [23, 6]. Given the face region, a mapping is constructed between HOG-based region descriptors and head poses. Since face localization, e.g., owing to a face detector, may be erroneous, we propose a method which maps HOG descriptors onto the union of head poses and position offsets, such as to optimally align an image region with a face. The method is evaluated onto two public datasets [7, 8]. Experimental evaluations show that our head pose estimation method outperforms state-of-the-art methods.

Support from both the EU-FP7 ERC Advanced Grant VHIA (#340113) and STREP project EARS (#609645) is greatly acknowledged.

The remainder of the paper is organized as follows. Sec. 2 describes related head pose estimation methods. Sec. 3 outlines the proposed regression model and its extension to account for face localization errors. Experimental results are discussed in Sec. 4 and conclusions are drawn in Sec. 5.[1]

2. RELATED WORK

We briefly discuss the head pose estimation literature the most relevant to our work. The reader may refer to [4] for a detailed survey. Landmark-based approaches constitute a wide category of methods that rely on the detection of facial landmarks such as eyes, lip corners and nose tip. Flexible models like the active appearance model [9] and geometric methods [10] belong to this category. These methods use landmark configurations in order to estimate the head pose, either directly from the geometry or by further modeling a mapping, e.g., graph matching [11] or linear regression [12]. In both cases, however, the detection of the facial features is required and hence they are sensitive to landmark localization. Furthermore, facial landmarks localization is possible only for high resolution face images.

In presence of low resolution images, global face descriptors are more robust as they only require face localization [13]. The most common approach in this category is the appearance template method that work in a nearest-neighbor manner [14]. Subspace embedding also uses global information [15] but it solves template matching in a subspace, e.g., using PCA [16].

Non-linear regression methods learn a mapping from feature spaces onto one or more pose angles. While they mostly use global information, they can also learn a mapping from local features onto poses. The main drawback of these methods is the high dimension of the input space. PCA [17] or local features [18] may be used to reduce the dimensionality of the data. This presents the risk to map the input onto an intermediate low-dimensional space that does not contain the information needed to correctly predict the output. Neural network based approaches fall into this category as well [19, 20]. Deep networks [21] and random regression forests [7] have been also recently used for head pose estimation. While the above mentioned methods mainly use color information, the release of depth sensors has led to the development of methods that rely on depth [7] or on RGB-D data [22].

3. THE REGRESSION METHOD

This section summarizes the regression method of [6]. Let $\boldsymbol{X}$, $\boldsymbol{Y}$ be two random variables, such that $\boldsymbol{X} \in \mathbb{R}^L$ denotes the low-dimensional output (or response), e.g., head-pose ($L = 3$) and $\boldsymbol{Y} \in \mathbb{R}^D$ ($D \gg L$) denotes the high-dimensional input, e.g., the multi-scale HOG descriptor [13]. The goal is to predict a *response* $\boldsymbol{X}$ given an *input* $\boldsymbol{Y}$ and the model parameters, $p(\boldsymbol{X}|\boldsymbol{Y};\boldsymbol{\theta})$. The *inverse* mapping, possibly non-linear, from $\boldsymbol{X}$ to $\boldsymbol{Y}$, is modeled by a mixture of locally-linear transformations:

$$\boldsymbol{Y} = \sum_{k=1}^{K} \mathbb{I}(Z = k)(\mathbf{A}_k \boldsymbol{X} + \boldsymbol{b}_k + \boldsymbol{E}_k), \tag{1}$$

where $\mathbb{I}$ is the indicator function, Z is a missing variable such that $Z = k$ if and only if $\boldsymbol{Y}$ is the image of $\boldsymbol{X}$ by the affine transformation $\mathbf{A}_k\boldsymbol{X} + \boldsymbol{b}_k$, with $\mathbf{A}_k \in \mathbb{R}^{D\times L}$ and $\boldsymbol{b}_k \in \mathbb{R}^D$, and $\boldsymbol{E}_k \in \mathbb{R}^D$ is an error vector capturing both the observation noise and the reconstruction error due to the piecewise approximation. The missing variable Z allows us to write the joint probability of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as the following mixture:

