Gender Classification from Unconstrained Video Sequences[1]

Meltem Demirkus[1] Matthew Toews[2] James J. Clark[1] Tal Arbel[1]

[1] Centre for Intelligent Machines, McGill University, Montreal, Canada [2] Department of Radiology, Harvard Medical School, Boston, USA

1. Motivation and Problem Definition

Face classification has been receiving a wide amount of attention recently, especially in the context of video surveillance. However, it is a challenging task due to the joint occurrence of arbitrary head poses, face scale changes, non-uniform illumination conditions and partial occlusion present in real video surveillance images.

image courtesy of [5]

Even though several approaches claim to perform successful face classification from unconstrained environments, these algorithms require pre-processing steps which are difficult to achieve in real-world, unconstrained environments, such as the requirement for face alignment (e.g. with no large variability in head pose permitted), or the requirement for specific facial regions to track (e.g. no occlusion is allowed). 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 scene is unrestricted in terms of facial expression, head viewpoint change, occlusion and illumination.

2. Methodology

Training Phase

Still Face Image Database Collected Under Controlled Conditions (e.g. FERET)

OCI Model Learning[2]

Manual facial axis labelling

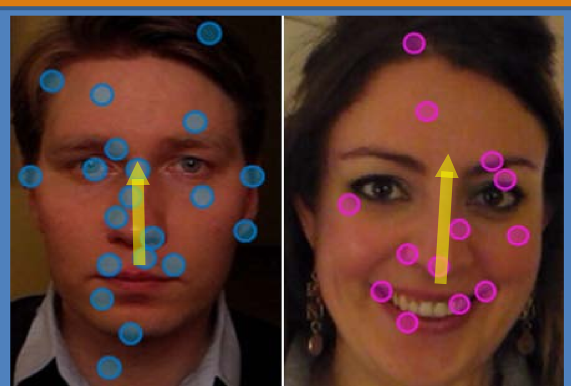
$t=1$ $t=2$ $t=3$ $t=4$ $t=5$

sample images from FERET database

Local (e.g. SIFT[3]) feature extraction

Clustering features based on appearance (gradient histogram), geometry (scale and orientation) and relative location according to facial axis

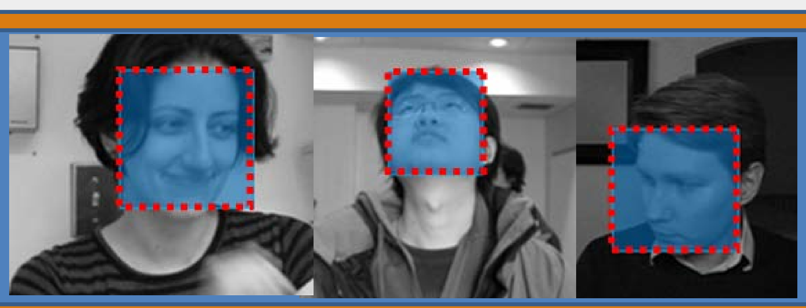
Obtain a ***feature bank*** containing discriminative features of facial characteristics, which will be used for both face detection and gender classification



Sample SIFT features associated with gender class (McGill Face Video Database)

Bayesian Gender Classifier Training

Local feature ($F_t = \{ f_{t,1}, \ldots, f_{t,K} \}$) extraction and face detection



Compare each feature ($f_{t,i}$) to the ***feature bank*** and obtain a binary feature occurrence map

Learn the following gender-feature probabilities:

$p(f_{t,i} \mid female)$ and $p(f_{t,i} \mid male)$

Learn joint occurrences of local feature probabilities:

$p(f_{t,i} \mid f_{t-1,i}, female)$ and $p(f_{t,i} \mid f_{t-1,i}, male)$

Testing Phase

Unconstrained Face Video Database

Frame at time t

LAST FRAME?

Y → Finish

N → Local feature extraction and face detection from frames at time t and $t-1$

Compare each local feature ($f_{t,i}$ and $f_{t-1,i}$) to the ***feature bank*** and obtain a binary feature occurrence map

Apply the proposed **Temporal Bayesian Classifier** to obtain gender class probability at time t

GO TO NEXT FRAME

[1] M. Demirkus, M. Toews, J. Clark, T. Arbel, "Gender Classification from Unconstrained Video Sequences", IEEE Workshop on Analysis and Modelling of Faces and Gestures (AMFG) held in conjunction with IEEE CVPR 2010, San Francisco, CA, June 2010.
[2] 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, Vol. 31, Issue 9, pp. 1567-1581, 2009.
[3] D. G. Lowe, " Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, Vol. 60, Issue 2 , pp. 91–110, 2004.
[4] G. Shakhnarovich, P.A. Viola, and B. Moghaddam, "A unified learning framework for real time face detection and classification", Int. Conf. on Automatic Face and Gesture Recognition, 2002.
[5] M. Demirkus, K. Garg and S. Guler, " Automated person categorization for video surveillance using soft biometrics", SPIE Biometric Technology for Human Identification VII, April 2010.
[6] P. Viola and M. J. Jones, "Robust real-time face detection", International Journal of Computer Vision, Vol. 57, Issue 2, pp. 137–154, 2004.

