TINY HEAD POSE CLASSIFICATION BY BODILY CUES

Irtiza Hasan[⋆]*, Theodore Tsesmelis*[†]*, Fabio Galasso*[⊙]*, Alessio Del Bue*[†]*, Marco Cristani*[⋆]

University of Verona[⋆], Istituto Italiano di Tecnologia[†], Corporate Innovation OSRAM GmbH[⊙]

ABSTRACT

The head pose is an important cue for computer vision. Traditionally considered in human computer interaction applications, it becomes very hard to model in surveillance scenarios, due to the tiny head size. Additionally, no public dataset contains continuous head pose annotations in open scenery, making the challenge even harder to face. Here we present a framework based on Faster RCNN, which introduces a branch in the network architecture related to the head pose estimation. The key idea is to leverage the presence of the people body to better infer the head pose, through a joint optimization process. Additionally, we enrich the Town Center dataset with head pose labels, promoting further study on this topic. Results on this novel benchmark and ablation studies on other task-specific datasets promote our idea and confirm the importance of the body cues to contextualize the head pose estimation.

***Index Terms*—** Head pose estimation, surveillance, person detection

1. INTRODUCTION

The head pose is an important visual cue for several computer vision applications. In surveillance videos, the joint attention of people towards a direction can signal a particular event is happening [1]. In social signal processing, the head orientation is necessary to infer group formations [2] and capture social roles, such as leaders/followers [3]. Most recently, the head pose has been used for novel marketing strategies and architectural design, as a proxy to personal interest in goods, impact of adverts and space utilization [4].

The head pose estimation (HPE) problem is challenging in particular when people are captured at far and not yet addressed "in the wild". In many practical problems, such as video surveillance, HPE input is a head region as small as a 24×24 head pixel. This information alone is not enough to obtain reliable performance in HPE [5], and multi-view camera setting are necessary [6].

This paper proposes to increase HPE performance by leveraging information from the entire body of the person instead of using the head information only.

Specifically, we enrich the recent Faster RCNN [7] architecture with a branch specialized on the yaw modeling of the head pose (in this paper, we focus on yaw, keeping the modeling of pitch and roll as future goals), called Head Pose Network (HPN). The idea is to jointly optimize the pedestrian detection and the HPE tasks, in order to establish and exploit a structural connection between the appearance of the body and the head pose. Secondly, we manually label the Town Center dataset [8], which nicely portrays a surveillance scenario where 71,446 heads are imaged on 24×25 pixel patches.

The experiments, on this dataset and on standard benchmarks (oracle head detections are provided) show the net potentialities of our approach; additional ablation studies confirm that the body estimation, even if noisy, greatly improve the head pose estimation.

2. RELATED WORK

Most literature on HPE has considered high resolution images [9], which does not apply to surveillance videos. More recently, HPE from low resolution images [10, 11, 5] has emerged to address the surveillance camera viewpoint. Here several state-of-the-art works leverage SVM [11], deep neural networks [12, 13] and random forest [14]. Differently from this, we consider the joint HPE estimation and the person detection and we argue for the virtues of their joint training.

Our work further relates to literature on people detection, which can be widely grouped into integral channel features+boosting [15], deformable part model [16] and deep neural network techniques [17, 18]. Interestingly, only recently CNN techniques have achieved the state-of-the-art [19] on the Caltech benchmark [20] but this dataset has images of people that differs consistently from a video surveillance scenario as the one in the Town Center scenario.

3. HEAD POSE NETWORK (HPN)

Our goal is to automatically predict the head pose of the pedestrians in addition to their bounding boxes. To this end, we propose a new network branch called the head pose classification network (HPN) as shown in Figure 1. The network is based on Faster RCNN [7] but with novel additions and modifications to the network structure. Similarly to Fast

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 676455.

