Gender Classification Using Pyramid Segmentation for Unconstrained Back-facing Video Sequences

Hao Tang, Hong Liu,* and Wei Xiao
Key Laboratory for Machine Perception (Ministry of Education)
Engineering Lab on Intelligent Perception for Internet of Things (ELIP)
Shenzhen Graduate School, Peking University, China
haotang@sz.pku.edu.cn; hongliu@pku.edu.cn; xiaoweipkusz@pkusz.edu.cn

ABSTRACT

This paper presents a pioneering study on gender classification from unconstrained back-facing video sequences in natural scenes. In many cases, classifying gender simply via faces or other biometric cues may fail when the video only contains back-facing people. To address this problem, we propose a novel approach to classify the gender according to back-facing video sequences. For this task, a novel Pyramid Segmentation approach is proposed to divide video sequence into a suite of equal time-length sleeves with different scales. Moreover, a heuristic approach is used to compute weights for different features from each sleeve. Finally, a framework of gender classification based on video sequences is presented. To validate our approach, we introduce a new dataset, called BackFacing dataset, featured by 720 annotated back-facing human video sequences. To our knowledge, this is the first dataset only containing back-facing video shots. Experiments demonstrate that the proposed approach achieves competitive results on VidTIMIT, Cohn-Kanade, CASIA Gait and BackFacing datasets.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition Applications *Computer vision* ; I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision Applications

Keywords

gender classification; pyramid segmentation; back-facing; unconstrained video sequences

1. INTRODUCTION

Gender classification is one of the most interesting topics in pattern classification and computer vision. It aims at determining whether a given person in an image or video is male or female. This task has important applications, such as intelligent user interface, video surveillance, demographic statistics collection, people counting, etc. In video surveillance systems, gender classification is a challenging problem, e.g. people always appear in different scales, poses and viewpoints in uncontrolled environment, and frames usually contain back-facing people, low resolution, and variations in illumination and scale. Recent gender classification works are mainly based on face [1], gait [13, 23], iris, lip, posture [19], and age [6], etc. However, it is hard to obtain these discriminative cues due to illumination variation, occlusion and low resolution. Departing from previous works, this paper considers another situation where a video sequence only contains back-facing people. For video sequence, Zhao et al. propose a block-based method which divided face sequences into several non-overlapping block volumes (Below 'volumes' will be called directly) to deal with specific dynamic events (e.g. facial expressions analysis) [24]. The introduction of spatial feature within still images has also been primarily handled with spatial pyramids [10]. However, block-based method and spatial pyramids only provide a rough spatial description, because the pose of the person may vary significantly. To fix this problem, a semantic pyramid approach has recently been proposed for feature extraction [18]. Spatial pyramids can also be used in, for example, scenes categorization [12].

In this paper we propose a novel Pyramid Segmentation (PS) approach, which divides video sequence into a suite of equal time-length sleeves with different scales, show in Figure 2. The inspiration of PS approach comes from literatures of re-identification [21], dynamic texture recognition [24], and spatial pyramids [10, 18, 12]: different regions of the video sequence contain visually distinct feature repre-

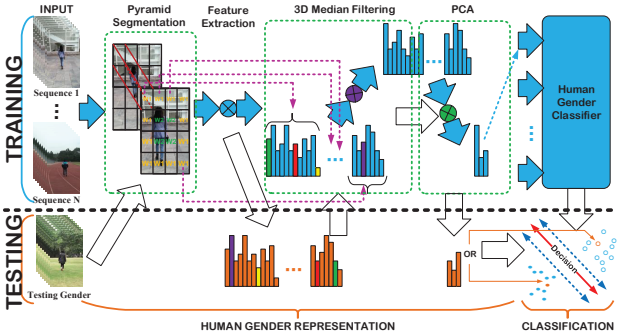

Figure 1: The proposed gender classification framework.

*Corresponding author.

MM'15, October 26-30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806312.

