Gender Classification from Unconstrained Video Sequences

Meltem Demirkus
Centre for Intelligent Machines
McGill University, Montreal, Canada
demirkus@cim.mcgill.ca

Matthew Toews
Department of Radiology
Harvard Medical School, Boston, USA
mt@bwh.harvard.edu

James J. Clark
Centre for Intelligent Machines
McGill University, Montreal, Canada
clark@cim.mcgill.ca

Tal Arbel
Centre for Intelligent Machines
McGill University, Montreal, Canada
arbel@cim.mcgill.ca

Abstract

This paper presents the first investigation into the classification of faces from unconstrained video sequences in natural scenes, i.e., with arbitrary poses, facial expressions, occlusions, illumination conditions and motion blur. To overcome difficulties from individual frames, a novel Bayesian formulation is proposed to estimate the posterior probability of a face trait at a specific time, conditional on features identified in previous frames of a video sequence. A Markov model is used to represent temporal dependencies, and classification involves determining the maximum a posteriori *class at a given time. Showing the robustness of the proposed system, the Bayesian framework is first trained on a database collected under controlled conditions, and then applied to the previously unseen faces obtained from an unconstrained video database. The Markovian temporal model results in a gender classification rate of 90% by the last video frame, and is shown to outperform alternative approaches previously introduced in the literature.*

1. Introduction

Classifying face images in terms of soft biometric traits, such as gender, age or ethnicity, has been receiving a wide amount of attention in the recent computer vision literature, especially in the context of video surveillance. Performing face detection and facial trait classification in realistic scenarios presents significant challenges [26], particularly in terms of achieving robustness over large changes in a person's viewpoint (head pose), various face scales, non-uniform illumination conditions, partial occlusion (see Figure 1) and overcoming potentially noisy images or false face detection. To date, the face classification, particularly gender classification, literature has focused on relatively restricted scenarios with still images of frontal/near frontal, upright faces from controlled databases [13, 17, 10, 11, 4, 12, 3, 8, 20]. Relatively few formulations have addressed classification of face images acquired from multiple viewpoints [7, 9, 14, 22, 15] or from video sequence data [7, 9, 18, 21]. Even though some works claim to introduce algorithms for face classification from unconstrained environments, these algorithms have several stages which are not compatible with real-world unconstrained environments, such as the need for good face alignment (e.g. no extreme head pose is allowed) and/or the requirement for specific facial regions to track (e.g. no occlusion is allowed). Furthermore, such approaches were developed without any consideration of joint occurrence of arbitrary facial occlusion, arbitrary and non-uniform lighting, arbitrary viewpoints and arbitrary facial expression. Rather, they [13, 9, 18, 17] focus on analyzing images (or video frames) with limited degrees of freedom. For instance, some work well on multi-view (e.g. specific, non-arbitrary) face images, but with optimal and fixed indoor lighting, whereas some others work well on video frames from only frontal faces, arbitrary background clutters and optimal indoor illumination, but without any occlusions, or non-uniform lighting (e.g. video databases obtained from TV broadcasts news).

The proposed methodology in this paper presents the first attempt to achieve gender classification from face images acquired from *totally unconstrained* video sequences where the person is unrestricted in terms of facial expression, viewpoint, illumination, occlusion, etc. (see Figure 1). We present a Bayesian framework for classifying faces in video sequences acquired from realistic unconstrained viewpoints and where cluttered background, arbitrary and non-uniform illumination, and free movement of the sub-

978-1-4244-7028-0/10/$26.00 ©2010 IEEE

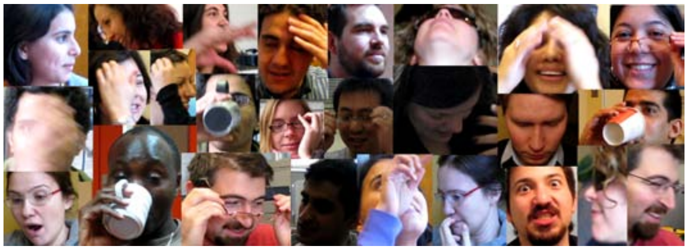


Figure 1. Sample faces from video frames in our free-form (unconstrained), in-house video database. Note that the original image-scales are preserved.

jects are permitted in the scene. As a precursor to classification, we first require a robust face detector to work under these conditions. To this end, we found that the viewpoint invariant face detection algorithm found in [22] worked better than the one of [24] under such arbitrary conditions. SIFT [16] features obtained from training images are clustered -according to the algorithm in [22]- to learn the most robust features differentiating between female and male. The face class is modeled - on binary presence/absence of these SIFT features - as a generative Markov model, where features in a detected and tracked face are assumed to arise from a single person of a particular class.

