Robust Face Identification with Small Sample Sizes using Bag of Words and Histogram of Oriented Gradients

Mahir Faik Karaaba, Olarik Surinta, L. R. B. Schomaker and Marco A. Wiering
Institute of Artificial Intelligence and Cognitive Engineering (ALICE),
University of Groningen, Nijenborgh 9, Groningen 9747AG, The Netherlands

Keywords: Face Recognition, Histogram of Oriented Gradients, Bag of Words, Small Sample Problem.

Abstract: Face identification under small sample conditions is currently an active research area. In a case of very few reference samples, optimally exploiting the training data to make a model which has a low generalization error is an important challenge to create a robust face identification algorithm. In this paper we propose to combine the histogram of oriented gradients (HOG) and the bag of words (BOW) approach to use few training examples for robust face identification. In this HOG-BOW method, from every image many sub-images are first randomly cropped and given to the HOG feature extractor to compute many different feature vectors. Then these feature vectors are given to a K-means clustering algorithm to compute the centroids which serve as a codebook. This codebook is used by a sliding window to compute feature vectors for all training and test images. Finally, the feature vectors are fed into an L2 support vector machine to learn a linear model that will classify the test images. To show the efficiency of our method, we also experimented with two other feature extraction algorithms: HOG and the scale invariant feature transform (SIFT). All methods are compared on two well-known face image datasets with one to three training examples per person. The experimental results show that the HOG-BOW algorithm clearly outperforms the other methods.

1 INTRODUCTION

Face recognition is an important skill which we humans perform without much effort. Computers, on the other hand, still do not perform good enough to be fully trusted in real-world applications. There are two distinct application fields which are both generally called face recognition. One is face identification, in which the question is to whom a given face image belongs, the other is face verification that tries to answer the *same/not same* question given two face images. While face identification is basically a multi-class classification task and requires a reference training image dataset for identity registration, face verification is a binary classification task and does not require a reference training set containing the identity of persons. In this paper, we focus on the face identification problem.

Face identification is an active research field due to different important possible applications and several difficulties which are not yet solved (Jafri and Arabnia, 2009). Some of these difficulties have to do with pose variances and facial expressions, which arise from the capability we have to move our head and to express ourselves with our faces. Being able to move our heads in various angles results in very different poses of the face of the same person (Zhang and Gao, 2009). If we tilt our heads clockwise or counter clockwise, a simple geometrical alignment procedure is enough to transform the face image to its frontal position. On the other hand, if we turn our head to the left, right, up or down, then without a complex 3d interpolation technique (Chu et al., 2014), geometrical normalization is very difficult, which in turn causes significant performance losses for a face recognition algorithm. Another difficulty is the non-rigidity of the face because we can change the appearance of our faces significantly (opening and closing of mouth and eye, etc). Yet another difficulty is related to occlusions which can be caused by different objects such as glasses, hands we can bring to our face, and shawls (Azeem et al., 2014).

There are many face recognition algorithms that rely on a large amount of training data to work optimally. Since more data will include more variances, the trained classifiers can generalize better to the unknown distribution of the test images. However, in a variety of application fields such as forensic research, data collection is very difficult and the obtained reference data set may not include more than a couple of

Karaaba, M., Surinta, O., Schomaker, L. and Wiering, M.
Robust Face Identification with Small Sample Sizes using Bag of Words and Histogram of Oriented Gradients.
DOI: 10.5220/0005722305820589
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 582-589
ISBN: 978-989-758-175-5

images per person. This is called the small sample problem (SSP). Many research attempts target SSP (Yan et al., 2014; Lu et al., 2013; Su et al., 2010), and in this paper we also propose a new algorithm to deal with few training examples for face identification.

Related Work. The first successful face recognition algorithm, called Eigenfaces (Turk and Pentland, 1991), was based on the nowadays well-known subspace method principal component analysis. Another often used method is Fisherfaces (Belhumeur et al., 1997) that uses linear discriminant analysis. These methods can perform well if a large amount of correctly aligned and normalized face data is available. However, since they directly use pixel intensities as input data, pose variances and alignment errors can easily deteriorate the performance of these algorithms.

