Sparse Representations For Facial Expressions Recognition via l_1 optimization

Stefanos Zafeiriou and Maria Petrou
Department of Electrical and Electronic Engineering
Imperial College London, UK
{s.zafeiriou,maria.petrou}@imperial.ac.uk

Abstract

In this paper, the principles of sparse signal representation theory are explored in order to perform facial expressions recognition from frontal views. Motivated by the success such methods have demonstrated in the face recognition problem, we formulate the feature extraction procedure in order to achieve facial expression recognition as an l_1 optimization problem. We show that the straightforward application of these methods to expressive images imposes certain difficulties. The use of difference images (i.e., the images that are derived from the subtraction of the neutral image from the expressive one) for sparse facial expression representations is justified. The use of expressive facial grids for similar tasks is also studied. Finally, the robustness of the proposed representations under facial image occlusion is shown and the efficacy of the proposed method in a series of experiments is demonstrated.

1. Introduction

The procedure followed by the different levels of the human visual system in order to process and formulate images constitutes a very active research topic in neuropsychology, neurophysiology, psychophysics, signal processing and computer vision. Many of the recent theoretical and experimental studies support the idea that the visual system performs object detection and recognition in a hierarchical and parsimonious way in which neurons become selective (i.e., they are selective for various stimuli such as color, texture, orientation etc) to process progressively more complex features of the image structure. Such models of the human visual system have initiated studies for image-based object recognition using sparse image representations.

In computer vision, sparse coding corresponds to object representation by using bases with components that are spatially distributed without any connectivity. The representation acquired using sparse bases was the first step towards the implementation of a part-based representation [16, 14, 9]. As shown in [16], the linear sparse coding of natural images yielded features qualitatively very similar to the receptive fields of simple-cells in the primary visual cortex. Subsequently, the very closely related model of Independent Component Analysis (ICA) [11] was introduced to provide similar results [16]. However, the above mentioned models [16] allow the existence of negative entries to the acquired representation. This is in contrast with the fact that the firing rates of the simple-cells in the primary visual cortex are nonnegative [14, 9]. The nonnegativity of the firing rates and the fact that the representation of an object by its parts is more naturally coded using only additions between the different bases [17, 19, 18, 1, 15], lead to the introduction of the Nonnegative Matrix Factorization (NMF) algorithm, proposed in [14].

Recently, another sparse representation for object representation and recognition was proposed in the seminal work [20] based on principles of compressed sensing [7]. In [20] an attempt to find an object representation using a sparse linear combination of an overcomplete dictionary was made. More specifically, a facial image is represented as a sparse linear combination of the training facial images. It is shown that, if a sufficient number of training samples are available for each facial class, it is possible to represent the test facial image samples as a linear combination of only those training samples that belong to the same facial class. The resulting optimization problem penalizes the l_1- norm of the coefficients of the linear combination. It was proven that the representation is indeed sparse, involving only a small fraction of the overall training database. It is also demonstrated that the sparsest representation constitutes a mean for performing a discriminant analysis between the facial classes. This means, intuitively, that the test image is most likely to belong to the facial class with the most nonzero coefficients.

The sparse representation approaches that were produced through NMF [14, 3] and the l_1 norm optimization in [20] have certain differences in modelling. That is, in NMF-based approaches both bases and weights are calculated. An example of NMF basis images can be found in Figure 1a. Then, an image is represented as a nonnegative

978-1-4244-7030-3/10/$26.00

linear combination of the calculated bases. An example of an image decomposition using the NMF bases can be found in Figure 1b. In many cases, the NMF bases contain components and features that are spatially distributed without any connectivity (we will refer to them as sparse bases from now onwards). On the other hand, sparse approaches in [20] and [10], which are based on l_1 optimization, attempt only to find the weights of the linear combination assuming that the bases are the training images. We shall refer to such bases as dictionaries from now and on. Thus, they estimate a sparse weighting vector with larger values for those training images which bear a resemblance to the test image.

