Head Pose Estimation in First-Person Camera Views

Stefano Alletto, Giuseppe Serra, Simone Calderara and Rita Cucchiara
Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
Via Vignolese 905, 41125 Modena - Italy

***Abstract*—In this paper we present a new method for head pose real-time estimation in ego-vision scenarios that is a key step in the understanding of social interactions. In order to robustly detect head under changing aspect ratio, scale and orientation we use and extend the Hough-Based Tracker which allows to follow simultaneously each subject in the scene. In an ego-vision scenario where a group interacts in a discussion, each subject's head orientation will be more likely to remain focused for a while on the person who has the floor. In order to encode this behavior we include a stateful Hidden Markov Model technique that enforces the predicted pose with the temporal coherence from a video sequence. We extensively test our approach on several indoor and outdoor ego-vision videos with high illumination variations showing its validity and outperforming other recent related state of the art approaches.**

I. Introduction

With the advent of wearable cameras, such as Google Glass and Vuzix Smart Glass, there has been an increasing interest of the research community toward the automatic analysis of video captured from a first-person perspective. First-person videos present many new challenges, such as background clutter, large ego-motion and extreme transitions in lighting, but they also have some unique advantages [1]. Egocentric videos are recorded from the same person over time, there is no need to place multiple fixed cameras on the environment; furthermore occlusions are less likely because the wearer moves naturally to provide a clear view. An important element in ego-vision applications is to be able to provide to its users real-time additional information regarding the surrounding world. These tools should be able to perform on commercial wearable devices with limited hardware resources.

Initial efforts have been made on the definition of methodologies to automatically understand human actions, objects and gestures in egocentric vision [2], [3], [4]. Furthermore a preliminary work [5] has been proposed for the detection and recognition of social interactions in first-person view categorizing into three subtypes: dialogue, discussion and monologue. This kind of problem is particularly relevant, because the presence or absence of social interaction is an important cue as to whether a particular event is likely to be memorable. Moreover the study of patterns of attention of subjects in a social environment is an important hint to recognize leaders in a group or categorize the amount of attention each subject receives. In this scenario face analysis is particular relevant due to the importance that facial expression, head movements and gaze direction play in human interactions.

Even if great strides have been made in face analysis, it is still challenging to obtain reliable evaluations of head movements in an unconstrained scenario. To address this issue, in this paper we present a method for head pose real-time estimation in complex scenarios such as ego-vision videos. In particular, the Hough-Based Tracker is extended in order to cope with multiple subjects, addressing the detection problem which can still be a challenge in scenarios where extreme noises due to lightning, camera motion, strongly variable subject distance and background clutter are present. In our approach we use linear SVMs classifiers that provide good classification speed which suits well real-time needs of ego-vision. Furthermore in order to include temporal coherence and improve the prediction performance, a Hidden Markov Model technique is incorporated in our approach. Experimental results shows that our method is robust to illumination variations of indoor and outdoor scenarios and outperforms state of the art techniques over unconstrained videos. Two examples are shown in Fig. 1. In order to test our method and encourage other researchers to tackle this new challenging problem we release an annotated video dataset: EGO-HPE.

Fig. 1. Head pose estimation results of our approach showing the output from the tracker and the predicted orientation.

This paper makes these main contributions:

- We propose a method capable of real-time head pose estimation in unconstrained scenarios that includes temporal consistency and outperforms related state of the art approaches in this context.
- We present a fully labeled indoor/outdoor ego-vision dataset of different social interaction events, containing several videos and more than 3400 frames fully annotated with head pose for each subject in the scene.

II. Related work

Head pose estimation has been widely studied in computer vision. Already existing approaches can be roughly divided in two major categories, whether their aim is to estimate the head pose on still images or video sequences.

Among the most notable solutions for still images problem, Ma et al. [6] proposed a multi-view face representation based on Local Gabor Binary Patterns (LGBP) extracted on many subregion in order to obtain spatial information. Wu et al. [7] presented a two-level classification framework: the first level has the objective of deriving pose estimates with some uncertainty; the second level minimizes this uncertainty by analysing finer structural details captured by the bunch graphs. While being very accurate on several publicly available datasets (e.g. [7] achieves a 90% accuracy over the Pointing 04), these works suffer significant performance losses when applied to less constrained environments like the ones typical of ego-vision. State of the art head pose estimation on still images in the wild is achieved by [8], which models every facial landmark as a part and uses global mixtures to capture topological changes due to viewpoint. However this technique has high computational costs, resulting in up to 40 seconds per image, excessively demanding for the real-time requirements of an ego-vision based framework. Recently a comprehensive study that has summarized the head pose estimation methods and systems published over the past 14 years has been presented in [9].

