Soft Biometric Trait Classification from Real-world Face Videos Conditioned on Head Pose Estimation

Meltem Demirkus[1], Doina Precup[2], James J. Clark[1] and Tal Arbel[1]
[1]Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
[2]School of Computer Science, McGill University, Montreal, Canada
demirkus@cim.mcgill.ca, dprecup@cs.mcgill.ca, {clark, arbel}@cim.mcgill.ca

Abstract

Recently, soft biometric trait classification has been receiving more attention in the computer vision community due to its wide range of possible application areas. Most approaches in the literature have focused on trait classification in controlled environments, due to the challenges presented by real-world environments, i.e. arbitrary facial expressions, arbitrary partial occlusions, arbitrary and non-uniform illumination conditions and arbitrary background clutter. In recent years, trait classification has started to be applied to real-world environments, with some success. However, the focus has been on estimation from single images or video frames, without leveraging the temporal information available in the entire video sequence. In addition, a fixed set of features are usually used for trait classification without any consideration of possible changes in the facial features due to head pose changes. In this paper, we propose a temporal, probabilistic framework first to robustly estimate continuous head pose angles from real-world videos, and then use this pose estimate to decide on the appropriate set of frames and features to use in a temporal fusion scheme for soft biometric trait classification. Experiments performed on large, real-world video sequences show that our head pose estimator outperforms the current state-of-the-art head pose approaches (by up to 51%), whereas our head pose conditioned biometric trait classifier (for the case of gender classification) outperforms the current state-of-the-art approaches (by up to 31%).

1. Introduction

As the cost of the cameras has decreased in recent years, the size of the available real-world, e.g. surveillance, video data and its range of possible applications has increased substantially. Some of these application areas include face recognition/verification, human tracking, human computer interaction and electronic customer management. Considering the size of available surveillance data, optimizing and automating such applications is required. For this, soft biometric traits, such as gender, age, height, weight, eye color and ethnicity, can be used [14, 11]. Soft biometric traits can be used for video indexing to reduce the search space or to boost the human tracking across difference cameras.

Face classification from unconstrained environments is not a trivial task considering the challenges presented by real-world environments (Figure 1). Despite the wide literature on soft biometric trait classification [30, 22, 9, 24, 13, 28, 4, 16, 32, 29, 6, 23, 5, 10, 19, 33] and head pose estimation [25, 31, 1, 26, 3, 8], most of these approaches are not built for unconstrained environments (see Section 2 for details). Humans, on the other hand, are good at such classification/estimation tasks in real-world environments since they take into consideration not only the facial features, but also the conditions under which these features are collected, such as the head pose for the case of biometric trait classification. Thus, in this paper, we argue that it would be better to perform trait classification from real-world face videos conditioned on head pose estimates.

The methodology introduced in this paper is developed in the context of arbitrary variations in the scene, namely: (i) arbitrary face scales, (ii) non-uniform illumination conditions, (iii) arbitrary partial occlusions, (iv) motion blur, (v) background clutter, (vi) wide variability in image quality, (vii) subject variability, and (viii) the existence of frames where face detection fails, i.e. no facial features are detected. To achieve such a framework, we represent face images with facial codebooks which are learned from local scale invariant features extracted from the detected faces in the training database. Once the faces in the training and testing databases are represented by a codebook, we use codeword statistics to achieve robust head pose estimation and soft biometric trait classification. The proposed framework is a two-stage Bayesian approach: (1) temporal head pose estimation, and (2) temporal soft biometric trait classification conditioned on estimated head pose. In the first stage (Section 3.1), the proposed head pose estimator lever-

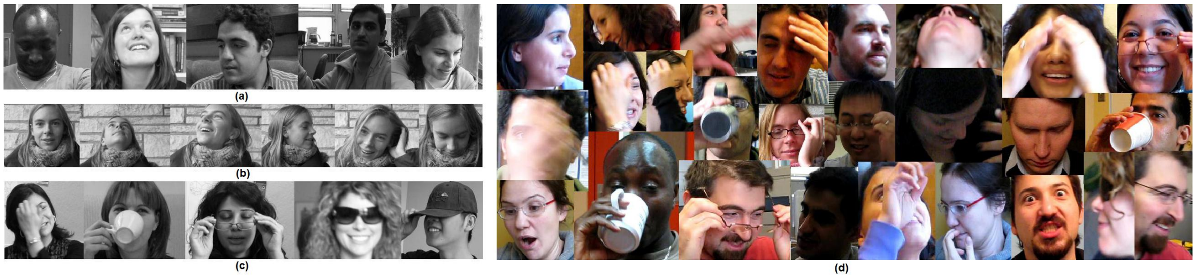

Figure 1. Some of the challenges of real-world environment shown on the McGill Real-World Face Video Database: (a) various illumination conditions and background clutter, (b) arbitrary face poses and scales, (c) arbitrary partial occlusions. (d) More example video frames showing the challenges mentioned in (a), (b) and (c). Note that to fit all the images to this figure, we cropped the facial area.

ages the temporal information available in a video sequence by building generative models of pose variation. This permits accumulating evidence in the pose estimate probabilistically over the test video sequence. The proposed pose estimator is different from the current state-of-the-art head pose approaches since it is a temporal framework which provides posterior probabilities on the continuous head pose (yaw) angles, and it is developed for real-world videos. In the second stage (Section 3.2), a Bayesian trait classifier temporally estimates the probability of each trait hypothesis over the entire video sequence, conditioned only on those frames with high confidence pose estimates obtained from the previous stage. At each selected frame, trait classification is done on the corresponding pose dependent codebook which are learned from the training images with the corresponding pose label. The proposed soft biometric trait classifier is different from the current state-of-the-art approaches, since it is a temporal, probabilistic trait classification scheme that uses pose-specific features rather than a general, viewpoint-invariant representation [30, 17, 9].

