Face Recognition from Low Resolution Images: A Comparative Study

Marc Babtist and Sebastian Wehkamp

Abstract—Surveillance cameras generally have a poor resolution. The face images captured by these cameras tend to be of low quality, in various poses and with differing illumination. Recognizing faces under these conditions is a difficult problem. In recent years, several new solutions for face recognition in low resolution images have been introduced. These solutions compensate for the quality, poses and illumination using different methods. This paper performs a comparative study of four of these state of the art solutions, along with the baseline super resolution method. The four methods to compare differ strongly in ways of approaching the problem. The techniques covered in this paper consist of a method that uses Multi-HOG, one that utilizes sparse coding of local features, another which uses multidimensional scaling based on tensors and finally an enhancement of super resolution using relationship learning. Over the course of the paper, we will explain how these different methods work, what the drawbacks and benefits to each of them are and how they compare to one another in terms of accuracy. The final conclusion of our paper indicates that RSLR has the best performance, followed by MDS and SCLF.

Index Terms—Face recognition, very low resolution (VLR), low-resolution matching, super resolution, multi-HOG, sparse coding of local features, tensor-based multidimensional scaling, relationship-learning-based super resolution.

1 Introduction

In face recognition, one seeks to match a set of gallery images to a probe image. The performance of automated face recognition systems deteriorates significantly when the resolution of input images degrades. Surveillance camera footage tends to be of low quality; often face images are made with a low resolution (LR) like 12x12, with varying illumination conditions and poses. These LR probe images then need to be matched with high resolution (HR) gallery images. A real world example case where this occurs would be criminal investigations, in which we want to compare HR mug-shots to LR face images captured by surveillance cameras. Some of these methods make the image enhancements as seen in fictional crime dramas an actual reality.

In recent years there have been some significant developments in solving this problem. In this paper we will review some state of the art methods and compare the performance. In order to provide the reader with a good overview of the different techniques, we will discuss the benefits and limitations for each algorithm.

The commonly used approach for matching a LR image with an HR image is to use **super resolution** (SR). SR creates a higher resolution image from the probe image with the primary goal of creating a good visual reconstruction. The problem is that the various SR methods are not designed with a recognition perspective in mind. Since SR is still a common approach, we consider it as the baseline method to which we will compare other methods.

The first method we cover is the **Multi-HOG** method. This algorithm is based on a combination of different histograms of oriented gradients (HOG). Two distance measures are proposed to calculate the distance between the LR probe image and the HR gallery image. To cope with the various poses a new technique is provided for finding the most similar region. A combination of the Multi-HOG, *distance measure*, and the *most similar region* techniques promises to deliver state of the art results. [8]

The second paper extracts local features from the LR probe image and projects them in a feature space. Linear regression is used to compare the projected features with the HR image. In this paper, we will refer to this method as **Sparse Coding of Local Features** or SCLF for short. [12]

The third algorithm utilizes multidimensional scaling to transform both the HR and LR images. Tensor analysis is used to localize facial landmarks in order to compute the features to scale to. We will call this manner of LR face recognition **Tensor-based Multidimensional Scaling** or TMS. [3]

The last paper in the comparison improves on the SR method by introducing relationships. The paper proposes an approach which learns the relationship between the HR image space and the LR image space using SR. This algorithm will be referred to as **Relationship-learning-based Super Resolution** or RLSR. [13]

After a more in-depth explanation of these methods in Chapter 2, we explain the datasets used by the various papers in Chapter 3. In the same chapter, we compare the results the original authors obtained on similar datasets. To compare the papers we will lay out the performance data of the methods described, while keeping the characteristics of the datasets they were run on in mind. This should provide a good view on what their strengths and weaknesses are. In Chapter 4, we show the results of these comparisons, while we draw conclusions about them in Chapter 5. Finally, in Chapter 6, we discuss on how to improve the algorithms and further research that can be done.

2 Methods

In this section we will discuss all of the methods together with their respective benefits and limitations.