A lot of research has been conducted regarding facial expressions recognition in the past fifteen years. The facial expressions under examination were defined by psychologists as a set of six basic facial expressions (anger, disgust, fear, happiness, sadness, and surprise) [8]. An example of expressive images is depicted in Figure 2. An interested reader may refer to [13, 21] and in the references therein, regarding the various technologies developed for facial expression recognition. The application of sparse representations produced by NMF has been extensively studied for the analysis of expressive images [3, 4] and subsequently for facial expression recognition.

In this paper we study the facial expressions recognition problem using principles from the sparse representation theory [20, 7, 6]. We demonstrate that the straightforward application of the method proposed in [20] does not produce a meaningful sparse representation for facial expressions and prove that this approach is tuned for the recognition of facial identity. Afterwards, we study the use of the differences images (calculated by subtracting the neutral image intensity values from the corresponding values of the fully expressive facial expression image) for the creation of sparse representations for the decomposition of expressive images. The difference images tend to emphasize the facial parts that are moved and eliminate in that way the identity of the facial image. Some examples of difference images are shown in Figure 3.

Moreover, we show that facial grids can be also used for sparse facial expression representation. Examples of facial expressive grids are depicted in Figure 2. Facial grids, which are graphs consisting of nodes placed at prespecified fiducial facial points (like lips, eyebrows etc), describe facial expressions in a person-independent way. Thus, they eliminate the problems imposed by facial images for sparse representation of facial expressions.

The rest of the paper is organized as follows. In Section 2 we describe the method proposed in [20] to acquire a sparse representation for face recognition. Afterwards we present the difficulties aroused by the application of such a method in the facial expression recognition problem. Furthermore, we propose ways to overcome these difficulties. In Section 3 we present the experimental results which support our theoretical arguments. Finally, conclusions are drawn in Section 4.

2. Motivating Sparse Representations for Facial Expressions Representation

2.1. Face Recognition via l_1 sparse representation

Let a set of N training facial images be separated to K different facial identity classes. Each of the facial images was scanned row-wise in order to form an F-dimensional vector. Afterwards, dimensionality reduction methods are applied, like Principal Component Analysis (PCA) [12], in order to form from each each image a vector $\mathbf{x}_i \in \Re^f$, with $f \ll F$, which is normalized in such a way that $||\mathbf{x}_i|| = 1$ (from now and on we shall refer to images as vectors with magnitude one). Let the dictionary (matrix) $\mathbf{X}$ be defined as $\mathbf{X} = [\mathbf{x}_1|\dots|\mathbf{x}_N] \in \Re^{f\times N}$. Let also a test image $\mathbf{y} \in \Re^f$. In [20], a method for feature extraction in order to achieve face recognition was proposed via a sparse decomposition. That is, they motivated the use of an l_1 optimization problem in order to find a sparse vector of weights $\mathbf{w}$ which depicts the contribution of each facial training image $\mathbf{x}_i$ in the formation of the test facial image $\mathbf{y}$. Let be a threshold ϵ. Then, the optimization problem for finding the sparse vector $\mathbf{w}$ is the following:

$$\tilde{\mathbf{w}} = \arg\min ||\mathbf{w}||_1 \quad \text{subject to} \quad ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 < \epsilon. \quad (1)$$