Literature focusing on video streams for head pose estimation can be further divided in whether it uses any kind of 3D information or not. If such information can be used, a significant accuracy improvement can be achieved as in [10], which uses a stereo camera system to track a 3D face model in real-time, or [11] where the 3D model is recovered from different views of the head and then the pose estimation is done under the assumption that the camera stays still. Wearable devices used for ego-vision video capture, being aimed to more general purpose users and being on a mid-low price tier, usually lack the ability to capture 3D information; furthermore due to the unpredictable motion of both the camera and the object a robust 3D model is often hard to recover from multiple images. Rather then using a 3D model, Huang et al. [12] utilized a computational framework for robust detection, tracking, and pose estimation of faces captured by video arrays. To estimate face orientations they presented and compared respectively two algorithms based on MLKalman filtering and multi-state CDHMM models. Orozco et al. [13] proposed a technique for head pose classication in crowded public space under poor lighting condition on low-resolution images using mean appearance templates and multi-class SVM.

To our knowledge, our solution is the first work to utilize egocentric videos in order to estimate the head pose motions. Some aforementioned works have started addressing the head pose estimation "in the wild" using still images from social networks, but these methods do not consider and model several peculiarities of the unconstrained ego-vision videos.

III. Proposed method

To deal with the complexity of the face analysis in real and unconstrained ego-vision scenarios where even face detection can be challenging, our method relies on the use of a tracker in order to follow the subject's head between frames, starting with an initial detection and then tracking the person. In our approach we use the Hough-Based Tracker by Godec et al. proposed in [14] (HBT), that aims at tracking non-rigid targets in a discriminative classifier with segmentation of the target itself. The head given in the first frame is used to train the target and the background models that use as features first and second derivatives of x- and y-direction and a histogram of gradients in the Lab-color space. Finally, the support of the target is located through back projection from a Hough Forest. A Hough Forest is an extension of a Random Forest [15] by a vector, measuring for positive samples the displacement of each patch to the target center, similar to the R-table [16] in Generalized Hough transforms. The sparse pixels voting, obtained by back projection, are used to segment the target using the GrabCut algorithm [17].
This tracker has been proven particularly robust to abrupt changes in camera viewpoint, target aspect ratio, scale, and orientation [18], factors that must be considered in ego-vision settings. In order to track in real-time simultaneously an arbitrary number of subjects we extend and parallelize it. The tracker is automatically initialized using Viola-Jones face detector as soon as a frontal detection is possible. This assumption, while being quite restrictive in certain environments, can be reasonably explained as in a video based on an egocentric point of view the people who are interacting with the person recording the video will soon be watching at him, and the face detector will be able to initialize the tracker.

In order to estimate head pose a preprocess step is taken before calculating the head descriptor: a mask obtained from the segmentation provided by the tracker is applied and the background is removed from the tracked head. The obtained image is converted to grayscale, resized to a fixed size 100×100 and, eventually, histogram equalization is applied to it. A dense HOG descriptor is then extracted using 16 bins per cell and 64 cells, resulting in a 1024 dimensional feature vector $\mathbf{x}$ for each face.
Feature normalization techniques have shown, especially in image classification, the potential to increase the overall performance of the classification step. In particular using a linear SVM that relies on dot-product there are recent approaches that modify their features representation using power normalization. For such reason, we treat our feature vector with this technique by applying the following function:

$$f(\mathbf{x}) = sign(\mathbf{x})|\mathbf{x}|^{\alpha} \quad with \quad 0 < \alpha < 1. \tag{1}$$

Based on initial observations we fix $\alpha = 0.5$. By optimizing this value, the performance could slightly improve but it would lead to a data-dependent tuning, a situation in contrast with

