Comparison of Recent Machine Learning Techniques for Gender Recognition from Facial Images

Joseph Lemley
Computer Science Department
Central Washington University
Ellensburg, WA, USA

Sami Abdul-Wahid
Computer Science Department
Central Washington University
Ellensburg, WA, USA

Dipayan Banik
Computer Science Department
Central Washington University
Ellensburg, WA, USA

Răzvan Andonie
Computer Science Department
Central Washington University
Ellensburg, WA, USA
and
Electronics and Computers Department
Transilvania University
Braşov, Romania

Abstract

Recently, several machine learning methods for gender classification from frontal facial images have been proposed. Their variety suggests that there is not a unique or generic solution to this problem. In addition to the diversity of methods, there is also a diversity of benchmarks used to assess them. This gave us the motivation for our work: to select and compare in a concise but reliable way the main state-of-the-art methods used in automatic gender recognition. As expected, there is no overall winner. The winner, based on the accuracy of the classification, depends on the type of benchmarks used.

Introduction

A major goal of computer vision and artificial intelligence is to build computers that can understand or classify concepts such as gender in the same way humans do.

Automatic classification of gender from frontal face images taken under contrived conditions has been well studied with impressive results. The variety of methods published in the literature show that there is not a unique or generic solution to the gender classification problem.

Applications of gender classification include, image search, automatic annotation of images, security systems, face recognition, and real time image acquisition on smart phones and mobile devices.

The state-of-the-art gender classification methods generally fall into the following main categories: Convolutional Neural Networks (CNN), Dual Tree Complex Wavelet Transform (DTCWT) + a Support Vector Machine (SVM) classifier, and feature extraction techniques such as Principal Component Analysis (PCA), Histograms of Oriented Gradients (HOG) and others with a classifier (SVM, kNN, etc). The SVM approach is natural, since we have a two class problem. The CNN is related to the well-known deep learning paradigm. The DTCWT provides approximate shift invariance and directionally selective filters (properties lacking in the traditional wavelet transform) while preserving the usual properties of perfect reconstruction and computational efficiency with good well-balanced frequency responses (Kingsbury 2001).

To assess gender classification techniques, two types of benchmarks may be used: standard posed datasets (with well defined backgrounds, lighting and photographic characteristics) and datasets containing "In the wild" images that exhibit the diversity of subjects, settings, and qualities typical of everyday scenes.

The diversity of the methods and benchmarks makes a comparison between gender classification a challenging task, and this gave us the motivation for our work. We compare state-of-the-art methods used in automatic gender recognition on two benchmarks: the most popular standard dataset Facial Recognition Technology (FERET) (Phillips et al. 2000) and a more challenging data set of "in the wild" images (Adience) (Eidinger, Enbar, and Hassner 2014).

We only compare the accuracy of the classification and not other performance measures (precision, recall, F1 score, etc). The main reason is that the misclassification cost in this particular problem is the same, regardless if we misclassify a male or a female. We also do not compare the running time, since the experiments are performed on different computer architectures (the CNN is implemented on a GPU).

Related work: recent gender classification methods

Classifiers such as SVMs and feedforward NNs are often used to classify images after the faces have been cropped out from the rest of the image, and possibly aligned and normalized. Various feature extraction methods such as Principal Component Analysis (PCA), independent component analysis, Fischer linear discriminants (Belhumeur, Hespanha, and Kriegman 1997) (Wu et al. 2015), and edge detection algorithms can be used to encode useful information from the image that is fed into the classifier, leading to high levels of accuracy on many benchmarks. Other approaches use hand-crafted template features to find facial keypoints such as nose, eyes etc, while also using edge detection methods (Sobel) and line intensities to separate facial edges from wrinkles. The resulting feature information, when fed into a feedforward neural network, allows age and gender to be classified with overall 85% accuracy on two test sets with a total of 172 images in the FERET and FGNET databases (Kalansuriya and Dharmaratne 2014).