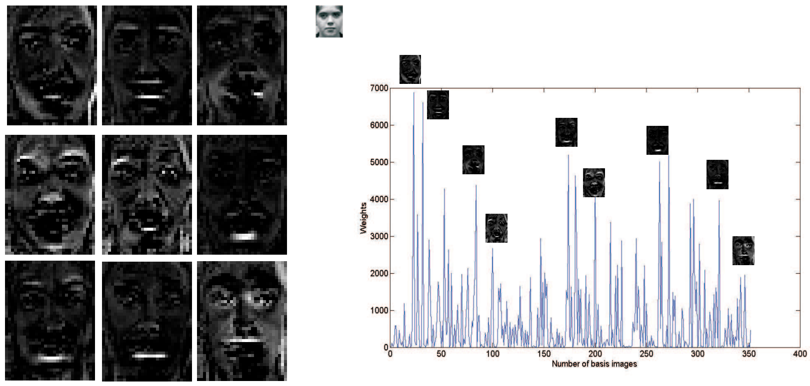
After the calculation of the optimal vectors $\tilde{\mathbf{w}}$ according to (1), we attempt to classify image $\mathbf{y}$ to one of the K facial identity classes. Let $\delta_k(\tilde{\mathbf{w}})$ be a new vector whose only nonzero entries are the entries in $\tilde{\mathbf{w}}$ that are associated with class k. Using only the coefficients associated with the k-th facial identity class, one can approximate the given test sample $\mathbf{y}$ as $\tilde{\mathbf{y}}_k = \mathbf{X}\delta_k(\tilde{\mathbf{w}})$. Image $\mathbf{y}$ is classified based on these approximations to the object class that minimizes the residual between $\mathbf{y}$ and $\tilde{\mathbf{y}}_k$:

$$l(\mathbf{y}) = \arg\min_k r_k(\mathbf{y}) = ||\mathbf{y} - \tilde{\mathbf{y}}_k||_2. \quad (2)$$

Moreover, in order to model pixel corruptions and deal with the presence of occlusion, an error vector $\mathbf{e}$ has been taken into consideration in the optimization problem. Assuming that the error vector $\mathbf{e}$ has sparse nonzero entries with respect to the natural pixel coordinates we create the dictionary $\mathbf{X}_e = [\mathbf{X}, \mathbf{I}] \in \Re^{f\times(N+f)}$ where $\mathbf{I}$ is the identity matrix. Then, we seek a vector $\mathbf{w}_e = \begin{bmatrix} \mathbf{w} \\ \mathbf{e} \end{bmatrix}$ from the optimization problem:

$$\tilde{\mathbf{w}}_e = \arg\min ||\mathbf{w}_e||_1 \quad \text{subject to} \quad ||\mathbf{X}_e\mathbf{w}_e - \mathbf{y}|| < \epsilon. \quad (3)$$

For solving the optimization problem (1) and (3) we used the l_1-magic software package in [5].



(a) (b)

Figure 1. a) NMF basis images; b) a decomposition of an image using NMF basis images and nonnegativity weights.

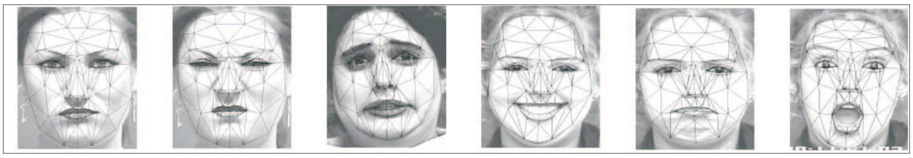


Figure 2. Expressive images and their corresponding facial grids. From left to right the depicted expressions are Anger, Disgust, Fear, Happiness, Sadness and Surprise.

Figure 3. The differences images that correspond to the expressive images.

Once the sparse solution $\tilde{\mathbf{w}}_e = \begin{bmatrix} \tilde{\mathbf{w}} \\ \tilde{\mathbf{e}} \end{bmatrix}$ is computed from (3), then the clean from corruption or occlusion facial image $\mathbf{y}_r = \mathbf{y} - \tilde{\mathbf{e}}$ is used for classification using the rule:

$$l(\mathbf{y}) = \arg\min_k r_k(\mathbf{y}) = ||\mathbf{y}_r - \mathbf{X}\delta_k(\tilde{\mathbf{w}})||_2. \quad (4)$$

Intuitively, if the database comprises K facial identity classes then vector $\tilde{\mathbf{w}}$ should contain high valued coefficients that correspond to the facial identity class to which image $\mathbf{y}$ belongs and very low (or probably zero) values for all other images. An example is shown in Figures 4a and b, where two different facial images are decomposed having in the dictionary images of the same facial identity class. As it can be seen, a sparse representation is acquired where high peaks are calculated for images inside the same facial identity class and very low responses for most of the other images.