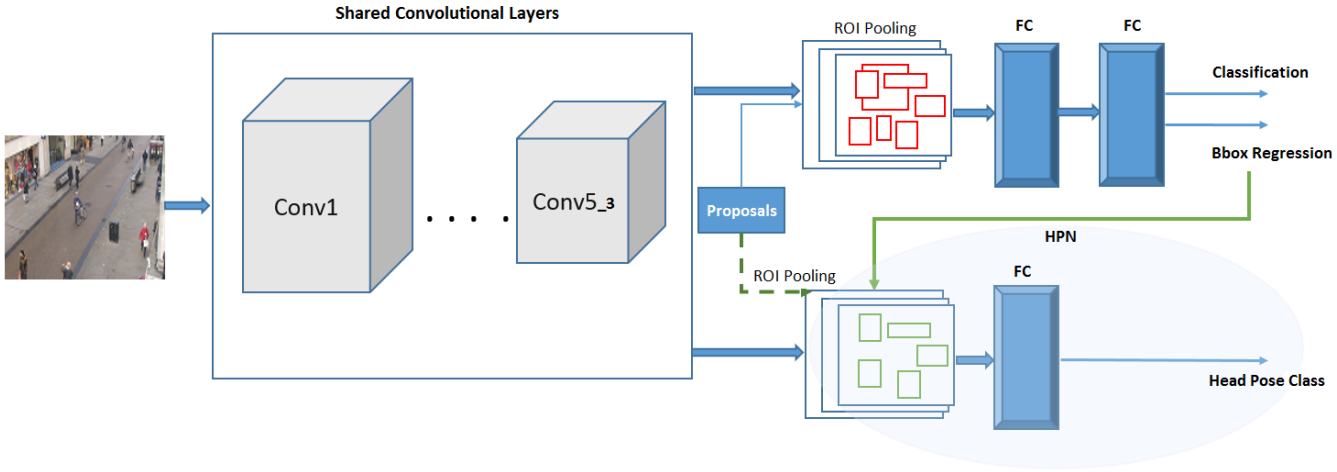

Fig. 1. Network Architecture. The figure illustrates the proposed Head Pose Classification Network (HPN). The green dotted-line represents the filtered proposals at the training time and green solid represents the pedestrian detections at testing time.

RCNN, HPN has also two modules: a fully convolutional region proposal network (RPN) that provides class-agnostic object proposals and a Fast RCNN [21] approach classifying the incoming proposals into pre-defined object classes.

In our HPN approach, we add an additional branch to the Faster RCNN network after the last shared convolutional layer (*i.e.* conv5_3), parallel to the classification and regression layers of the Faster RCNN. HPN includes also its own ROI pooling layer, a fully connected layer with sigmoid activation, and a K-way softmax layer for view-frustum classification for K discrete classes.

3.1. Training

We keep the alternative optimization approach as described in the Faster RCNN approach [7] which iteratively trains the RPN and Fast RCNN stages. Related to the RPN optimization, we keep the shared convolutional layers of Faster RCNN in their original form. Moreover, the default Fast RCNN specific layers remain unchanged. The ROI pooling layer of the original Fast RCNN takes each object proposal as input and extracts a fixed-length feature vector from the entire feature map which is then fed into a couple of fully connected layers (fcs). Our new ROI pooling layer of HPN works in the same way, except it takes only filtered region proposals at the input. This is important since we want to learn the head pose of the pedestrian proposals without being distracted by the pedestrian false-positives. To select the examples for training the HPN, we use the standard Jaccard overlap of greater than or equal to 0.5 between the ground-truth bounding boxes and the region proposals.

Adding this parallel branch (HPN) in the Fast RCNN framework essentially extends the multi-task loss of Fast RCNN to penalize the view-frustum of the person bounding box. This allows us to learn jointly both detection and head pose classification tasks. Following the same naming conventions as Fast RCNN paper, our multi-task loss for jointly training pedestrian detection and head pose is given by,

$$\begin{aligned} L(p, u, t^u, v, h, g) = & L_{cls}(p, u) \\ & + \lambda[u = 1] L_{loc}(t_u, v) \\ & + \gamma[u = 1] L_{hp}(g, h) \end{aligned} \quad (1)$$