To cope with the noise caused by illumination and pose variances, edge and local feature extraction based methods have been proposed. Some of the best known of these are Gabor filters (Jemaa and Khanfir, 2009), the histogram of oriented gradients (HOG) (Dalal and Triggs, 2005), the scale invariant feature transform (SIFT) (Lowe, 2004) and local binary patterns (LBP) (Ahonen et al., 2004). These methods have been shown to yield better performances than the use of Eigenfaces or Fisherfaces. However, without additional preprocessing on the input data and a sufficient number of training images, they cannot very well handle pose differences or alignment errors.

To cope with pose differences and alignment problems, the bag of words (BOW) method (Csurka et al., 2004), which has been successfully applied for different computer vision problems (Shekhar and Jawahar, 2012; Montazer et al., 2015), was proposed for the face recognition problem (Li et al., 2010; Wu et al., 2012). In this method, input images are treated non-holistically by their many sub-images. These sub-images are processed by a clustering algorithm to create a codebook (the bag of words) and this codebook is then used to extract feature vectors from images which are finally given to the classifier.

Similarly to the BOW approach, in (Simonyan et al., 2013), many sub-images processed by the SIFT descriptor are used to train gaussian mixture models to compute improved Fisher vectors (Perronnin et al., 2010) for face verification. The results reported in their paper are comparable with the results of state-of-the-art face verification papers.

As for classifiers used for face recognition, k-nearest neighbour (K-NN), support vector machines (SVM) (Vapnik, 1998) and artificial neural networks (ANN) have been shown to be successful. If classifier speed is important and features from face images are selected robustly, then K-NN can be a good choice. Since no training is required for using the K-NN classifier, it is practical for fast face recognition applications, in which possibly new people are continuously added to the dataset. However, if accuracy is more important than speed, then an SVM (Wei et al., 2011) and an ANN can be preferable, even though they need retraining in case the dataset is augmented with new people and images.

Convolutional neural networks (CNNs), as a powerful feature extractor and classifier, are currently considered by researchers as one of the state-of-the-art machine learning algorithms. CNN is a special kind of multi-layer perceptron, which has many specialized layers used for feature extraction and classification. In a recent CNN based face verification study (Parkhi et al., 2015), a novel database construction and a CNN architecture are presented. Here, they construct a face database with 2.6K subjects composing of total 2.6M images from Internet, removing the duplicate images by employing a state-of-the face recognition application as well as a group of human annotators. After the database construction, they optimize a relatively simpler new CNN which integrates a combination of the most efficient features of the state-of-the-art CNNs proposed recently for face recognition.

The SVM has also several varieties. Although it was first proposed as a linear classifier, non-linear models have been proposed to classify data sets, which are not separable with the standard linear SVM. Another popular SVM algorithm is the L2-norm regularized SVM (L2-SVM) (Koshiba and Abe, 2003; Deng et al., 2012). It is used to tackle the problem that occurs when the size of the feature vectors is very long (e.g. more than 2,000 items) which cannot be handled very efficiently by the standard SVM.

2 FACE RECOGNITION BY THE HOG-BOW METHOD

Contributions. In this paper, as our main contribution, a bag of words (BOW) algorithm is proposed that uses feature vectors extracted with the histogram of oriented gradients (HOG) to recognize faces under small sample per person conditions (SSPP). Although the HOG and BOW algorithms are well-known algorithms, to the best of our knowledge, the combination of them is not evaluated for face recognition, especially in the case of SSPP.

In our method, a K-means clustering algorithm is used to compute the visual codebook from feature vectors extracted by HOG from many randomly cropped sub-images. Then this codebook is used to

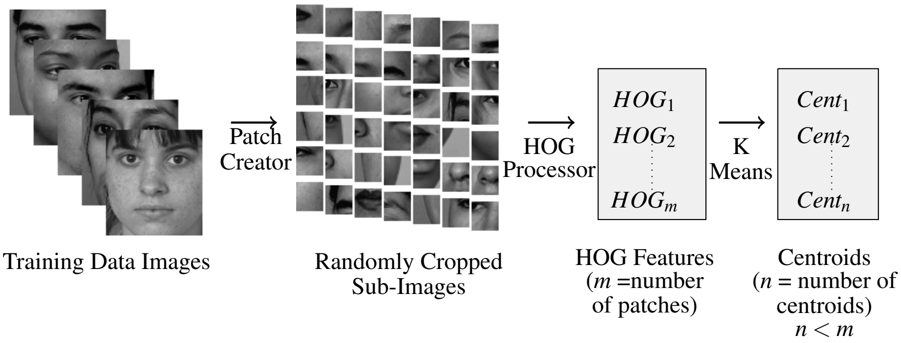

Patch Creator
HOG Processor
HOG_1
HOG_2
HOG_m
K Means
$Cent_1$
$Cent_2$
$Cent_n$
Training Data Images
Randomly Cropped Sub-Images
HOG Features (m =number of patches)
Centroids (n = number of centroids) $n < m$

Figure 1: Graphical depiction of the codebook construction in *HOG-BOW*.

compute feature vectors from all images in the training and test set. The computed feature vectors and the labels from the training images are subsequently fed into an L2-SVM classifier to learn the model which is used to classify faces.