2.1 Super Resolution

Here, we discuss the Super Resolution (SR) method. SR is a commonly used approach for solving the very low resolution face recognition problem and is considered a baseline method.

2.1.1 Basic method

The basic idea behind SR is to combine non-redundant information of multiple LR images into a single HR image. Closely related to SR is the image interpolation approach in which the resolution of an image is increased using a single image. However, due to the lack of extra information, the quality of the HR image is relatively low. Therefore, when performing SR we use multiple LR images. Super resolution methods can be categorized into two separate approaches. The first approach is based on statistics. Both the movement and the blurring can be regarded as stochastic variables. Therefore the SR

- *Marc Babtist is a Master student of Computing Science at the Rijksuniversiteit Groningen, E-mail: marcbabtist@gmail.com.*
- *Sebastian Wehkamp is a Master student of Computing Science at the Rijksuniversiteit Groningen, E-mail: sebastianwehkamp@gmail.com.*

reconstruction can be cast to a full Bayesian framework in which you try to maximize conditional probability. [15][13]

The second approach is example-based in which you try to make a linear combination of the LR images. First the weights are determined using a set of HR training images by trying to minimize the error. Using these weights, LR images are combined into a single HR image. [13]

2.1.2 Benefits

Although SR is commonly used in LR face recognition, the method has very few benefits without modifications. It is one of the earliest methods which tries to solve the LR face recognition problem, which is why we consider it the baseline method in this comparative study.

2.1.3 Weaknesses and limitations

The first problem of the SR method is image registration. Image registration is the process of transforming separate images into the same coordinate system. This is a fundamental image processing problem which is very hard to solve. The problem is made even more difficult because of the images having a low resolution. Traditional SR techniques treat image registration as a distinct process from the reconstruction. Therefore the quality of the reconstructed image depends on the quality of the previous step. [15]

Another problem is that the reconstructed face might not look like the original face or have serious artifacts (errors due to compression). Both the statistical and example based approach employ two constraints, the first of which is the data constraint φ_D. φ_D ensures that the reconstructed image is similar to the LR image. The statistical approaches use it to model the probability and the example based approaches use it implicitly to determine the weights. However, because of the limited information being carried by the LR images this constraint does not suffice. This is why the second, algorithm-specific constraint φ_S exists. φ_S ensures that the reconstructed image is a face image. However φ_S is designed for generic faces rather than for a specific individual. This might result in artifacts or unrecognizable faces. [13]

2.2 Multi-HOG

The method detailed in this section is Multi-Histogram of Oriented Gradients (Multi-HOG). [8]

2.2.1 Basic method

In order to explain this method, we first need to explain the histogram of oriented gradients (HOG). HOG is a feature extraction technique that uses gradient detectors in order to calculate oriented gradients, which are then saved in a histogram. The mathematical description of the HOG method is as follows: G_x and G_y represent the horizontal and vertical components of the gradients. These are computed using the intensity of the pixel I at (x,y), as per the following equations:

$$G_x = I(x+1,y) - I(x-1,y) \tag{1}$$

$$G_y = I(x,y+1) - I(x,y-1) \tag{2}$$

These components are then used to calculate the magnitude M and angle θ of the gradient at location x,y:

$$M(x,y) = \sqrt{G_x^2 + G_y^2} \tag{3}$$

$$\theta_{x,y} = \tan^{-1}\frac{G_y}{G_x} \tag{4}$$

Now that we have defined the oriented gradients, we need to create a histogram. To do so, we define the value bins of the histogram as follows:

$$V(b_\theta) = \sum_{y=1}^{y_{max}} \sum_{x=1}^{x_{max}} M_{b_\theta}(x,y) \tag{5}$$

In this equation, M_{b_θ} is defined as follows:

$$M_{b_\theta}(x,y) = \begin{cases} M(x,y) & \text{if } b_\theta = \lceil \frac{\theta_{x,y}B}{2\pi} \rceil \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where b_θ is the bin for the angle θ and B is the size of the bins. This translates to $V(b_\theta)$ containing the sum of all magnitudes that fit in bin b_θ. With the normal HOG method, the image is divided into sub-images, for each of which a HOG will be constructed. After construction, these are concatenated and normalized, resulting in the feature vector for the image.