2.2. Facial Expressions Recognition via l_1 sparse representation

Let the database be separated in 6 different facial classes one for each facial expression to be recognized (anger, disgust, fear, happiness, sadness and surprise). The data used were created from the Cohn-Kanade (CK) database. All the available subjects were taken into consideration to form the database for the experiments (a total of 352 videos). For the expressive images we considered all the last frames of the CK video sequences. The first frames of the sequences were considered to be the neutral state images. All the images from the database were used to create a dictionary. A collection of expressive images from the CK database is shown in Figure 2.

Let us test algorithm (3) using a real facial expression database. Let us assume that a facial image $\mathbf{y}$ is available and has to be classified into one of the six facial expression classes. Using the above framework we have to find a sparse decomposition, by means of a sparse weighting vector, of image with respect to the dictionary $\mathbf{X}$ which contains all the available expressive images. Intuitively, this vector $\mathbf{w}$ should have high valued responses for the expressive images of the same facial expression class of the image $\mathbf{y}$ and low valued responses for all the other classes. As we discuss next this framework cannot be applied in a straightforward way for the decomposition and the classification of expressive images to facial expressions.

The image that is decomposed in Figure 4b is an expressive facial image (disgust). In this Figure we also placed delimiters in order to denote the coefficients that correspond to different facial expression classes (A holds for anger, D for disgust, F for fear, H for happiness, Sa for sadness, Su for surprise and N for neutral). As it can be seen in the decomposition, we achieved high responses for images of the same person that depict a different facial expression. Let us remove the images of the same person from the dictionary. In Figure 4c the decomposition of the same image from Figure 4b is depicted but now the dictionary does not contain images of the same person and of the neutral expression. As it can be seen, the decomposition is not as sparse as the one in Figure 4d and it does not produce high valued responses for the coefficients that correspond to the correct facial expression class (i.e., failed to achieve decomposition using images from the same expression class).

As it has been shown, the direct use of the expressive images results in a rather difficult decomposition in terms of facial expression classes using the algorithm defined in (1) or in (3). This is due to the fact that the features of the same facial identity influence the result to a greater extent than the features of the same facial expression class. It is clear that in order to use such algorithms we need to alleviate the contribution of the facial identity in the description of expressions. That is, we should seek as much as possible person-independent descriptions of expressions. Such a representation can be found using the difference images. Examples of difference images calculated for every facial expression are shown in Figure 3. An example of the decomposition using the difference images is shown in Figure 4d. In this Figure the decomposition of the expressive difference image (disgust) of the image depicted in Figure 4c having in the dictionary difference images of the same facial identity class, is shown. As it can be seen: 1) the decomposition did not result in high valued responses for difference images of the same person 2) the decomposition, using the difference images produced a meaningful sparse representation where high valued responses were calculated for the difference images of the same facial expression class (disgust).

We propose that we can increase the recognition performance of the difference images by breaking the image into smaller blocks and then fusing the overall result. In a first experiment conducted, we partitioned each of the training images into 3 blocks of size $a_i \times b_i$, thus producing a set of matrices $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}$. In Figure 6 the three different blocks acquired are depicted. The test expressive image is accordingly partitioned into three blocks $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)}$. The l-th block of the test image is written as $\mathbf{y}^{(l)} = [\mathbf{X}^{(l)}, \mathbf{I}]\mathbf{w}^{(l)}$. The sparse vector $\mathbf{w}^{(l)}$ is recovered by solving the optimization problem (3). We apply the classifier (4) within each block and then aggregate the results by voting.