Additionally, we compared the HOG-BOW method to two other well-known methods, namely HOG and the scale invariant feature transform (SIFT), both using a standard-SVM with the radial basis function (RBF) kernel as the classifier since the feature vectors created by these methods are relatively shorter in size than those of the HOG-BOW method. We performed experiments using two datasets, namely FERET (Phillips et al., 1998) and LFW (Huang et al., 2007) with one, two and three training images per person. The results show that the HOG-BOW method clearly outperforms the other methods.

Paper Outline. The rest of the paper is organized as follows: In Section 2, the proposed face recognition algorithm is described. In Section 3, experimental settings and the results are presented. In Section 4, the conclusion and future work are given.

The idea of the bag of visual words (BOW) is that, just as a text is composed of many words, an image is composed of many sub-images which resemble visual words that can be present in an image (Csurka et al., 2004). In our proposed HOG-BOW method, the bag of words model is constructed by using features extracted by HOG from sub-images, instead of directly using pixel intensities. We will now explain the codebook construction, the computation of the activity matrix of visual words on the entire image, and the final creation of the feature vector containing visual word activities per block. Note that we use the L2-norm regularized SVM as classifier, but we will not explain it because it is a well-known supervised learning algorithm.

2.1 Codebook Construction

Random cropping is used to extract a large number of sub-images (in our experiment we used 500,000 sub-images) from the training set. Then these sub-images are processed by the HOG filter and the extracted feature vectors are given to a K-means clustering algorithm that computes the centroids which serve as the visual words and make up the codebook. For the graphical illustration of the codebook construction, see Figure 1.

2.2 Creating Activity Matrix

After the codebook is constructed, the activities of all visual words are calculated per image. These activities denote the presence of different visual words in the image. For this, sub-images are obtained using a sliding window approach using a stride of 1 pixel. To compute the activities the soft assignment approach is adopted in our system. Soft assignment schemes have previously been shown to outperform hard assignment schemes where one sub-image only activates the winning cluster (or visual word). We will now explain in detail how the activities a_{ij} of the activity matrix A are computed for a single image, where i is the cluster index, and j is the index of the sub-image (patch). Our method used the soft assignment scheme proposed in (Coates et al., 2011):

$$a_{ij} = max\{0, \bar{d} - d_{ij}\} \tag{1}$$

where $\bar{d}$ is the mean of the elements of d_{ij} and d_{ij} is the Euclidean distance between a cluster c_i and an image patch p_j:

$$d_{ij} = \left\| p_j - c_i \right\|_2 \tag{2}$$

Note that p_j is the HOG filtered sub-image vector and c_i is a cluster centroid computed from feature vectors extracted by HOG.

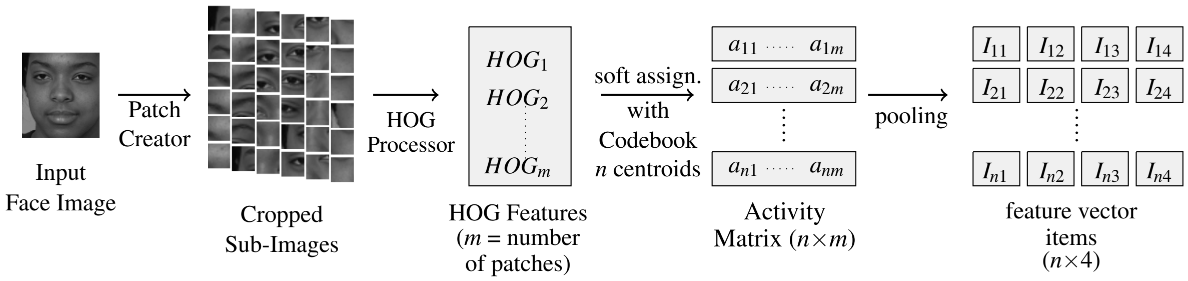
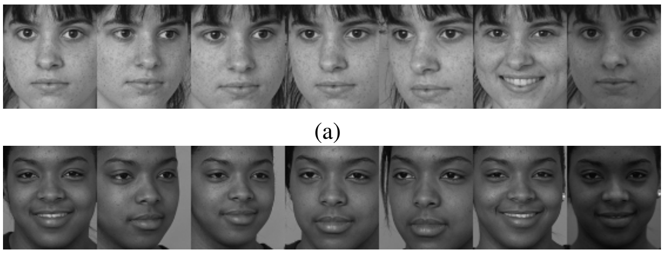