Multi-HOG expands upon this by creating multiple sets of sub-images, where each set has different grid dimensions and bin sizes. Then, each of these sets of HOGs are compared to the corresponding HOG for the other image by calculating the Euclidean distance between the two. These distances are then concatenated into a distance vector. A graphical explanation can be seen in figure 1.

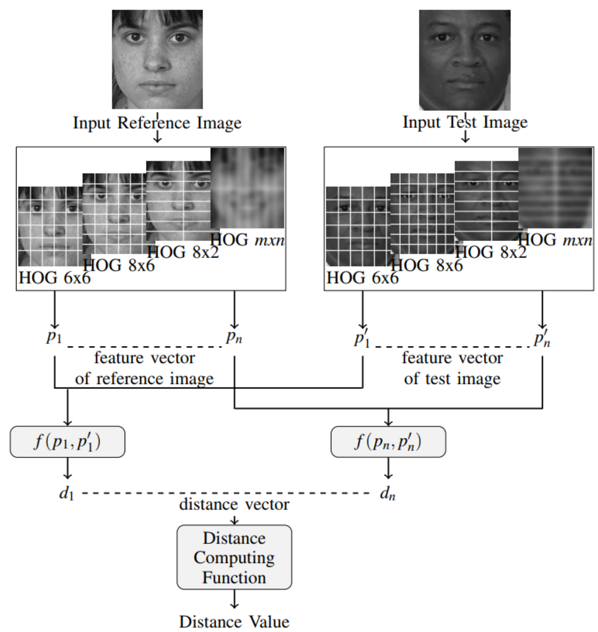

Fig. 1: Multi-HOG method of computing the distance between two faces. $f()$ is the Euclidean distance function. [8]

The resulting distance vector is then given to a distance computation function, which calculates a single distance value between the two images. After this, a 1-nearest neighbor classifier is used to conclude which face is the most similar based on the distance value. The original article experimented with two different distance computation functions: mean minimum distance (MMD) and multi-layer perceptron distance (MLPD).

During the paper's experiments, the authors also performed several general operation that increased performance. These included adding mirrored versions of the reference images to the dataset, correcting illumination differences by adjusting brightness and contrast to a fixed mean and standard deviation, and selecting the most similar regions of faces for one person.

2.2.2 Benefits

Since the HOG descriptor works on local cells, geometric and photometric transformations do not affect it, except for object orientation. This makes it fairly robust against changes in image quality. [4]

2.2.3 Weaknesses and limitations

One of the drawbacks of this method is that it is as of yet untested on image resolutions lower than 72 × 80. This means that performance on very low resolutions is unknown for this method. Besides that, calculating multiple HOGs per image makes this solution computationally intensive. [9] This is mitigated somewhat by calculating the HOGs from the reference images in advance and only once.

2.3 Sparse Coding of Local Features

In this section the approach of using SCLF for solving the LR face recognition problem will be discussed. [12]

2.3.1 Basic method

The SCLF method consists of four steps: preprocessing, extraction of local features, sparse coding, and classification. The first step is to apply some preprocessing operations to the face images. The top-hat filter is used to enhance bright objects of interest in a dark background while removing poor-contrast features. It is an excellent tool for correcting the effect of non-uniform illumination. Bottom-hat filtering is the opposite of top-hat filtering and enhances dark features of interest. The authors propose to add the difference between the top-hat and bottom-hat filtered outputs to the image. Mathematically this can be written as follows:

$$I_{CE} = I + I_{TH} - I_{BH} \tag{7}$$

where I_{TH} and I_{BH} are the top and bottom-hat filtered images respectively. The resulting image I_{CE} is a contrast enhanced version of the original image I. An example of the contrast enhanced image can be seen in Figure 2.