In the experimental results section, we provide qualitative and quantitative evaluations of both steps, i.e. head pose estimation and soft biometric trait classification. Even though the proposed method is applicable to any trait, in this paper, we choose to test the framework on the task of gender classification due to the recent attention it has received (e.g. [30, 9]. We compare the performances of these two steps to the current state-of-the-art approaches ([3, 1, 8] for head pose and [30, 9] for gender) on the McGill Real-World Face Video Database (Figure 1). The experimental results show that the proposed video-based head pose estimator and gender classification steps significantly outperforms the current state-of-the-art approaches despite the large number of video frames containing no reliable face information (i.e. face detection failure, and very few detected features).

2. Related Work

Most approaches in the gender classification literature have focused on controlled environments, due to the challenges presented by real-world environments, i.e. arbitrary facial expressions, arbitrary partial occlusions, arbitrary and non-uniform illumination conditions and arbitrary background clutter [22, 2, 18, 24, 12, 13, 28]. These approaches have incompatible stages with real-world environments, such as the need for good face alignment (no extreme head pose is allowed), and the requirement for specific facial regions to track (no occlusion is allowed). Furthermore, such approaches focus on analyzing images with limited degrees of freedom. For instance, some work well on multi-view (e.g. specific, non-arbitrary) face images, but with optimal indoor lighting whereas some others work well on images with arbitrary background clutters, but without any occlusions, or non-uniform lighting.

Due to the wide range of possible applications for real-world, e.g. surveillance, videos, recently gender classification from real-world environments has been receiving more attention [30, 17, 9]. Except for the work by [9], these approaches are single-image based approaches which do not leverage the temporal information available in a video sequence. Kumar et al. [17] achieved robust face verification from real-world frontal face images using several facial attributes, e.g. gender, ethnicity, age, hair color, face shape. The facial attribute classification was done using an SVM (similar to [24, 28]) on appearance features from the detected faces. The approach by Toews and Arbel [30] created a viewpoint-invariant appearance model for face detection purpose. The model used local invariant features, i.e. SIFT [21], to probabilistically create a geometrical model robust to various transformations based on which faces are detected and localized. Later, the model features were used for gender classification from multi and arbitrary viewpoints. Demirkus et al. [9], on the other hand, modeled the gender trait temporally using a Bayesian sequential approach. Estimation of the posterior probability of a face trait at a specific time was achieved via the viewpoint-invariant model in [30]. Later, a Markov model was used to model temporal dependencies. It was shown that such a temporal framework outperformed both the alternative single image-based trait classification methods.

The literature on head pose estimation from 2D images can be divided into several groups (see the survey by [25]): appearance template methods, manifold/subspace embedding methods, geometric (facial landmark) methods, and tracking methods. Most of these approaches assume that the entire set of facial features typical for frontal poses is always visible. Facial features are often manually labeled in the testing data, rather than automatically extracted. However, many of these requirements and assumptions are not feasible in the context of real-world videos.

Estimation of head pose from uncontrolled environments has recently been receiving more attention [31, 1, 26, 8]. Orozco et al. [26] and Tosato et al. [31] addressed the problem of head pose estimation in *single, low resolution video frames* of crowded scenes under poor lighting, where they treated the problem as a multi-class discrete pose *classification* problem. Demirkus et al. [8], on the other hand, proposed spatial and probabilistic pose templates which are obtained from local codewords. Overall, most approaches treat the head pose estimation problem as a classification problem, that is, assigning a face image to one of discrete poses, rather than perform continuous head pose estimation. One exception is the work by Aghajanian and Prince [1] proposing a patch-based regression framework to estimate continuous head pose from single images. Finally, all the approaches mentioned are developed for single face images, and do not attempt to leverage the relative head pose information available between consecutive video frames.

3. Methodology

The proposed framework works as follows: First, we run a face detection algorithm to detect faces in the training and testing images. Once the faces are detected, local invariant features extracted from the detected faces are mapped into a *codebook*, which can be learned by sophisticated clustering methods (e.g. [30]). The motivation behind the use of a codebook is its high degree of robustness to various transforms, such as the changes in scale, viewpoint, rotation and translation. Next, we estimate pose probabilities for each single frame using the association between the codeword statistics and head pose. Afterwards, we develop a novel temporal and probabilistic pose estimation scheme on these estimated pose probabilities. Later, we train on the biometric trait for each estimated pose to get a pose specific codebook. Online, we condition trait classification on the most confident pose estimates, and do trait classification temporally over the entire video sequence.