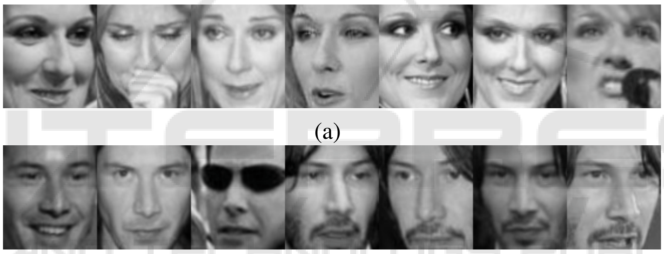
Figure 2: Graphical illustration of the feature vector creation from the codebook in *HOG-BOW*.



(a)

(b)

Figure 3: Sample aligned face images of two subjects from the FERET dataset.



(a)

(b)

Figure 4: Sample aligned face images of two subjects from the LFW dataset.

2.3 Image Partitioning and Feature Vector Construction

After the centroid activities are computed for each sub-image, each row of the activity matrix (which corresponds to centroid activities for all sub-images) is summed up per image block. We will use B blocks to partition each image and to better keep the spatial relations between activated visual words. For this we compute visual word activities I_{ib} for each cluster i and each block b:

$$I_{ib} = \sum_{j} a_{ij}, j \in block(b) \tag{3}$$

After this the size of the resulting feature vector is $B \times n$. These feature vectors are then given to a classifier. In our experiments we use 4 blocks of equal size. For the feature vector creation, see Figure 2.

3 EXPERIMENTAL SETTINGS AND RESULTS

In this section, we first briefly explain the datasets used in the experiments, the alignment of the face images, and the selected parameters. After that the results are presented and discussed.

3.1 Datasets

In our experiments we use two datasets, namely FERET (Phillips et al., 1998) and Labeled Faces in the Wild (LFW) (Huang et al., 2007). We divide each dataset into train and test sets by selecting from 1 up to 3 reference images randomly as training data and the rest is used as test data.

The FERET dataset was created by the defence advanced research projects agency (DARPA) and the national institute of standards and technology (NIST) to evaluate face recognition algorithms. We selected a subset of this dataset to use in our experiments, in total 196 subjects are used with 7 face samples per subject. This subset has basically 3 features: illumination, pose and expression variances which present challenges for the performance of a typical face recognition system. For example face photos of FERET, see Figure 3.

The LFW dataset is introduced in (Huang et al., 2007) to evaluate face recognition algorithms under unconstrained conditions. It contains approximately 13,000 images of around 6,000 subjects. These images are mainly collected from news web sites. In the experiments, we have selected 150 subjects each of which contains at least 7 samples. For example face photos of the LFW dataset, see Figure 4.

For both datasets, we adopted a similar experimental setup as described in (Yan et al., 2014). The differences between our and their protocols are briefly given as follows: For the FERET, (Yan et al., 2014) uses 200 subjects for which we could find only 196 in the copy of our FERET dataset folder with the same subset specification they defined. The second difference is that while in (Yan et al., 2014) LFW subjects are chosen as 10 samples per subject where even some of these samples are chosen from subject folders which contain more than 10 samples, we choose the subject folders which contains at least 7 subjects and without a maximum number limit.

3.2 Alignment

We use an eye-coordinate based 2D alignment for all the face images before the experiments. In this method, eye centers are used to compute the roll angle of the face. Then the face is rotated to roll-normalized position as described in (Karaaba et al., 2015). All eye coordinates are obtained from the dataset directories, except for some images (of each subject) of the FERET dataset for which we used an automatic alignment algorithm.

3.3 Selected Parameters

In this section, we will present the selected parameters that worked best in our experiments. For all the train and test images, we use 80×88 as the image resolution. For SIFT, we used 40×44 as the patch size which corresponds to 4 sub-images for each face image. Then for each sub-image by applying the standard SIFT algorithm, we obtained a feature vector with size $(128\times4) = 512$.