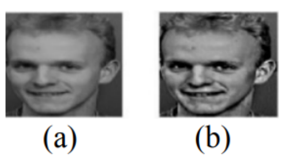

Fig. 2: (a) Original image (b) Contrast enhanced version of the original image [12]

The second step is to extract local features from the face images. Due to the face images being captured in an unconstrained environment, mistakes caused by pose variations, varying illumination, and different face expressions can occur. A robust solution to this problem is to look at descriptors of image patches: local features. The image is split in sub-images using spatial-frequency analysis which provides useful information about the image's structure at various scales and orientations. The key idea is to use Gabor wavelets as a multiscale operator using both the spatial and the frequency domain.

The third step is to create a sparse representation from each face image. Sparse representation has proven to be an efficient technique for face recognition. For every face a sparse representation is made with a linear combination of other training samples. The paper proposes a new way to create such a sparse representation by finding an optimal projection matrix, which maps local features to a sparse low-dimensional feature space.

After projecting, the last step is to perform classification using linear regression. This is done by calculating the residual errors of representing a testing feature in terms of training features. The image is assigned to the class with the lowest reconstruction error. A flow chart of the proposed approach can be found in Figure 3. [12]

2.3.2 Benefits

In general, the advantage of classification using sparse representation is that it gives very good results when images are heavily corrupted by noise or occlusions. This allows for good results in practical face recognition problems. [7] Another advantage of the proposed method is that it is more concise. Only the sparse representations of the images are needed for comparison instead of the complete image, meaning this method requires relativlely little storage space. [14]

2.3.3 Weaknesses and limitations

The training images must be carefully controlled and sufficient training samples of each class must be present or else the performance is affected badly. [7]

2.4 Tensor-based Multidimensional Scaling

In this section the approach of using MDS for solving the LR face recognition problem will be discussed. [3]

2.4.1 Basic method

The article [3] proposes a multidimensional scaling based approach which transforms features from both the LR and HR images to a common space such that distances can be calculated. The features are retrieved from the images using the Scale Invariant Feature Transform (SIFT) which is robust to changes of scale and rotations. These features should be transformed to a common space using a transformation. In this common space the distance D_t between the transformed HR image and the transformed LR image should equal the distance between the images had they been the same resolution and pose. To determine this transformation, training data consisting of frontal HR images is used. Tensor (vector) analysis is performed on the LR probe image to estimate the locations of facial landmarks. A flow chart of the proposed approach can be found in Figure 4. [3]

2.4.2 Benefits

Due to the automatic feature localization using tensor analysis, the algorithm performs very well in tracking and recognition in surveillance videos. This can be done simultaneously with a couple of modifications. Because of the tensor analysis which localizes features, the algorithm should also be able to handle changes in pose well. [3] [7]

2.4.3 Weaknesses and limitations

SIFT is based on the HOG of each pixel which makes it computationally intensive. Although the paper proposes to use PCA in combination with SIFT to alleviate this problem. [9]

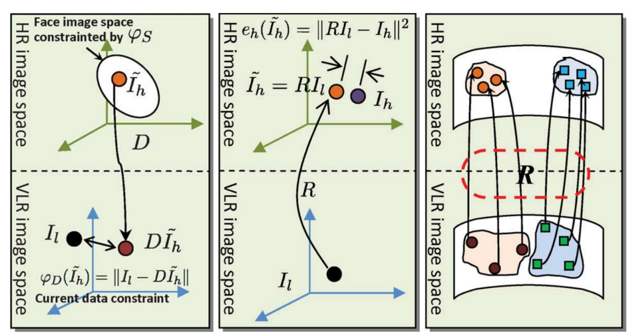

Fig. 5: Comparison of error calculation methods for human-based recognition. Left is the error calculation of existing approaches. The middle is the error calculation of the proposed approach. The right is the machine-based recognition approach. [13]