Now we shall consider the facial grids in order to acquire meaningful sparse representations for facial expression recognition. The facial grid used in this paper was the well-known Candide wireframe model [13]. At the first frame of the video sequence (depicting the neutral state), certain points of the Candide grid are matched against the facial features of the actual face image. Grid node tracking is performed by a pyramidal variant of the well-known KanadeLucasTomasi (KLT) tracker available in [2]. The Candide grid comprises 114 nodes (104 of them are used in reality to form the grid). In order to use the Candide grid for the description of expressions we created a vector $\mathbf{x}_g \in \Re^{104\times 2=208}$ which contains the concatenated x and y displacement coordinates of the grid [13]. Let us create a dictionary which contains the expressive grids, called $\mathbf{X}_g$. We shall produce expressive grid vectors from the dictionary and apply the algorithm provided in (1). Two such examples, anger and surprise, are shown in Figures 5 a and b, respectively. By inspecting these Figures one can see that the use of grids ensures the creation of meaningful sparse representations.

3. Experimental Results

For the experiments, we applied a leave one out principle where one expressive sample (image or grid) is left out to be used as a test sample and the remaining are used to build the dictionary. Then, the test sample is classified into one of the 6 facial expression classes. We originally applied Principal Component Analysis (PCA) in order to reduce the dimensionality and form the overcomplete dictionary $\mathbf{X}_{\text{train}} \in \Re^{f\times N}$ which in our experiments is formed by the N available in the database vectors, which have at most $f = 200$ dimensions (and $f = 208$ for grids). We observed that by keeping more dimensions and as $f \to N$ then, the system of approximating the test sample $\mathbf{y}_{\text{test}} = \mathbf{X}_{\text{train}}\mathbf{w}$ is not underdetermined any more and vector $\mathbf{w}$ is not sparse. Moreover, in order to demonstrate the robustness of the proposed method to occlusion and corruption, we conducted additional experiments using images under eyes' occlusion. Some of the occluded images can be seen in Figure 8 where artificial glasses were manually superimposed on the images. For the partially occluded grids we substituted the original coordinates, that correspond to the nodes in the occluded area, with the value $(0, 0)$. Moreover, in order to demonstrate the robustness of the algorithm in case of data