3.1. Temporal Modeling of Head Pose Over a Video Sequence

3.1.1 MRF-based Head Pose Temporal Model

The goal of our pose framework is to estimate an entire set of pose probability density functions throughout a video. Assume that we have a codebook with N codewords. For each codeword, we define some statistics $\vec{f}$. Let $F = \{\vec{f_1}, \vec{f_2}, \ldots, \vec{f_N}\}$ be a vector of *codebook statistics*. Each $\vec{f_i}$ has the following attributes: $\{o_i, l_i, a_i\}$ where o_i is the occurrence statistic of the *i*-th codeword, a_i is the anatomical region labeling and l_i is the location on the face image. $\theta = \{\phi_1, \phi_2, \ldots, \phi_T\}$ is the set of possible head pose angles. The observation from a video sequence $\mathbf{F}$ is defined as $\mathbf{F} = (F_1, F_2, \cdots, F_M)$ for M video frames, and the *configuration* of the underlying head pose in a video sequence Θ is defined as $\Theta = (\theta_1, \theta_2, \cdots, \theta_M)$.

Our goal is to calculate the posterior distribution $p(\Theta|\mathbf{F})$, where $p(\Theta|\mathbf{F}) = \frac{p(\Theta,\mathbf{F})}{p(\mathbf{F})}$. It is evident that $p(\mathbf{F})$ is a normalization constant Z with respect to Θ, such that $p(\Theta|\mathbf{F}) = \frac{1}{Z}p(\Theta, \mathbf{F})$. Note that, if Z can not be calculated directly, $p(\Theta, \mathbf{F})$ becomes an approximation to the posterior distribution $p(\Theta|\mathbf{F})$. We wish to estimate the most likely configuration of the posterior distribution Θ^*. Computing Θ^* can be difficult without any approximations [15]. Thus, we use a graphical model to model the head pose over a video sequence Θ. Now, we can express the posterior distribution as an MRF with pairwise interactions:

$$p(\Theta|\mathbf{F}) = \frac{1}{Z}\left(\prod_{i=1}^{M}\vartheta(\theta_i, F_i)\right)\left(\prod_{i=1}^{M-1}\varphi(\theta_i, \theta_j)\right) \qquad (1)$$

where $\vartheta(\theta_i, F_i)$ is the unary compatibility function accounting for local evidence (likelihood) for θ_i and $\varphi(\theta_i, \theta_j)$ is the pairwise compatibility function between θ_i and θ_j (which corresponds to the horizontal edges of the model and $j = i + 1$).

3.1.2 Inference through Belief Propagation

One way to estimate the most likely head pose configuration is by calculating the MAP estimate, i.e. $\Theta^* = \text{argmax}_{\Theta} p(\Theta|\mathbf{F})$, which can be achieved through Belief Propagation (BP) [27]. BP is an inference method developed for graphical models, which can be used to estimate the *marginals* or the most likely *states*, e.g. MAP. In our experiments, we adapt the "sum-product" BP algorithm which estimates the probability distributions. BP provides the exact solution if there is no loop (cycle) in the graph, i.e. if the graph is a chain or a tree [27], which is the case here.In order to estimate the marginal distributions, the BP algorithm creates a set of message variables which are updated iteratively via passing between neighbors. $m_{ij}(\theta_j)$ corresponds to the message sent from node i to node j about the degree of its belief that node j should be in state θ_j.The BP algorithm updates the messages according to:

$$m_{ij}^{(t+1)}(\theta_j) = \frac{1}{Z_j}\sum_{\theta_i}\varphi(\theta_i,\theta_j)\vartheta(\theta_i,F_i)\prod_{k\in N(i)\backslash j} m_{ki}^{(t)}(\theta_i) \tag{2}$$

where $\frac{1}{Z_j} = \sum_{\theta_i} m_{ij}^{(t+1)}(\theta_i)$ is a normalization factor, and the set of nodes in the neighborhood of i is denoted by $N(i)$. $(t+1)$ and (t) represent the iteration indices. The initial messages $m_{ij}^{(0)}(.)$ are typically initialized to uniform positive values. In a general graph, the update procedure is repeated iteratively until the messages converge to a consensus, then the *marginals* (*beliefs*) are calculated (Equation 3). Since our graph here is acyclic, two passes are sufficient to compute all messages, making the algorithm efficient.