2.5 Relationship-learning-based Super Resolution

In this section the approach of using RLSR for solving the LR face recognition problem will be discussed. [13]

2.5.1 Basic method

The RLSR approach tries to find the relationship between the HR and LR image by learning it in the training phase. This relationship is used to reconstruct the HR images from the LR images in the testing phase. The first step in the training phase is to create linear clusters from LR and HR images. This means that the relationship in each cluster can be represented by a matrix. In the next step a linear regression model is used to learn this relationship with two different constraints. The first constraint aims at human-based recognition. This method first maps the LR image to an HR image and then calculates the error. This

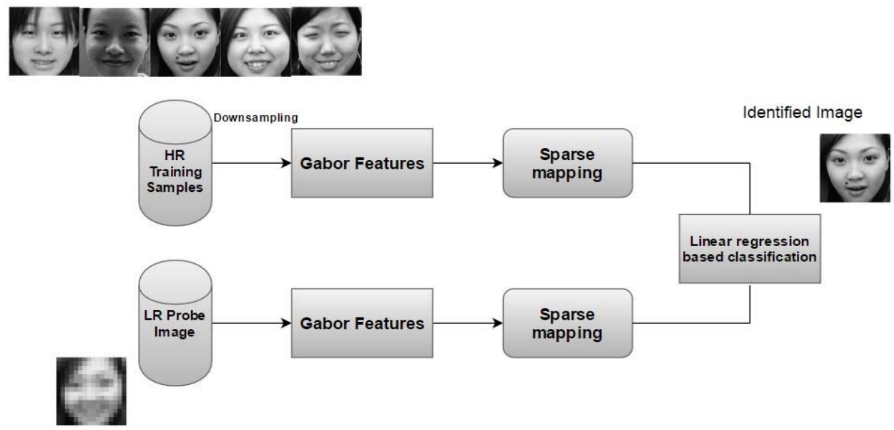

Fig. 3: Flow chart of the proposed approach. [12]

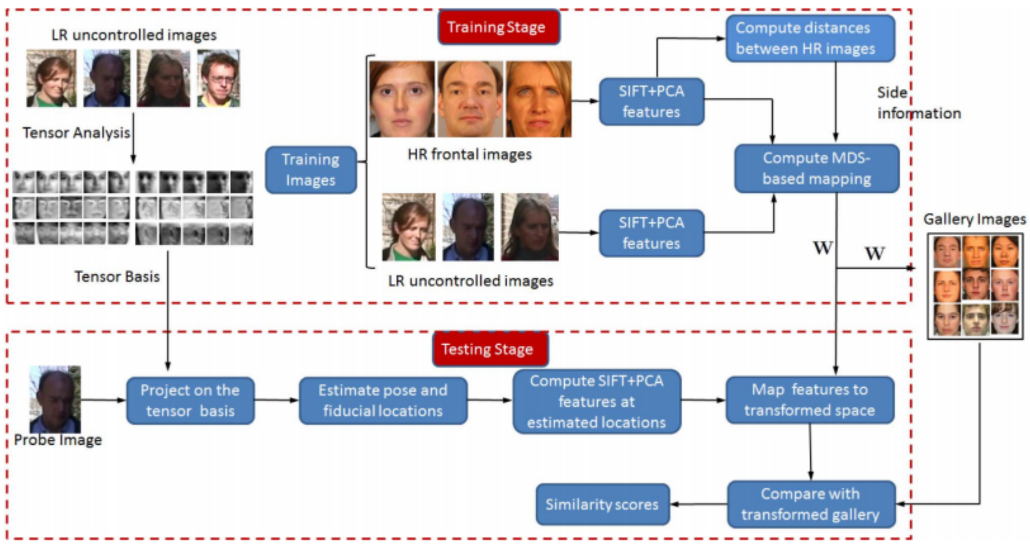

Fig. 4: Flow chart of the proposed approach. [3]

is different from existing methods, which calculate the error in the LR image space. Examples of both the error calculation in existing methods and the proposed method can be seen in Figure 5. [13]