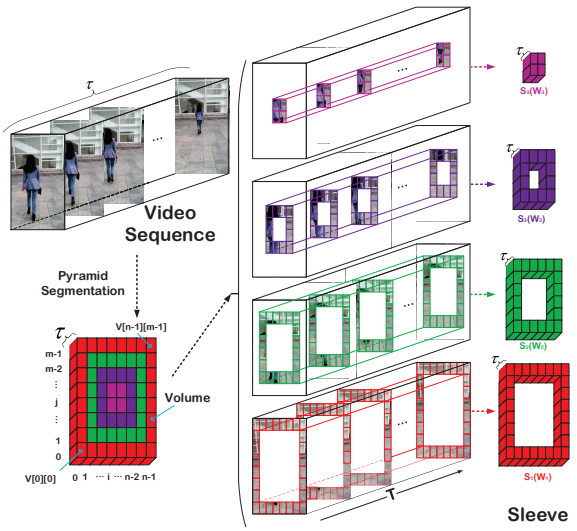

Figure 2: Pyramid segmentation for gender representation.

Algorithm 1 Pyramid segmentation and weights selection.

Require: N training video sequences, each sequence is divided into n by m volumes;
Ensure: The weight $W = \{W_1, W_2, ..., W_t\}$ for each pyramid sleeve $S = \{S_1, S_2, ..., S_t\}$;
1: **for** $t = 1$ to $min(\lceil \frac{n}{2} \rceil, \lceil \frac{m}{2} \rceil)$ **do**
2: **for** $j = 0$ to $m - 1$ **do**
3: $S_t.add(V[t-1][j])$, $S_t.add(V[n-t][j])$;
4: **end for**
5: **for** $i = 1$ to $n - 2$ **do**
6: $S_t.add(V[i][t-1])$, $S_t.add(V[i][m-t])$;
7: **end for**
8: $m \leftarrow m - 1$, $n \leftarrow n - 1$;
9: **end for**
10: Obtain $S = \{S_1, S_2, ..., S_t\}$, features extracted from each sleeve, produce corresponding histogram $h = \{h_1, h_2, ..., h_t\}$, respectively;
11: By computing the classification rates for each histogram $h = \{h_1, h_2, ..., h_t\}$ separately, get t rates $R = \{R_1, R_2, ..., R_t\}$;
12: Using Formula (1) to normalize R;
13: Using Formula (2) to calculate W;
14: **return** W.

sentation. Some approaches use a single rectangle to capture the global appearance in still images [15], and others utilize more complicated structural representation [5]. Nevertheless, these approaches are applied to a still image of cropped body. Instead we choose a representation using PS to distinguish and reinforce the contributions of the features from each sleeve: the greater contribution it has, the higher weight it will be assigned. Those partitions with higher weight are considered as more important sleeves. For selecting the weights, we adopt the method in [24], which employs the weighted method to the three orthogonal planes of LBP-TOP [24]. Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [24] consists of three orthogonal planes: XY, XT and YT, and concatenates local binary pattern co-occurrence statistics in these three directions. However, in this paper, we apply the weighted method to each sleeve and use weighed sleeves to construct gender representation. The flowchart in Figure 1 shows the processing details of our gender classification framework based on video sequences.