LDA (Linear Discriminant Analysis) based approaches to the face recognition task promise invariance to differing illuminations (Belhumeur, Hespanha, and Kriegman 1997). This has been further studied in (Bekios-Calfa, Buenaposada, and Baumela 2011). Fisher linear discriminant max-

imizes the ratio of between- class scatter to that of within-class scatter. Independent component analysis has been used on a small subset (500 images) of the FERET dataset, leading to 96% accuracy with an SVM classifier (Jain, Huang, and Fang 2005). Likewise, PCA has been used in conjunction with a genetic algorithm that eliminated potentially unnecessary features. The remaining features were then fed to a feedforward neural network for training, and an overall 85% accuracy was obtained over 3 data sets (Sun et al. 2002). Various information theory based metrics were also fused together to produce 99.13% gender classification accuracy on the FERET (Perez et al. 2012). To overcome the challenge of inadequate contrast among facial features using histogram analysis, Haar wavelet transformation and Adaboost learning techniques have been employed, resulting in a 97.3% accuracy on the Extended Yale face database which contains 17 subjects under 576 viewing conditions (Laytner, Ling, and Xiao 2014). Another experiment describes how various transformations, such as noise and geometric transformations, were fed in combination into a series of RBFs (Radial Basis Functions). RBF outputs were forwarded into a symbolic decision tree that outputs gender and ethnic class. 94% classification accuracy was obtained using the hybrid architecture on the FERET database (Gutta, Wechsler, and Phillips 1998).

HOG (Histogram of Oriented Gradients) is commonly used as a global feature extraction technique that expresses information about the directions of curvatures of an image. HOG features can capture information about local edge and gradient structures while maintaining degrees of invariance to moderate changes in illumination, shadowing, object location, and 2D rotation. HOG descriptors, combined with SVM classifiers, can be used as a global feature extraction mechanism (Torrione et al. 2014), while HOG descriptors can be used on locations indicated by landmark-finding software in areas such as facial expression classification (Déniz et al. 2011). One useful application of variations in HOG descriptors is the automatic detection of pedestrians, which is made easier in part because of their predominantly upright pose (Dalal and Triggs 2005). In addition, near perfect results were obtained in facial expression classification when HOG descriptors were used to extract features from faces that were isolated through face-finding software (Carcagnì et al. 2015).

A recent technique proposed for face recognition is the DTCWT, due to its ability to improve operation under varying illumination and shift conditions when compared to Gabor Wavelets and DWT (Discrete Wavelet Transform). The Extended Yale B and AR face databases were used, containing a total 16128 images of 38 human subjects under 9 poses and 64 illumination conditions. It achieved 98% classification accuracy in the best illumination condition, while low frequency subband image at scale one (L1) achieved 100% (Sultana et al. 2014).

Recent years have seen great success in image related problems through the use of CNN, thereby seeing the proliferation of a scalable and more or less universal algorithmic approach to solving general image processing problems, if enough training data is available. CNNs have had a great deal of success in dealing with images of subjects and objects in natural non-contrived settings, along with handling the rich diversity that these images entail. One investigation of CNN fundamentals involved training a CNN to classify gender on images collected on the Internet. 88% classification accuracy was achieved after incorporating L2 regularization into training, and filters were shown to respond to the same features that neuroscientists have identified as fundamental cues humans use in gender classification (Verma and Vig 2014). Another experiment (Levi and Hassner 2015) uses a convolutional neural network on the Adience dataset for gender and age recognition. They used data augmentation and face cropping to achieve 86% accuracy for gender classification. This is the only paper we know of that uses CNN on Adience.

A method recently proposed by (Eidinger, Enbar, and Hassner 2014) uses an SVM with dropout, a technique inspired from newer deep learning methods, that has shown promise for age and gender estimation. Dropout involves dropping a certain percent of features randomly during training. They also introduce the Adiance dataset to fulfill the need for a set of realistic labeled images for gender and age recognition in quantities needed to prevent overfitting and allow true generalization (Eidinger, Enbar, and Hassner 2014).