The second method is aimed at machine-based recognition. This method was designed because visual quality is not the only criterion for SR. When looking from a machine-based perspective discriminability should also be considered. The goal is to create HR images with discriminative features in contrast to creating an HR image with a small error. This is done by using their Discriminative SR (DSR) algorithm which optimizes the discriminability of different images using class label information. In Figure 5 one can see that the small differences in LR space are enhanced in HR space. [13]

2.5.2 Benefits

One of the benefits of this approach is that the authors thought of both machine-based and human-based recognition. Since the results of human-based recognition ought to be better than the results of standard SR, this method can be used instead of SR in other use cases than LR face recognition. The machine-based approach should emphasize the discriminative features compared to SR so that it performs better than standard SR. [13]

2.5.3 Weaknesses and limitations

The problem of image registration in the standard SR method still holds for this method. This means that the quality of the SR algorithm depends on the quality of the image registration.

3 Datasets and Experimental Setup

This section describes the various datasets used and which experiments were performed.

3.1 Datasets

In this section, we will perform a short comparison between the various datasets used in the articles. To see which dataset was used in which article, refer to Figure 6.

Dataset	Multi-HOG	SCLF	MDS	RSLR
ORL		x	x	
Yale B		x		x
CAS-PEAL		x		x
CMU-PIE			x	x
FERET	x			

Fig. 6: Table showing what datasets were used in each article.

3.1.1 ORL

The Olivetti Research Laboratory [2] (ORL) database from AT&T in collaboration with Cambridge University is a set of four hundred images, containing ten different images of each of forty subjects. Some of the images were taken at different times, varying the lighting, facial expressions and accessories such as glasses. There is no variation in background or position.

3.1.2 Yale B

The Extended Yale B database [6] contains images of 28 subjects under 9 angles and 64 illumination conditions, for a total of 16128 images. Facial expressions do not differ between images.

3.1.3 CAS-PEAL

The Chinese Academy of Sciences - Pose, Expression, Accessory and Lighting database and its subset CAS-PEAL-R1 [5] contain many images of a large amount of individuals (99,594 of 1040 and 30,900 of 1040 respectively). The set boasts a large amount of different poses, expressions, accessories and lighting situations, including combinations of the four. One thing to note is that this dataset contains mostly people of Chinese ancestry, while the other databases tend to contain more Caucasian faces.

3.1.4 CMU-PIE

The Carnegie Mellon University Multi-Pose, Illumination, Expression Face Database (CMU-PIE) [1] is an improvement over the earlier PIE database. It contains 337 subjects, each captured under 15 view points and 19 illumination conditions in four recording sessions, for a total of more than 750,000 images. Of these images, 6152 are provided with labels for feature points of the faces.

3.1.5 FERET

The Facial Recognition Technology dataset [11] [10] was developed by DARPA specifically for the evaluation of facial recognition algorithms. It contains 1564 sets of images for a total of 14,126, with varying illumination, facial expressions, angles and accessories.

3.2 Performance Comparisons

The goal of our paper is to select the best performing face recognition method in different situations. The problem is that the four papers selected different datasets to test their algorithms. To be able to select the best algorithm we will perform three comparisons. The first one is comparing Sparse Coding of Local features with Tensor-based Multidimensional Scaling using the Yale-B and the CAS-PEAL datasets. The second is between MDS and RLSR using the Multi-PIE dataset. The last comparison is between Multi-HOG, SCLF, and MDS, using ORL and FERET. The datasets of both FERET and ORL can be considered similar enough to directly compare on, since they both contain images with varying illumination, facial expressions, angles, and accessories.

4 Results

In this chapter, we will compare the performance of the different methods, keeping in mind the datasets the performance was measured on. All of the data was obtained from the corresponding papers. In all comparisons SR is included as a baseline method.