$$p(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}, Z = k;\boldsymbol{\theta}) \\ p(\boldsymbol{X} = \boldsymbol{x}|Z = k;\boldsymbol{\theta})p(Z = k;\boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\theta}$ denotes the model parameters and $\boldsymbol{y}$ and $\boldsymbol{x}$ denote realizations of $\boldsymbol{Y}$ and $\boldsymbol{X}$ respectively. Assuming that $\boldsymbol{E}_k$ is a zero-mean Gaussian variable with covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{D\times D}$, we obtain that $p(\boldsymbol{y}|\boldsymbol{x}, Z = k;\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}; \mathbf{A}_k\boldsymbol{x} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k)$. If we further assume that $\boldsymbol{X}$ follows a mixture of Gaussians via the same assignment $Z = k$, we can write that $p(\boldsymbol{x}|Z = k;\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ and $p(Z = k;\boldsymbol{\theta}) = \pi_k$, where $\boldsymbol{c}_k \in \mathbb{R}^L$, $\boldsymbol{\Gamma}_k \in \mathbb{R}^{L\times L}$ and $\sum_{k=1}^{K} \pi_k = 1$. Note that this representation induces a partition of $\mathbb{R}^L$ into K regions $\mathcal{R}_k$, where $\mathcal{R}_k$ is the region where the transformation $(\mathbf{A}_k, \boldsymbol{b}_k)$ is most likely invoked. This model, described by the parameter set $\boldsymbol{\theta} = \{\boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \pi_k, \mathbf{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, can be learnt using a training set $\{\boldsymbol{y}_n, \boldsymbol{x}_n\}_{n=1}^N$ via a closed-form EM algorithm [23], which alternates between (i) the update of the responsibilities $p(Z|\boldsymbol{X}, \boldsymbol{Y};\boldsymbol{\theta})$ given the current parameters and (ii) the estimation of the parameters that maximize the expected complete-data log-likelihood $\mathrm{E}[\log p(\boldsymbol{X}, \boldsymbol{Y}, Z|\boldsymbol{\theta})]$. Initial responsibilities are obtained by fitting a K-component GMM to the low-dimensional data $\{\boldsymbol{x}_n\}_{n=1}^N$.

Once $\boldsymbol{\theta}$ is estimated, one can easily obtain the low-to-high prediction $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ as well as the high-to-low prediction:

$$p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}) = \sum_{k=1}^{K} \frac{\pi_k^* \mathcal{N}(\boldsymbol{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K} \pi_j^* \mathcal{N}(\boldsymbol{y}; \boldsymbol{c}_j^*.\boldsymbol{\Gamma}_j^*)} \mathcal{N}(\boldsymbol{x}; \mathbf{A}_k^*\boldsymbol{y} + \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*) \tag{3}$$

which is a Gaussian mixture parameterized by the set

$$\boldsymbol{\theta}^* = \{\boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*, \pi_k^*, \mathbf{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*\}_{k=1}^K. \tag{4}$$

[1] Supplementary material for this paper can be found at `https://team.inria.fr/perception/research/head-pose/`

A prominent feature of this model is that $\boldsymbol{\theta}^*$ can be expressed analytically from $\boldsymbol{\theta}$ (the reader is referred to [6, 23] for further details). As a consequence, we can easily use the expectation of (3) to obtain, e.g., the desired pose $\boldsymbol{x}$ given a descriptor $\boldsymbol{y}$:

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}] = \sum_{k=1}^{K} \frac{\pi_k \mathcal{N}(\boldsymbol{y};\boldsymbol{c}_k^*,\boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{y};\boldsymbol{c}_j^*,\boldsymbol{\Gamma}_j^*)} (\mathbf{A}_k^* \boldsymbol{y} + \boldsymbol{b}_k^*). \quad (5)$$

3.1. Iterative Prediction with GLLiM

The above model assumes that a bounding box is perfectly centered onto a face. In practice, face detectors yield localization errors. To compensate such errors, we consider an augmented response variable $\boldsymbol{x}' = [\boldsymbol{x}, \boldsymbol{du}]^\top$, where $\boldsymbol{x}$ is the pose vector as above, and $\boldsymbol{du} \in \mathbb{R}^2$ is the 2D offset between an *observed* face bounding-box and the true one. We train the mapping from the HOG-based descriptor $\boldsymbol{y}$ to the augmented vector $\boldsymbol{x}' \in \mathbb{R}^5$ allowing our model to predict both a head pose and an image offset. If this prediction is run multiple times, it follows that the bounding-box position is refined incrementally to eventually converge to an optimal alignment between the predicted bounding-box and the true bounding-box. Eventually, this provides accurate head-pose estimates because the face descriptors better match those used for training. Alg. 1 outlines this iterative prediction scheme. While a few iterations may be sufficient, one can run the algorithm until the norm of the offset $\boldsymbol{du}$ vanishes.

Algorithm 1 Iterative prediction with GLLiM.