2. PYRAMID SEGMENTATION AND GENDER CLASSIFICATION FRAMEWORK

We introduce a novel pyramid segmentation approach considering spatial feature in different area of video sequence, which divides video sequence into a suite of equal time-length sleeves with different scales. Assuming that a back-facing video sequence is divide into the n by m volumes evenly, as we can see from Figure 2, each unit volume denoted as $V[i][j]$, where $i = 0, 1, ..., n-1$ and $j = 0, 1, ..., m-1$. We concatenate the volumes which have the same scales as a sleeve, we denote each sleeve as S_t, where $t = min(\lceil \frac{n}{2} \rceil, \lceil \frac{m}{2} \rceil)$. Illustrated in Figure 2, video sequence is divided into 8 by 8 volumes, i.e. 64 volumes. Then $t = min(\lceil \frac{n}{2} \rceil, \lceil \frac{m}{2} \rceil) = 4$, we select $V[0][j]$, $V[7][j]$, $V[i][0]$ and $V[i][7]$, 28 volumes in total as the first sleeve, where $i = 1, 2, ..., 6$ and $j = 0, 1, ..., 7$. The second sleeve is consisted of $V[1][j]$, $V[6][j]$, $V[i][1]$ and $V[i][6]$, where $i = 2, 3, ..., 5$ and $j = 1, 2, ..., 6$, totally 20 volumes. The third sleeve is made up of 12 volumes, i.e. $V[2][j]$, $V[5][j]$, $V[i][2]$ and $V[i][5]$ (where $i = 3, 4$ and $j = 2, 3, ..., 5$). Finally, the last sleeve is composed of the last 4 volumes $V[3][j]$ and $V[4][j]$ (where $j = 3, 4$). Hence the video sequence is divided into 4 sleeves. Feature extracted from each sleeve. However different sleeves have different importance contributions for feature representation. Therefore, we assign different weights for different sleeves. For weights selection, we adopt the method proposed in [24], which assigns weights of three orthogonal planes (XY, XT and YT) in LBP-TOP [24]. In this paper, instead, we use this method for selecting the weights of different sleeves. First, by computing the classification rates for each sleeve separately, we obtain t rates $R = \{R_1, R_2, ..., R_t\}$. Based on the assumption that the higher the rate is, the better gender representation becomes, we compute the weights as follows:

$$T = \frac{R - min(R)}{(100 - min(R))/10}. \quad (1)$$

Finally, considering that the weight of the lowest rate is 1, the other weights can be obtained according to a linear relationship of their differences to that with the lowest rate. The final step is written as:

$$\begin{aligned} T1 &= round(T) \\ T2 &= \frac{T \times ((max(T1) - 1))}{max(T)} + 1 \\ W &= round(T2) \end{aligned} \quad (2)$$

in which W is the weight vector corresponds to each sleeve. Pyramid segmentation and weights selection is summarized in Algorithm 1. Given N video sequences for training, each sequence is divided into n by m volumes equally. The n by m volumes are partitioned t sleeves $\{S_1, S_2, ..., S_t\}$ (step 1 – 9). Then feature extracted from each sleeve S_t generate corresponding histogram h_t (step 10). Therefore, obtain t rates R_t via computing the classification rates of h_t respectively (step 11). Finally, using Formula (1) and (2) to calculate the W (step 12 and 13). In the end of the algorithm, return the weight vector W (step 14).

The gender classification framework is based on video sequence and composed of two parts, training and testing, which is summarized in Algorithm 2. During training stage,

Algorithm 2 Gender classification framework.

Require: N video sequences for training, corresponds to the gender label $N.Gender_{label}$ and each sequence divided into n by m volumes; Testing video sequences T;
Ensure: $T.Gender_{label}$;
1: TRAINING STAGE:
2: $W = \{W_1, W_2, ..., W_t\} \leftarrow$ Algorithm 1;
3: **for** $i = 1$ to N **do**
4: $h^i = \{h_1, h_2, ..., h_t\} \leftarrow$ from each sleeve $S = \{S_1, S_2, ..., S_t\}$ using LBP-TOP, respectively;
5: $h^i * W = \{W_1 * h_1, W_2 * h_2, ..., W_t * h_t\} \leftarrow$ weighted for each histogram;
6: Training gender descriptors $H^i = [W_1 * h_1, W_2 * h_2, ..., W_t * h_t]$ are constructed and concatenated from each sleeve $h^i * W$;
7: **end for**
8: Perform 3D median filtering operation of vector H;
9: Carried on the PCA to vector H;
10: SVM classifier $\leftarrow H \cup N.Gender_{label}$;
11: TESTING STAGE:
12: Obtain gender representation H_T for testing T use the same method as training stage;
13: Made a decision by classifier after calculation;
14: **return** $T.Gender_{label}$.

