SPATIAL AND PROBABILISTIC CODEBOOK TEMPLATE BASED HEAD POSE ESTIMATION FROM UNCONSTRAINED ENVIRONMENTS

Meltem Demirkus, Boris Oreshkin, James J. Clark, Tal Arbel

Department of Electrical and Computer Engineering, McGill University

ABSTRACT

In unconstrained environments, head pose detection can be very challenging due to the joint and arbitrary occurrence of facial expressions, background clutter, partial occlusions and illumination conditions. Despite the wide range of head pose literature, most current methods can address this problem only up to a certain degree, and mostly for restricted scenarios. In this paper, we address the problem of head pose classification from real world images with large appearance variation. We represent each pose with a probabilistic and spatial template learned from facial codewords. The inference of the best template representing a test image is achieved probabilistically and spatially at the codebook. The experimental results are obtained from 5500 video frames collected under different illumination and background conditions. Our probabilistic framework is shown to outperform the current state-of-the-art in head pose classification.

Index Terms— Head pose, unconstrained environment, uncontrolled environment, codebook, local invariant feature.

1. INTRODUCTION

In recent years, the range of applications for video surveillance technology has increased substantially. Considering the massive size of surveillance data collected from unconstrained environments, fast and accurate automatic systems are in high demand. In this paper, we consider the particular problem of automatically inferring head pose from face images from video surveillance data collected in unconstrained environments. Developing an automatic head pose classification system for video surveillance data is not a trivial task, given the challenges of unconstrained conditions present in the real world, which include: arbitrary face scale, nonuniform illumination conditions, arbitrary partial occlusions and background clutter as well as a wide variability in possible face image quality (see Figure 1 and Figure 5).

Currently head pose estimation methods from 2D images can be divided into several groups, namely appearance template methods [1, 2], manifold embedding methods [3, 4], tracking methods [5], and geometric methods [6] (for details see the survey by [7]). However, most of the approaches in the literature are not built for unconstrained environments. For example, most approaches assume that the entire set of facial features from a frontal view is visible. Most approaches are trained and tested on images which do not exhibit any kind of appearance variation, such as facial expressions and illumination. The databases tested on mostly contain images with solid or constant background, limited facial expression, no random illumination, and with limited or no facial occlusion. Estimation of head pose from uncontrolled environments has recently been receiving some attention. Orozco et al. [1] and Tosato et al. [4] address the problem of head pose classification in low-resolution video images of crowded scenes under poor lighting, where they treated the problem as a multi-class discrete pose classification problem. The current state-of-the-art for estimating head pose from a higher resolution single 2D face image from uncontrolled environment is provided by Aghajanian and Prince's probabilistic patch-based framework [8]. Our experimental results indicate that the approach in [8] requires a good face localization to be able to avoid background clutter and correctly divide a face image into non-overlapping patches.

Fig. 1. Sample images (with original face scales preserved) from in-house unconstrained environment face database. Color coding shows the assigned head pose: -90° (red), -45° (green), 0° (blue), $+45^\circ$ (yellow), $+90^\circ$ (magenta).

In this paper, we propose a novel probabilistic approach to infer discrete head pose from 2D face images obtained from an in-house video data set (see Figure 1 and 5) collected under different unconstrained environments. We represent the

face images with a *codebook*, i.e. a set of local invariant features, namely *codewords*. The motivation behind the use of local invariant features is due to their high degree of robustness to various transforms, such as the changes in scale, viewpoint, rotation, translation, and occlusion. The proposed methodology learns a "probabilistic spatial codebook template" for each head pose (Figure 2). These templates are inspired by the anatomical face regions (e.g. nose, mouth, ear and eyes) since the spatial distribution of the anatomical face regions is very unique for each head pose (Figure 2). Our methodology obtains pose information not only from codewords, but also from the inferred anatomical regions. Our novel approach differs from current methods in several aspects. Current "templates" usually consist of a set of training images (or tailor made deterministic images) with corresponding pose labels to which a test image is compared to by image-based comparison techniques. Thus, they are not well suited for the images from unconstrained environments. The proposed *probabilistic templates*, on the other hand, are spatial codebook maps learned from training data where each codeword contains the probability density functions for head pose class and anatomical labeling (see Figure 2). To our knowledge, the proposed approach provides the first adaptation of the codebook representation, prevalent in object detection and classification, to the problem of head pose inference. Furthermore, as opposed to the common bag-of-features approach, our approach assigns spatial and functional (anatomical labeling) information to each codeword. In addition, the proposed codebook-based Bayesian formulation allows arbitrary partial occlusions unlike most approaches available in the current literature. Over a large dataset of 5500 unconstrained video frames, our approach provides a higher accuracy rate in head pose classification compared to the current state-of-the-art [8].