The *belief* (b_i) is an estimate of the marginal distribution, derived from converged message variables as follows:

$$b_i(\theta_i) = \frac{1}{\tilde{Z}_i}\vartheta(\theta_i,F_i)\prod_{k\in N(i)} m_{ki}(\theta_i) \tag{3}$$

where $\tilde{Z}_i$ is a normalization factor guaranteeing that $\sum_{\theta_i} b_i(\theta_i) = 1$. Since our graph does not have loops, the *beliefs* are guaranteed to be the true marginals $p(\theta_i|\mathbf{F})$. Note that in the case of "sum-product" BP, the belief is an estimate of marginals whose maximal point indicates the most likely state. We define the unary compatibility function for each node *i*, i.e. $\vartheta(\theta_i,F_i)$, as the joint distribution $p(\theta_i,F_i) = p(\theta_i|F_i)p(F_i)$ where $p(F_i)$ is assumed to be uniform. We make this assumption because F is sparse, and it is difficult to obtain a more informed prior. $p(\theta_i|F_i)$ is obtained via the approach explained in Section 3.1.3. Furthermore, the pairwise compatibility function $\varphi(\theta_i,\theta_j)$ is assumed to be a Gaussian distribution $N(\mu,\Delta)$ with mean μ and covariance matrix Δ.

3.1.3 Continuous Head Pose Estimation from a Single Video Frame

We first summarize the previously introduced approach in [8] to obtain samples from the head pose distribution given a set of observed codewords, i.e. $p(\theta|F)$ estimated from a single image. Next, we explain how we take these pose samples to the continuous pose space in order to use them later in the Belief Propagation as the unary compatibility function $\vartheta(\theta,F)$.

The approach in [8] first learns five $(-90°,-45°,0°,+45°,+90°)$ spatial and probabilistic codebook pose templates from the training database. Each template provides a probabilistic representation of the head pose class and anatomical labeling distribution, and each of them will be used to estimate the probability of observing the related head pose for the given face image. The general Bayesian MAP classification task is to infer the most probable pose angle $\widehat{\phi}$, such that $\widehat{\phi} = \max_{\phi\in\theta} p(\phi|F) = \max_{\phi\in\theta}\left\{\frac{p(F|\phi)p(\phi)}{p(F)}\right\}$. Since the denominator $p(F)$ is just a normalizing factor, one can write: $\widehat{\phi} \propto \max_{\phi\in\theta}\{p(F|\phi)p(\phi)\}$. Here, $p(\phi)$ is the *a priori* probability density function on the pose class value and $p(F|\phi)$ is the likelihood for the class, conditioned on the codewords observed in the image. Furthermore, one can assume the *conditional* independence of observed codewords given ϕ since *(i)* there is a the strong possibility of occlusion, and *(ii)* an individual codeword is not necessarily providing any information about another codeword given the pose, i.e. $p(F|\phi) = \prod_{i=1}^{N} p(\vec{f}_i|\phi)$. Finally, using the definition of $\vec{f}_i$ and the chain rule:

$$p(\phi|F) \propto \prod_{i=1}^{N} p(a_i|l_i,o_i,\phi)p(l_i|o_i,\phi)p(o_i|\phi)p(\phi) \tag{4}$$

where $p(o_i|\phi)$ models the probability density describing the probability of observing the *i*-th codeword for a specific pose ϕ. $p(l_i|o_i,\phi)$ is the spatial density of features around location l_i given ϕ, where the *i*-th codeword (o_i) occurs. $p(a_i|l_i,o_i,\phi)$ 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. To be able to estimate the probabilities $p(l_i|o_i,\phi)$ and $p(a_i|l_i,o_i,\phi)$, we need to learn the spatial density of features and the probability distribution of the anatomical regions over training images for each head pose class ϕ, namely head pose specific probabilistic codebook templates. Furthermore, as suggested in [8], we used histogram estimation followed by kernel density smoothing in the vicinity of codeword location while obtaining $p(l_i|o_i,\phi)$ and $p(a_i|l_i,o_i,\phi)$.

Note that, unlike the 5 anatomical regions used in the original formulation in [8], here we modeled each visible anatomical region adding up to 23 unique anatomical regions over 5 bins of angles $(-90°,-45°,0°,+45°,+90°)$. For example, right eye from $-90°$, mouth from $0°$ and left ear from $+45°$. Next, we estimate the entire pose density $p(\theta|F)$ in the range $[-90°,+90°]$. To be able to achieve this, we tested a number of parametric and non-parametric density estimators, namely Gaussian, Cauchy and kernel-based (gaussian kernel), on our validation data. We observed that, among all, Gaussian model fitting provides the best results (see Section 4.3). Gaussian models perform well due to their ability to smooth over false positives.

3.2. Temporal Biometric Trait Classification

Our goal now is to infer the most probable soft biometric trait value (c^*) for a video sequence using *(i)* the facial local invariant features obtained from each video frame, and *(ii)* MAP estimation of head pose for each video

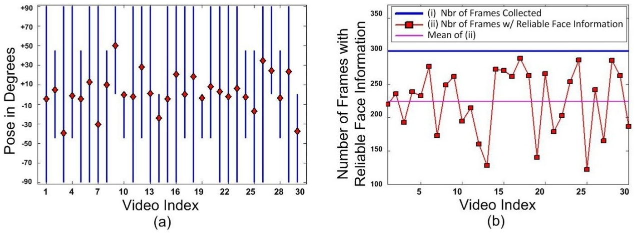

Figure 2. Head pose statistics for the test database of 30 unique subjects (videos): (a) The change in the yaw angle. Each ◇ shows the mean viewpoint angle for a specific subject, and the bars show pose range for each subject. (b) The number of frames (out of 300 frames per video) which contain reliable face information. Note that out of 9000 video frames, only 6732 of them have available face information (i.e. face reliably detected), so we need to interpolate the head pose estimation of 2268 frames (25.2%) using our temporal model.