Require: Trained GLLiM for mapping $\boldsymbol{y} \rightarrow \boldsymbol{x}' = [\boldsymbol{x}, \boldsymbol{du}]^\top$, initial face center $\boldsymbol{u}_0$ and face size $N_1 \times N_2$, $ITER_{MAX}$.

1: $i \leftarrow 1$
2: **repeat**
3: Build $\boldsymbol{y}_i$ (HOG descriptor) from the $N_1 \times N_2$ bounding-box centered at position $\boldsymbol{u}_{i-1}$
4: Predict $\boldsymbol{x}_i'$ from (5) and extract predicted offset $\boldsymbol{du}_i$
5: Update face position as $\boldsymbol{u}_i \leftarrow \boldsymbol{u}_{i-1} + \boldsymbol{du}_i$
6: $i \leftarrow i + 1$
7: **until** $i = ITER_{MAX}$
8: **return** head pose $\boldsymbol{x}_i$ and center position $\boldsymbol{u}_i$

4. RESULTS

In order to evaluate the performance of the presented method, we use two widely known datasets for head pose estimation, the Prima head pose image dataset [8] and the Biwi Kinect head pose dataset [7], referred as Prima and Biwi respectively. The Prima dataset is composed of static images, while the Biwi dataset contains videos. Both datasets are fully annotated thus providing head pose angles in degrees and face position (pixel location in images) for every image/frame. The pose $\boldsymbol{x}$ is expressed with two angles (yaw, pitch) in Prima and three angles (yaw, pitch, roll) in Biwi, i.e., $\boldsymbol{x} \in \mathbb{R}^2$ and $\boldsymbol{x} \in \mathbb{R}^3$ respectively. To build the face HOG descriptor $\boldsymbol{y}$, we resize the annotated area into an image of size 64×64 pixels, and we split it into cells using 8×8 and 4×4 grids. Then, the HOG features of all grid cells are stacked into a vector of dimension $D = 1856$.

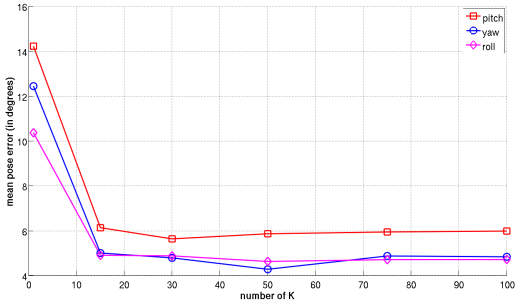

Fig. 2: Pose estimation error as a function of the number of affine transformations K (Biwi dataset, "unseen faces").

We consider two evaluation protocols. A typical one, where the data are split into training and testing sets of equal size, and a leave-one-out protocol, where the test set consist of all the images of one person, while the training set contains all the other persons in the dataset. We refer to these protocols as "seen faces" and "unseen faces" respectively. In all cases, the pose estimation error of a test sample is quantified by the mean absolute error over the estimated angles [4], and this value is further averaged over the entire test set.

4.1. The Biwi Dataset

The Biwi dataset is accurately and automatically annotated.[2] This allows to evaluate the performance of the head pose estimation method and to quantify the influence of the number of affine transformations K, i.e., the optimal (minimum) number of transformations needed by GLLiM to model the non-linear mapping between the high- and low-dimensional variables. Due to space limitations, we only show the influence of K in the challenging case of "unseen faces".

Fig. 2 shows the error curve for the three angles as a function of K. As can be seen, the error smoothly decreases with K and reaches a steady state. There is a bound ($K = 30$) after which the error is not decreasing anymore. The average error floor of the three curves is about 5 degrees. Table 1 shows that the proposed method performs well in comparison with state-of-the-art methods for both evaluation protocols. Note that only Ahn *et al.* [21] use color information. Fanelli *et al.* [7] use depth data while Wang *et al.* [22] exploit both modalities. For the "seen faces" protocol, our method performs similarly to [21]. While for the more challenging "unseen faces"

[2] http://www.faceshift.com

Method	yaw	pitch	roll
	seen faces		
Ahn *et al.* [21]	2.6 ± 2.5	3.4 ± 2.9	2.8 ± 2.4
Our method	$\mathbf{2.6 \pm 2.3}$	$\mathbf{2.9 \pm 3.1}$	$\mathbf{2.3 \pm 2.5}$
	unseen faces		
Fanelli *et al.* [7]	$\mathbf{3.5 \pm 5.8}$	$\mathbf{3.8 \pm 6.5}$	5.4 ± 6.0
Wang *et al.* [22]	8.8 ± 14.3	8.5 ± 11.1	7.4 ± 10.8
Our method	4.9 ± 4.1	5.9 ± 4.8	$\mathbf{4.7 \pm 4.6}$

Table 1: Pose estimation error (in degrees) on Biwi data-set.

case, our method's performances are similar to [7] and better than [22], which benefit from 3D depth data.