3. Temporal Bayesian Classifier

Given binary occurence map of N local features from video frame at time t: $F_t = \{ f_{t,1}, \ldots, f_{t,N} \}$

Find $p(C \mid F_t, F_{t-1}, \ldots, F_1)$ where C represents female (c) or male ($\bar{c}$) class.

Optimal Bayes classifier at time t is:
$$c^* = \arg\max_c \left\{ \log \frac{p(C=\bar{c} \mid F_t, F_{t-1}, \ldots, F_1)}{p(C=c \mid F_t, F_{t-1}, \ldots, F_1)} \right\} \quad \text{(Eq.1)}$$

To find the optimal classifier we need to calculate the following formulation for both c and $\bar{c}$.

$$p(c \mid F_t, F_{t-1}, \ldots, F_1) = \frac{p(F_t, F_{t-1}, \ldots, F_1, c)}{p(F_t, F_{t-1}, \ldots, F_1)} = \frac{p(F_t \mid F_{t-1}, \ldots, F_1, c)}{p(F_t \mid F_{t-1}, \ldots, F_1)} p(c \mid F_{t-1}, \ldots, F_1) \quad \text{(Eq.2)}$$

Applying 1st-order Markovian assumption on Eq. 2

$$p(c \mid F_t, F_{t-1}, \ldots, F_1) \propto \frac{p(F_t \mid F_{t-1}, c)\, p(c \mid F_{t-1}, \ldots, F_1)}{p(F_t \mid F_{t-1}, \ldots, F_1)} \quad \text{(Eq.3)}$$

Considering the log ratio in Eq.1, $p(F_t \mid F_{t-1}, \ldots, F_1)$ term can be ignored

$$p(c \mid F_t, F_{t-1}, \ldots, F_1) \propto p(F_t \mid F_{t-1}, c)\, p(c \mid F_{t-1}, \ldots, F_1) \quad \text{(Eq.4)}$$

Sequential Update !

where

$$p(F_t \mid F_{t-1}, c) = \prod_{i=1}^{N} p(f_{t,i} \mid f_{t-1,i}, c) \quad \text{(Eq.5)}$$

$$p(F_1 \mid c) = \prod_{i=1}^{N} p(f_{1,i} \mid c) \quad \text{where} \quad p(f_{t,i} \mid c) \propto \frac{k(f_{t,i}, c)}{p(c)} + d_t \quad \text{(Eq.6)}$$

(d_t is the Dirichlet parameter and k is the frequency function)

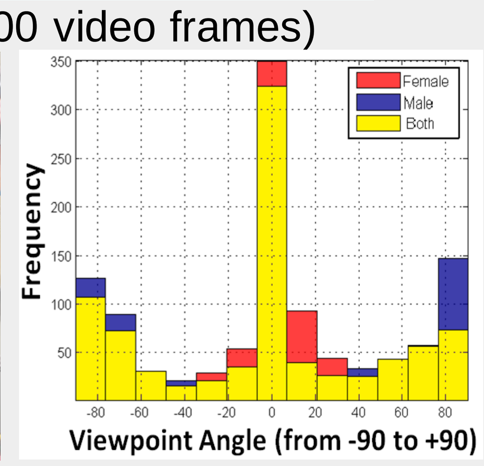
Probabilities Obtained at Bayesian Classifier Training Phase

4. Experimental Results and Conclusion

Local Features: SIFT[3] features are used in our experiments due to their robustness to various illuminations, face scales, head poses (change both in rotation and translation) and partial occlusions.
Training Database : 4450 FERET Images (890 Unique Subjects with 5 viewpoint images per subject)

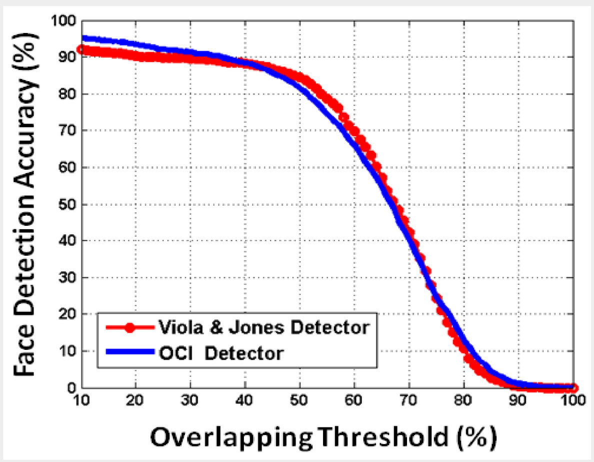
McGill Fully Unconstrained Face Video Database

