Instantaneous real-time head pose at a distance

CONFERENCE PAPER · SEPTEMBER 2015

READS
6

3 AUTHORS, INCLUDING:

Rolf Baxter
Heriot-Watt University
12 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Neil M. Robertson
Heriot-Watt University
55 PUBLICATIONS **257** CITATIONS
SEE PROFILE

INSTANTANEOUS REAL-TIME HEAD POSE AT A DISTANCE

Sankha S. Mukherjee, Rolf H. Baxter, and Neil M. Robertson

Visionlab, Heriot-Watt University, Edinburgh, UK
visionlab.eps.hw.ac.uk
{sm794, r.h.baxter, n.m.robertson}@hw.ac.uk

ABSTRACT

In this paper we focus on robust, real-time human head pose estimation in low resolution RGB data without any smoothing motion priors e.g. direction of motion. Our main contributions lie in three major areas. First, we show that a generative Deep Belief Network model can be learned on human head data from multiple types of data sources. These sources have similar underlying data that are not necessarily labelled or have the same kind of ground truth. Second, we perform discriminative training using multiple disparate supervisory labels to fine tune the model for head pose estimation. Third, we present state-of-the-art results on two publicly available datasets using this new approach. Our implemetation computes head pose for a head image in 0.8 milliseconds, making it real-time and highly scalable.

***Index Terms*—** Head Pose, Gaze, Surveillance , Deep Belief Network, Deep Learning, Unsupervised Learning

1. INTRODUCTION

Automatic gazing direction estimation has become an important feature for applications of computer vision to surveillance and human behaviour inference [2]. Human head pose is the most important factor in determining focus of attention [3] and provides important information for group detection, gesture, interaction detection, and scene understanding [4].

There remains a significant gap in the current methods for unconstrained head-pose estimation in low resolution. This work addresses the need for computing low-resolution gaze estimators without reliance on motion priors to smooth the estimate and presents a demonstrably more robust method using deep learning. In summary, the main scientific contributions of this paper are:
(a) Learning a generative human head model in an abstract head space that can reconstruct heads from low resolution, noisy inputs; (b) Discriminating between head pose angles from the input image without other prior information using multi label discriminative training using various loss functions; (c) We report state-of-the art results on two publicly available datasets when compared to the state-of-the-art approaches. Figure 1 illustrates the output of our system on a typical surveillance dataset.

1.1. Related work

In visual surveillance the resolution of detected heads can be very small so head-pose is often estimated in coarse discrete directional bins of the azimuthal angle [5]. See for example the eight classification bins used in this paper in Figure 1. Walking direction is often used as priors [6, 7, 8]. which reduces mean squared error, but also attenuates the pure information content of the head pose signal. As shown in Fig 2, an analysis of gazing behaviour in several datasets demonstrates that most people look where they are going. However, the cases that are of more interest are when people deviate from this behaviour (i.e. look somewhere else), as this information could be useful for anomaly detection or improving tracking [9].

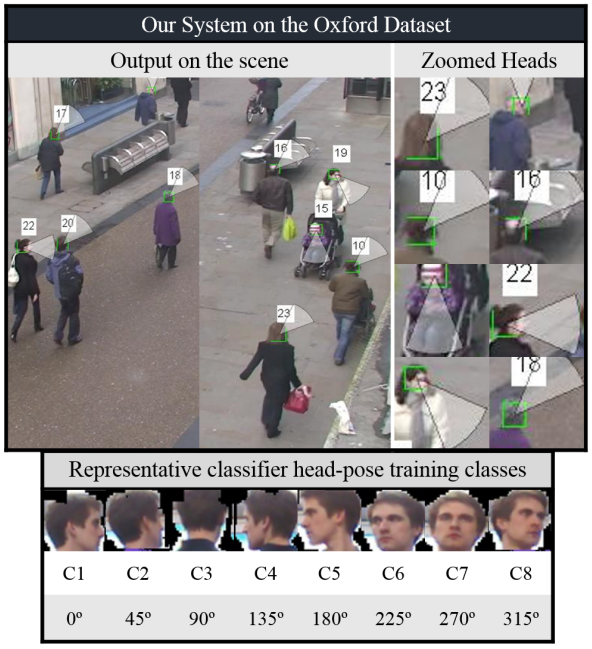

Fig. 1. Example output of our system showing Head pose estimation in the Oxford town center dataset[1]. Individual heads have been zoomed in. The row on the bottom shows the Head pose classes used for classification.