4.2. The Prima Dataset

The Prima dataset was manually annotated so that the provided face localization may not be very accurate. We focus again on the "unseen faces" scenario. The convergence of the pose estimation error in terms of the number of transformations K is shown in Fig. 3. The error reaches the floor of 7.5 degrees for both yaw and pitch, when at least $K = 50$ transformations are used. Unlike Biwi, the pose space of the Prima dataset is uniformly sampled on a 2D grid (yaw and pitch), hence the higher error floor and the larger low bound of K.

To study the contribution of the iterative prediction, we proceed as follows. We keep the size of the bounding-box fixed and we manually add Gaussian noise to the box position on both training and testing images. Based on Sec. 3.1, we learn the mapping from $\boldsymbol{y}$ to the augmented vector $\boldsymbol{x}'$ by drawing offsets $\boldsymbol{du}$ from a zero-mean Gaussian distribution with deviation σ_p. Fig. 4 shows how the pose error increases with σ_p when the prediction is done once. Note that the value of σ_p simulates the magnitude of the face localization error, e.g., owing to a face detector. Fig. 5 depicts the error curves for both pose and position estimation as function of the iterations, when the iterative prediction is considered (Alg. 1). As can be seen, a lower pose error is achieved after 3 to 4 iterations. No further improvements were observed after 9 iterations. Note also the smooth position error decay (Fig. 5 right).

Table 2 compares our model's performances with state-of-the-art methods. For the challenging "unseen faces" case,

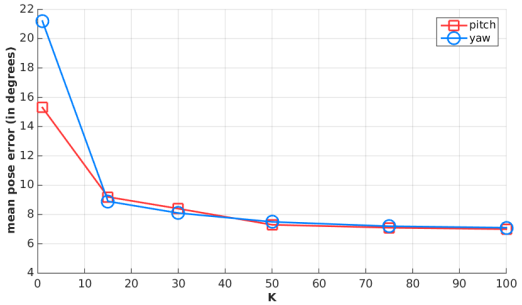

Fig. 3: Pose estimation error as a function of the number of affine transformations K (Prima dataset, "unseen faces").

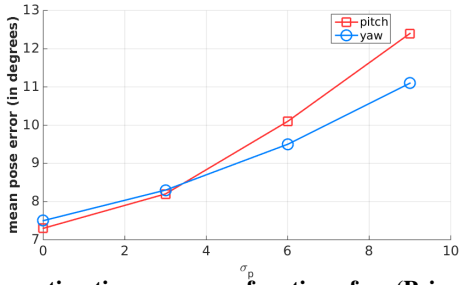

Fig. 4: Pose estimation error as a function of σ_p (Prima dataset, $K = 50$, "unseen faces").

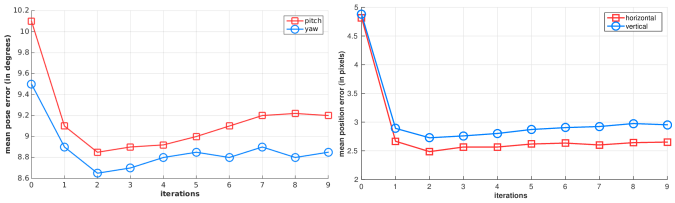

Fig. 5: Pose (left) and position (right) estimation error as a function of the number of iterations (Prima dataset, $K = 50$, $\sigma_p = 6$, "unseen faces")

results obtained without and with the iterative prediction are reported. For both evaluation protocols, our method yields accurate results and outperforms the existing methods.

5. CONCLUSION

We presented a head pose estimation method based on learning a mixture of linear regression model that maps high-dimensional HOG-based descriptors onto the low-dimensional space of head poses. Most prominently, our method is able to recover from face localization errors, which are common with low-resolution faces. We evaluated the method on publicly available datasets: our model achieves state-of-the-art results and often outperforms existing techniques. In order to address head pose estimation in videos, we plan to extend the proposed regression with a temporal model in order to smoothly track head-pose parameters.