For HOG, 10×11 is used as the patch size (8×8 = 64 patches) and the number of bins is chosen as 24. Hence 8×8×24 is used and the size of the feature vector is 1,536.

For HOG-BOW, 600 centroids are used. For the FERET dataset, 15×15 is selected as the patch size and for the LFW dataset we selected 20×20 as the patch size, which worked better for LFW. The reason different patch sizes were found to work best can be due to differences in the resolution of the two datasets. For both datasets, 4 block partitions are used resulting in a feature vector with size (600×4) = 2,400. For all methods a linear L2-norm regularized SVM is used, for which the C parameter is tuned using cross validation.

3.4 Experiments and Results

In our experiments, 10-fold cross validation is used. We randomly select $(t = 1,2,3)$ samples for each subject from the training set and the rest of the samples is used as the test data. It should be noted that in (Yan et al., 2014), 20-fold cross validation is employed.

Tables 1, 2, and 3 show the results [1]

(average accuracy and standard deviation) on FERET and LFW for $t = 1$, $t = 2$ and $t = 3$, respectively. The results show that the HOG-BOW method obtains the best performances for both datasets, except for LFW without mirrored images with $t = 3$. Especially when the available training data is the smallest in number, the HOG-BOW method shows a significant performance gain (9% and 18% for FERET, and 4% and 1% for LFW for the mirrored and non-mirrored case respectively) compared to the HOG method, which performs second best. The average performance gain over all 12 experimental results of HOG-BOW compared to HOG is slightly more than 5%.

As for the mirrored image samples, a significant performance improvement is obtained for the FERET dataset, especially where $t = 1$. The improvement becomes smaller when more original training data is provided. For instance, while the performance difference is only around 1% for $t = 3$ for almost all the methods, for $t = 1$ this is 4% for the HOG-BOW method and even 13% for the HOG and SIFT methods. This shows that mirrored data sampling is a powerful way to boost the face identification performance for the FERET dataset when there are only one or two training examples per person. On the other hand, for the LFW dataset, mirrored images, except for the HOG-BOW method, do not provide any significant performance gains and even decrease the performance in some cases (e.g. the HOG method with $t = 1$). This

[1] Note that results of DMMA and MS-CFB are referenced from the same source (Yan et al., 2014).

Table 1: Face Recognition Results on FERET and LFW ($t = 1$).

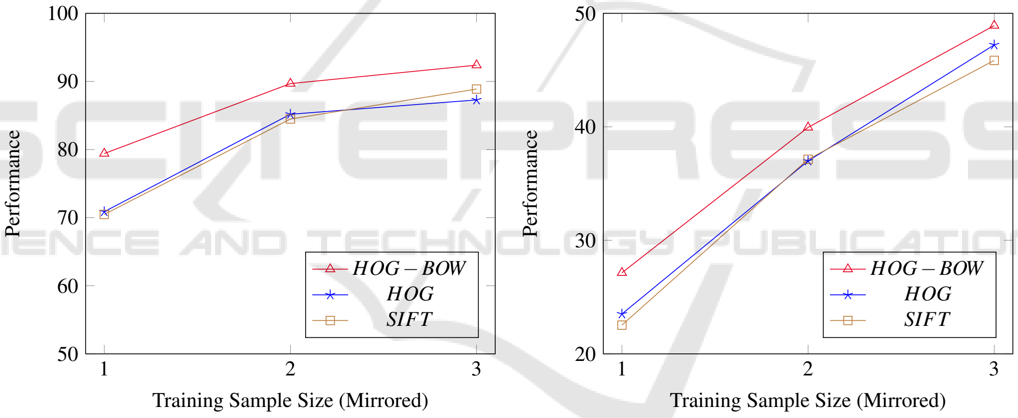
Methods	FERET		LFW	
	Mirrored	Non-Mirrored	Mirrored	Non-Mirrored
HOG	70.87±1.3	57.62±0.7	23.51±0.6	23.73±0.8
SIFT	70.47±1.2	56.51±1.2	22.53±1.0	21.56±0.9
HOG-BOW	**79.41**±3.3	**75.97**±1.1	**27.14**±1.0	**24.68**±0.8
DMMA (Yan et al., 2014)	-	65.24±2.0	-	22.17±2.8
MS-CFB (Yan et al., 2014)	-	66.60±2.1	-	21.15±2.9

Table 2: Face Recognition Results on FERET and LFW ($t = 2$).

