Improving Face Gender Classification By Adding Deliberately Misaligned Faces To The Training Data

M. Mayo, E. Zhang

Dept. of Computer Science, University of Waikato
Hamilton, New Zealand.

Email: mmayo@cs.waikato.ac.nz

Abstract

A novel method of face gender classifier construction is proposed and evaluated. Previously, researchers have assumed that a computationally expensive face alignment step (in which the face image is transformed so that facial landmarks such as the eyes, nose, chin, etc, are in uniform locations in the image) is required in order to maximize the accuracy of predictions on new face images. We, however, argue that this step is not necessary, and that machine learning classifiers can be made robust to face misalignments by automatically expanding the training data with examples of faces that have been deliberately misaligned (for example, translated or rotated). To test our hypothesis, we evaluate this automatic training dataset expansion method with two types of image classifier, the first based on weak features such as Local Binary Pattern histograms, and the second based on SIFT keypoints. Using a benchmark face gender classification dataset recently proposed in the literature, we obtain a state-of-the-art accuracy of 92.5%, thus validating our approach.

Keywords: Gender classification, face detection, face alignment, face classification, Spatial Pyramid, Local Binary Pattern, SIFT keypoints, Support Vector Machines, image classification, machine learning.

1 Introduction

Automatically assigning a gender to faces detected in a scene is a challenging pattern recognition problem. Whereas face detection can be performed with high accuracy [1,2], gender assignment given segmented face images is much more difficult. Despite this, accurate face gender classification has many useful potential applications. For example, in digital video libraries, there is a need to detect the current speaker and annotate videos with this information [3]. Gender classification could be useful here by reducing ambiguities, for example matching only female voices to female faces. Face gender classification may also be useful in more general face recognition systems, where an accurate predication of gender could eliminate false identifications on the basis of mismatched gender.

An open question, and the main question relevant to this paper, is whether automatic face alignment is needed for gender classification. Face misalignments occur, for example, when the face detection algorithm does not perfectly detect the bounding box around a face – the face may be shifted a little off centre. Misalignments may also occur simply because people have different sized faces, or their face is tilted slightly.

Automatic face alignment therefore involves searching the detected face image for basic facial features such as the eyes, nose, mouth, and chin. The face image is then transformed so that the detected facial features line up in all images. This, according to its proponents, should lead to better gender classification performance.

Mäkinen and Raisamo [2] evaluated the effects of face alignment on automatic gender classification performance. They used four different methods of facial alignments, three of which were based on Active Appearance Models (AAMs) [4] and one based on profile alignment [2]. These authors also built and made available a standard dataset for face gender classification experiments, which consists of a set of faces detected using Viola & Jones' face detector [1] applied to a subset of images from the FERET face recognition database [6,7]. The faces are not aligned or registered, and only the bounding box around the face as determined by the face detection algorithm is available. The face are categorized into two classes, male and female, with a training set consisting of 304 images (half male, half female) and a test set consisting of 107 images (60 male and 47 female).

What Mäkinen and Raisamo found was that all alignment methods currently lead to either a loss in accuracy or no change in accuracy compared to classifying detected images without any alignment. Their average results were 84.6% without alignment compared to 82.1% when alignment was used. Manually aligning the faces (i.e. using a human to indicate the location off facial features), however, led

978-1-4244-2582-2/08/$25.00 ©2008 IEEE

to a slight increase in accuracy – to 87.1% on the test data. This prompted the authors to conclude that current automatic face alignment methods are of insufficient quality to be useful in gender classification.

Besides the apparent drop in performance that automatic alignment methods lead to, there is also the computational overhead to consider, even if alignment methods do improve accuracy eventually. A face detector with built-in automatic alignment of faces is likely to be much more complex and slower at testing time than a straightforward classifier that operates directly on the detected face images themselves. Real-time performance under these conditions therefore may be hindered.

Our main contribution therefore, is to show that automatic alignment methods are not needed if the machine-learning based gender classification approach is made robust to misalignments. Rather than attempting to align faces as a preprocessing step for classification of new face images, we propose instead to artificially extend the training set used to build the classifier with examples of deliberately misaligned faces. By increasing the diversity of images in the training set in this way, a robust classifier can be built that can cope with any non-extreme variation due to misalignment. Effectively, the "alignment" step is shifted from testing time to the training phase.

To illustrate our approach, we use two different methods of artificially extending the dataset: (i) copying the original training images and cropping the copies along one border (a different border for each copy) and then adding the cropped copies back to the dataset (which corresponds to translation misalignments); and (ii) adding both cropped and rotated copies of original images to the dataset (translation and rotation misalignments).