Model parameters can be estimated off-line from a set of still face image pairs acquired from nearby viewpoints about the same face, thereby approximating adjacent views of a video sequence. Thus, the proposed Markovian temporal model is first trained on a well-established still face image database (the FERET[1]) of uniform illumination, noise-free, fixed face viewpoints, and the absence of major occlusion. The model is then evaluated on faces detected from in-house free-form (unconstrained) video sequences - with no restrictions on the person's movements, background cluttering, occlusion or lighting- by determining the maximum *a posteriori* class of a face given a collection of image features associated with an individual face. Evidence in the trait class (e.g. gender) is accumulated probabilistically over the entire video sequence, thus permitting quick and robust classification over time. As the face classification amid viewpoint changes in video sequences is relatively a new area of research, public datasets are rare, and we demonstrate the effectiveness of our approach on a new, proprietary video dataset. We compiled a free-form face video database of 9000 video frames from 30 unique subjects (15 males and 15 females), exhibiting a high degree of appearance variability over time due to changes in viewpoint, illumination, facial expression, degree of occlusion (sunglasses, cups, hands), etc. Our experimental results provide a detection rate of 90% by the last video frame and show that the proposed classifier significantly outperforms alternative approaches [22, 21, 20, 5] (Section 4.4) and the "bag-of-frames" (Section 4.3), an approach that classifies each frame of a video sequence separately and then considers the results from the complete set of frames from the sequence when asserting the final classification.

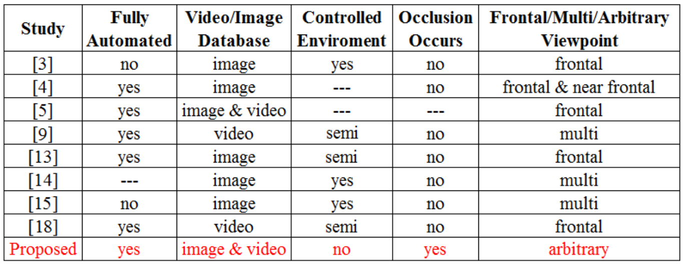
2. Related Work

Moghaddam and Yang [20] used the SVM in conjunction with pixel intensity values, and reported one of the highest classification accuracies in the literature. However, their reported results were based on experiments on the FERET database [1], which was collected under controlled conditions. Furthermore, since they only used frontal face images, a common geometric shape alignment and masking were possible to apply on all face images. However, such an alignment and masking processes are not applicable to the faces from our unconstrained video database.

Shakhnarovich et al.[21], to our knowledge, were the first to classify gender from video frames. Their automatic face detection algorithm was based on that of Viola&Jones' [24]. Shakhnarovich adopted the Haar-like features [24] and used Adaboost for feature selection purpose. For experimental purpose, an image database from World Wide Web (WWW) and a video database from 30 subjects were collected. However, the images with faces more than $+/-$ 30 degree off frontal orientation were removed from the WWW image database. The video database, on the other hand, captured faces with different poses and expressions, but under controlled illumination conditions. For the image data set, the SVM with RBF kernel was used on Adaboost selected Haar features for the gender classification purpose whereas for the video database, the fusion formula in Equation 10 was utilized (see Section 4.4).

The object class invariant (OCI) approach of Toews and Arbel [22] created a viewpoint-invariant appearance model for face detection. The model used local features due to their high degree of invariance to various transforms, i.e. SIFT [16], to probabilistically model a robust geometrical model based on which faces are detected and localized. Later, the model features were used for gender classification from multi and arbitrary viewpoints. However, comparing the approach in [22] to the proposed one, their work presents preliminary work examining gender detection from static-face images -not from video sequences- with no motion blur or arbitrary occlusions and with no discussion as to how to accumulate information over single frames. Their results were based on a very small database of 132 faces (highly biased towards males, i.e. 100 males). For video sequences presented in this paper, our approach outperforms the approach in [22] by 20% (Section 4.3).

Figure 2 summarizes several other approaches introduced in the literature on the topic.



Study	Fully Automated	Video/Image Database	Controlled Enviroment	Occlusion Occurs	Frontal/Multi/Arbitrary Viewpoint
[3]	no	image	yes	no	frontal
[4]	yes	image	---	no	frontal & near frontal
[5]	yes	image & video	---	---	frontal
[9]	yes	video	semi	no	multi
[13]	yes	image	semi	no	frontal
[14]	---	image	yes	no	multi
[15]	no	image	yes	no	multi
[18]	yes	video	semi	no	frontal
Proposed	yes	image & video	no	yes	arbitrary

Figure 2. Comparison of different approaches in gender classification literature.