2. METHODOLOGY

The proposed methodology (Figure 3) first detects the faces in our training and testing images via a face detection algorithm, and then learns the local invariant feature based codebook which represents face from different poses. The terms used in the proposed formulation are as follows:

$F = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_N\}$ is a codebook containing N codewords. Each $\vec{f}_i$ has the following attributes: $\{o_i, l_i, a_i\}$ where o_i is the occurrence of i-th codeword, a_i is the anatomical region labeling and l_i is the location on the face image.

$Y = \{y_1, y_2, \ldots, y_K\}$ are $K \ll N$ measurements, local invariant features extracted from a test image. We perform the inference task based on these measurements.

$\Phi = \{\phi_1, \phi_2, \ldots, \phi_T\}$ is the set of possible head pose angles.

Assume that there is a function g which maps measurements from test image to the learned codebook, $Y \mapsto F$. Thus, rather than building the head pose inference on Y, we can build it on the codebook F. The function g identifies correspondences between elements of Y and F. If $y_j \in Y$ is matched to $\vec{f}_i \in F$ we have $o_i = 1$. In this work, we used the algorithm in [9] to obtain the function g.

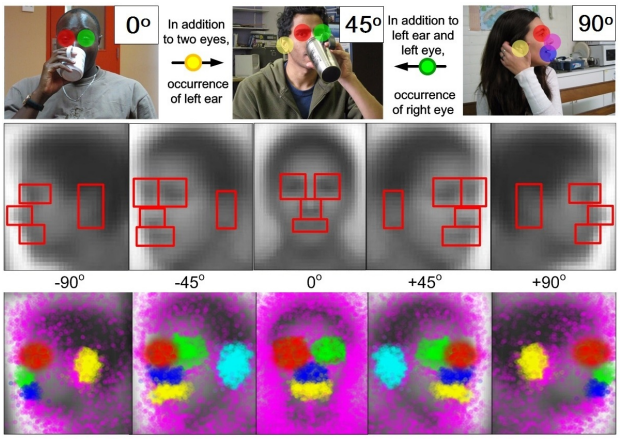

Fig. 2. The use of anatomical regions for head pose. The first row: how the anatomical regions effect head pose even in the presence of the occlusion. The second row: average face images from training database (FERET) for the five pose classes and the manual labeling of anatomical regions. The third row: the inferred head pose templates where each color is assigned to a different anatomical region. Note that the higher the brightness is, the higher the probability $p(a_i|l_i, o_i, \phi)$ is.

2.1. Head Pose Inference

The posterior probability of the pose class given the observed codebook from an image, i.e. $p(\phi|F)$, can be written as:

$$p(\phi|F) = \frac{p(F|\phi)p(\phi)}{p(F)} \tag{1}$$

where the general Bayesian MAP classification task is to infer the most probable pose angle based on:

$$\widehat{\phi} = \max_{\phi \in \Phi} p(\phi|F). \tag{2}$$

Since the denominator in Equation (1) is just a normalizing factor, one can write:

$$p(\phi|F) \propto p(F|\phi)p(\phi) \tag{3}$$

where $p(\phi)$ is the a priori probability on the pose class value, $p(F|\phi)$ is the likelihood for the pose class over all the codewords observed in the image. Given the strong possibility of occlusion and the fact that an individual codeword

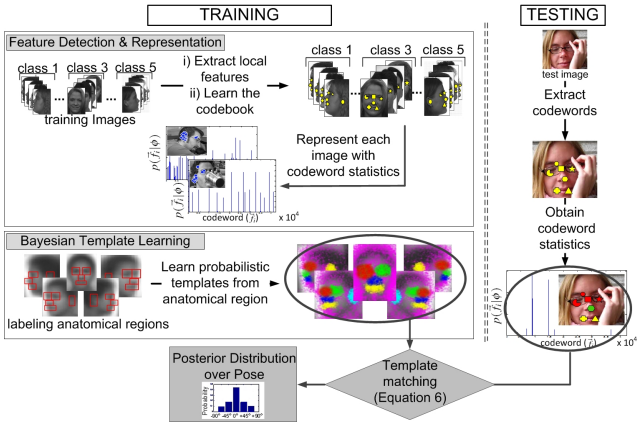

Fig. 3. Flowchart of the algorithm.