We also use two different methods of face image classification, a "weak" approach and a "strong" approach. Weak approaches involve computing many low-level statistical features that are chosen a priori from the training images, and using these as the basis for classification. The features used here are Spatial Pyramids and Local Binary Pattern frequency histograms

Strong approaches, on the other hand, involve the computation of more sophisticated features. Unlike weak features, they cannot be chosen a priori, because the features are derived only after analysis of the training images. Strong features are therefore considered to be more informative features, and fewer of them are needed as a result. The strong features we consider in this paper are the presence/absence of frequently occurring SIFT keypoints [7] in the training images. In both cases, Support Vector Machines (SVM) [12, 13] and Random Forests [14] are used as the classifiers.

Our results show that our proposed method is highly effective: the best accuracy on the gender classification dataset achieved is 92.5%, representing a new state-of-the-art performance on this dataset.

2 Face Gender Classification

Before describing the experiment and analysis, a more detailed overview of the features and classification methods used in this study is given.

2.1 Weak Feature-Based Classification

Weak features are low level statistical measures derived directly from image pixels. Four simple but effective examples of weak features are the mean, variance, skew and kurtosis of an image's intensity histogram, all of which are used in this study.

Another frequently used image feature is the Local Binary Pattern (LBP) histogram [8]. Although originally used in the domain of texture recognition, this feature type has also recently been found to be effective for face recognition (e.g. [9]).

Briefly, a LBP is a property of a pixel. The pixel's n circular neighbours at distance r are examined (bilinear interpolation being used if necessary), and a binary string of length n is constructed such that the ith bit of the string is 1 if the neighbour's intensity exceeds that of the pixel, and 0 otherwise. Neighbours must be equally spaced around the perimeter of the circle. If, upon a circular transversal of the bit-string, there are two or less 0 to 1 or 1 to 0 transitions, then the LBP is considered "uniform" and therefore assigned to a category specified by the number of 1s in the string.

Figure 1 gives an example of a LBP that is uniform. The bit-string for the pixel (starting with topmost neighbour and running clockwise) is 11100001, yielding a uniform LBP of category 4. On the other hand, a pixel with a bit string such as 01100110 is not uniform and therefore the pixel would not be considered to have a valid LBP.

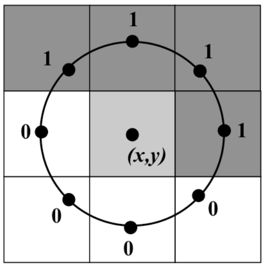

Figure 1: A LBP calculated for a pixel (x,y).

For any given grey scale image, a histogram of uniform LBP frequencies can be constructed with n+1 bins. They tend to capture the edges, curves, peaks and troughs in an image.

Low level or weak features can be computed either globally or by image region. One effective method that captures both global and local variability is the spatial pyramid [10]. Spatial pyramids repeatedly subdivide an image, computing all features repeatedly for all progressively smaller sub-images.

The first image is always the global image, and then the image is divided into 2x2 sub-images, and features are computed from each of those. The image may then be further subdivided, this time 4x4 regions, and so on. For a spatial pyramid with l levels, the maximum granularity will be a division of an image into 2l-1x2l-1 sub-images. Figure 2 illustrates a division of a face image into sub-images for an l=3 pyramid.

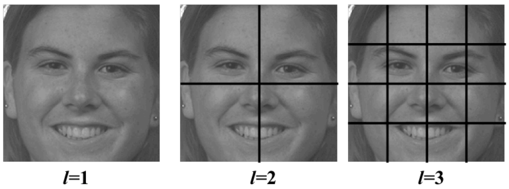

Figure 2: Sub-images for a spatial pyramid.

The primary advantage of spatial pyramids is that they capture the spatial distribution of features at the finer resolutions, while also maintaining the global features that are in themselves highly effective features for classification.

We used all of these weak features (i.e. histogram statistics, LBPs, and spatial pyramids) as the input to our SVM and Random Forest classifiers.

2.2 Strong Feature-Based Classification

The second classification method is a novel approach based on local invariant feature descriptors, specifically SIFT keypoints [7]. SIFT keypoints are computed from image regions that are detected by an interest point detection process. Each keypoint descriptor contains 128 attributes that describe that region in a scale and orientation invariant way. We can simply see SIFT keypoints as the fingerprint of images, where each fingerprint identifies a unique feature of an image and hence enables us to discover similar features across different images.