frame which is provided by provided by Section 3.1, i.e $\hat{\phi} = \underset{\phi \in \theta}{max} \{p(\phi|F)\}$. For this purpose, we define a Bayesian classifier over the posterior probability of the soft biometric trait class C given the observed codebook statistics obtained in all previous frames until time t, $\mathbf{F} = \{F_t, F_{t-1}, \ldots, F_1\}$ where $F_i = \{f_1, f_2, \ldots, f_N\}$ and codeword $f_j = o_i$:

$$c^* = \underset{c}{max} \left\{ log \frac{p(C = c|F_t, F_{t-1}, \ldots, F_1)}{p(C = \bar{c}|F_t, F_{t-1}, \ldots, F_1)} \right\} \quad (5)$$

Since, in this paper, we consider the binary trait of gender (i.e. male or female), $C = c$ or $C = \bar{c}$, where c and $\bar{c}$ are opposing genders. One can define the posterior probability density function in Equation 5, such that:

$$p(c|F_t, F_{t-1}, \ldots, F_1) = \frac{p(F_t, F_{t-1}, \ldots, F_1|c)}{p(F_t, F_{t-1}, \ldots, F_1)} p(c) \quad (6)$$

where $p(c)$ is the *a priori* probability on the class trait value c, which is set to be uniform, and $p(F_t, F_{t-1}, \ldots, F_1)$ is the joint probability density function over all the features, which is not needed to be calculated since it is omitted due to the ratio in Equation 5. Furthermore, to model the likelihood function $p(F_t, F_{t-1}, \ldots, F_1|c)$, one can assume the conditional independence of observed codebooks given c, i.e. $p(F_t, F_{t-1}, \ldots, F_1|c) = p(F_t|c)p(F_{t-1}|c) \ldots p(F_1|c)$. Such an assumption is reasonable should we only choose a subset of frames, i.e. confident frames decided based on a confidence measure which will be explain later in this section, from which to estimate the biometric trait. Modelling the trait likelihood function $p(F_i|c)$ requires the codebook to be learned from an appropriate set of facial features, so that the facial local invariant features extracted from test images are mapped to the appropriate codeword. Our observations show that the codebook changes dramatically when the head pose changes. Thus, we propose to use for each frame pose specific codebook statistics (F_ϕ) which are obtained from a codebook learned from the training images with pose ϕ. To achieve this, we first obtain the MAP estimation of head pose for the *ith* frame, i.e. $\hat{\phi}_i$, using the algorithm in Section 3.1, and then using codebook statistics $F_{\hat{\phi}}$ obtain the trait class likelihood function for the *ith* frame, i.e. $p(F_{\hat{\phi}_i}|c)$. Thus,

$$c^* = \underset{c}{max} \left\{ log \frac{p(F_{\hat{\phi}_t}|c)p(F_{\hat{\phi}_{t-1}}|c) \ldots p(F_{\hat{\phi}_1}|c)p(c)}{p(F_{\hat{\phi}_t}|\bar{c})p(F_{\hat{\phi}_{t-1}}|\bar{c}) \ldots p(F_{\hat{\phi}_1}|\bar{c})p(\bar{c})} \right\} \quad (7)$$

To model the likelihood function $p(F_{\hat{\phi}_i}|c)$, we can assume the conditional independence of observed codewords given c since (i) there is a the strong possibility of occlusion, and (ii) an individual codeword is not necessarily providing any information about another codeword given the gender: $p(F_{\hat{\phi}_i}|c) \propto \prod_{j=1}^{N} p(f_{\hat{\phi}_i}|c)$. One can learn the probability $p(f_{\hat{\phi}_i}|c)$ via the frequencies obtained from the training database.

In our formulation we considered five poses, i.e. $\theta = (-90°, -45°, 0°, +45°, +90°)$. One can increase the number of head poses to better model the relationship between the gender and the head pose; however, the time cost of learning a codebook for each pose should also be considered.

The conditional independence assumption that we make on observed codebooks allows us to model each frame independently in the video sequence, and select only a subset of frames whose confidence measure is higher than a threshold T. Here, we define this confidence measure as $p(\theta|F)$ obtained from Section 3.1. Thus, in Equation 7, we use only the frames whose head pose estimation probability is higher than $T_{p(\theta|F)}$.

4. Experiments

4.1. Experimental Setup

Although there are several face detectors developed for unconstrained environments [34], we used the OCI model [30] to detect faces and create a SIFT [21] based face codebook since it was shown to robustly model and detect facial features in a viewpoint invariant manner in cluttered scenes. For training purposes, we built a database from 3500 FERET images from 700 unique subjects (350 female and 350 male) containing an equal number of images from each of the five head poses. The motivation behind learning on a clean (i.e. no occlusion) and controlled database is to be able to increase the number of samples for each codeword in the codebook as much as possible. We noted empirically that when training on a subset of the test dataset during cross-validation experiments, the number of codeword samples available during training was substantially reduced, indirectly leading to reduced trait distinctiveness. This training database was used to *(i)* learn the OCI model [30] to localize faces, *(ii)* learn a viewpoint invariant face codebook representation, and the spatial and anatomical region probabilistic pose templates [8] to obtain a robust head pose distribution (see Section 3.1.3), and *(iii)* learn the pose specific gender codebook representation, and the corresponding codebook statistics (F_ϕ) for gender detection purpose (see Section 3.2).