As we can see, most of the state-of-the-art methods for gender classification fall into the categories described in Section .

Data sets

A number of databases exist that can be used to benchmark gender classification algorithms. Most image sets that contain gender labels suffer from insufficient size, and because of this we chose two of the larger publicly available datasets: Color-FERET (Phillips et al. 2000) and Adience (Eidinger, Enbar, and Hassner 2014).

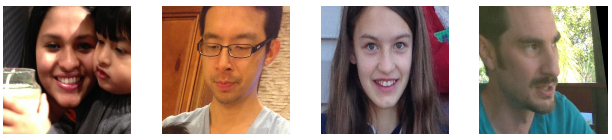

Figure 1: Randomly selected images from the Adience dataset illustrating the wider range of photographic conditions found.

Color FERET Version 2 was collected between December 1993 and August 1996 and made freely available with the intent of promoting the development of face recognition algorithms. Images in the FERET Color database are 512 by 768 pixels and are in PPM format. They are labeled with gender, pose, name, and other useful labels.

Although FERET contains a large number of high quality images in different poses and with varying face obstructions (beards, glasses, etc), they all have certain similarities in quality, background, pose, and lighting which make them very easy for modern machine learning methods to correctly

Figure 2: Randomly selected images from the FERET dataset show similarities in lighting, pose, subject, background, and other photographic conditions.

classify. We used all 11338 images in FERET for which gender labels exist in our experiments.

As machine learning algorithms are increasingly used to process images of varying quality with vast differences in scale, obstructions, focus, and which are often acquired with consumer devices such as web cams or cellphones, benchmarks such as FERET have become less useful. To address this issue, datasets such as LWS (labeled faces in the wild) and most recently Adience have emerged. LWS lacks gender labels but it has accurate names from which gender can often be deduced automatically with reasonable accuracy.

Adience is a recently released benchmark that contains gender and approximate age labels separated into 5 folds to allow duplication of results published by the database authors. It was created by collecting Flickr images and is intended to capture all variations of pose, noise, lighting, and image quality. Each image is labeled with age and gender. It is designed to mimic the challenges of "real world" image classification tasks where faces can be partly obscured or even partly overlapping (for example in a crowd or when an adult is holding a child and both are looking at the camera). Eidlinger, et al published a paper where they used a SVM and filtering to classify age and gender along with the release of Adience (Eidinger, Enbar, and Hassner 2014).

We used all 19370 of the aligned images from Adience that had gender labels, to create our training and testing sets for all experiments that used Adience.

Using the included labels and meta data in the FERET and Adience datasets, we generated two files containing reduced size 51x51 pixel data with values normalized between 0 and 1, followed by a gender label. We choose to resize the images to 51x51 because this produced the best quality images after Anti-Aliasing

Classification methods

We compare the accuracy of CNN and several SVM based classifiers. We limit ourselves to methods involving these two approaches because they are among the most effective and most prevalently used methods reported in the literature for Gender Classification.

Gender classifications with SVM perform on the raw image pixels along with different well known feature extraction methods, namely DTCWT, PCA, and HOG. Training is done separately on two widely differing datasets consisting of gender labeled human faces: Color FERET, a set of images taken under similar conditions with good image quality, and Adience, a set of labeled unfiltered images intended to be especially challenging for modern machine learning algorithms. Adience was designed to present all variations in appearance, noise, pose, and lighting, that can be expected of images taken without careful preparation or posing. (Eidinger, Enbar, and Hassner 2014)

We use the following steps in conducting our experiments:

- Uniformly shuffle the order of images.
- Use 70% as training set. 30% as testing set.
- Train with training set.
- Record correct classification rate on testing set.

Steps 1-4 are repeated 10 times for each experiment using freshly initialized classifiers.