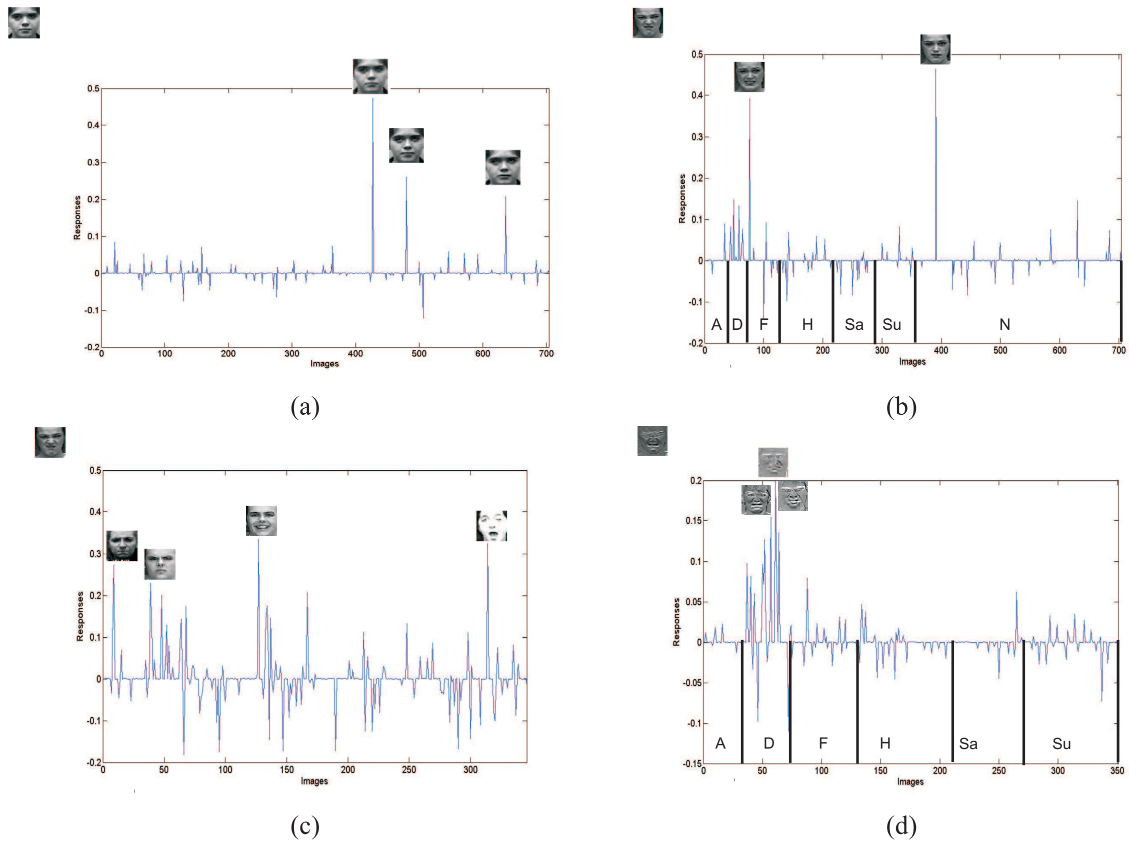


Figure 4. a) Sparse decomposition of a facial image including images of the same person in the dictionary; b) Sparse decomposition of an expressive facial image including images of the same person in the dictionary; c) Sparse decomposition of the same image excluding images of the same facial class from the dictionary; d) Sparse decomposition of the differences images.

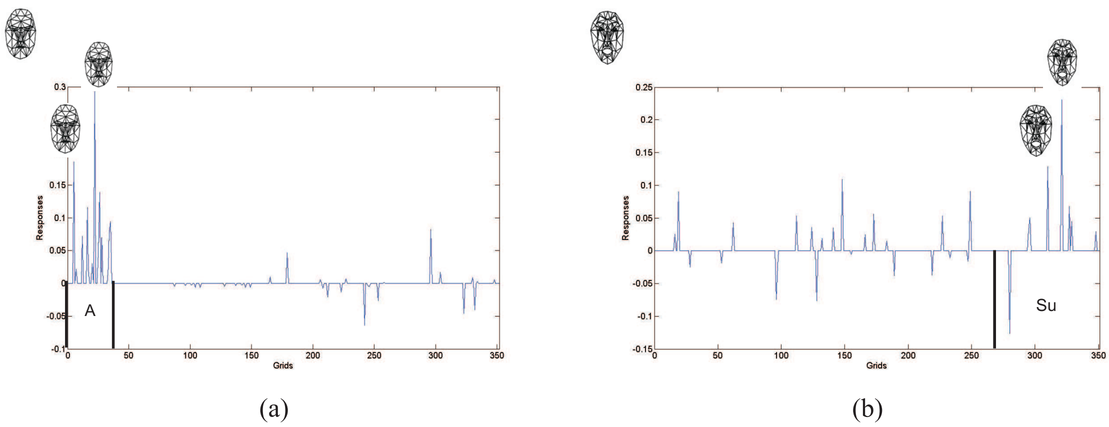


Figure 5. Sparse decomposition of facial grids: a) anger; b) surprise.

corruption, we randomly corrupted $10\% - 60\%$ of the nodes (by replacing the (x, y) values with $(0, 0)$).