Method	yaw	pitch
	seen faces	
Stiefelhagen [19]	10.4	10.6
Stiefelhagen (Mirror) [19]	9.5	9.7
Gourier *et al.* [15]	7.3	12.1
Our method (localization annotation)	$\mathbf{6.7 \pm 8.3}$	$\mathbf{7.2 \pm 8.1}$
Our method ($\sigma_p = 6$)	$\mathbf{8.4 \pm 8.0}$	$\mathbf{8.5 \pm 10.1}$
	unseen faces	
Gourier *et al.* [15]	10.3	15.9
Ricci & Odobez [13]	9.1	10.5
Our method (localization annotation)	$\mathbf{7.5 \pm 7.28}$	$\mathbf{7.3 \pm 8.8}$
Our method ($\sigma_p = 3$)	$\mathbf{8.3 \pm 7.75}$	$\mathbf{8.2 \pm 9.4}$
Our method ($\sigma_p = 6$)	$\mathbf{9.5 \pm 8.63}$	$\mathbf{10.1 \pm 10.9}$
Our method ($\sigma_p = 6$, 3 iterations)	$\mathbf{8.7 \pm 8.0}$	$\mathbf{8.85 \pm 9.97}$

Table 2: Pose estimation error (in degrees) on Prima data-set.

6. REFERENCES

[1] S. Sabanovic, M.P. Michalowski, and R. Simmons, "Robots in the wild: Observing human-robot social interaction outside the lab," in *The 9th IEEE International Workshop on Advanced Motion Control*, 2006, pp. 596–601.

[2] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: A survey," *Found. Trends Hum.-Comput. Interact.*, vol. 1, no. 3, Jan. 2007.

[3] Er. Murphy-Chutorian, A. Doshi, and M. M. Trivedi, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation," in *IEEE Intelligent Transportation Systems Conference*, 2007, pp. 709–714.

[4] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.

[6] A. Deleforge, F. Forbes, and R. Horaud, "High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables," *Statistics and Computing*, 2014.

[7] A. Deleforge, *Acoustic Space Mapping: A Machine Learning Approach to Sound Source Separation and Localization*, Ph.D. thesis, Université Grenoble Alpes, Grenoble, November 2013.

[8] 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*, vol. 101, no. 3, pp. 437–458, 2013.

[9] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures*. FGnet Cambridge, UK, 2004, pp. 1–9.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[11] T. Horprasert, Y. Yacoob, and L. S. Davis, "Computing 3d head orientation from a monocular image sequence," in *The 25th Annual AIPR Workshop on Emerging Applications of Computer Vision*, 1997, pp. 244–252.

[12] N. Krüger, M. Pötzsch, and C. von der Malsburg, "Determination of face position and pose with a learned representation based on labelled graphs," *Image and Vision Computing*, vol. 15, no. 8, pp. 665–673, 1997.

[13] T.F Cootes, G.V Wheeler, K.N Walker, and C.J Taylor, "View-based active appearance models," *Image and Vision Computing*, vol. 20, no. 9-10, pp. 657–664, 2002.

[14] Elisa Ricci and Jean-Marc Odobez, "Learning large margin likelihoods for realtime head pose tracking," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 2593–2596.

[15] J. Ng and S. Gong, "Composite support vector machines for detection of faces across views and pose estimation," *Image and Vision Computing*, vol. 20, no. 5-6, pp. 359–368, 2002.

[16] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head pose estimation on low resolution images," in *Multimodal Technologies for Perception of Humans*, pp. 270–280. Springer, 2007.

[17] S. J. McKenna and S. Gong, "Real-time face pose estimation," *Real-Time Imaging*, vol. 4, no. 5, pp. 333–347, 1998.

[18] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 300–305.

[19] Y. Ma, Y. Konishi, K. Kinoshita, S. Lao, and M. Kawade, "Sparse bayesian regression for head pose estimation," in *The 18th International Conference on Pattern Recognition*, 2006, vol. 3, pp. 507–510.

[20] R. Stiefelhagen, "Estimating head pose with neural networks – results on the Pointing'04 ICPR workshop evaluation data," in *Pointing'04 ICPR Workshop*, 2004.

[21] M. Osadchy, Y. Le Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *Journal of Machine Learning Reasearch*, vol. 8, pp. 1197–1215, 2007.

[22] B. Ahn, J. Park, and I. S. Kweon, "Real-time head orientation from a monocular camera using deep neural network," in *The 12th Asian Conference on Computer Vision*, 2014.

[23] Bingjie Wang, Wei Liang, Yucheng Wang, and Yan Liang, "Head pose estimation with combined 2D SIFT and 3D HOG features," in *The IEEE Seventh International Conference on Image and Graphics*, 2013, pp. 650–655.