We report the results of 18 experiments, 16 of which use SVMs and two of which use CNN.

SVM classification

Both linear and RBF kernels were used, with each constituting a separate experiment using the SVC implementation included as part of scikit-learn(Pedregosa et al. 2011) with $C = 100$ parameter set.

In one experiment, raw pixels are fed into the SVM. Other experiments used the following feature extraction methods: PCA, HOG, and DTCWT. Feature extraction was applied to images uniformly without using face finding software to isolate and align the face.

Histogram of Oriented Gradients

HOG descriptors, combined with SVM classifiers, can be used as a global feature extraction mechanism (Torrione et al. 2014), while HOG descriptors can be used on locations indicated by landmark-finding software in areas such as facial expression classification (Déniz et al. 2011).

One application of HOG descriptors is the automatic detection of pedestrians, which is made easier in part because of their predominantly upright pose (Dalal and Triggs 2005). We use the standard HOG implementation from the scikit-image library (van der Walt et al. 2014).

For every image in the Adience and FERET databases, HOG descriptors were uniformly calculated. 9 orientation bins were used, and each histogram was calculated based on gradient orientations in the 7x7 pixel non-overlapping cells. Normalization was done within each cell (i.e., 1 x 1). The result was fed into a SVM (SVC class from scikit-learn).

Training and testing on both Adience and FERET was performed separately. 30% of images in each database were used for testing, and the rest for training. For each database, after reading the data into arrays, the arrays were shuffled and then the testing and training set were separated. Training and testing were repeated 10 times with freshly shuffled data.

Principal Component Analysis

PCA is a statistical method for finding correlations between features in data. When used on images of faces the resulting

images are often referred to as Eigenfaces. PCA is used for reducing dimensionality of data by eliminating non-essential information from the dataset and is frequently used in both image processing and machine learning.

To create the Eigenfaces we used the RandomizedPCA tool within scikit-learn, which is based on work by (Halko, Martinsson, and Tropp 2011) and (Martinsson, Rokhlin, and Tygert 2011). The resulting Eigenfaces were then used in a linear and RBF SVM.

Convolutional Neural Network

For the learning stage we used a convolutional neural network with 3 hidden convolutional layers and one softmax layer. The training was done using a GTX Titan X GPU using the Theano based library Pylearn2 and CUDNN libraries. Stochastic gradient descent was used as the training algorithm with a momentum of 0.95, found by trial and error. Learning rates under 0.001 did not show any improvement. Increasing the learning rate above around 0.005 results in decreased classification accuracy.

A general outline of the structure of our CNN is:

- Hidden layer 1: A Rectified Linear Convolutional Layer using a kernel shape of 4x4, a pool shape of 2x2, a pool stride of 2x2 and 128 output channels. Initial weights are randomly selected with a range of 0.5.
- Hidden layer 2: A Rectified Linear Convolutional Layer using a kernel shape of 4x4, a pool shape of 2x2, a pool stride of 2x2 and 256 output channels. Initial weights are randomly selected with a range of 0.5.
- Hidden layer 3: A Rectified Linear Convolutional Layer using a kernel shape of 3x3, a pool shape of 2x2, a pool stride of 2x2 and 512 output channels. Initial weights are randomly selected with a range of 0.5.
- Softmax layer: Initial weights randomly set between 0 and 0.5. Output is the class (male or female).

Experimental results

Tables 1 and 2 summarize the classification accuracy of each approach on each data-set after random shuffling and separation into 70% training and 30% testing sets. For each method the grayscale pixels were used as the features, either directly to the classifier, or to the filter mentioned. For example, HOG+SVM[RBF] indicates that we use the pixels as input to a HOG filter, the output of which is used as the input to a SVM with an RBF kernel.