is not necessarily providing information about another codeword given the pose, one can make the *conditional* independence assumption of observed codewords given ϕ:

$$p(\phi|F) \propto \prod_{i=1}^{K} p(\vec{f}_i|\phi)p(\phi). \quad (4)$$

Using the definition of $\vec{f}_i$, one can write $p(\vec{f}_i|\phi)$ as $p(a_i, l_i, o_i|\phi)$. This leads to:

$$p(\phi|F) \propto \prod_{i=1}^{K} p(a_i, l_i, o_i|\phi)p(\phi). \quad (5)$$

Using the chain rule one can further expand it as:

$$p(\phi|F) \propto \prod_{i=1}^{K} p(a_i|l_i, o_i, \phi)p(l_i|o_i, \phi)p(o_i|\phi)p(\phi). \quad (6)$$

Here $p(o_i|\phi)$ models the probability of observing the i-th codeword for a specific pose ϕ. This probability is estimated by observing the number of times the i-th codeword appears when every training image with the given ϕ is subjected to procedure. $p(l_i|o_i, \phi)$ is the spatial density of features around location l_i for all training images with the given ϕ in which i-th codeword has been detected. $p(a_i|l_i, o_i, \phi)$ models the probability of observing an anatomical label a_i around location l_i in all training images with the given ϕ in which i-th codeword has been detected. Obtaining $p(l_i|o_i, \phi)$ and $p(a_i|l_i, o_i, \phi)$ requires learning the spatial density of features and the probabilistic distribution of the anatomical regions over a face image for each head pose class ϕ, namely head pose specific probabilistic codebook maps, i.e. "templates" (see Figure 2 and Figure 3). In this work we consider five head poses, $\Phi = \{-90°, -45°, 0°, +45°, +90°\}$, and thus we create five templates. Figure 2 shows the learned codebook-based probabilistic anatomical region models for each pose. To calculate the probabilistic models $p(a_i|l_i, o_i, \phi)$ and $p(l_i|o_i, \phi)$, we used histogram estimation followed by the kernel density smoothing in the vicinity of codeword location.

3. EXPERIMENTS

Experimental Setup: For testing purposes, we collected 5500 video frames from 50 videos each of which was collected from a unique subject and under different illumination and background conditions, where each subject was free in his/her movements, resulting in various face expressions, viewpoints, scales and occlusions (Figure 1 and 5). Although there are several face detectors developed for unconstrained environments, we used the *Object Class Invariant Model* [9] to detect faces and create SIFT [10] based face codebook since it was shown to robustly model and detect facial features in a viewpoint invariant manner in cluttered scenes. (We thank Dr. Toews for providing assistance in adapting the OCI model into our system.) For training purposes, we built a face database from 1000 FERET images (Figure 2) from 200 unique subjects containing equal number of images from each of the five head poses. It is crucial to note that this database was collected under controlled illumination conditions and subjects presented only the head pose change in the yaw angle whereas in the testing the proposed approach has to be able to tackle small angular changes in pitch and roll (Figure 5). The training database was used to learn 1) the OCI model to localize faces and create codebook, and 2) the spatial and anatomical region probabilistic pose templates.

Head pose classification via alternative approaches: On the 5500 test images from our in-house test data set (see Figure 1 and 5), we first compare the performance of the proposed head pose classification approach against the work by Aghajanian and Prince [8] which is the current state-of-the-art for our problem in this paper. We use the implementation and the training parameters learned from 10,900 "real world" training images, which are provided by the authors of [8]. Thus, our results were not affected by any error in implementation or in algorithm learning step. Since the algorithm in [8] provided continuous estimation of the head pose, we needed to discretize their pose space into 5 head pose classes to be able to compare their accuracy with the proposed approach. As done in [8], test images were transformed to a 60x60 template using a Euclidean warp. Setting the patch grid resolution to 10x10, we tested the algorithm in [8] with three different values of σ (standard deviation): 11.25, 45 and 90 degrees, as suggested by authors. The best average accuracy (47.91%) was obtained by 10x10 grid resolution and $\sigma = 11.25$. The corresponding confusion matrix is shown in Figure 4(a).