where L_{cls} and L_{loc} are the original loss functions for background *vs* pedestrian classification and bounding-box regression respectively. We refer the reader to original paper [21] for more details on these terms. The L_{hp} term refers to the loss for the head pose of the pedestrian. We are using the softmax loss over K discrete directions of the head pose. Here g is the ground-truth label of the head pose class and $h = (h_1, h_2, ...h_K)$ is the output vector of softmax probabilities. Hence, $L_{vf} = -\log h_g$, is the negative log loss for the true view-frustum class g. As mentioned earlier, we train only for positive head pose classes and do not introduce any background class. This is given by the Iversion bracket indicator function $[u = 1]$. This means the two losses L_{loc} and L_{vf} are only used when the region proposals correspond to the pedestrian class. These losses are ignored for the background proposals. The weights λ and γ of the later two tasks are hyper-parameters which are set to 1.0 in our experiments.

3.2. Testing

At test time, our approach works in three stages. First the RPN outputs object proposals and passes them on to Fast RCNN detection network as usual. Note that this procedure

basically is the Faster RCNN framework where we keep the pedestrian detections of the Faster RCNN with the confidence score 0.5 or greater. Finally, our HPN predicts the view-frustum class for each of these incoming detections.

4. EXPERIMENTS

We first show the behavior of our technique in detecting and classifying head poses starting from raw frames. At the same time, we include ablation studies analyzing performance on head pose estimation. The latter test assumes that the head has been already detected by an oracle. The comparative approaches will be introduced later in the section.

4.1. The Town Center dataset

The Town Center dataset [8] has 4,500 frames portraying a crowded scenario with an average of 16 pedestrians per frame. The average size of the heads is about 24×25 pixels. We enrich the pedestrian bounding boxes labels by manually annotating the head direction. Towards this goal, we developed a software with a point-click interface that allows the annotator to inspect few frames of the dataset, selecting the direction where the pedestrian is looking at. From the annotation, we extract quantized head pose directions, namely, 4 and 8. We then divide the sequence into a training and a testing subsequence of length 3,000 and 1,500 frames respectively.

4.2. Head pose estimation in the wild

The protocol for evaluating the pose estimation in the wild assumes that the algorithm takes a frame as input, and provides pedestrian bounding boxes plus the head orientation, initially evaluated over 4 classes (north, east, west, south). Results are in Table 2. Figures of merits are LAMR (Log Average Miss Rate) [20] and AP (average precision) [22] for monitoring the pedestrian detection performance. It is worth noting that, in the head pose estimation accuracy, missed heads are counted as wrong detections: in this way, false negatives in the pedestrian detection flow down and impact in the final score. False positives are captured by LAMR and AP scores.

As competitors, we evaluate the Faster R-CNN [7] directly as head pose estimator in the wild, trained over pedestrian bounding boxes associated to 5 classes (4 head directions and a background class, *FR-CNN 5-class* in the table 2). This will help us in showing the added value of our HPN branch in the joint optimization, which is absent here. The poor LAMR score (78%) contrasts the rather positive AP score 0.81%. The pose estimation accuracy, based on the whole body, achieves a reasonable 66%.

The second alternative approach is composed by a recent head detector, the Face detection with Aggregate Channel Features (FACF) [23], which has shown to work pretty good on raw images, plus a head pose estimator, the Random Projected Forest (RPF) [14], which takes as input head bounding boxes, *FACF + RPF*. Both of them have been trained on the training partition of our dataset. As visible, performance is dramatically inferior, since obviously the head patches are very tiny and hard to catch without the body context.

The third approach wants to fill this gap, adding a pedestrian detection to constraint the head detector to work on pedestrian bounding boxes. In this case, we consider the Local Decorrelation Channel Features detector (LDCF) [24], giving rise to the *LDCF + FACF + RPF* pipeline. Results on Table 2 show that performances are higher, but still inferior than FR-CNN 5-class.

We further question the importance of face detection by testing *LDCF + HRCNN + RPF*, where a CNN-based head detector (HRCNN [25]) replaces the FACF. Reasonably, HRNN improves the head detection considerably, 10% LAMR and 12% AP (cf. Table 2), resulting in a better but still poor head pose estimation score of 50%. We conclude from this that the face, when so tiny, is not sufficient to estimate the pose estimation alone.