4.1 Comparison 1

First we compared SCLF and MDS on Yale-B and CAS-PEAL. The results can be found in Figure 7 and Figure 8. The SCLF results were obtained by using LR images of 20x18 and gallery images of 60x55. The MDS results were obtained by using 20x20 images for CAS-PEAL and 16x16 for Yale-B. We used the results obtained by using four training samples. The testing conditions are very comparable, which means that the results are, too. In both comparisons SCLF and MDS perform very similar but SCLF has a slight edge over MDS.

4.2 Comparison 2

This experiment compares RLSR and MDS using the CMU-PIE dataset. The RLSR results were obtained using a very low resolution of 7x6 and the gallery had a resolution of 56x48. These results were taken from rank 6. This rank indicates that the result came from the sixth worst training/test set, in this case out of ten. This was chosen because we feel it best represents the capabilities of the method.

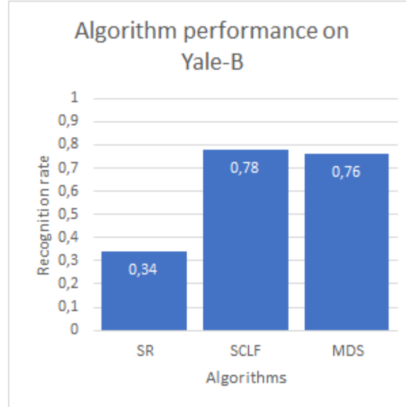

Fig. 7: Performance of various algorithms on the Yale-B dataset.

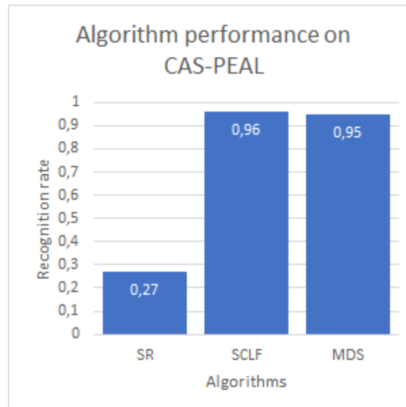

Fig. 8: Performance of various algorithms on the CAS dataset.

The MDS results are obtained using LR images of 13x11 while the gallery images have a resolution of 60x55. The results can be found in Figure 9. RLSR clearly performs significantly better than MDS even though the LR images of RLSR have a lower resolution than the LR images of MDS.

4.3 Comparison 3

The last comparison is done on both the ORL and FERET datasets, which are very similar to each other. The MDS and SCLF results come from the ORL dataset while the Multi-HOG results come from FERET. Multi-HOG uses gallery images of 80x88 but it is unclear what the resolution of the LR images is. We used the results of the best performing variant of Multi-HOG, which is Multi-HOG MSRS-MLPD. Both the MDS and SCLF results are obtained using gallery images of 60x55 and LR images of 16x16 using 4 training samples. In Figure 10 one can see that, as in comparison 1, MDS and SCLF perform very similarly with SCLF coming slightly ahead. They both perform better than Multi-HOG. The addition of mirrored images does give a minor increase to the performance of Multi-HOG.

5 Conclusion

All of the tested methods perform significantly better than the baseline SR method. It is clear that MDS and SCLF perform very similarly, with SCLF coming slightly ahead of MDS. We only have results of one dataset for Multi-HOG, but it comes in last in that case by a margin, though it should be noted that the exact testing conditions such as the resolution of the LR images remain unclear. From the Multi-HOG results we can conclude that adding mirrored faces does increase the performance. From comparison 2 it becomes clear that RLSR performs much better than MDS. Judging from the gap between RLSR

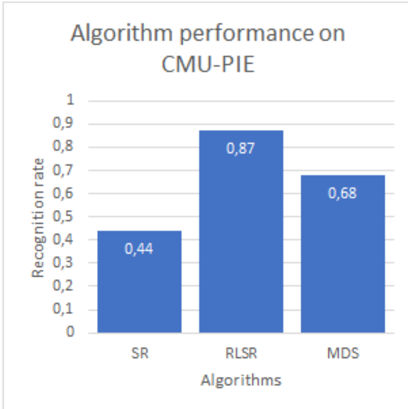

Fig. 9: Performance of various algorithms on the CMU-PIE dataset.