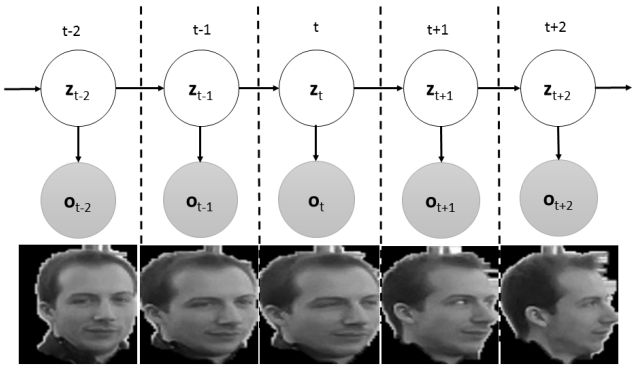

Fig. 2. Hidden Markov Models showing the latent variable and a set of observations in our approach.

the highly diversified characteristics of unconstrained ego-vision scenarios. Using these features, the head pose is then predicted using a linear SVM classifier. When dealing with high dimensional feature vectors, the linear SVM has proven competitive w.r.t its kernelized version while requiring less computational resources coping well with low tier ego-vision devices [19].

Typically, in a social scenario where the three or more subjects' activity revolves around a discussion or any kind of similar social interaction, each head orientation state will be more likely to remain unchanged for a while. For example one subject is talking, then a new subject takes the floor and becomes the new focus of the group's attention. In this situation the head pose of the subjects will have a transition to the state that allows them to focus their line of sight on the new talking subject.

In order to model this behavior and considering the temporal coherence that derive from a video sequence, a stateful Hidden Markov Model technique is included in our pipeline. Hidden Markov Models are statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be considered the simplest type of dynamic Bayesian network and, for their simplicity, have been effectively applied in the context of sequential data analysis, speech and action recognition. The HMM is a first order Markov chain built upon a set of time-varying unobserved variables/states $\mathbf{z}_t$ and a set of observations $\mathbf{o}_t$, as depicted in Fig. 2. In our case, we set the latent variables to coincide with the possible head poses while the observed variable are the input images.

Following the graphical model in Fig. 2, the probability distribution of $\mathbf{z}_t$ depends on the state of the previous latent variable $\mathbf{z}_{t-1}$ through a conditional distribution $p(\mathbf{z}_t|\mathbf{z}_{t-1})$, namely the transition probability distribution; while the conditional pdf that involves both observed and hidden variable is referred as the *emission function*, $p(\mathbf{o}_t|\mathbf{z}_t)$. In a discrete setting, with an enumerable number of states, the conditional distribution corresponds to a matrix denoted by $\mathbf{A}$, where the elements are transition probabilities among the states themselves. They are given by $A = \{a_{jk}, j, k = 1 \dots K\} \equiv p(z_{tk} = 1|z_{t1,j} = 1)$ so that the matrix $\mathbf{A}$ has $K(K-1)$ independent parameters. During learning, out of the box techniques like the Baum Welch training algorithm can be used to train the Hidden Markov Model. Nevertheless, whenever applicable, transition probabilities among discrete states can be directly set by an expert in order to impose constraint on the possible transitions. In practice, we fixed the $\mathbf{A}$ values in order to encode the context of ego-vision videos. In particular, we set in the state transition matrix a high probability of remaining in the same state, a lower probability for a transition to adjacent states and a very low probability for a transition to the not adjacent states. This leads our approach to have continuous transitions between adjacent poses. Furthermore, this also allows the removal of most of the impulsive errors that are due to wrong segmentation or to the presence of a region of background in the calculation of the descriptor. This translates in a smooth transition among possible poses that is what conventionally happens during social interaction among people in ego-vision settings.

We set the initial state $\mathbf{z}_0$, being under the assumption that the tracked sequence starts with a frontal face, as the frontal pose. Following the graphical model factorization, the joint probability distribution over both latent and observed variables results:

$$p(\mathbf{z}_t, \mathbf{o}_t) = p(\mathbf{z}_0) \prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{z}_t) p(\mathbf{z}_t|\mathbf{z}_{t-1}). \tag{2}$$

Our method combines the likelihood $p(\mathbf{z}_t|\mathbf{o}_t)$ of a measure $\mathbf{o}_t$ to belong to a pose $\mathbf{z}_t$ provided by the SVM classifier with the previous state $\mathbf{z}_{t-1}$ and the transition matrix $\mathbf{A}$ derived from the HMM, obtaining the predicted pose likelihood which is the final output.