The basic idea is that the presence or absence of a particular keypoint is used as an attribute in the feature vector for a face image. And instead of building the feature vector from the entire collection of keypoints that can be calculated from the training data, our approach is to include only keypoints that frequently occur in the images – since less frequently occurring keypoints are more likely to be irrelevant. The primary advantage of this strong approach is that many less features should be needed to achieve the same levels of accuracy as in the weak case.

The rule we use for matching two keypoints is to compute the χ^2 distance between their descriptors. If this distance falls below a certain fixed threshold, they are considered to be the same keypoint, otherwise they are considered different keypoints. The χ^2 distance metric for two keypoints, *x* and *y*, is shown in equation (1).

$$d\chi^2(x,y) = \sum_{i=1}^{128} \frac{(x_i - y_i)^2}{x_i + y_i} \qquad (1)$$

Although keypoint-based approaches in computer vision are most frequently used for object recognition in cluttered scenes, we were interested in testing this approach in the completely different domain of gender classification.

3 Experiment

In the experiments we performed, we used the two different methods of classification (weak and strong) with three different versions of the training dataset, the original version used in [2], and two new versions created by automatically expanding the original training set with the addition of deliberately misaligned copies of the original images.

For weak classification, we used the four intensity statistics (i.e., mean, variance, skew, and kurtosis) as features, along with a LBP histogram with *r*=2 and *n*=16. This gave a total of 21 features per image region. We then extracted features using spatial pyramids of three different sizes, specifically *l*=2, *l*=3 and *l*=4. This yielded a total feature count per image of 106, 442, and 1786 features respectively.

Two versions of an SVM classifier [12, 13] were used to perform the classification, specifically an SVM with a linear kernel ($\text{SVM}_{\text{linear}}$) and an SVM with a quadratic kernel (SVM_{quad}). A Random Forests classifier [14] with 200 random trees (RF_{200}) was also used. The implementations of all of these classifiers were those available in Weka version 3.5.6 [11], and to avoid parameter tuning to the test set, all the other parameters were left at default values.

For the strong classification experiment using frequent keypoints, we choose the number of selected frequent keypoints to be either 50 or 200. We also used the same $\text{SVM}_{\text{linear}}$, $\text{SVM}_{\text{quadratic}}$, and RF_{200} classifiers as in the weak classification case.

The artificially expanded versions of the datasets were generated as follows. First, the original training set (Train1), consisting of 304 images, was expanded by copying every image four times, and cropping 16 pixels off each of the four copies along one of the four different borders, either top, bottom, left or right. This yielded a second version of the training dataset (Train2) with 304*5=1,520 unique images. The third artificially expanded dataset was constructed by duplicating all of the images in Train2 twice, with one copy being rotated by 5°, and the other by -5°. This yielded a dataset (Train3) with 1,520*3=4,560 unique images.

Consequently, for every original image from Train1, Train2 contained four misaligned images in addition to the original image; and Train3 contained twelve misaligned versions in addition to the original image.

The only image preprocessing performed prior to feature extraction was the application of two averaging filters in order to smooth the images. All classifiers were tested on the same set of 107 test images regardless of which training set or classification method they used.

4 Results

Using the weak classification method with intensity statistics, LBPs, spatial pyramids and SVMs, the best accuracy achieved was 92.5%. A more detailed breakdown of results by training set, spatial pyramid size, and classifier is given in Table 1.

Table 1: Classification accuracy on the test set by classifier, training set, and spatial pyramid size (number of features).

		Spatial Pyramid Size		
Classifier	**Training**	***l*=2 (106)**	***l*=3 (442)**	***l*=4 (1,786)**
SVM$_{\text{linear}}$	**Train1**	79.4	84.1	88.7
	Train2	82.2	91.6	91.6
	Train3	86.0	90.6	91.6
SVM$_{\text{quad}}$	**Train1**	80.4	85.0	89.7
	Train2	82.2	89.7	90.6
	Train3	81.3	**92.5**	90.6
RF$_{200}$	**Train1**	84.1	89.7	86.0
	Train2	84.1	88.7	87.8
	Train3	78.5	86.9	87.8

An examination of Table 1 reveals that in most cases, the addition of deliberately misaligned versions of the training images to the original training set for the SVM classifiers significantly improves prediction accuracy. Furthermore, the improvement is most pronounced when the number of features is low.

For example, when *l*=2 (only 106 extracted features), the performance of SVM$_{\text{linear}}$ jumps from 79.5% given the original training set, to 86% after being trained on Train3. When more features are extracted, such as the *l*=3 case (where there are 442 features extracted), the performance of SVM$_{\text{linear}}$ given Train2 jumps to 91.6%.