we first de-compose the training videos into frames, so that each training sample contains τ back-facing frames. Then we obtain weight W corresponds to the sleeve S using Algorithm 1 (step 2). Next we use LBP-TOP [24] to extract feature from each sleeve, producing the corresponding histogram h (step 4). Each histogram h is multiplied by their weights (step 5). Training descriptors H_i (where $i = 1, 2, ..., N$) are constructed and concatenated from each weighted histogram (step 6). In the end of the iteration, we get the training representation vector H (step 7). Therefore, we perform 3D median filtering for the vector H (step 8). It is time-consuming because H is a high-dimensional. Furthermore, we use the PCA [4] to reduce the dimension (step 9). Then H and corresponding labels are feed to a SVM classifier (step 10). During testing stage, testing gender representation is obtained in the same way as training stage (step 12). Thereby the trained classifier is used to predict the gender label $T.Gender_{label}$ (step 13).

3. EXPERIMENTS AND ANALYSIS

To our best knowledge, there is no publicly available dataset for gender classification based on video sequences of human back-facing. To bridge the gap between theoretical analysis and practical testing, a dataset of video sequences collected for further studying the problem of gender classification from back-facing sequences. This dataset named BackFacing which consists of 5760 video frames of 640 by 360 resolution from 720 video sequences, each of which recorded from a different subject (30 males and 30 females) in 12 different natural environment (room, corridor, elevator, stair, restaurant, supermarket, playground, parking, road, lawn, square and ATM). Challenging factors in this dataset is each video sequence is shotted under arbitrary illumination conditions and background clutter. Furthermore, people are completely free in their movements (walking or running), leading to arbitrary body scales, motion blur, and local or global occlusions. Figure 3 show two sample sequences of female and male walking on the corridor and lawn, respectively.

Figure 3: Back-facing video sequences.

$LBP\text{-}TOP_{8,8,8,1,1,1}$ [24] is used to extract feature from each sleeve. Note that this work does not focus on finding optimal features. Many other descriptors, e.g. EVLBP [7], 3D-SIFT [17] and VLBP [24] can be used as well, and may further improve performance. We use an RBF kernel binary SVM as the classifier and the implementation is provided by LIBSVM [3]. Specifically, we adopted a 5-fold cross validation test scheme by dividing the 720 sequences into five groups and using the data from four groups for training and the left group for testing. The sequence of a particular subject appear only in one group. We repeated this process ten times and report the average classification rates (A_{vg}). The implementation of 3D Median Filtering (3DMF) provided by Matlab is used. For PCA, we set the number of principle component coefficients retained as 90. Furthermore, LBP-TOP [24] + SVM is selected as the baseline method on BackFacing dataset.

3.1 Experiments on Public Datasets

To evaluate the performance of our gender classification framework, public datasets of face and gait video sequence, are adopted in our experiments.

VidTIMIT & Cohn-Kanade datasets. The VidTIMIT dataset [16] consists of face video sequence (with a resolution of 512×384 pixels) recordings of 43 people (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus. Cohn-Kanade dataset [9] is for research in automatic facial image analysis and synthesis and for perceptual studies. We randomly segment the datasets and extract over 6400 video shots of 8 frames each. Table 2 shows our approach has the best performance comparing with EVLBP + AdaBoost, LBP + SVM and Pixels + SVM [7] on the VidTIMIT and Cohn-Kanade datasets.

Table 2: Gender classification performance on VidTIMIT and Cohn-Kanade datasets.