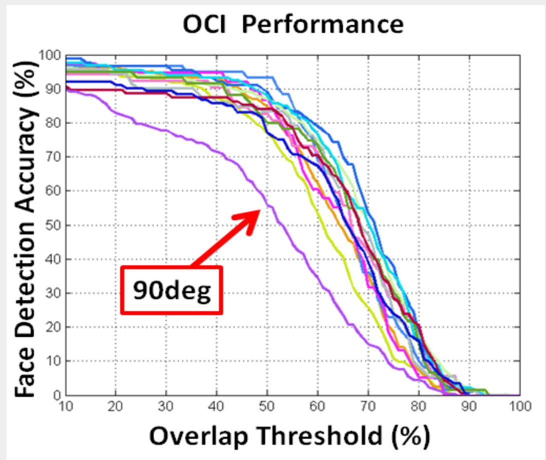
30 unique Subjects with 300 video frames per subject (9000 video frames)

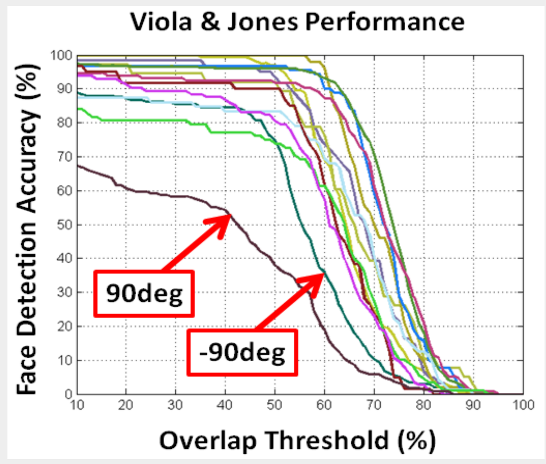


Face Detection in Unconstrained Video Sequences

Comparison of Viola&Jones[6] and OCI[2] Face detectors over 9000 video frames:



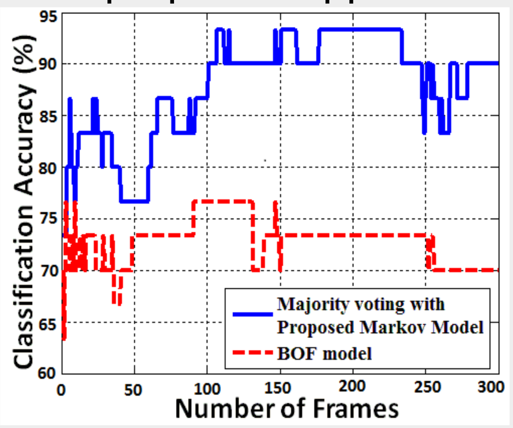


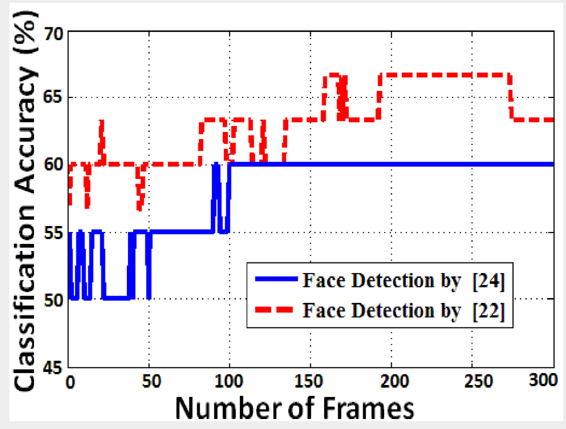


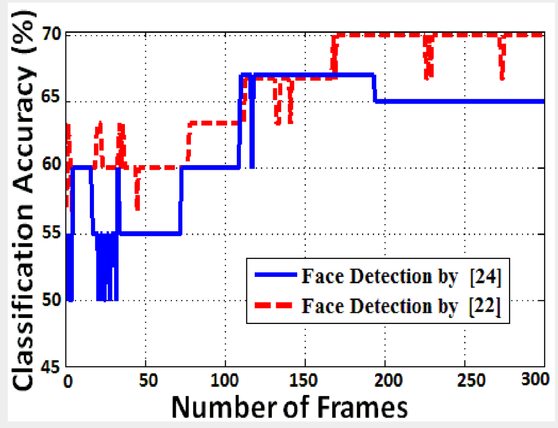
Similar detection accuracy Viola&Jones → larger variance over different head poses

Proposed Method vs. Other Approaches

Bag of frames method vs. proposed approach | SVM + pixel intensity values | SVM + pixel intensity values + Shakhnarovich's fusion[4]







Note: Face Detectors [24] → Viola &Jones Face detector and [22] → OCI Face detector

The proposed Markovian temporal model *(i)* achieved high gender classification performance (90%) considering the fact that it was trained on still face image database collected under controlled environment (FERET), and (ii) achieved a superior classification performance compared to its alternative approaches, reaching a performance increase of up to 30%.

We intend to extend our classification formulation and our unconstrained video database to explicitly account for uncertainty in detection and tracking, in order to classify faces in crowded scenes.