We mark as "ours" in the table the combination of pedestrian detection and pose estimation, jointly optimized within our model, cf. Eq. 1. As seen from Table 2, in the Town Center dataset a Faster-R-CNN person detector performs on par with the person specific LDCF [24]. More interestingly, using the whole body for the estimation of pose greatly improves performance by 18%, resulting in the best technique, HPN, which we propose. This resonates with the baseline Faster-R-CNN 5-classes in the first row, also based on the whole body.

4.3. Ablation study: head pose classification

The ablation studies serve to evaluate how our approach works in the case of a correct person detection. To enrich the analysis, in Table 3 we consider different numbers of pose quantization, namely 4 and 8 classes, in which the quantization has been obtained by uniformly dividing 360 degrees. As competitors, we consider *RPF* [14], the FR-CNN N-class (N refers to the quantization bins), and 2 different versions of our approach. The variation we want to analyze (***Ours disjoint optimization***) does the following thing: as in the proposed version, the complete body is used for head pose estimation but the optimization terms for object classification L_{cls} and bounding box regression L_{loc} are set to zero. In practice, this breaks up the joint optimization and let the system operate as two separate modules, where the object detection loss does not contribute to the head pose classification training.

Ours Joint Model is the proposed methodology where as explained above pedestrian detection module and the head pose estimator are jointly optimized. Table 3 illustrates the robustness of our approach in regards to the granularity level of head pose. Secondly, pedestrian detection and head pose estimation are related task, therefore when posed as a joint

Table 1. Comparison of head pose classification accuracy in regard to image scale variation.

Methods	Dataset	*HIIT*			*QMUL*			*QMULB*		
	Image Size	15x15	20x20	50x50	15x15	20x20	50x50	15x15	20x20	50x50
Frobenius [5]		82.4	89.6	95.3	59.5	82.6	94.3	54.5	76.5	92
CBH [5]		84.6	90.4	95.7	59.8	83.2	94.9	57	76.9	92.2
RPF [14]		97.6	97.6	97.6	94.1	94.3	94.3	91.9	92.1	92.2
PSMAT [11]		-			-		82.3	-		64.2
ARCO [10]		-			-		93.5	-		89
HPN		**98.4**	**98.9**	**99.01**	**97.4**	**97.9**	**98**	**95.3**	**95.9**	**94.7**

optimization problem performance for head pose estimation gets boosted.

In Table 1 the second ablation study stresses the ability of our approach in estimating the head poses by starting from correct head bounding boxes. For this purpose we train HPN over head images and pose it as a classification problem. Except for the QMULB [11] dataset, which has an additional background class, in that case we train HPN to have a cascaded output, first distinguish between person and non-person and then classifying only persons for the head poses. This procedure is consistent to our proposed joint model. Results have been computed on the datasets HIIT [5], QMUL [11] and its extension with background class QMULB [11]. *HIIT* dataset has 24,000 images with 6 head poses and a static background. *QMUL* dataset contains 15,660 images that has 4 different head poses with varying illumination and occlusion. *QMUL* dataset with additional 3,099 background images is referred to as *QMULB*.

As visible, our approach is capable of overcoming, in terms of average accuracy, all of the competitors at each resolution.

Table 2. Head pose estimation in the wild. For LAMR, lower is better.

Pipeline	**Pedestrian Detection**		**Head Detection**		**Head Pose Est. Accuracy**
	LAMR	AP	LAMR	AP	
FR-CNN [7] 5-class	78	0.81	N/A	N/A	0.66
FACF [26] + RPF [14]	N/A	N/A	90.67	0.336	0.3
LDCF [24] + FACF [26] + RPF [14]	54.99	0.83	96.37	0.2087	0.27
LDCF [24] + HRCNN [25] + RPF [14]	54.99	0.83	84.36	0.31	0.5
Ours	**55**	**0.86**	N/A	N/A	**0.68**

Table 3. Head pose classification accuracy on oracle.