The occlusion experiments were conducted as follows. The dictionary was built using the original not occluded images or not occluded grids. Then, by following the leave one out scenario we removed one sample out of the dictionary and used for testing its occluded version. The PCA dimensionality reduction that is used in this paper is calculated using only the training not occluded dictionary. This is a very realistic scenario since for training we usually have not occluded images and grids but for testing we can not actually guarantee that the image or the grid will not be occluded or

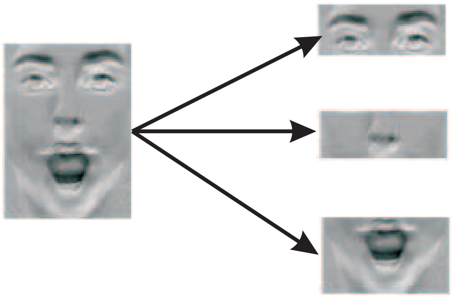


Figure 6. The 3 blocks of the differences image.

corrupted.

In Figure 7a the recognition error, i.e. the number of misclassified samples, is plotted versus the number of PCA dimensions kept using the original images, the difference images, the block-based difference images and the occluded difference images. Let us abbreviate the method that uses the original images, PCA for dimensionality reduction and optimization problem (3) for classification as PCA-Or-L_1. In a similar manner let us abbreviate the method using the difference images and PCA as PCA-Diff-L_1, the difference images and PCA and the block-wise method as PCA-BDiff-L_1 and the occluded difference images and PCA-block wise as PCA-BDiff-OC-L_1. As it can be seen, using the original images the recognition error (the percentage of misclassified samples) exceeds 90%. That is, only 10% of the images were classified to the correct expressions. This demonstrates the fact that the straightforward application of the method described in Section 2.1 for face recognition totally failed in the case of facial expression recognition.

As it can be seen in the same Figure, by using the difference images and PCA or PCA using the block based method, we achieved a recognition error of about 40% and 19% (60% and 81% recognition rate), respectively. Finally, by applying PCA and the block based method on the occluded difference images, we achieved a recognition error of about 21.3% which is a first indicator of the robustness of the proposed method to partial occlusion.

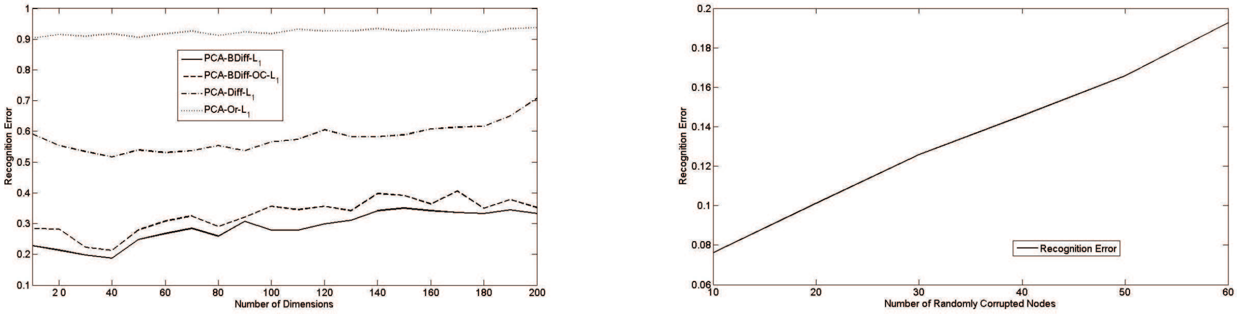
Moreover, we believe that by dividing the difference image into blocks and applying the sparse representation in (3) we could also perform Facial Action Units (FAUs) recognition. Then, the facial expression that corresponds to the difference image can be determined by the use of the recognized FAUs.

The second set of experiments was conducted using the facial grids. The application of the algorithm in (3) at the grids resulted to 7.6% recognition error (abbreviated in Table 1 as Grids-l_1). We also tested linear Support Vector Machines (SVMs) [13] for the classification of the expressive grids and we achieved a recognition error of 10% (abbreviated in Table 1 as Grids-SVMs). By using the partial occluded grids that correspond to eye-occlusion shown in Figure 6, which correspond to a corruption of 30% of the nodes, we achieved a recognition error of 10.7% (abbreviated in Table 1 as Grids-l_1-OC). The application of linear SVMs to the above corrupted grids achieved a recognition error of more than 24.5% (abbreviated in Table 1 as Grids-SVM-OC). The lowest Recognition Error (RE) are summarized in Table 1.