3. Proposed Methodology

Consider a sequence of video frames of a face moving in an uncontrolled manner through a scene. At any individual frame, the person can be found at arbitrary viewpoints (e.g. due to the head movements) with respect to the camera, or partially occluded, etc. The task of the system is to classify the gender of the person in the scene based on the acquired collection of frames. The key to the success of the algorithm described here is to first detect and localize faces in the scene and to acquire a set of features from the faces under such arbitrary conditions. This pre-processing step is not the focus of the work described here and other approaches can be examined (see Section 4.3). Once acquired, these features can then be used to classify the face according to traits, like gender. We now define a Bayesian formulation for the classification of faces over image sequences.

Let $F_t = \{f_{t,1}, \ldots, f_{t,N}\}$ represent a set of N features extracted from an image frame at time t. Each single feature, $f_{t,j}$, is a binary feature representing the occurrence or non-occurrence of a potentially class-related feature. Let C denote the random variable for a face trait class. The posterior probability of the face class C at time t given features extracted in all previous frames $F_t, \ldots, F_1$ is:

$$p(C|F_t, F_{t-1}, \ldots, F_1), \tag{1}$$

where C can generally take on any face trait class value. In this paper, we consider the binary trait of gender (i.e. male or female). Therefore $C = c$ or $C = \bar{c}$, where c and $\bar{c}$ are opposing genders. The optimal Bayes classification at time t is:

$$c* = \underset{c}{\text{argmax}} \left\{ \log \frac{p(C = c|F_t, F_{t-1}, \ldots, F_1)}{p(C = \bar{c}|F_t, F_{t-1}, \ldots, F_1)} \right\}. \tag{2}$$

One can define the posterior probability density function over frames, such that:

$$p(c|F_t, F_{t-1}, \ldots, F_1) = \frac{p(F_t, F_{t-1}, \ldots, F_1|c)}{p(F_t, F_{t-1}, \ldots, F_1)} p(c), \tag{3}$$

where $p(c)$ is the *a priori* probability on the class trait value c, $p(F_t, F_{t-1}, \ldots, F_1|c)$ is the likelihood for the class trait over all the features in the video sequence, and $p(F_t, F_{t-1}, \ldots, F_1)$ is the joint probability density function over all the features. Alternatively, utilizing the Chain rule, one can define the posterior probability density function recursively over frames, such that:

$$p(c|F_t, F_{t-1}, \ldots, F_1) = \frac{p(F_t|F_{t-1}, \ldots, F_1, c)}{p(F_t|F_{t-1}, \ldots, F_1)} p(c|F_{t-1}, \ldots, F_1). \tag{4}$$

In Equation (4), the posterior density function $p(c|F_{t-1}, \ldots, F_1)$ for the class given the features from all the previous frames therefore acts as a prior for the current frame at time t.

Assuming that the face moves slowly relative to the video frame rate, significant dependencies will exist between facial feature sets F_t and F_{t-1}.

We make the following first-order Markov assumption:

$$p(c|F_t, F_{t-1}, \ldots, F_1) \propto p(F_t|F_{t-1}, c) p(c|F_{t-1}, \ldots, F_1). \tag{5}$$

The implication of the Markov assumption is that the feature set F_t is conditionally independent of features in all previous frames $F_{t-2}, \ldots, F_1$ given features F_{t-1} in the most recent frame and the class value c. The density $p(F_t|F_{t-1}, c)$ can be learned from the training set. Given the fact that the speed of the person's movement between frames is unknown and given the strong possibility of occlusion, one can make further simplifications by invoking the Naive Bayes assumption regarding the relationships between the features. Namely, that individual features at time t, $f_{t,i}$, are conditionally independent from other features in the previous frame, given trait value c. This leads to:

$$p(F_t|F_{t-1}, c) = \prod_{i=1}^{N} p(f_{t,i}|f_{t-1,i}, c). \tag{6}$$

During training, one can then define a probability lookup table which compiles the occurrence frequencies for all possible combinations of the feature pairs $f_{t,i}, f_{t-1,i}$ for a given trait value c over the entire database. The assumption of feature independency is reasonable for our model since it is based on spatially separated SIFT features, not on pixels. This assumption allows us to properly address the commonly occurring problem of partial facial occlusion, where many detection and classification frameworks fail to handle.

Embedded within the recursive formulation of the Equation (5) for the posterior is the likelihood term for the first frame $p(F_1|c)$. In general, under the assumption that individual features in a frame t, $f_{t,i}$, are conditionally independent given trait value c, one can estimate $p(F_t|c)$ as:

$$p(F_t|c) = p(f_{t,1}, \ldots, f_{t,N}|c) = \prod_{i=1}^{N} p(f_{t,i}|c), \qquad (7)$$

where $p(f_{t,i}|c)$ defines the probability of the particular feature in frame t given a class, c. It can be estimated off-line during training as in [22]:

$$p(f_{t,i}|c) \propto \frac{k(f_{t,i}, c)}{p(c)} + d_t, \qquad (8)$$

where $k(f_{t,i}, c)$ is the count of the joint occurrence event $(f_{t,i}, c)$ and $p(c)$ is the probability of occurrence of trait value c. d_t is the Dirichlet regularization parameter required to compensate for the sparsity of the feature occurrences. As a uniform prior is assumed, d_t is thus constant for all t.