Method	**Classification Accuracy (4 classes)**	**Classification Accuracy (8 classes)**
RPF [14]	0.6	0.31
FR-CNN N-class [7]	0.71	0.32
Ours (Disjoint Optimization)	0.72	-
Ours (Joint Model)	0.74	0.33

5. CONCLUSIONS

We proposed a CNN pipeline that copes simultaneously with pedestrian detection and head pose estimation, in surveillance scenarios. We demonstrated that the joint model performs competitively with the state-of-the-art, beating up-to-date serial pipelines composed by pedestrian detectors, head detectors and head pose estimators. At the same time, we confirmed that the body information is an important cue to increase performance of head pose estimation, especially when the head patch size is small.

6. REFERENCES

[1] Shaogang Gong, Tao Xiang, and Somboon Hongeng, "Learning human pose in crowd," in *Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*, New York, NY, USA, 2010, MPVA '10, pp. 47–52, ACM.

[2] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino, "Social interaction discovery by statistical analysis of f-formations.," in *BMVC*, 2011, vol. 2, p. 4.

[3] Basil G Englis, "The role of affect in political advertising: Voter emotional responses to the nonverbal be-

havior of politicians," *Attention, attitude, and affect in response to advertising*, pp. 223–247, 1994.

[4] C. Djeraba, A. Lablack, and Y. Benabbas, *Multi-Modal User Interactions in Controlled Environments*, Multimedia Systems and Applications. Springer US, 2010.

[5] Diego Tosato, Mauro Spera, Marco Cristani, and Vittorio Murino, "Characterizing humans on riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1972–1984, Aug. 2013.

[6] Anoop Kolar Rajagopal, Ramanathan Subramanian, Elisa Ricci, Radu L Vieriu, Oswald Lanz, Nicu Sebe, et al., "Exploring transfer learning approaches for head pose classification from multi-view surveillance images," *International journal of computer vision*, vol. 109, no. 1-2, pp. 146–167, 2014.

[7] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[8] Ben Benfold and Ian Reid, "Guiding visual surveillance by tracking human attention," in *Proceedings of the 20th British Machine Vision Conference*, September 2009.

[9] Erik Murphy-Chutorian and Mohan Manubhai 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.

[10] Diego Tosato, Michela Farenzena, Mauro Spera, Vittorio Murino, and Marco Cristani, "Multi-class classification on riemannian manifolds for video surveillance," in *European conference on computer vision*. Springer, 2010, pp. 378–391.

[11] Javier Orozco, Shaogang Gong, and Tao Xiang, "Head pose classification in crowded scenes.," in *BMVC*, 2009, vol. 1, p. 3.

[12] Davide Conigliaro, Paolo Rota, Francesco Setti, Chiara Bassetti, Nicola Conci, Nicu Sebe, and Marco Cristani, "The s-hock dataset: Analyzing crowds at the stadium," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2039–2047.

[13] CAI Ying, Meng-long Yang, and LI Jun, "Multiclass classification based on a deep convolutional," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, no. 11, pp. 930–939, 2015.

[14] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh, "Fast and accurate head pose estimation via random projection forests," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1958–1966.

[15] Piotr Dollar, Ron Appel, Serge Belongie, and Pietro Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[16] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester, "Cascade object detection with deformable part models," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 2241–2248.

[17] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan, "Scale-aware fast r-cnn for pedestrian detection," *arXiv preprint arXiv:1510.08160*, 2015.

[18] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning strong parts for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1904–1912.

[19] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He, "Is faster r-cnn doing well for pedestrian detection?," in *European Conference on Computer Vision*. Springer, 2016, pp. 443–457.

[20] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 304–311.

[21] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[23] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li, "Aggregate channel features for multi-view face detection," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*. IEEE, 2014, pp. 1–8.

[24] Woonhyun Nam, Piotr Dollár, and Joon Hee Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[25] Huaizu Jiang and Erik G. Learned-Miller, "Face detection with the faster R-CNN," *CoRR*, vol. abs/1606.03473, 2016.

[26] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li, "Aggregate channel features for multi-view face detection," *CoRR*, vol. abs/1407.4023, 2014.