4.2. McGill Real-World Face Video Database

This test database consists of 30 unconstrained (real-world) videos from 30 unique subjects (15 female and 15 male). Each video was collected with different illumination conditions and backgrounds, and each subject was free in his/her movements, resulting in arbitrary face scales, expressions, viewpoints, local and/or global occlusions (due to closed eyes, glasses, hand, coffee cup, scarf or hat) (see Figure 1). For each subject, a 60-second video with 30 fps at 640x480 resolution was recorded. The face scale changed (on average from 113x104 to 222x236) not only from one video to another, but also within the same video sequence. The sub-sampling of frames was empirically set to 5 frames per second, leading to $300 \times 30 = 9000$ video frames in total. Each of these frames are labeled with the correct gender class and the closest pose angle from $\{-90°, -45°, 0°, +45°, +90°\}$. The individual frames in the McGill Database exhibit wide variability in head pose in terms of angle, yaw and partial occlusions. For instance, 36.7% of the frames are beyond the range $[-45^o, +45^o]$ of which 36.5% is either -90^o or $+90^o$ (see Figure 2(a)). Furthermore, each subject in the video database has a broad variety of viewpoints, so that our experimental results were not biased by any specific subject (see Figure 2(a)).Please note that, the database we use in this paper is the latest version, which is publicly available.

4.3. Evaluation of the Head Pose Estimation

In this section, we aim to show that the algorithm described in Section 3.1 is a robust head pose estimator, thus suitable for the later use in the gender detection phase. Thus, we tested our pose estimation algorithm explained in Section 3.1 and compared the results with several other “state-of-the-art” approaches. Here, we relied on the manual pose labels provided in Section 4.2 which served as coarse “ground truth” for the experiments. Although we estimated a continuous head pose density function at each frame in the sequence, we were bounded by the precision of the manual labeling of the database. Thus, we could only evaluate the performance of the proposed and the state-of-the-art algorithms in terms of classification accuracy, which in this context is defined as the number of times the estimated pose angle fell into the correct pose bin. For the proposed framework, the MAP of the probability density function served as the estimated angle.

	Accuracy (%)
BenAbdelkader [3]	7.80
Aghajanian and Prince [1]	25.5
Demirkus et al. [8]	43.7
Proposed pose approach	***58.8***

Table 1. Comparison of the head pose classification accuracies over 9000 video frames from McGill Real-World Database.

Over the 9000 real-world face images, we evaluated the performance of the approaches mentioned in [1], [8] and [3]. During the training procedure of [1], we used the same training parameters described in [1]. Following the approach presented by the authors, we transformed the detected face images to a 60x60 template using a Euclidean warp. The best average accuracy of 25.5% was obtained for 10x10 grid resolution and $\sigma = 11.25$. Next, we evaluated the algorithm in [8]. During the training, we used 1000 FERET images (Section 4.1) to learn the spatial and anatomical region pose templates, as was performed in [8]. Furthermore, we observed that we could obtain a better pose representation once we defined 23 anatomical regions rather than 5, as specified in the paper [8] (see Section 3.1.3 for details). Since the framework is probabilistic, we used the MAP over $p(\phi|F)$ (see Equation 4) to estimate the pose class. This led to an accuracy of 43.7%. The results we obtained via BenAbdelkader’s supervised manifold-based approach [3] were the lowest in terms of classification accuracy: 7.8% for a 2D manifold. We had tried different embedding dimensions, i.e. 2D, 3D, 8D and 20D (50D and more was leading to unstable manifolds). To eliminate the possibility that this low accuracy was due to our implementation, we tested our implementation on the same FacePix [20] database used in [3] and achieved similar results to those reported.

Finally, we tested the proposed approach over the 30 videos containing 9000 video frames. Note that we empirically set the parameters of $\varphi(\theta_i, \theta_j)$ in BP as $\mu = [91\ 91]^T$, $\sigma_x^2 = 2500$, $\sigma_y^2 = 5000$ and $\rho = 0.8$ (correlation coefficient). As shown in Table 1, using the temporal framework, a classification accuracy of 58.8% was achieved, which is significantly higher than the other approaches. This relatively high accuracy was obtained despite the large number of frames (i.e. 25.2%) which contain no reliable face information (see Figure 2(b)). The low accuracies of the *image-based* state-of-the-art approaches were mainly a result of frames containing no reliable face information (see Figure 2(b)). Because the proposed framework treats each video as an MRF and uses the BP-based algorithm to infer the entire sequence of poses in a video, it can robustly estimate the poses even with frames containing no reliable face information, and to correct for inconsistent head pose labels. Moreover, as shown in Figure 3, the proposed approach is able to robustly classify pose with small angular changes in pitch and roll, even when only presented with head pose variation in the yaw angle in the training set.