In order to calibrate a confidence level to a probability in a SVM classifier, so it can be used as a observation for our HMM, we trained a set of Venn Predictors (VP) [20], on the SVM training set. We have the training set in the form $S = \{s_i\}_{i=1\dots n-1}$ where s_i is the input-class pair $(\mathbf{x}_i, y_i)$. Venn predictors aim at estimating the probability of a new element $\mathbf{x}_n$ belonging to each class $Y_j \in \{Y_1 \dots Y_c\}$. The prediction is performed assigning each one of the possible classification Y_j to the element $\mathbf{x}_n$ and dividing all the samples $\{(\mathbf{x}_1, y_1) \dots (\mathbf{x}_n, Y_j)\}$ into a number of categories based on a taxonomy. A taxonomy is a sequence $Q_n, n = 1, \dots, N$ of finite partitions of the space $S^{(n)} \times S$, where $S^{(n)}$ is set of multisets of S of length n. In the case of multi class SVM the taxonomy is based on the largest SVM score, therefore each example is categorized using the SVM classification in one of the c classes.
After partitioning the element using the taxonomy, the empirical probability of each classification Y_k in the category τ_{new} that contains (x_n, Y_j) is:

$$p^{Y_j}(Y_k) = \frac{|\{(\mathbf{x}^*, y^*) \in \tau_{new} : y^* = Y_k\}|}{|\tau_{new}|} \tag{3}$$

This is the pdf for the class of $\mathbf{x}_n$ but after assigning all possible classification to it we get:

$$P_n = \{p^{Y_j} : Y_j \in \{Y_1, \dots, Y_c\}\} \tag{4}$$

that is the well-calibrated set of multi probability prediction of the VP used as the emission function of Eq. 2.

Fig. 3. Examples taken from the two datasets: in the upper row, Pointing'04, in the second row, EGO-HPE

IV. Experimental Results

In order to evaluate our approach we used two different datasets: Pointing'04 [21] and EGO-HPE, a set of first-person videos that we recorded and that is publicly available[1]. Pointing'04 provides 2790 images annotated with head pose of 15 subjects across 13 yaw angles and 9 pitch angles. This dataset contains a large set of orientation poses, but each image is captured in a controlled environment. In fact, it does not present neither particular illumination conditions nor background clutter. EGO-HPE dataset provides a set of ego-vision video and a total of 3474 frames with different subjects, each video containing a combination of them. This dataset presents both indoor and outdoor sequences under significant background changes, different illumination conditions and occasional poor image quality due to camera motion. Each video is annotated at frame level for five yaw angle orientations (-75, -45, 0, 45, 75) with respect to the subject wearing the camera. Few yaw angles are considered since social interaction analysis does not require a finer subdivision of the orientations. Figure 3 shows few examples from both datasets. Performances are reported in terms of accuracy, where each prediction has been treated as correct only when resulting in the groundtruth label; without considering correct adjcent classifications as in [9].

A baseline method for estimating the head pose is proposed in order to define a fair and true technique for comparing our approach in ego-vision videos. A simplified version of HBT has been used to track the head by the means of a fixed bounding box, initialized with Viola-Jones face detector. In each frame, our HOG descriptor has been extracted from the bounding box without any step of background subtraction nor illumination normalization. This baseline is tested on the publicly available Pointing'04 dataset. Table I presents the accuracy results over the yaw angle of the evaluation of our baseline and recently proposed head pose estimation methods. In order to be comparable to state of the art works we performed classifications over each one of the 13 yaw classes instead of using the 5 classes quantization of our method. Note that, while not outperforming current state of the art works, our baseline provides comparable results.

[1]http://imagelab.ing.unimore.it/files/EGO-HPE.zip

TABLE I. Comparison of our baseline with recent related approaches on Pointing'04.