To obtain an unbiased classifier we estimate head pose from the image alone by learning to represent human heads in an unsupervised fashion. Blanz et al. [10] use a generative morphable 3D model of human faces in an abstract face-space that can generate human faces with different shapes, colours and expressions. We learn a representation that is valid for human heads under different poses and is invariant to expressions, occlusions, hair, hats, and glasses. The power of a generative model, as shown by Tang et al. in [11], lies in being able to reconstruct original images under noise or heavy occlusions and so we use Deep Belief Networks [12]. These have been successfully applied to image and voice recognition [13]. Convolutional Neural Networks [14] have mostly replaced the DBNs in terms of accuracy in large labelled datasets like the imagenet, they are completely supervised so cannot learn from unlabelled data as

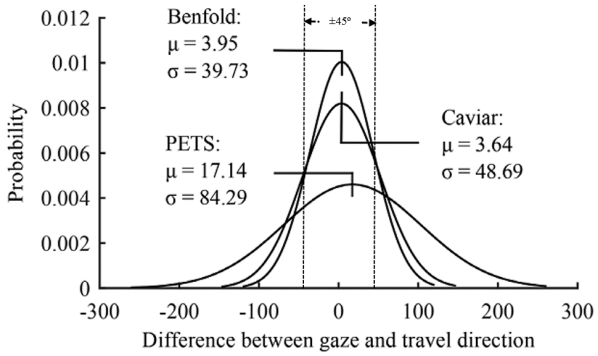

Fig. 2. Head pose deviation from walking direction as a Probability Density Function in various datasets [9]

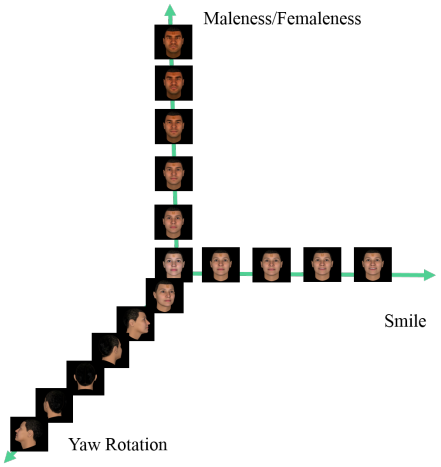

Fig. 3. Conceptual diagram showing different parameters controlling the appearance of the head

we do here.

The pioneering work on low resolution head pose estimation by Robertson and Reid [5] used a detector based on template training to classify head poses in 8 directional bins. This technique was extended to allow colour invariance by Benfold et al. [7], who proposed a randomized fern classifier for hair face segmentation before template matching. A few non-linear regression approaches such as Artificial Neural Networks [15, 16] and High-dimensional manifold based approaches [17, 18] try to estimate the head poses in a continuous range. These techniques however are more suited to high resolution human computer interaction cases where the head is more or less constrained to near frontal poses. Chen and Odobez [8] proposed the state-of-the-art method for unconstrained coupled head pose and body pose estimation in low resolution surveillance videos. They used multi-level HOG for the head and body pose features and extracted a feature vector for adaptive classification using high dimensional kernel space methods. Coupling of head pose with such priors results in a head pose signal that is less informative: these techniques perform very well in the range indicated in Figure 2, but perform poorly when the head pose is not aligned to the priors. We stress this point because it is important for the head pose estimation to provide robust information that can be further exploited (e.g. improving tracking, anomaly detection, group detection, behaviour analysis) and achieving this goal is what this paper demonstrates.