Approaches	A_{vg}(%)	Approaches	A_{vg}(%)
LBP + SVM	91.0	EVLBP + AdaBoost	81.5
Pixels + SVM	89.4	Ours	**97.92**

CASIA Gait Set B. The CASIA Gait Set B [22] is one of the largest gait datasets in the gait-research community currently. The dataset consists of 124 subjects aged between 20 and 30 years, of which 93 are male and 31 are female, and 123 are Asian and 1 is European. Three variations, namely view angle, clothing and carrying condition changes, are separately considered. We randomly select half of the dataset for training and half for testing when dividing the data into training and testing sets. Table 3 validates the superiority of our method on CASIA Gait Set B comparing with state-of-the-arts.

Table 1: . Gender classification rates on *BackFacing* dataset.

$n \times m$	baseline	T_b(s)	+PCA	+3D filtering	t	R	W	+Weighted	T_o(s)
1×1	69.00	0.520	64.31	76.11	1	[76.11]	[1]	76.11	0.235
2×2	74.22	1.776	71.51	85.64	1	[85.64]	[1]	85.64	0.288
3×3	76.40	4.634	75.14	87.51	2	[86.71, 81.44]	[3, 1]	90.31	0.429
4×3	77.91	9.221	77.94	89.10	2	[88.64, 85.98]	[2, 1]	91.18	0.570
4×4	78.06	10.270	77.27	90.03	2	[88.03, 84.60]	[2, 1]	91.35	0.630
5×4	78.49	19.153	78.44	90.14	2	[86.92, 89.50]	[1, 2]	**92.74**	0.647
5×5	77.34	21.674	80.14	90.02	3	[88.32, 88.77, 80.65]	[4, 4, 1]	91.69	0.709
6×5	76.56	32.151	78.49	89.15	3	[87.89, 89.10, 84.43]	[2, 3, 1]	91.70	0.735
6×6	76.30	37.253	79.55	88.94	3	[91.18, 88.58, 88.06]	[3, 1, 1]	91.52	0.819
7×6	77.60	52.936	**81.14**	90.09	3	[88.06, 90.83, 89.45]	[1, 2, 2]	91.14	0.852
7×7	79.07	59.567	80.69	89.76	4	[75.43, 76.82, 76.13, 60.38]	[4, 4, 4, 1]	90.14	0.926
8×7	77.85	78.772	80.67	**90.33**	4	[91.35, 90.66, 88.41, 83.56]	[5, 5, 3, 1]	92.00	0.981
8×8	76.82	87.376	79.74	90.22	4	[88.43, 85.29, 90.87, 83.91]	[3, 2, 5, 1]	91.83	**1.083**

Table 3: Classification results on CASIA dataset.

Approaches	Ours	[13]	[11]	[23]	[20]	[8]	[14]
A_{vg}(%)	**99.51**	98.0	93.3	95.79	96.0	98.39	98.4

3.2 Experiments on BackFacing Dataset

The results on BackFacing are shown in Table 1 and Figure 4 (a). The performance of the baseline can be noticeably improved by adding PCA, 3DME and Weight. t, R and W are mentioned in the Section 2. In our evaluation, Table 1 and Fig 4 (b) reports the average sum of training and testing time[1] per sequence. We can calculate that our approach is about 38 times faster than the baseline method on average. In addition we compare our approach with state-of-the-arts. The results, shown in Table 4, which indicate that our approach is superior to these approaches on BackFacing dataset. Finally, it is evident that our proposed PS approach and gender classification framework consistently outperforms other methods on there four datasets. The main reason lies in, PS strategy fully exploits the intrinsic feature of the video sequence, and further enhances its contribution by weighted.

Table 4: Comparison to state-of-the-arts on BackFacing dataset.