DTCWT was both the second best method (after CNN) and the very worst method we examined; its performance has the greatest degree of variability depending on the dataset. It performs very well when objects are consistent in location and scale. CNN outperformed all methods. Even the worst CNN experiment on the most difficult dataset performed better than the best of any other method on the easiest dataset. This is not a surprising outcome. We wanted to see if HOG alone was sufficient to increase classification accuracy as a filter. We found that HOG filters with SVM, without the usual additional models, provide no benefit on their own over raw pixel values for this experimental setup.

Table 1: Mean Classification accuracy and Standard Deviation for different methods on the Adience dataset over 10 runs. 70% of images used for training and 30% used for testing.

Method	Mean	SD
CNN	96.1%	0.0029
PCA+SVM[RBF]	77.4%	0.0071
SVM[RBF]	77.3%	0.0046
HOG + SVM[RBF]	75.8%	0.006
HOG+SVM[linear]	75%	0.0053
PCA+ SVM[linear]	72 %	0.0032
SVM[linear]	70.2%	0.0052
DTCWT on SVM[RBF]	68.5%	0.0059
DTCWT on SVM[linear]	59%	0.0046

Table 2: Mean Classification accuracy and Standard Deviation for different methods on the FERET dataset over 10 runs. 70% of images used for training and 30% used for testing.

Method	Mean	SD
CNN	97.9%	0.0058
DTCWT on SVM[RBF]	90.7%	0.0047
PCA+SVM[RBF]	90.2%	0.0063
SVM[RBF]	87.1%	0.0053
HOG+SVM[RBF]	85.6%	0.0042
HOG+SVM[linear]	84.6%	0.0024
DTCWT on SVM[linear]	83.3%	0.0047
PCA+SVM[linear]	81%	0.0071
SVM[linear]	76.5%	0.0099

PCA ties with DTCWT on the best performance on FERET but performs better than DTCWT on Adiance. As expected RBF methods performed better than linear SVM classifiers, however unexpectedly this did not hold true for Adiance, where differences in filters were enough to cancel out the effect of RBF in some cases. Every time we used a filter on FERET RBF was better than linear with filters. This did not hold for Adience. None of the filters worked particularly well on Adience, with only PCA slightly outperforming raw pixels for the RBF classifier.

On the FERET dataset DTCWT is better (90% vs 86%). On Adience, it is worse ($\tilde{6}7\%$ vs 77%). This would lend support to the idea that DTCWT seems to work better (in theory) on images that are more similar to FERET (uniform lighting, no complex backgrounds, no extreme warping, pixelation, or blurring).

Using an initial momentum of 0.95 tended to promote fast convergence without getting stuck in local minimum. We use a momentum of 0.95 and a learning rate of 0.001.

Using this setup we have achieved an average valid classification rate of 98% on FERET and 96% on Adience which is better than the previous highest reported results according to (Levi and Hassner 2015) on Adience, but we do not recommend direct comparison of our results with theirs because of different experimental protocols used.

One of our aims is to investigate the use of the dual tree

complex wavelet transform (DTCWT) on the face feature classification task. Several recent papers report success in using DTCWT in gender recognition from frontal face images citing the benefits of partial rotation invariance. It is somewhat unclear how to best use this for "In the wild" images.

Conclusions

Much of the previous work on automatic gender classification use differing datasets and experimental protocols that can make direct comparisons between reported results misleading. We have compared nine different machine learning methods used in gender recognition on two benchmarks, using identical research methodology to allow a direct comparison between the efficacies of the different classifiers and feature extraction methods. In addition to providing updated information on the effectiveness of these algorithms, we provide directly comparable results.

The aim of our study was to explore gender classification using recent learning algorithms. We carried out experiments on several state-of-the-art gender classification methods. We compared the accuracy of these methods on two very different data sets ("In the wild" verses posed images).

To the extent of our knowledge, this is the first use of DTCWT on a large $\geq 15,000$ database of "in the wild" images, specifically addressing gender classification. We have achieved an average accuracy of 98% (FERET) and 96% (Adience), which is better than the previous highest reported results (according to (Levi and Hassner 2015)) on Adience using a CNN.