Method	Accuracy
Tu (PCA) [22]	0.552
Stiefelhagen [23]	0.520
Our Baseline	**0.519**
Gourier (Associative Memories) [24]	0.500
Tu (High-order SVD) [22]	0.492

Table II reports the accuracy results of our proposed approach, our baseline method and two state of the art head pose estimation methods [8], [25] over the EGO-HPE dataset. We present the overall accuracy and the results for each video and each subject. In our solution the linear SVM classifier has been trained by using the Pointing'04 dataset, since it provides several different subjects with large yaw variations. Furthermore this dataset includes significant pitch variations that grant good robustness to our method while facing steep pitch poses. All the variants of our method have been tested on 480x270 images; our preliminary experiments showed how increasing resolution does not impact significantly on accuracy while drastically reduces the speed. Note how our baseline method, while being able to maintain good performance on the constrained environment of Pointing'04, achieves poor results on the ego-vision dataset. This is mainly due to the differences in lightning conditions between indoor and outdoor scenarios and the presence of significant background clutter, which leads to a very noisy descriptor.

In order to be effective in ego-vision scenarios, our approach exploits a segmentation technique and a robust tracker that enhances the accuracy of the head localization reducing the background clutter as described in Section 3. This approach leads to better performance while maintaining an average speed of 5 frames per second. In particular we can observe that the usage of the power normalization and the temporal consistency provided by HMM stateful model further improves the performance in all videos. In particular the HMM permits to remove a significant amount of errors due to the presence of noise in our descriptor caused by a not precise localization of the tracker bounding box or the inclusion of a portion background in the segmentation. Figure 4 presents a comparison between the two variants of our method. It can be noted that the improvement of the HMM is clearly visible in sequences where the subject stays still for a while in a certain pose (see e.g. frames between 400 and 500).

Finally, we compared our method to two state of the art methods that publicly release their codes [8], [25]. As it has already been stated, the aim of an ego-vision approach should be to be able to perform on a low tier device in real-time, in fact all the following results have been obtained using a Intel Core I7 Q720 @ 1.6GHz.

In particular, our method is compared with [8]. Their model is based on a mixture of trees with a shared pool of parts, where each part represents a different facial landmark. Then, a global mixture is used to recognize topological changes due to different orientations. We used the Matlab code provided by the authors[2]. In order to achieve fair and comparable results in terms of speed, the number of levels per octave