4. Experiments

We have conducted several experiments on the collected free-form video database: (1) Evaluation of the robustness of the utilized face detector [22] and the cascaded face detector [24] over random viewpoints of face; (2) Evaluation of the performance of the methodology (SVM with pixel intensity) in [20] which reported the highest gender classification accuracy in the literature; (3) Evaluation of the previously introduced classifier fusion methods for gender classification by Shakhnarovich et al. [21] and by Castrillon-Santana et al. [5]; (4) Evaluation of the proposed Markovian temporal model, and comparison of this approach and the method introduced in [22].

4.1. Experimental Setup

For training purposes, we built a face database from 445 female and 445 male FERET [1] subjects of various ethnicities under controlled illumination conditions, with/without glasses, etc. As a pre-processing step, color FERET images with resolution of 256x384 pixel were first converted to grey scale. For each of the 890 subjects, 5 viewpoint images were used (2 profile, 2 quarter and 1 frontal), for a total of 4450 images (see Figure 3). This database is used by the OCI face detector [22] and by the proposed Markovian temporal model to learn gender under controlled conditions. Our gender classifier should then prove to be robust to variations from these ideal conditions when tested under arbitrary conditions.

Since there is no standard annotated face video database in the public domain, for testing purposes, we collected a video database of 30 unique subjects (15 females and 15 males) using a Canon PowerShot SD770 camera. For each subject, a 60-second video with 30 fps (1800 frames per subject) and 640x480 resolution was recorded. The subsampling of frames was empirically set to 5 frames per second, leading to 300 × 30 = 9000 video frames in total to evaluate the classifiers in Section 4.4 and Section 4.3. The major guideline we followed during the video collection was not to restrict our subjects to any kind of controlled motion or environment. Unlike many video or face image databases used in the face classification literature, each video sequence was shot under different illumination and background conditions. Furthermore, our subjects were free to move as they wanted, resulting in arbitrary face scales, expressions, viewpoints, local and/or global occlusions (due to closed eyes, glasses, hand, coffee cup, scarf or hat). Figure 1 illustrates such challenges depicted by our in-house video database. The only restriction was made to have a single person in the scene of each video clip.

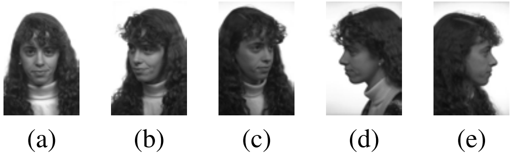


Figure 3. Viewpoint images of a training subject: (a) frontal, (b) left quarter, (c) right quarter, (d) left profile, and (e) right profile.

Both in training and in testing databases, the male:female ratio was kept 1:1, so as to avoid biasing any gender class (i.e. $p(c)$ in Equation (3) and Equation (8) is uniform). Similarly, it is important that the distributions of male and female data over viewpoint in the video database do not demonstrate any significant gender-related bias. To this end, we manually labeled the viewpoint angle in all 9000 video frames of our test data. The viewpoints are represented by one of the following angles: [-90 (left profile), -75, -60, -45, -30, -15, 0 (frontal) , 15, 30, 45, 60, 75, 90 (right profile)]. Our observations show that the video database has nearly equal proportions of female and male faces in each viewpoint. Furthermore, we have a broad variety of viewpoints for each subject in our video database. Thus, our experimental results are not biased by any specific viewpoint or subject or gender class.