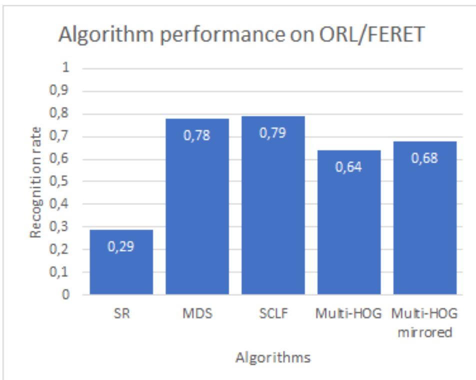

Fig. 10: Performance of various algorithms on the ORL and FERET dataset.

and MDS in comparison 3, and MDS and SCLF performing very similarly, could indicate that RLSR also performs better than SCLF.

6 Discussion

Here, we will suggest general improvements to the various algorithms as well as propose further research that can be performed on the subject of low resolution face recognition.

6.1 General Improvements

Some of the articles used in this paper contain improvements that can work on all algorithms. These include the following:

- adding mirrored faces [8],
- correcting the illumination by adjusting brightness and contrast to fit a mean and standard deviation [8],
- top- and bottom-hat filtering [12].

By performing these preprocessing operations, the set of reference images can be enhanced and expanded, providing increased accuracy.

6.2 Further research

The biggest flaw in the manner of comparison between the various face recognition algorithms is that we were unable to test the various methods on the same datasets due to time constraints. This means that the raw accuracy numbers obtained might not be strictly comparable. Therefore, we suggest a follow-up article that compares the various prescribed methods by running them in the same manner on the same datasets. This should provide numbers that can be compared as-is, without further alteration. However, one should take care to pick a greater variety of datasets, as certain algorithms might outperform others in specific situations.

Another subject of research could be combining the various algorithms. These could either be naively combined by voting, or by more complicated methods such as using the automatic feature localization from [3] to select the maximum similarity region for Multi-HOG. We feel that by combining the various strengths of the different methods covered in this paper a new algorithm can be created that is better than the sum of its parts.

References

[1] The cmu multi-pie face database. http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html. Accessed: 2018-04-11.

[2] AT&T Laboratories Cambridge. The database of faces. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, 2002. Accessed: 2018-04-11.

[3] S. Biswas, G. Aggarwal, and P. J. Flynn. Pose-robust recognition of low-resolution face images. *Cvpr 2011*, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *International Conference on Computer Vision & Pattern Recognition*, 1:886–893, June 2005.

[5] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Trans. on System Man, and Cybernetics (Part A)*, 38(1):149–161, Jan. 2008.

[6] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[7] S. Habeeba and P. Simon. Review on sparse based face recognition methods in uncontrolled environment. *International Journal on Computer Science and Engineering (IJCSE)*, 9(8), 2017.

[8] M. Karaaba, O. Surinta, L. Schomaker, and M. A. Wiering. Robust face recognition by computing distances from multiple histograms of oriented gradients. *2015 IEEE Symposium Series on Computational Intelligence*, 2015.

[9] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004.

[10] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The feret evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.

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

[12] M. S. Shakeel and Kin-Man-Lam. Recognition of low-resolution face images using sparse coding of local features. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.

[13] W. W. Z. Wilman and P. C. Yuen. Very low resolution face recognition problem. *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2010.

[14] J. Wright, A. Y. Yang, Arvind, S. S. Sastry, and Y. Ma. Face recognition via sparse representation, Mar 2008.

[15] J. Yang and T. K. Huang. Image super-resolution: Historical overview and future challenges. 2010.