We discuss the theory behind generative and discriminative modelling and how we exploit both in order to solve the problem. We then discuss the results on two publicly available datasets and show the benefits of our approach compared to prior work.

2. THEORY AND METHODOLOGY

In this paper we do not concern ourselves with the problem of detecting heads. Instead we can adapt the output of any head detector and normalize the heads to a resolution of 32×32 as input to our algorithm. Figure 4 shows the overall architecture of the DBN used.

2.1. Parametric Human Head Space

The underlying motivation of this work comes from the theory that human heads lie in a parametric space. This was first shown to be working by Blanz et al. [10] where they derived the basis of this space by a linear combination of shape and texture information of high resolution 3d head scans of 200 adult faces. Hence by using 400 shape and texture parameters they derived a morphable model that could be used to synthesize new faces or estimate a model from 2D images of a given face. However, the human head space is much more complicated because aside from low resolution, surveillance data contains other complicating factors such as varying hair styles, facial hair, and occlusions (e.g. hats, glasses). This requires a much larger parameter space. However, for headpose, which is a very big factor in appearance (and hence has a big eigenvalue in the pca subspace), fewer parameters are needed. Figure 3 shows how a parametric head-space can generate various human heads with different identity, expression and pose. The head pose datasets are limited in the number of examples per person and image quality. Hence, we consolidated many different datasets not necessarily ground truthed for head pose into an unsupervised framework in a generative model. Deep Belief Networks [12] are very well suited for this purpose.

2.2. Deep Belief Networks (DBN)

A DBN is constructed from unsupervised, greedily trained stacks of restricted boltzmann machines (RBMs). RBMs are a form of energy based generative model in which the energy functions can be written as follows:

$$E(v,h) = -b'v - c'h - h'Wv \tag{1}$$

Where b,c and W are the parameters θ and v and h are the visible and hidden units of the model. The model is trained with contrastive divergence that estimates the gradients of the energy function with respect to the model parameters given the training data $\mathbf{X}$.

$$\frac{\partial E(\mathbf{X},\theta)}{\partial \theta} = \frac{\partial \log \mathbf{Z}(\theta)}{\partial \theta} - \left\langle \frac{\partial \log f(x,\theta)}{\partial \theta} \right\rangle \tag{2}$$

where $\mathbf{Z}(\theta)$ is the partition function defined as

$$\mathbf{Z}(\theta) = \int f(x,\theta)\, dx \tag{3}$$

Where $f(x,\theta)$ is the underlying distribution we are trying to model. It is not easy to find the derivative of the partition function because we do not know the underlying representation. It can be