4.2. Evaluation of Face Detection in Free-form Video Sequences

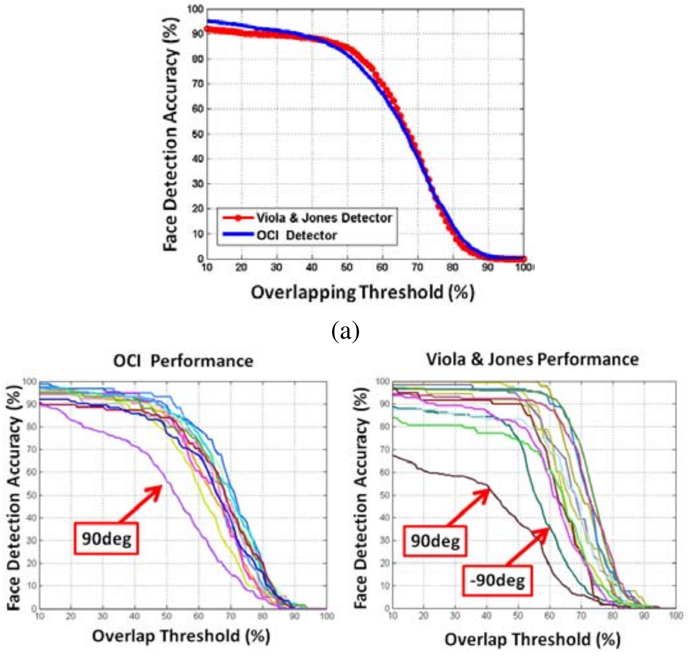
Prior to the gender classification phase, face detection and facial feature extraction steps are required, however they are not the focus of the work presented in this paper. Considering our test video database, we need to detect and localize the face in the scene and extract the gender features robustly regardless of the face viewpoint changes, partial occlusions and global/local illumination changes. Face localization and detection can be addressed via various techniques [25, 26, 24, 23], whereas the crucial step is deciding on which features to use. A variety of different image features have been for used modeling the face, including global features, such as principal components [23, 20] or independent components [11], and local features, such as Haar wavelets [24, 21] or scale-invariant features [19, 16]. In general, global features are sub-optimal in terms of detection performance, while local features can be robustly identified amid factors, such as the occlusion and face variability. Haar wavelets are not invariant to image rotation.

In terms of face detection from free-form video sequences, we chose to examine two face detectors. We first examine the Viola & Jones' cascaded detector [24], one of the best and most well-known detectors available, due to its high speed, accuracy and reliable open source implementation. We next examine the OCI face detector [22], a viewpoint-invariant detector which can locate faces in natural images by estimating a face reference frame from local features (e.g. SIFT [16]). The viewpoint-invariant face model (OCI) uses the most commonly occurring SIFT features in the face images to represent the image regions containing faces.

We evaluated the OCI [22] and the cascaded [24] face detectors on the in-house free-form video database. The purpose of this experiment is to i) select a face detector that is robust to any changes in facial pose, and ii) to ensure that our final classification results are not biased by a dominant false detection of a specific facial pose.

The implementation of the Viola & Jones' face detector provided by OpenCV 1.0 [2] is used. Different cascaded detector models obtained from different viewpoint training databases are employed. The decision as to which possible detection to use for a frame is done by using the ground truth. We basically choose the detection which maximizes the intersection area between the ground truth bounding box and the detected bounding box. Note that while this resolution strategy for detector responses cannot be used in a practical setting without ground truth; here it allows us to consider a generous, best-case scenario for Viola & Jones' detection [24].

The comparison of two face detectors is shown in Figure 4. If the evaluation is done without considering the variations in facial pose, the OCI and Viola & Jones approaches seem to have a very comparable detection accuracy (see Figure 4 (a)). However, it is crucial to note that the OCI detector is only trained on 890 random viewpoint FERET images, whereas the cascaded detector is trained over thousands of images. The shape of the accuracy curves are very similar due to the fact that the collected in-house video sequence database is not biased by any viewpoint (see Section 4.1).



(a)

(b)

Figure 4. Comparison of the OCI and the cascaded face detectors: (a) average face detection performance over the whole in-house video sequence database, (b) face detection performance evaluated separately for each facial pose.

More interestingly, as seen in Figure 4 (b), Viola's detection has a large variance in accuracies over different viewpoints. Here we can see the detector accuracy against the overlap threshold used for detection, the latter defined as the cutoff for the percentage of overlap between the ground truth and the detected facial regions. For a 60% threshold, for example, the difference between the maximum and minimum detection rate for different poses goes up to 75% for the Viola detector. On the other hand, the OCI detector has a more concentrated graph in Figure 4 (b), which shows that this face detector performs consistently and robustly over viewpoint. As a point of comparison, the difference between the best and the worst detection accuracy for a 60% threshold for the OCI is only 45%. The only viewpoint angle for which the OCI detector [22] performance drops is at 90 deg profile images, which is also the viewpoint with the lowest accuracy for the Viola cascaded detector [24]. Furthermore, it should also be noted that Viola & Jones' detector performs well for frontal and near frontal face images, which makes it the optimal face detector when using only frontal face image databases [21, 17].

4.3. Experiments with Our Methodology

For the reasons cited above, we chose to use the viewpoint-invariant appearance (OCI) model [22]. The parameters of the OCI model used in this work are based on those found in [22]. Once the OCI model is robustly learned from the training dataset, it can be used on video frames in the test database to detect and localize face regions and obtain the facial features. Furthermore, facial features extracted from training images are used to estimate the probabilities mentioned in Section 3.