Approaches	A_{vg}(%)	Approaches	A_{vg}(%)
VLBP [24] + SVM	84.95	LDA [2]	67.82
LBP [1] + SVM	64.07	PCA [2] + SVM	79.98
Raw + SVM [6]	79.40	PCA + LDA [2]	59.60
$LBP\text{-}TOP_{[5,2,1]}$ [24]	82.18	Ours	**92.74**

4. CONCLUSIONS

In this paper, we have addressed the relatively open problem of back-facing gender classification by proposing a novel Pyramid Segmentation approach based on video sequence that reinforces the weights of sleeve for the gender representation. Moreover, a framework of gender classification based on video sequence is presented. Experimental results show that the proposed approach is outperforms the competitive methods on VidTIMIT, Cohn-Kanade, CASIA Gait and BackFacing dataset.

5. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China(NSFC, No.61340046), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20130331144631730).

[1]We run our experiments on a PC with 3.40 GHz four-Core CPU and 8 GB RAM.

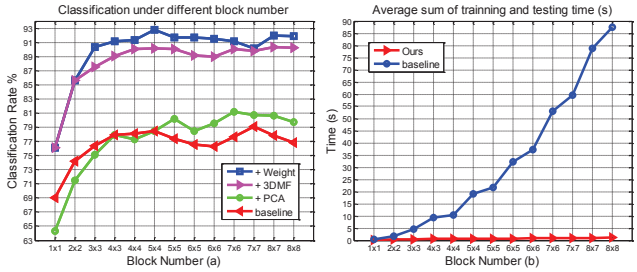

Figure 4: (a) Average gender classification rates (%) and (b) average sum of training and testing time (s) on BackFacing dataset.

6. REFERENCES

[1] T. Ahonen et al. Face description with local binary patterns: Application to face recognition. *IEEE TPAMI*, 2006.
[2] J. Bekios-Calfa et al. Revisiting linear discriminant techniques in gender recognition. *IEEE TPAMI*, 2011.
[3] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM TIST*, 2011.
[4] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE TMM*, 2000.
[5] N. Gheissari et al. Person reidentification using spatiotemporal appearance. *in CVPR*, 2006.
[6] G. Guo et al. Is gender recognition affected by age? *in ICCVW*, 2009.
[7] A. Hadid and M. Pietikäinen. Combining appearance and motion for face and gender recognition from videos. *Elsevier PR*, 2009.
[8] M. Hu et al. Gait-based gender classification using mixed conditional random field. *IEEE TSMC Part B*, 2011.
[9] T. Kanade et al. Comprehensive database for facial expression analysis. *in FG*, 2000.
[10] F. S. Khan et al. Coloring action recognition in still images. *Springer IJCV*, 2013.
[11] W. Kusakunniran et al. Support vector regression for multi-view gait recognition based on local motion feature selection. *in CVPR*, 2010.
[12] S. Lazebnik et al. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *in CVPR*, 2006.
[13] J. Lu et al. Human identity and gender recognition from gait sequences with arbitrary walking directions. *IEEE TIFS*, 2014.
[14] D. S. Matovski et al. The effect of time on gait recognition performance. *IEEE TIFS*, 2012.
[15] B. Prosser et al. Multi-camera matching under illumination change over time. *in ECCVW*, 2008.
[16] C. Sanderson and K. K. Paliwal. Noise compensation in a person verification system using face and multiple speech features. *Elsevier PR*, 2003.
[17] P. Scovanner et al. A 3-dimensional sift descriptor and its application to action recognition. *ACM MM*, 2007.
[18] F. Shahbaz et al. Semantic pyramids for gender and action recognition. *IEEE TIP*, 2014.
[19] S. Wuhrer et al. Posture invariant gender classification for 3d human models. *in CVPRW*, 2009.
[20] D. Xu et al. Human gait recognition using patch distribution feature and locality-constrained group sparse representation. *IEEE TIP*, 2012.
[21] Y. Yang et al. Salient color names for person re-identification. *in ECCV*, 2014.
[22] S. Yu et al. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. *in ICPR*, 2006.
[23] S. Yu et al. A study on gait-based gender classification. *IEEE TIP*, 2009.
[24] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE TPAMI*, 2007.