Head pose classification via the proposed approach: We tested the proposed approach on 5500 images from our

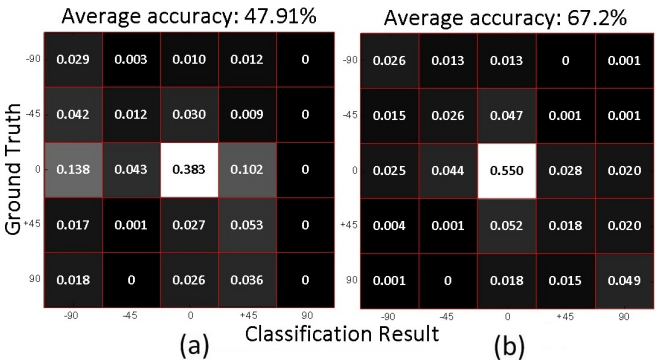

Fig. 4. Confusion matrix obtained by (a) the algorithm in [8], (b) the proposed approach.

in-house test data set, and achieved an average head classification accuracy of 67.2%, which is higher than the current state-of-the-art. The corresponding confusion matrix for our approach can be found in Figure 4(b). The obtained classification accuracy is due to the multi-inference approach embedded in the Bayesian formulation. To make a decision, the proposed approach not only uses the head pose information attached to the individual codewords ($p(o_i|\phi)$), but also the head pose information available in the spatial and anatomical region probabilistic pose templates ($p(a_i|l_i, o_i, \phi)$ and $p(l_i|o_i, \phi)$). Comparing the confusion matrices in Figure 4, we see that the proposed approach outperforms the current state-of-the-art for most of the pose classes. We observed that this could be due the distinctiveness of the ear in performing head pose classification. That is, observing any feature associated with an ear is a good evidence of non-frontal (non-0°) pose. Furthermore, it is observed that the proposed approach was able to robustly classify images with decent variations in pitch and roll angles (see images in Figure 5). However, our algorithm needs to be improved to better distinguish between half ($\{-45^\circ, +45^\circ\}$) and frontal ($\{0^\circ\}$). Most of the wrong classifications are due to the images with in-between poses between half and frontal.

4. CONCLUSIONS AND FUTURE WORK

We have proposed a novel probabilistic approach to infer head pose from 2D face images collected under unconstrained environments. The experimental results have shown that our probabilistic and spatial codebook based head pose representation significantly outperforms the state-of-the-art. We are currently investigating how to extend our classification formulation in order to do continuous head pose inference. Furthermore, our in-house unconstrained face database will be available for academic use soon. To obtain a copy of the data set, please visit authors' website or send an e-mail to *demirkus@cim.mcgill.ca*.

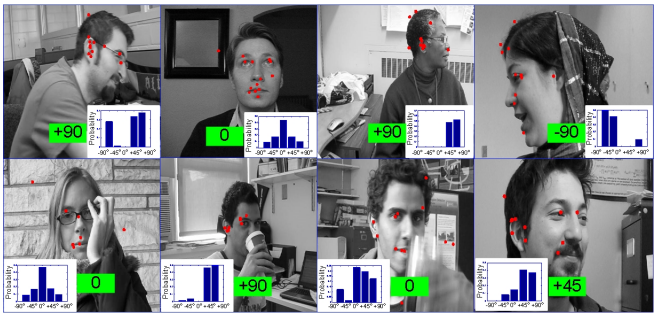

Fig. 5. Sample video frames from the unconstrained face video database, and the corresponding codewords (red dots) for each frame which are used by the proposed approach to infer the posterior pose distribution $p(\phi|F)$ (shown as a plot).

5. REFERENCES

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

[2] J. Sherrah and S. Gong, "Fusion of perceptual cues for robust tracking of head pose and position," *Pattern Recognition*, vol. 34, no. 8, pp. 1565–1572, 2001.

[3] C. BenAbdelkader, "Robust head pose estimation using regression-based supervised manifold learning," in *ECCV*, 2010.

[4] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani, "Multi-class classification on riemannian manifolds for video surveillance," in *ECCV*, 2010, pp. 378–391.

[5] Gangqiang Zhao, Ling Chen, Jie Song, and Gencai Chen, "Large head movement tracking using sift-based registration," in *IEEE ICME*, 2007, pp. 807–810.

[6] J.-G. Wang and E. Sung, "Em enhancement of 3d head pose estimated by point at infinity," *Image and Vision Computing*, vol. 25, no. 12, pp. 1864–1874, 2007.

[7] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE PAMI*, vol. 31, pp. 607–626, 2009.

[8] J. Aghajanian and S.J.D. Prince, "Face pose estimation in uncontrolled environments," in *BMVC*, 2009, pp. 1–11.

[9] M. Toews and T. Arbel, "Detection, localization and sex classification of faces from arbitrary viewpoints and under occlusion," *IEEE PAMI*, 2008.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.