The DTCWT seems to work better ($\approx 90\%$) on images that are more similar to FERET (uniform lighting, no complex backgrounds, no extreme warping, pixelation, or blurring).

The Adience and FERET data sets are relatively large and this may explain why the CNN method generally outperforms other methods: it is known that deep learning performs well when large training sets are being used. It is interesting to determine in this particular application what "large" actually is.

References

Bekios-Calfa, J.; Buenaposada, J. M.; and Baumela, L. 2011. Revisiting linear discriminant techniques in gender recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33(4):858–864.

Belhumeur, P. N.; Hespanha, J. P.; and Kriegman, D. J. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19(7):711–720.

Carcagnì, P.; Del Coco, M.; Leo, M.; and Distante, C. 2015. Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus* 4(1):1–25.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, 886–893. IEEE.

Déniz, O.; Bueno, G.; Salido, J.; and De la Torre, F. 2011. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters* 32(12):1598–1603.

Eidinger, E.; Enbar, R.; and Hassner, T. 2014. Age and gender estimation of unfiltered faces. *Information Forensics and Security, IEEE Transactions on* 9(12):2170–2179.

Gutta, S.; Wechsler, H.; and Phillips, P. J. 1998. Gender and ethnic classification of face images. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 194–199. IEEE.

Halko, N.; Martinsson, P.-G.; and Tropp, J. A. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* 53(2):217–288.

Jain, A.; Huang, J.; and Fang, S. 2005. Gender identification using frontal facial images. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 4–pp. IEEE.

Kalansuriya, T. R., and Dharmaratne, A. T. 2014. Neural network based age and gender classification for facial images. *ICTer* 7(2).

Kingsbury, N. 2001. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis* 10(3):234 – 253.

Laytner, P.; Ling, C.; and Xiao, Q. 2014. Robust face detection from still images. In *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2014 IEEE Symposium on*, 76–80. IEEE.

Levi, G., and Hassner, T. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 34–42.

Martinsson, P.-G.; Rokhlin, V.; and Tygert, M. 2011. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis* 30(1):47–68.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Perez, C.; Tapia, J.; Estévez, P.; and Held, C. 2012. Gender classification from face images using mutual information and feature fusion. *International Journal of Optomechatronics* 6(1):92–119.

Phillips, P. J.; Moon, H.; Rizvi, S. A.; and Rauss, P. J. 2000. The feret evaluation methodology for face-recognition algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(10):1090–1104.

Sultana, M.; Gavrilova, M.; Alhajj, R.; and Yanushkevich, S. 2014. Adaptive multi-stream score fusion for illumination invariant face recognition. In *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2014 IEEE Symposium on*, 94–101. IEEE.

Sun, Z.; Yuan, X.; Bebis, G.; and Loui, S. J. 2002. Neural-network-based gender classification using genetic search

for eigen-feature selection. In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, volume 3, 2433–2438. IEEE.

Torrione, P. A.; Morton, K. D.; Sakaguchi, R.; and Collins, L. M. 2014. Histograms of oriented gradients for landmine detection in ground-penetrating radar data. *Geoscience and Remote Sensing, IEEE Transactions on* 52(3):1539–1550.

van der Walt, S.; Schönberger, J. L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J. D.; Yager, N.; Gouillart, E.; Yu, T.; and the scikit-image contributors. 2014. scikit-image: image processing in Python. *PeerJ* 2:e453.

Verma, A., and Vig, L. 2014. Using convolutional neural networks to discover cogntively validated features for gender classification. In *Soft Computing and Machine Intelligence (ISCMI), 2014 International Conference on*, 33–37. IEEE.

Wu, Y.; Zhuang, Y.; Long, X.; Lin, F.; and Xu, W. 2015. Human gender classification: A review. *arXiv preprint arXiv:1507.05122*.