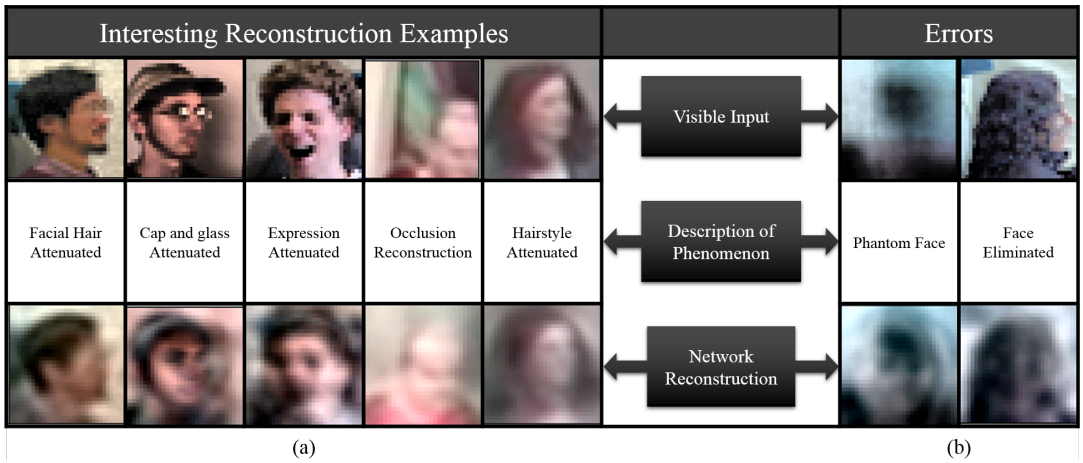

Fig. 5. What the network sees. This figure shows a reconstruction of the input image in top row with their reconstruction from network parameters in the last layer in the bottom row. Sub-figure (a) on the left suggests that for head pose the eye and mouth region is very important whereas facial hair, hairstyle and facial expressions are attenuated. The network has learned to handle occlusions and shift. Sub-figure (b) on the right shows some interesting errors made by the network. Under extreme low resolution or noisy input on the left the network sees a face where none exist. On the right the middle the face is eliminated. However even in these extreme low resolution cases the network can estimate parameters.

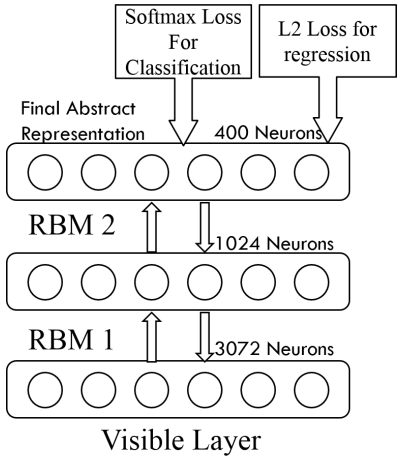

Fig. 4. Here we show the hierarchical DBN architecture which creates a 400 dimensional head representation that is then discriminativly trained on various datasets based with varying ground truths ranging from basic front/back classification to full real valued angle regression with interchangeable softmax and L2 loss functions

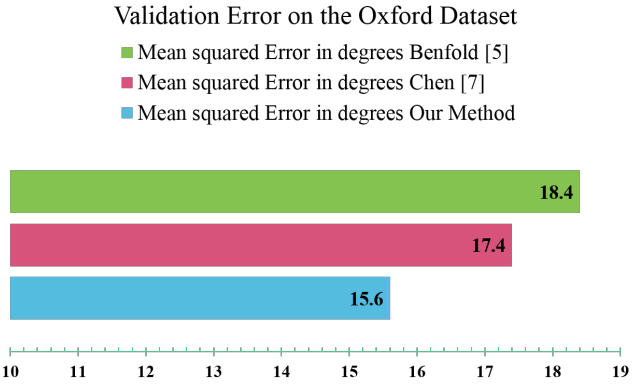

Fig. 6. This graph compares our algorithm in terms of MSE with the Benfold [6] and Cheng algorithm [8]

suitably derived by using Markov Chain Monte Carlo sampling from the training data and given sufficient examples it should converge to the real derivative, however, this is not computationally tractable. The parameter update equation derived from just one step of Markov Chain Monte Carlo sampling from the training data has empirically proven to be effective by Hinton et al [12]. It can be written as:

$$\theta_{t+1} = \theta_{t+1} + \eta \left(\left\langle \frac{\partial \log f(x,\theta)}{\partial \theta} \right\rangle_{\mathbf{x}^0} - \left\langle \frac{\partial \log f(x,\theta)}{\partial \theta} \right\rangle_{\mathbf{x}^1} \right) \quad (4)$$

Where η is the training rate. The layers of RBMs are trained in an unsupervised fashion layer by layer to form a Deep Belief Network. Conceptually, by changing the number of neurons in each subsequent hidden layer the representation of the underlying data can be learned in a hierarchical fashion. Figure 4 shows the architecture used for our system. We use only two layer because more layers degraded performance with the amount of data present for training.

3. EXPERIMENTS AND VALIDATION

We use multiple datasets to train our system and we validate our approach on two publicly available datasets as discussed below.