Methods	FERET		LFW	
	Mirrored	Non-Mirrored	Mirrored	Non-Mirrored
HOG	85.18±0.7	77.78±1.3	36.99±1.2	37.25±1.0
SIFT	84.48±0.8	75.75±0.8	37.14±1.0	36.14±1.1
HOG-BOW	**89.68**±0.6	**86.13**±1.3	**39.95**±1.3	**39.10**±1.1
MS-CFB (Yan et al., 2014)	-	80.60±1.4	-	37.17±1.8

Table 3: Face Recognition Results on FERET and LFW ($t = 3$).

Methods	FERET		LFW	
	Mirrored	Non-Mirrored	Mirrored	Non-Mirrored
HOG	87.28±0.8	86.44±0.9	47.22±1.6	**48.25**±1.2
SIFT	88.88±0.6	85.93±1.1	45.85±1.3	46.02±1.3
HOG-BOW	**92.39**±0.6	**92.62**±0.8	**48.92**±1.6	47.16±0.7
MS-CFB (Yan et al., 2014)	-	84.72±1.3	-	43.10±1.5



(a) FERET dataset

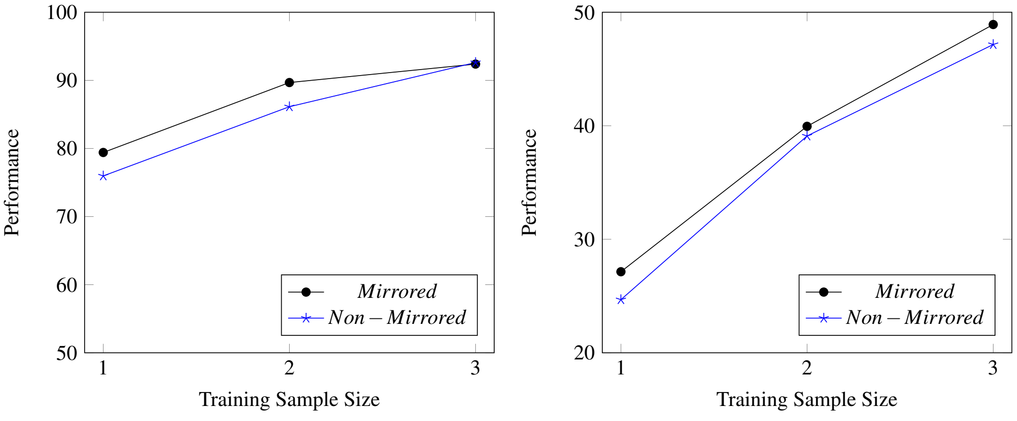
(b) LFW dataset

Figure 5: Average recognition performance of different methods versus different number of training samples per person on the FERET (a) and LFW (b) datasets with mirrored face images.

might be due to the nature of the LFW dataset where low resolution, occlusions and a high-degree of pose differences are prevalent.

The HOG-BOW method also significantly outperforms two state-of-the-art face recognition algorithms for the non-mirrored case with few training examples. These methods are the multi-subregion based correlation filter bank (MS-CFB) (Yan et al., 2014) with the cosine similarity metric and discriminative multi-manifold analysis (DMMA) (Lu et al., 2013), which were specially designed for face recognition problems with few examples.

We also show two additional figures drawn from the results to obtain more insights. The first one is the comparison of the methods in relation to the training sample size, see Figure 5. The second one is to see the performance effect when mirrored data is added, see Figure 6. As can be seen from the method comparison figures, the HOG-BOW method is always better than the other methods for each training data size if the images are mirrored and its performance stays a large margin above the performances of the other methods. Figure 6 shows that adding mirrored data helps to increase the performance of HOG-BOW the most when the training data size is the smallest ($t = 1$), although in most cases it improves the results.



(a) FERET dataset

(b) LFW dataset

Figure 6: Average recognition performance of HOG-BOW method with and without mirrored data versus different number of training samples per person on the FERET (a) and LFW (b) dataset.

4 CONCLUSION

In this paper, we described a new face identification algorithm, namely a bag of visual words using extracted features of histogram of oriented gradients (HOG-BOW). This method is designed to cope with small sample sizes in the training set, which is a challenge for obtaining good performances. We compared the HOG-BOW method with two other algorithms: the scale invariant feature transform (SIFT) and HOG, both with a standard SVM as classifier.