The most pronounced gain however is for SVM$_{\text{quad}}$ when *l*=3, which goes from 85% accuracy given the original training set to 92.5% given Train3 as a training set – an increase in accuracy of 7.2%, and the best recorded accuracy on this dataset. As the number of features increases to *l*=4, the accuracies of the classifiers given only the original training images also increases, for example to 89.7% for SVM$_{\text{quad}}$, but this is at significant additional computational cost (specifically, extracting 1,786 features per image compared to 442 features when *l*=3).

Interestingly, the Random Forest classifier does not benefit from artificial data expansion, and actually performs consistently worse given Train3 than it does given Train1.

With regards to the strong classification method using SIFT keypoints, the results by training set are given in Table 2.

Table 2: Classification accuracy on the test set by classifier and number of selected keypoints/features.

		Num. Selected Keypoints	
Classifier	**Training**	***n=50***	***n=200***
SVM$_{\text{linear}}$	**Train1**	68.2	73.8
	Train2	68.1	73.6
	Train3	67.5	74.0
SVM$_{\text{quad}}$	**Train1**	71.0	72.0
	Train2	70.5	71.5
	Train3	70.5	72.1
RF$_{200}$	**Train1**	70.1	**79.4**
	Train2	69.8	77.6
	Train3	70.3	78.4

As can be observed, there is little difference in the performance of the keypoint-based classification approach given expansion of the training set, and the best accuracy of this method of 79.4% is far below that of the weak approach.

5 Conclusion

The main focus of this paper is the question of how a face gender classifier should cope with misaligned face images – are expensive automatic face alignment methods needed? Or can classifiers be adapted to cope with misalignments? We have argued in this paper that it is better to build the classifier to be robust to misalignments than it is to add a computationally expensive face alignment step to the face classification phase.

Furthermore, we have shown a method by which this may be achieved: artificially expanding the training data with deliberately misaligned faces. The misalignments we choose were translation (via cropping) and small rotations, but other possible deliberate misalignments may bear even better results, for example small random affine transformations. The results bear witness to the effectiveness of this approach, with an overall best accuracy of 92.5%.

An unexpected benefit of this research is that artificially expanding the training dataset as we have described actually makes the features themselves, in the weak classification case, much more informative and therefore less are needed. For example, only a three level spatial pyramid is required to achieve accuracies of approximately 90% or above if the training set is artificially expanded, whereas a four level spatial pyramid is required to get the same accuracy if only the original training images are used.

Overall, this experiment has validated our new proposed approach to face gender classification.

6 Acknowledgements

Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

[1] Viola P. & Jones M. 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proc IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 511-518.

[2] Mäkinen E. & Raisamo R. 2008. Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 30(3) pp. 541-547.

[3] Wu Y., et. al. 2007. Robust Speaking Face Identification for Video Analysis. In. *Advances in Multimedia Information Processing - PCM 2007,* Lecture Notes in Computer Science 4810*,* Springer, pp. 665-674.

[4] Cootes T. & Taylor C. 2001. *Statistical Models of Appearance for Medical Image Analysis and Computer Vision.*

[5] Phillips P., Weschler H., Huang J., Rauss P. 1998. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J.* 16(5) pp. 295-306.

[6] Phillips P., Moon H., Rizvi S., & Rauss P. 2000. The FERET Evaluation Methodology for Face Recognition Algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22, v. 1090-1104.

[7] Lowe D. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), pp. 91-110.

[8] Mäenpää T & Pietikäinen M. 2005. Texture analysis with local binary patterns. In Chen C. & Wang P. (eds) *Handbook of Pattern Recognition and Computer Vision,* 3rd ed, World Scientific, 197-216

[9] Ahonen T., Hadid A. & Pietikäinen M. 2006. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. on Pattern Recognition and Machine Intelligence* 28(12) 2037-41.

[10] Lazebnik S., Schmid C. & Ponce J. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* - Volume 2, pp. 2169-2178.

[11] Witten I. and Frank E. 2005. Data Mining: Practical machine learning tools and techniques (2nd Edition). Morgan Kaufmann, San Francisco.

[12] Platt J. 1998. Machines using Sequential Minimal Optimization. In Schoelkopf B., Burges C. and Smola A. (Eds.) Advances in Kernel Methods -- Support Vector Learning.

[13] Keerthi S., et al. 1999. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13(3) 637--649.

[14] Breiman, L. 2001. Random Forests. Machine Learning 45 (1):5-32.