3.1. Datasets

To maximise the training corpus, we gathered data from multiple sources that had similar underlying distributions. Datasets annotated for unconstrained face recognition, facial landmark detection,

Fig. 7. Confusion matrices showing the output of (a) The Benfold algorithm [7] on the Oxford town center dataset, (b) Our DBN approach on the Oxford town center dataset, (c) Our DBN output on the Caviar dataset.

expression detection all have facial data under various poses. The different head pose datasets that we used are the Oxford town center dataset, the RGB data from Biwi Kinect headpose dataset [19], the Caviar shopping center dataset, the IIT Head Orientation dataset along with the IDIAP headpose dataset [20]. Furthermore our own dataset captured 46 people (32 males, 14 females) freely moving infront of a camera with a miniature wireless IMU sensor for head pose ground truth. Each person covered all possible head pose angles in a continuous manifold at a distance varying from 2m-8m from the camera. We gathered approximately 1500 frames per person giving a total of 68126 examples. It should be highlighted that the different datasets have different annotations; some of them have real valued ground truths, others have 6-8 classes spanning the 360°. The datasets vary in resolution from very high in the BIWI datset to very low in the Oxford town centre dataset. Furthermore, for regularisation of the network in the unsupervised phase we included the Multi-task Facial Landmark Dataset (MTFL) [21] and the Labelled Faces in the Wild [22] datasets as they have a wide range of poses, but these are not labelled for head pose.

3.2. Training

The network used two RBMs stacked to form a DBN as shown in Figure 4. The final output layer was interchanged for various headpose datasets depending on their ground truth. We normalized all the head images to 32×32 for input to the network. We also scaled the head bounding box to 0.8, 1, 1.5, 1.8, 2.0, and 2.5 scaled crops to achieve some scale invariance. To achieve translation invariance we also used scale 1 crops with strides of (3,3) pixels from the 1.5-2.5 scaled crops. The network was trained with 30% dropout and a decaying learning rate. For validation on the Caviar and the Oxford datasets we use a training-testing split of 70%-30%. Figure 5 shows the reconstructions of the image from the parameters estimated by the networks top most layer by back projection into the image space. This gives us a unique perspective into what the network actually found important for the problem feature selection.

4. RESULTS

We report our results on the Oxford and the Caviar datasets. In these datasets we classify the head pose into 8 equally spaced (45°) angular bins as shown in Figure 1. For comparison with [8] and Benfold [6] we use the Oxford dataset in which both have reported results. One consideration has to be made while comparing because [8] reported the mean square error (MSE) which they derived from a weighted combination of their 8 class classifier output multiplied with the bin angles as $\sum_{i=1}^{8} p_i \overrightarrow{\eta_{\theta_i}}$ where p_i is the classifier output value for the class i and $\overrightarrow{\eta_{\theta_i}}$ is the unit vector in that angular direction. Since our softmax layer gives probability, it is unclear how to interpret vector addition weighted by probability. But for the sake of comparison we derive our mean squared error (MSE) in the same way. Figure 6 shows the comparison between our method with the previous state of the art results. In terms of MSE we outperform the best results by 1.8°. The margin while comprehensive may not be representative of the true picture. We therefore present the confusion matrices on the Oxford and Caviar datasets. In terms of classification accuracy on the Caviar dataset we achieve 76.38% accuracy on the Caviar dataset. To our knowledge it is the best result on this dataset.

On the Oxford dataset, for comparison, we also show the output confusion matrix of the Benfold algorithm [7] along with our confusion matrix. Apart from the fact that we outperform the Benfold algorithm by a large margin, it is interesting to note that the Benfold algorithm shows some interesting biases connected to walking direction. The confusion matrix shows a large classifier bias in the C2 and C6 pose classes, which, as can be seen from Figure 1, coincides with the direction of the road. As most people are going up or down the road and generally looking where they are going (as can be seen from Figure 2) the algorithm seems to have learned this bias in the scene.