After learning the OCI face detection model from 890 FERET images, the rest of the database, i.e. $890 \times 4 = 3560$ images, was used to learn the proposed Bayesian classifier, namely the probabilities in Equation (8) and the probability lookup table. It is important to note that the Markov transition probabilities in the probability lookup table were estimated by mirroring the left profile (see Figure 3 (d)), the left quarter (see Figure 3 (b)) and the frontal (see Figure 3 (a)) images of each FERET subject. As the face images of the same person were approximately mirror-symmetric, image pairs obtained by mirroring here simulated the adjacent frames of a video sequence, from which the features of F_{t-1} and F_t were obtained.

4.3.1 Experimental Results

Once we obtained video sequences, we examined different metrics to determine the gender class for each frame. The simplest and probably the most widely accepted baseline in the literature would be to treat each frame individually and make the trait decision on frame by frame basis. For this purpose, we used the viewpoint-invariant classifier in [22] on each video frame as if it were a static image and obtain the decision, c^*, independently of all other frames:

$$c^* = \underset{c}{\operatorname{argmax}} \left\{ \log \frac{p(c|F_t)}{p(\bar{c}|F_t)} \right\}. \tag{9}$$

F_t is the set of face related SIFT features obtained at time $\texttt{t}$. The classification decision for 9000 video frames is achieved via applying the threshold at which the probabilities of misclassifying males and females are equal, i.e. the equal error rate (EER). The method introduced in [22] achieves an accuracy of 70% on 9000 video frames, which is relatively low for high accuracy needs in real-world biometric security applications, such as video surveillance.

As the next step, we accumulated evidence over individual frames probabilistically so as to increase the overall accuracy. To our knowledge, the gender classification literature does not discuss any gender classifier that accumulates viewpoint invariant-information throughout video sequences, either at a feature-level or at a decision-level. Thus, we conduct two experiments where we introduce and compare two different fusion schemes: (i) decision-level information fusion on classifier in Equation (9), here referred to as a bag-of-frames (BOF) model, and (ii) decision-level information fusion on the proposed classifier which probabilistically accumulates evidence (see Section 3). After the frame scores are obtained via the metric in Equation (9) and the proposed metric in Section 3, the majority voting operator is applied to them. Majority voting is achieved by looking at all the class decisions for frames at times $\{\texttt{1,...,t-1,t}\}$ and deciding the final class label as the one that occurs most often. The class decision for each frame (see Equation (2)) requires a fixed threshold which is determined empirically, e.g. the threshold of 61 is used for the experiment in (i) and 51 in (ii) where the normalized score range is $[0...100]$.

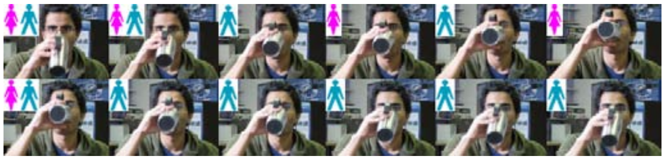


Figure 5. Consecutive video frames from a subject's short clip with the obtained class decisions via the proposed Bayesian classifier. Top left corner shows the signs for class decisions, i.e. female (pink) and male (blue). Note that this particular figure displays only frontal faces, however our video database contains any images containing any number of random viewpoints (see Section 4.1 and Figure 1).

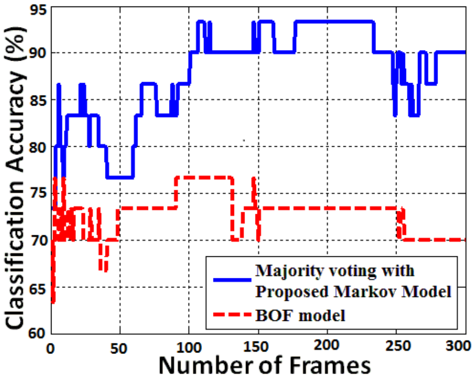


Figure 6. Classification accuracy comparison of different fusion schemes for 30 subjects with 300 frames per subject.

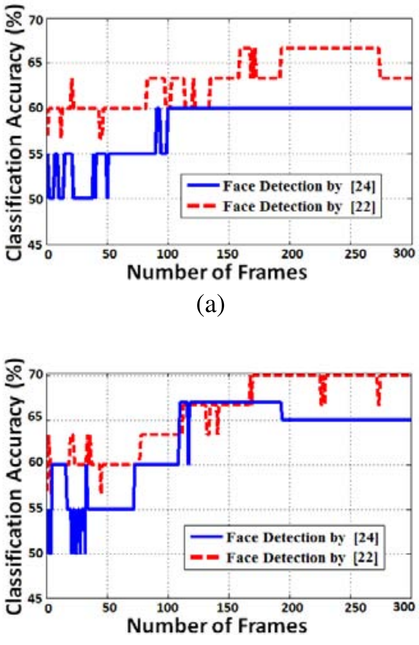
Figure 5 shows consecutive frames of a short video clip from our video database with corresponding class decisions obtained by the proposed approach. One can see that the class decisions for the first few frames are either inconclusive or inconsistent. However, as we accumulate more evidence over frames, the correct class decision is obtained repetitively regardless of the large occlusion that the coffee cup introduces to the system.