We have shown the effectiveness of the HOG-BOW method over the others. On the FERET dataset, for instance, it performs much better than the other methods for all the different selected small sample sizes of the training set. On the LFW dataset, except for $t = 3$ with the non-mirrored case, it also performs significantly better than the other methods. We also compared our results with two state-of-the-art face recognition algorithms by following similar dataset selections. From the results it can be seen that, HOG-BOW obtains state-of-the-art performances for face recognition with few training examples.

In future work, we plan to work on more datasets and we will further optimize the parameters of HOG-BOW to obtain higher accuracies. We are interested to use local binary patterns or features extracted with pre-trained convolutional neural networks (Krizhevsky et al., 2012) instead of HOG as the feature extraction scheme, and combine them with the bag of words approach. Finally, we want to experiment with other clustering algorithms which may work better than simple K-means clustering.

REFERENCES

Ahonen, T., Hadid, A., and Pietikinen, M. (2004). Face recognition with local binary patterns. In Pajdla, T. and Matas, J., editors, *Computer Vision - ECCV 2004*, volume 3021 of *Lecture Notes in Computer Science*, pages 469–481. Springer Berlin Heidelberg.

Azeem, A., Sharif, M., Raza, M., and Murtaza, M. (2014). A survey: face recognition techniques under partial occlusion. *Int. Arab J. Inf. Technol.*, 11(1):1–10.

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

Chu, B., Romdhani, S., and Chen, L. (2014). 3d-aided face recognition robust to expression and pose variations. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1907–1914.

Coates, A., Ng, A. Y., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 215–223.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

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, pages 886–893.

Deng, N., Tian, Y., and Zhang, C. (2012). *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*. Chapman & Hall/CRC, 1st edition.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Jafri, R. and Arabnia, H. R. (2009). A survey of face recognition techniques. *Journal of Information Processing Systems*, 5(2):41–68.

Jemaa, Y. B. and Khanfir, S. (2009). Automatic local Gabor features extraction for face recognition. *International Journal of Computer Science and Information Security (IJCSIS))*, 3(1).

Karaaba, M. F., Surinta, O., Schomaker, L. R. B., and Wiering, M. A. (2015). In-plane rotational alignment of faces by eye and eye-pair detection. In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, pages 392–399.

Koshiba, Y. and Abe, S. (2003). Comparison of L1 and L2 Support Vector Machines. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 2054–2059 vol.3.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Li, Z., Imai, J., and Kaneko, M. (2010). Robust face recognition using block-based bag of words. In *Pattern Recognition (ICPR), 20th International Conference on*, pages 1285–1288.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–110.

Lu, J., Tan, Y.-P., and Wang, G. (2013). Discriminative multi-manifold analysis for face recognition from a single training sample per person. volume 35, pages 39–51.

Montazer, G., Soltanshahi, M., and Giveki, D. (2015). Extended bag of visual words for face detection. *Advances in Computational Intelligence*, 9094:503–510.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg. Springer-Verlag.

Phillips, P. J., Wechsler, H., Huang, J., and Rauss, P. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306.

Shekhar, R. and Jawahar, C. (2012). Word image retrieval using bag of visual words. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 297–301.

Simonyan, K., Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2013). Fisher Vector Faces in the Wild. In *British Machine Vision Conference*.

Su, Y., Shan, S., Chen, X., and Gao, W. (2010). Adaptive generic learning for face recognition from a single sample per person. In *Computer Vision and Pattern Recognition (CVPR), the Twenty-Third IEEE Conference on*, pages 2699–2706.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Wei, J., Jian-qi, Z., and Xiang, Z. (2011). Face recognition method based on support vector machine and particle swarm optimization. *Expert Systems with Applications*, 38(4):4390 – 4393.

Wu, Y.-S., Liu, H.-S., Ju, G.-H., Lee, T.-W., and Chiu, Y.-L. (2012). Using the visual words based on affine-sift descriptors for face recognition. In *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–5.

Yan, Y., Wang, H., and Suter, D. (2014). Multi-subregion based correlation filter bank for robust face recognition. *Pattern Recognition*, 47(11):3487 – 3501.

Zhang, X. and Gao, Y. (2009). Face recognition across pose: A review. *Pattern Recognition*, 42(11):2876 – 2896.