[2]http://www.ics.uci.edu/ xzhu/face/

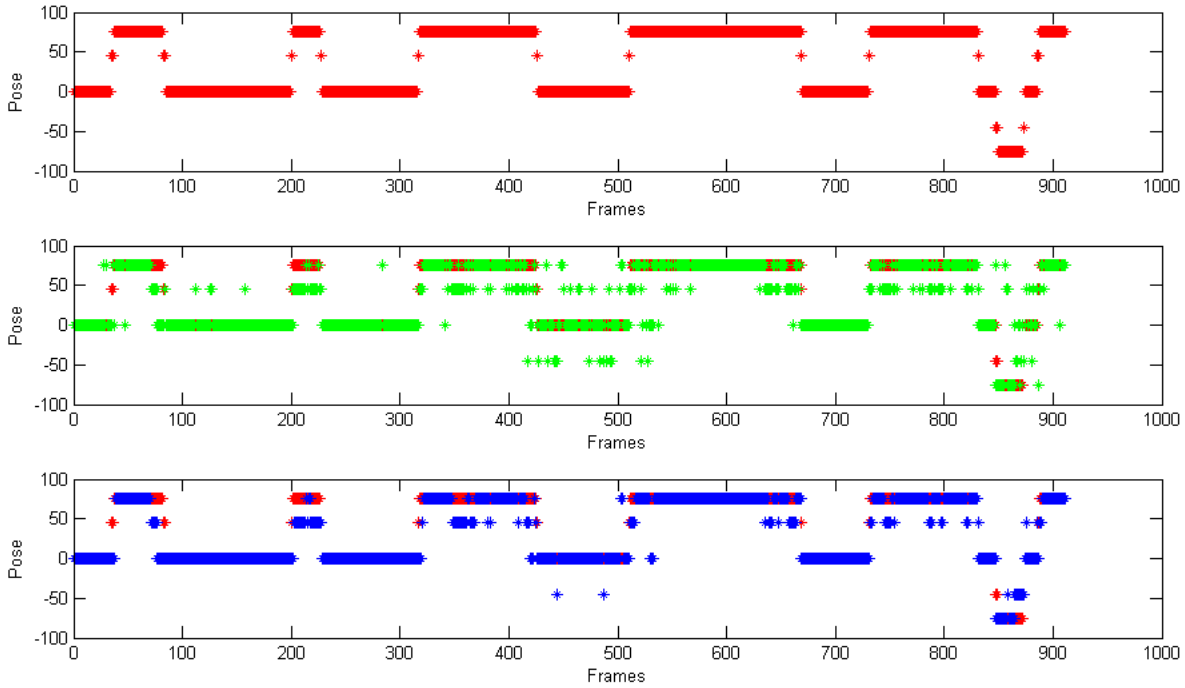

Fig. 4. Comparison between our method (with PN) and our method (with PN + HMM): the first graph shows the ground truth annotation for the given video sequence; the second graph plots the results of our method using power normalization and categorical output from SVM for classification; the third graph shows the improvements in accuracy gained by applying the HMM to our method.

TABLE II. COMPARISON BETWEEN OUR HEAD POSE ESTIMATION APPROACHES: BASELINE, OUR METHOD USING POWER NORMALIZATION (PN), OUR METHOD WITH POWER NORMALIZATION AND HMM STATEFUL MODEL AND TWO STATE OF THE ART METHODS ON EGO-HPE DATASET.

	EGO-HPE1			**EGO-HPE2**			**EGO-HPE3**			**EGO-HPE4**		
	Subject1	**Subject2**	**Mean**	**Subject3**	**Subject4**	**Mean**	**Subject3**	**Subject2**	**Mean**	**Subject3**	**Subject2**	**Mean**
Our Baseline	0.639	0.572	0.583	0.604	0.594	0.599	0.624	0.341	0.483	0.659	0.594	0.627
Our Method (with PN)	0.626	0.827	0.726	0.685	0.635	0.660	0.817	0.398	0.608	0.766	0.775	0.771
Our Method (with PN + HMM)	0.644	**0.857**	**0.750**	**0.692**	0.649	**0.670**	**0.875**	**0.461**	**0.668**	0.791	**0.850**	**0.821**
X. Zhu et al. [8]	**0.674**	0.696	0.685	0.454	**0.716**	0.585	0.206	0.423	0.315	0.755	0.786	0.771
M. Dantone et al. [25]	0.582	0.253	0.418	0.421	0.232	0.326	0.277	0.384	0.330	**0.844**	0.424	0.634

has been reduced to one and to cope with the relatively small dimension of faces in certain situations we used the trained model *face_p146_small*. Frames of a resolution double of the one we use in our method (960x540) have also been used in order to improve the detection rate, which resulted to be poor on lower resolutions. While having some detection issues on the provided images, [8] provides good performances that could be further increased using higher resolution images, though this is a situation that would significantly worsen its speed, which already is an average of 9.2 seconds per frame.

Fig. 5. A sequence of frames showing a critical situation due to strong changes in lightning conditions

As further comparison, we evaluated our method against [25]. Their method performs head pose estimation by training a regression forest over training data quantized in 5 poses as in our approach, then a continuous value is returned instead of the discrete head pose class label. Since their code does not yet perform the automatic parameter estimation of their work (landmark estimation and head localization), we used the open-source landmark detector *Flandmark* recently presented by [26]. To improve detection rate *Flandmark* has been combined with the Hough-Based Tracker; it has also been modified to infer the missing landmarks needed in the pose estimation phase. In both the landmarks and pose estimation, frames with a resolution of 960x540 have been used, slightly increasing detection performances against the resolution of 480x270 used in our approach. Both the landmark and pose estimation work in real-time on the given frames, but [25] achieves performance strictly related to the quality of the landmarks estimation. In particular, a missed detection by *Flandmark* has been treated as an error due to the impossibility of classifying the given face.

Though our method provides real-time results on both indoor and outdoor sequences, regardless of the background clutter and is robust to slight illumination changes, there are some situations in which our performance decreases, an example can be seen in the sequence of frames reported in Fig. 5: in the first part of the sequence, the lightning conditions are so extreme that extracting visual features from the image can be very difficult for both our method and the ones we compared to; in the last part of the sequence the lightning conditions change drastically and such extreme changes often involve the failure of our tracking and thus of our approach.