4.4. Evaluation of the Gender Estimation

We tested the proposed gender classification algorithm explained in Section 3.2 and compared the results with alternative state-of-the-art approaches (see Table 2). In our experiments, we examined the following possible methods: *(i)* the proposed temporal approach, *(ii)* the temporal gender classification approach by Demirkus et al. [9] based on pose-invariant features, *(iii)* the static image-based (non-temporal) approach by Toews and Arbel [30], and *(iv)* the static image-based (non-temporal) approach using majority voting with SVM classification on pixel intensity values.

Over the 9000 real-world face images, we applied the SVM classifier (using libsvm [7]) for gender detection purpose. The best classification accuracy is obtained by using no-normalization and downsampling detected face images to 24x24 (similar to the findings in [22, 9]). We trained the SVM classifier on the training database, i.e. FERET database. The best SVM parameters for the RBF kernel ($\gamma = 0.0078125, C = 32$) were obtained using a grid search. Later, for each of 30 videos, we obtained the gender class based on the majority voting leading to 62% accuracy. Next, we tested the single image-based gender classifier introduced in [30] over 9000 real-world face images. The approach in [30] uses the posterior probability of the gender trait given a pose-invariant codebook ($p(c|F)$) rather than using pose specific codebook ($p(c|F_\phi)$) -like the proposed approach does- for gender detection purpose, and its accuracy was limited to 66%. Next, we evaluated the temporal model introduced in [9] over 30 video sequences each of which has 300 frames. Similar to [30], this approach uses the pose-invariant codebook representation to obtain gender trait, i.e. $p(c|F)$. However, the algorithm in [9] is a temporal approach whereas the one in [30] is a single image-based approach. We used the original implementation, provided by authors, for [9], and at the end of 300 video frames ($t = 300$), we obtained 80% accuracy. Lastly, we evaluated the proposed temporal approach explained in Section 3.2. We observed that at the end of 300 video frames, we achieved a gender classification accuracy of 93% by setting the frame selection threshold $T_{p(\theta|F)}$ to 0.28 (see Table 2 and Figure 3). The value of $T_{p(\theta|F)}$ is empirically decided based on our validation data. The obtained classification accuracy was due to not only using the right set of codewords statistics (F_ϕ) for gender estimation, but also removing the frames which didn't provide reliable head pose estimation. We observed that such unreliable frames consist of ones which contained *(i)* no reliable face detection, *(ii)* the presence of false accepted codewords, e.g. codewords falsely detected outside of the face area.

	Accuracy (%)
SVM	62
Toews and Arbel [30]	66
Demirkus et al. [9]	80
Proposed temporal approach	***93***

Table 2. Comparison of the gender classification performance of the proposed and the current state-of-the-art approaches on McGill Real-World Face Video Database.

5. Conclusions

In this paper, we propose a two-stage temporal and probabilistic framework which first estimates the continuous head pose angle, and then uses this pose estimate to choose the frames with strong confidence in the pose estimate, and then condition trait classification on the pose by using the pose-specific codeword features. Experiments performed on a large, real-world video database show that the two stages of the proposed approach significantly outperforms the state-of-the-art approaches despite the existence of video frames with no reliable face information. One avenue for our future work is to investigate other features (e.g. SURF) to explore their ability in further improving the discrimination of soft biometric trait.

References

[1] J. Aghajanian and S. Prince. Face pose estimation in uncontrolled environments. In *BMVC*, 2009. 1, 2, 3, 6

[2] S. Baluja and H. A. Rowley. Boosting sex identification performance. *IJCV*, 2007. 2

[3] C. BenAbdelkader. Robust head pose estimation using supervised manifold learning. In *ECCV*, 2010. 1, 2, 6

[4] M. Cadoni, E. Grosso, A. Lagorio, and M. Tistarelli. From 3d faces to biometric identities. 1

[5] D. Cao, M. Chen, C.and Piccirilli, D. Adjeroh, T. Bourlai, and A. Ross. Can facial metrology predict gender? In *IJCB*, 2011. 1

[6] C. Chan, J. Kittler, and K. Messer. Multispectral local binary pattern histogram for component-based color face verification. In *BTAS*, 2007. 1

[7] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. 7

[8] M. Demirkus, B. Oreshkin, J. Clark, and T. Arbel. Spatial and probabilistic codebook template based head pose estimation from unconstrained environments. In *IEEE ICIP*, 2011. 1, 2, 3, 4, 6

[9] M. Demirkus, M. Toews, J. Clark, and T. Arbel. Gender classification from unconstrained video sequences. In *AMFG, IEEE CVPR 2010*, 2010. 1, 2, 7