As Figure 6 shows, gender classification accuracy increases as we have more frames regardless of which fusion scheme is used, leading to the accuracies of 70% for the BOF approach and 90% for the proposed approach at the end of the video sequences, i.e. 300th frame. At first few frames, both of the fusion schemes start with similar accuracies, whereas as we accumulate statistical information from the previous frames, majority voting starts to give much better accuracies for (ii) compared to the simpler fusion scheme of (i). We believe that the feature accumulation provides a better, much robust classification performance because the approach in Section 3 accumulates information over the entire set of facial features.

It is important to note that at frame #132, the first largest gap between the BOF approach in (i) and our methods in (ii) is observed: gender classification accuracies of 70% and 90% are obtained by the algorithms in (i) and (ii), respectively. The maximum classification accuracies are 76.6% for (i), whereas it is 93.3% for the proposed approach in (ii). Furthermore, it is observed that despite the constant increase in the classification accuracy for both schemes, both algorithms suffer from local accuracy drops for some frames (e.g. around frames #50, and #250). Our observations indicate that when the face localization is very off the ground truth in a number of consecutive video frames, - although the effect of poor localization is minimized by using local features- we may obtain poor classification results since the number of extracted SIFT features can be very small. We observe that certain viewpoints, such as the ones where the subject is looking down, lead to failure in face localization. It is due, in part, to a failure of the underlying assumptions of the OCI model. In addition, the training database does not capture all the viewpoint variations in the videos. This might be fixed by having a larger training database containing a broader variety of viewpoints beyond the limited viewpoints provided in the FERET[1] database.

4.4. Evaluation of the SVM Classification and Pixel Intensity-based Features in Free-form Video Sequences

As we mentioned in Section 2, there are several papers written on gender recognition, but none of them addresses the problem in this paper. Thus, we decided to select the approaches in pioneering papers, such as [20], [21] and [5], and apply them on our challenging database in order to compare their performances to ours. We first examined the approach in [20] for classifying the 9000 video frames from our in-house video database. Different image normalization techniques and image sizes are empirically investigated. The best gender classification accuracy is obtained by using no-normalization and downsampling detected face images to 24x24 (similar to the findings in [17]). We have trained the SVM classifier on the FERET database. The SVM implementation provided by LIBSVM [6] is used in our experiments. The best SVM parameters for RBF kernel ($\gamma = 0.000488$, $C = 8$) were obtained using a grid search and a 5-fold cross-validation. Average SVM classification accuracy over all 9000 video frames of 65.6% is obtained when the OCI face detector is used, whereas it is 60.2% for the cascaded face detector case. We have observed that regardless of which detection algorithm is used, the SVM classifier tends to tag males as females.



(a)

(b)

Figure 7. Performance of the SVM classifier with different temporal fusion methods and face detectors: (a) the SVM classifier with Majority voting, (b) the SVM classifier with fusion approach in equation 10.

We also examined majority voting to fuse gender class labels obtained from each frame [5]. As it is shown in Figure 7 (a), in terms of classification accuracy, the classifier that uses faces detected by OCI [22] outperforms the one based on the results of the cascaded face detector [24]. The the combination of {SVM+pixel intensity+Majority+OCI face detector} results in a gender classification rate of 63.3% by the last video frame (see Figure 7 (a)).

Finally, we adopted the proposed classification fusion method by Shakhnarovich et al. [21]:

$$D(T) = \frac{1}{T}\sum_{i=0}^{T} e^{-\alpha i} V(f(x_{t-i}))Q(x_{t-i}), \qquad (10)$$

where $D(t)$ is an exponentially weighted sum of classifier outputs, $f(t)$, fom the past T frames. $V(x)$ is a voting function and $Q(x)$ is a quality function. In our experiments, after investigating various voting and quality functions (such as degree of motion blur), a linear ramp, $V(f(x_{t-i})) = f(x_{t-i})$, and a uniform quality, i.e. $Q(x_{t-i}) = 1$, were used. The obtained result (Figure 7 (b)) is better than the one we obtain with using majority voting (Figure 7 (a)). Furthermore, none of the approaches examined manages to outperform our temporal method.