We out perform both the previous state of the art methods without using any kind of prior coupling as the confusion matrices in Figure 7 show very clearly. For completeness we show the MSE in Figure 6, as this metric is used in the papers against which we compare. The difference in MSE is not as dramatic as the confusion matrices shown suggests but nevertheless demonstrates a significant improvement. One feedforward pass through our DBN on a GPU for headpose estimation on a single 32×32 image takes 0.8 milliseconds. This makes our system real-time and it can be scale up massively but still maintain real-time performance.

5. CONCLUSION

In this paper we presented a data-driven semi-supervised approach to low resolution head pose estimation in the wild. We achieved state-of-the-art results on two publicly available datasets. The model fine tuned on head pose was able to select features that are invariant to occlusion and expression. In future we will consider making our model deeper and use convolution and pooling filters in the first few stages to improve spatial invariance and reduce the overall number of parameters in the net.

6. REFERENCES

[1] Ben Benfold and Ian Reid, "Stable multi-target tracking in real-time surveillance video," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp. 3457–3464.

[2] Benno Gesierich, Angela Bruzzo, Giovanni Ottoboni, and Livio Finos, "Human gaze behaviour during action execution and observation," *Acta Psychologica*, vol. 128, no. 2, pp. 324 – 330, 2008.

[3] StephenR.H. Langton, Helen Honeyman, and Emma Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction," *Perception & Psychophysics*, vol. 66, no. 5, pp. 752–771, 2004.

[4] John M. Henderson and Andrew Hollingworth, "High-level scene perception," *Annual Review of Psychology*, vol. 50, no. 1, pp. 243–271, 1999, PMID: 10074679.

[5] N.M. Robertson and I.D. Reid, "Estimating gaze direction from low-resolution faces in video," in *Proceeding of the 9th European Conference on Computer Vision, 2006*, 2006, vol. 3952/2006, pp. 402–415.

[6] Ben Benfold and Ian Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2344–2351.

[7] Ben Benfold and Ian Reid, "Colour invariant head pose classification in low resolution video," in *Proceeding of the British Machine Vision Conference*, 2008.

[8] Chen Cheng and J. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 1554–1551.

[9] R.H. Baxter, M.J.V. Leach, S.S. Mukherjee, and N.M. Robertson, "An adaptive motion model for person tracking with instantaneous head-pose features," *Signal Processing Letters, IEEE*, vol. 22, no. 5, pp. 578–582, May 2015.

[10] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1999, SIGGRAPH '99, pp. 187–194, ACM Press/Addison-Wesley Publishing Co.

[11] Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton, "Robust boltzmann machines for recognition and denoising," in *IEEE Conference on Computer Vision and Pattern Recognition, 2012, Providence, Rhode Island, USA*, 2012.

[12] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[13] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *ArXiv e-prints*, Sept. 2014.

[15] N. Gourier, J. Maisonnasse, D. Hall, and J.L. Crowley, "Head pose estimation on low resolution images," in *Proceeding of the 1st International Evaluation Conference on Classification of Events, Activities and Relationships*, 2006, pp. 270–280.

[16] R. Stiefelhagen, "Estimating head pose with neural network-results on the pointing04 icpr workshop evaluation data," in *Proceedings of the ICPR Workshop on Visual Observation of Deictic Gestures*, 2004.

[17] V. Balasubramanian, J. Ye, and S. Panchanathan, "Biased manifold embedding: a framework for person-independent head pose estimation," in *Proceeding of the IEEE Conference on Computer Vision and PatternRecognition*, 2007, pp. 1–7.

[18] C. BenAbdelkader, "Robust head pose estimation using supervised manifold learning," in *Proceeding of the 11th European Conference on Computer Vision*, 2010, pp. 518–531.

[19] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool, "Random forests for real time 3d face analysis," *Int. J. Comput. Vision*, vol. 101, no. 3, pp. 437–458, February 2013.

[20] D. Tosato, M. Spera, M. Cristani, and V. Murino, "Characterizing humans on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1972–1984, 2013.

[21] Zhanpeng Zhang, Ping Luo, ChenChange Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision - ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., vol. 8694 of *Lecture Notes in Computer Science*, pp. 94–108. Springer International Publishing, 2014.

[22] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.