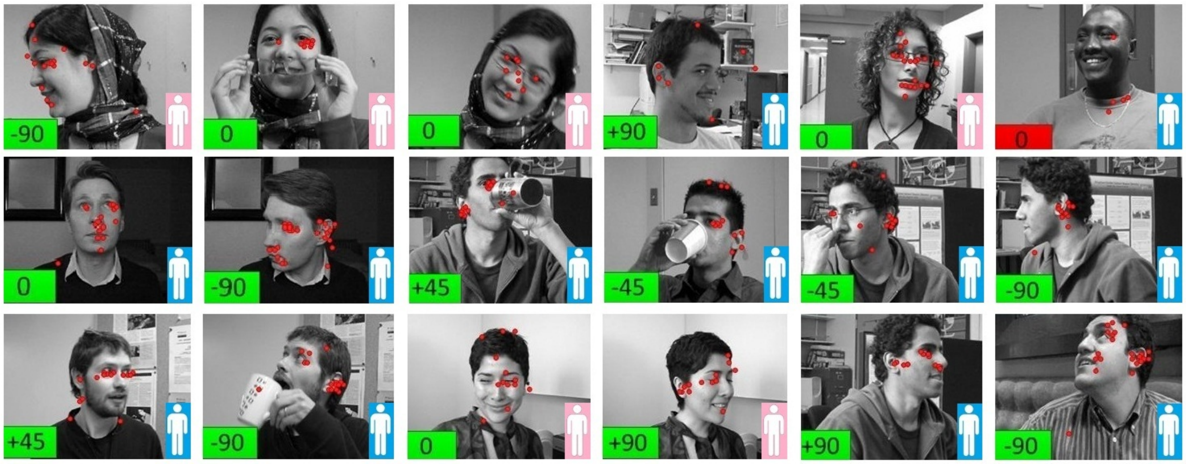

Figure 3. Sample video frames from the McGill Real-World Face Video Database, and corresponding codewords (red dots show their locations, i.e. l_i) and the head pose estimation results obtained by the proposed head pose estimator. Green and red boxes represent correct and false classification results, respectively. Bottom right corner shows the signs for gender decisions obtained by the proposed gender classifier, i.e. female (pink) and male (blue).

[10] T. Dhamecha, A. Sankaran, R. Singh, and M. Vatsa. Is gender classification across ethnicity feasible using discriminant functions? In *IJCB*, 2011. 1

[11] Y. Fu, G. Guo, and T. S. Huang. Soft biometrics for video surveillance. In *Intelligent Video Surveillance: Systems and Technology*, 2009. 1

[12] S. Gutta, H. Wechsler, and P. Phillips. Gender and ethnic classification of human faces using hybrid classifiers. In *IEEE FG*, 1998. 2

[13] A. Hadid and M. Pietikinen. Combining appearance and motion for face and gender recognition from videos. *Pattern Recognition*, 2009. 1, 2

[14] A. Jain, S. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition? In *SPIE*, 2004. 1

[15] D. Knill and W. Richards. *Perception as Bayesian inference*. Cambridge, 1996. 3

[16] K. Kollreider, H. Fronthaler, and J. Bign. Real-time face detection using illumination invariant features. In *SCIA*, 2007. 1

[17] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE PAMI*, 2011. 2

[18] A. Lapedriza, M. J. Maryn-Jimenez, and J. Vitria. Gender recognition in non controlled environments. In *ICPR*, 2006. 2

[19] X. Li, X. Zhao, Y. Fu, and Y. Liu. Bimodal gender recognition from face and fingerprint. In *IEEE CVPR*, 2010. 1

[20] G. Little, S. Krishna, J. Black, and S. Panchanthan. A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In *International Conference on Acoustics, Speech, and Signal Processing*, 2005. 6

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2, 5

[22] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. 2008. 1, 2, 7

[23] G. L. Marcialis, F. Roli, and D. Muntoni. Group-specific face verification using soft biometrics. In *Journal of Visual Languages and Computing*, 2009. 1

[24] B. Moghaddam and M. Yang. Learning gender with support faces. *IEEE TPAMI*, 2002. 1, 2

[25] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE PAMI*, 2009. 1, 3

[26] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *BMVC*, 2009. 1, 3

[27] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988. 3

[28] G. Shakhnarovich, P. A. Viola, and B. Moghaddam. A unified learning framework for real time face detection and classification. In *IEEE FG*, 2002. 1, 2

[29] V. Thomas, N. Chawla, K. Bowyer, and P. Flynn. Learning to predict gender from iris images. In *BTAS*, 2007. 1

[30] M. Toews and T. Arbel. Detection, localization and sex classification of faces from arbitrary viewpoints and under occlusion. *IEEE PAMI*, 2008. 1, 2, 3, 5, 6, 7

[31] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani. Multi-class classification on riemannian manifolds for video surveillance. In *ECCV*, 2010. 1, 3

[32] J. Yu, B. Bhanu, Y. Xu, and A. K. Roy-Chowdhury. Super-resolved facial texture under changing pose and illumination. In *IEEE ICIP*, 2007. 1

[33] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu. A study on gait-based gender classification. In *IEEE Trans. Image Processing.*, 2009. 1

[34] S. K. Zhou, R. Chellappa, and W. Zhao. *Unconstrained Face Recognition*. Springer-Verlag New York, Inc., 2005. 5