5. Discussion

We present a new Bayesian framework which, for the first time in the gender classification literature, achieves robust classification from free-form face video sequences. The proposed framework was first trained on FERET's face images and then evaluated on a video database presenting challenging classification scenarios. Video sequences were collected under different illumination and background conditions, where each subject was free in his/her movements, resulting in various face expressions, viewpoints, scales and occlusions. The proposed system achieved high gender classification performance (90%) considering the fact that it was trained on still face image database collected under controlled environment. Our Bayesian temporal model achieved a superior classification performance compared to its alternative approaches, reaching a performance increase of up to 30% depending on the selected alternative approach in Section 4.4 and 4.3. Our approach not only does a classifier level fusion, but also utilizes the occurrence statistics of the features, unlike any other alternative methods available in the literature. In this paper, we utilized SIFT features that were robust to the changes in scale, viewpoint, rotation, translation and occlusion. However, any other features could be tried without changing the underlying method.

We are currently investigating a number of avenues for future work. We intend to extend our classification formulation and our free-form video database to explicitly account for uncertainty in detection and tracking, in order to classify faces in crowded scenes. Moreover, we are currently compiling a comprehensive, annotated database of free-form video sequences with the goal of having 50 unique subjects, which will soon be publicly available.

References

[1] Color feret face database. http://face.nist.gov/colorferet/. 2, 4, 7

[2] Opencv. In *http://www.intel.com/technology/computing/opencv/*, 2006. 5

[3] S. Baluja and H. A. Rowley. Boosting sex identification performance. *IJCV*, 71(1):111–119, 2007. 1

[4] C. BenAbdelkader and P. Griffin. A local region-based approach to gender classification from face images. In *CVPR*, 2005. 1

[5] M. Castrillon-Santana, O. Dniz-Surez, J. Hernndez-Sosa, and D.-B. A. Identity and gender recognition using the encara real-time face detector. In *Proc. Conf. Assoc. Espaola para la Inteligencia Artificial (CAEPIA)*, 2003. 2, 4, 7

[6] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. In *http://www.csie.ntu.edu.tw/ cjlin/libsvm/*, 2001. 7

[7] M. Demirkus, K. Garg, and S. Guler. Automated person categorization for video surveillance using soft biometrics. In *Proc of SPIE, Biometric Technology for Human Identification VII*, 2010. 1

[8] S. Gutta, H. Wechsler, and P. Phillips. Gender and ethnic classification of human faces using hybrid classifiers. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 194–199, 1998. 1

[9] A. Hadid and M. Pietikinen. Combining appearance and motion for face and gender recognition from videos. *Pattern Recognition*, 42(11):2818 – 2827, 2009. 1

[10] Y. Hu, Y. Fu, U. Tariq, and T. S. Huang. Subjective experiments on gender and ethnicity recognition from different face representations. In *Int'l Conf on Multi-Media Modeling*, 2010. 1

[11] A. Jain, J. Huang, and S. Fang. Gender identification using frontal facial images. In *ICME*, pages 1082–1085, 2005. 1, 5

[12] H. Kim, D. Kim, Z. Ghahramani, and S. Y. Bang. Appearance-based gender classification with gaussian processes. *PRL*, 27:618–626, 2006. 1

[13] A. Lapedriza, M. J. Maryn-Jimenez, and J. Vitria. Gender recognition in non controlled environments. In *ICPR*, pages 834–837, 2006. 1

[14] J. Li and B. Lu. A framework for multi-view gender classification. In *Neural Information Processing*, pages 973–982, 2008. 1

[15] H. Lian and B. Lu. Multi-view gender classification using local binary patterns and support vector machines. In *Third International Symposium on Neural Networks*, volume 2, pages 202–209, 2006. 1

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 5

[17] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. 30(3):541–547, 2008. 1, 5, 7

[18] F. Matta, U. Saeed, C. Mallauran, and J.-L. Dugelay. Facial gender recognition using multiple sources of visual information. In *IEEE MMSP*, 2008. 1

[19] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 5

[20] B. Moghaddam and M. Yang. Learning gender with support faces. *IEEE TPAMI*, 24(5):707–711, 2002. 1, 2, 4, 5, 7

[21] G. Shakhnarovich, P. A. Viola, and B. Moghaddam. A unified learning framework for real time face detection and classification. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2002. 1, 2, 4, 5, 7

[22] M. Toews and T. Arbel. Detection, localization and sex classification of faces from arbitrary viewpoints and under occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. 1, 2, 4, 5, 6, 7

[23] M. Turk and A. P. Pentland. Eigenfaces for recognition. *CogNeuro*, 3(1):71–96, 1991. 5

[24] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 2, 4, 5, 7

[25] S. Yan, H. Wang, J. Liu, X. Tang, and T. S. Huang. Misalignment-robust face recognition. *IEEE Transactions on Image Processing*, 2009. 5

[26] S. K. Zhou, R. Chellappa, and W. Zhao. *Unconstrained Face Recognition*. Springer-Verlag New York, Inc., 2005. 1, 5