V. Conclusion

In this paper we have presented a real-time technique for head pose estimation that exploits tracking in order to be robust to challenging face detection situations. Hidden Markov Model technique is included in our pipeline to model the temporal coherence of sequence videos and to cope with the challenges provided by the egocentric camera view scenario. Furthermore, we showed how current state of the art methods, while providing satisfactory results on several publicly datasets, have difficulties when applied to the unconstrained ego-vision and we proposed a method capable of overcoming these issues. Finally, we released a fully annotated dataset of egocentric videos in order to encourage future research in this field.

References

[1] X. Ren and C. Gu, "Figure-ground segmentation improves handled object recognition in egocentric video," in *Proc. of CVPR*, 2010.

[2] A. Fathi and J. Rehg, "Modeling actions through state changes," in *Proc. of CVPR*, 2013.

[3] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. of CVPR*, 2012.

[4] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara, "Hand segmentation for gesture recognition in ego-vision," in *Proc. of ACMM International Workshop on Interactive Multimedia on Mobile and Portable Devices (IMMPD)*, 2013.

[5] A. Fathi, J. Hodgins, and J. Rehg, "Social interactions: A first-person perspective," in *Proc. of CVPR*, 2012.

[6] B. Ma, W. Zhang, X. C. S. Shan, and W. Gao, "Robust head pose estimation using lgbp," *Proc. of ICPR*, 2006.

[7] J. P. J. Wu, D. Putthividhya, D. Norgaard, and M. Trivedi, "A two-level pose estimation framework using majority voting of gabor wavelets and bunch graph analysis," *Proc. ICPR Workshop Visual Observation of Deictic Gestures*, 2004.

[8] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," *Proc. of CVPR*, 2012.

[9] 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, 2008.

[10] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky, "Real-time stereo tracking for head pose and gaze estimation," *Proc. of Automatic Face and Gesture Recognition*, 2000.

[11] S. Ohayon and E. Rivlin, "Robust 3d head tracking using camera pose estimation," *Proc. of ICPR*, 2006.

[12] K. Huang and M. Trivedi, "Robust real-time detection, tracking and pose estimation of faces in video streams," *Proc. of ICPR*, 2004.

[13] J. Orozco, S. Gong, and T. Xiang, "Head pose classification in crowded scenes," in *Proc. of BMVC*, 2009.

[14] M. Godec, P. M. Roth, and H. Bischof, "Hough-based tracking of non-rigid objects," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1245–1256, 2012.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] D. H. Ballard, "Readings in computer vision: Issues, problems, principles, and paradigms," 1987, ch. Generalizing the Hough Transform to Detect Arbitrary Shapes, pp. 714–725.

[17] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[18] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, p. 1, 2013.

[19] Y. Singer and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for svm," in *Proc. of ICML*, 2007.

[20] A. Lambrou, H. Papadopoulos, I. Nouretdinov, and A. Gammerman, "Reliable probability estimates based on support vector machines for large multiclass datasets," in *Artificial Intelligence Applications and Innovations*, 2012, vol. 382, pp. 182–191.

[21] D. N.Gourier and J.Crowley, "Estimating face orientation from robust detection of salient facial structures," *Proc. ICPR Workshop Visual Observation of Diectic Gestures*, 2004.

[22] J. Tu, Y. Fu, Y. Hu, and T. Huang, "Evaluation of head pose estimation for studio data," in *Multimodal Technologies for Perception of Humans*, 2007.

[23] R. Stiefelhagen, "Estimating head pose with neural networks–results on the pointing04 icpr workshop evaluation data," in *Proc. of ICPR Workshop Visual Observation of Deictic Gestures*, 2004.

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

[25] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. of CVPR*, 2012.

[26] M. Ui, V. Franc, and V. Hlav, "Facial landmarks detector learned by the structured output svm," in *Computer Vision, Imaging and Computer Graphics. Theory and Application*, ser. Communications in Computer and Information Science, G. Csurka, M. Kraus, R. Laramee, P. Richard, and J. Braz, Eds., 2013, vol. 359, pp. 383–398.