Head Pose Classification in Crowded Scenes

Javier Orozco
http://www.dcs.qmul.ac.uk/~orozco
Shaogang Gong
http://www.dcs.qmul.ac.uk/~sgg
Tao Xiang
http://www.dcs.qmul.ac.uk/~txiang

Queen Mary Vision Laboratory
School of Electronic Engineering and Computer Science
Queen Mary University of London
London E1 4NS, UK

Human head pose and gaze directions have traditionally been studied for expression and face recognition, and human computer interaction [6]. Accurate estimations of either head or gaze direction can provide useful information for the inference of a person's intent and behaviour. However, most existing techniques rely upon medium to high resolution images captured under well controlled conditions from a fairly close distance [3, 5, 7, 10]. Given high resolution images, most existing techniques deploy extensive feature extraction to capture detailed head/facial shape and texture information. However, this approach relies on accurate subtraction of head foreground region from the background which is not always feasible.

We propose a novel technique for head pose classification in crowded public space under poor lighting and in low-resolution video images. Unlike previous approaches, we avoid the need for explicit segmentation of skin and hair regions from a head image and implicitly encode spatial information using a grid map for more robustness given low-resolution images. Specifically, a new head pose descriptor is formulated using similarity distance maps by indexing each pixel of a head image to the mean appearance templates of head images at different poses. These distance feature maps are then used to train a multi-class Support Vector Machine for pose classification. Our approach is evaluated against established techniques [2, 8, 9] using the i-LIDS underground scene dataset under challenging lighting and viewing conditions. As in [8], a 360° head pose in panning angle is discretized into eight pose classes with 45° increment. The results demonstrate that our model gives significant improvement in head pose estimation accuracy, with over 80% pose recognition rate against 32% from the best of existing models.

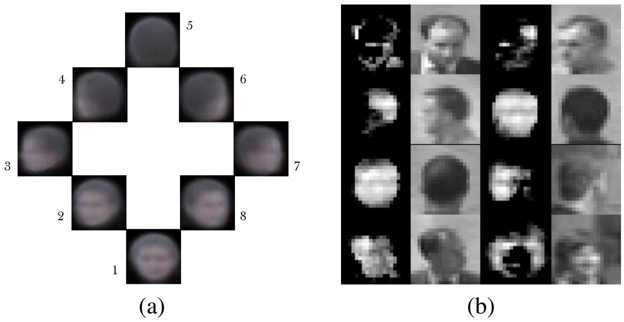

Figure 1: (a) The average (mean) head appearance templates of 360° head panning rotations quantised into eight discrete pose classes. (b) Examples of feature maps obtained by comparing input images and head templates.

Training and testing sets are collected from the i-LIDS dataset, by cropping and labelling heads into eight discrete pose classes $k45°$ where $k = 1,\ldots,8$. Different from [1, 8], neither skin nor hair pixels are distinguished. Low-resolution head images are assumed as single Gaussians in order to be able to compute a shape-free mean $\mathbf{M}^c$, per pose class (see Fig. 1 (a)). Subsequently, a given input image $\mathbf{N}$, is profiled by exhaustive comparison with all mean templates $\mathbf{M}^c$, where $c = 1,\ldots,8$. To that end, we measure the coefficient *Kullback Leibler divergence* (*KL*) between the input image $\mathbf{N}$ and each pose class in a pixel-wise fashion. The standard *KL* is defined as $D_{KL}(p||q) = p\left(\log\frac{p}{q}\right)$ which measures the divergence between two $p.d.f.$ What we measure is subtly different which we refer to as *KL* coefficients (δ_{KL}), computed as follows:

$$\delta_{KL}\left(m^c_{i,j}||n_{i,j}\right) = \max_{RGB}\left\{m^c_{i,j}\left(\log\frac{m^c_{i,j}}{n_{i,j}}\right)\right\} \qquad (1)$$

where $n_{i,j}$ and $m^c_{i,j}$ are pixel intensity values in the same RGB colour channel. Consequently, a similarity distance weighting map is constructed as a feature descriptor (2D matrix) containing the maximum divergence coefficients $\delta_{KL}(m^c_{i,j}||n_{i,j})$, from all possible pose templates and RGB channels at each pixel location:

$$x_{i,j} = \max_c\left\{\delta_{KL}(m^c_{i,j}||n_{i,j})\right\},\ \text{and } c = 1,\ldots,8 \qquad (2)$$

We impose an additional constraint so that $\delta_{KL}(m^c_{i,j}||n_{i,j}) = 0$ when $n_{i,j} \geq m^c_{i,j}$. This effectively removes those divergent pixels deem to be background.

In order to classify any input image by eight discrete head poses, we apply a Multi-class Support Vector Machine (SVM). We build a model where the $i-th$ SVM constructs a hyperplane between the class i-th and the $C-1$ remaining classes. Pose classification is determined by a majority vote among all eight classifiers. More specifically, we adopt a *one-against-rest* SVM strategy [4] using a polynomial kernel with the objective of finding a hyperplane capable of separating one pose class from the rest.

We demonstrate significant performance advantages of our proposed model compared to a state-of-the-art model and another established technique for head pose classification under challenging viewing conditions in crowded public space given by the UK Home Office i-LIDS dataset (see Fig. 2).

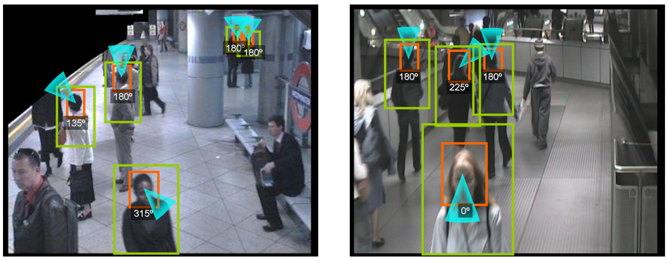

Figure 2: Examples of automated head poses classification of unknown multiple heads in two crowded underground stations. Head candidates are located automatically and head poses are estimated using the proposed method (red dial with pose value).

[1] B. Benfold and I. Reid. Colour invariant head pose classification in low resolution video. In *BMVC*, 2008.

[2] D. Beymer. Face recognition under varying pose. In *CVPR*, pages 756–761, 1994.

[3] L. Brown and Y. Tian. Comparative study of coarse head pose estimation. In *MOTION*, 2002.

[4] R. Debnath, N. Takahide, and H. Takahashi. A decision based one-against-one method for multi-class support vector machine. *PAA*, 7 (2):164–175, 2004.

[5] S. Gong, S. Mckenna, and J. Collins. An investigation into face pose distributions. In *FG*, 1996.

[6] E. Murphy-Chutorian and M.M. Trivedi. Head pose estimation in computer vision: A survey. *PAMI*, 31(4), 2009.

[7] S. Niyogi and W. Freeman. Example-based head tracking. In *FG*, 1996.

[8] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, pages 402–415, 2006.

[9] J. Sherrah, S. Gong, and E. Ong. Face distributions in similarity space under varying head pose. *IVC*, 19(12):807–819, 2001.

[10] H. Shimizu and T. Poggio. Direction estimation of pedestrian from multiple still images. In *IEEE Intelligent Vehicles Symposium*, 2004.