In order to test the method further we randomly changed the node coordinates until 60% of them were corrupted (placing $(0, 0)$ in place of the changed grid coordinates). In Figure 7 b the recognition error is plotted versus the percentage of the corrupted nodes. As it can be seen, by corrupting 60% of the nodes (i.e., keeping only 40% of the original nodes) we achieved a recognition error of about 22%. It is interesting to note here that the application of SVMs in the 60% corrupted grids achieved a error rate of more than 55%. The findings of our experiments demonstrate the potential of sparse representations based on l_1 optimization for facial expression recognition.

4. Conclusions

In this paper we proposed to exploit of sparse representations which are derived from l_1 optimization problems for facial expressions recognition. The use of such representations in computer vision applications is a very active and emerging topic since they exhibit great robustness to pattern occlusion and corruption. We showed that it is difficult to produce meaningful representations when using directly the original images. We argued that person-independent representations of facial expressions should be used in order to find efficient representations. Such representations include difference images and facial grids. Finally, we verified the findings of other works in sparse representations where it was shown that sparse representations could be handled uniformly and robustly within the same classification framework, something that is of great importance for the facial expressions recognition problem.



(a) (b)

Figure 7. a)Recognition error versus the number of dimensions for the original images, the difference images and occluded images using PCA and l_1 optimization problems; b) Recognition error versus the number of randomly corrupted nodes.

Table 1. Lowest (%) facial expressions Recognition Errors (RE)

	PCA-Or.-l_1	PCA-Diff-l_1	PCA-BDiffl_1	PCA-BDiff-OC-l_1	Grids-l_1	Grids-l_1-OC	Grids-SVMs	Grids-SVM-OC
RE	90%	51.2%	19.7%	21.3%	7.6%	10.7%	10%	24.5%

Figure 8. The eye-occluded expressive images.

5. Acknowledgment

This work was supported by the EPSRC project EP/E028659/1 Face Recognition using Photometric Stereo.

References

[1] I. Biederman. Recognition-by-components: a theory of human image understanding. *Phychol. Rev.*, 94:115147, 1987.

[2] J. Bouguet. Pyramidal Implementation of the Lucas-Kanade Feature Tracker: Description of the algorithm, OpenCV Documentation. Santa Clara, CA: Intel Corp., Microprocessor Research Labs, 1999.

[3] I. Buciu and I. Pitas. Application of non-negative and local non negative matrix factorization to facial expression recognition. In *ICPR*, pages 288–291, Cambridge, United Kingdom, 23-26 August 2004.

[4] I. Buciu and I. Pitas. A new sparse image representation algorithm applied to facial expression recognition. In *MLSP*, Sao Lus, Brazil, Sep. 29 - Oct. 1st 2004.

[5] E. Candes and J. Romberg. l1-magic: Recovery of sparse signals via convex programming. Available information in URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf, 2008.

[6] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2007.

[7] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[8] P. Ekman and W. V. Friesen. *Emotion in the Human Face*. Prentice Hall, New Jersey, 1975.

[9] P. O. Hoyer. Modeling receptive fields with non-negative sparse coding. *Neurocomputing*, 52-54:547–552, 2003.

[10] J. Huang, X. Huang, and D. Metaxas. Simultaneous image transformation and sparse representation recovery. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[11] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[12] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 12(1):103–108, Jan. 1990.

[13] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187, Jan. 2007.

[14] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[15] N. Logothetis and D. L. Sheinberg. Visual object recognition. *Annu. Rev. Neurosci.*, 19:577–621, 1996.

[16] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[17] S. E. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, (9):441–474, 1977.

[18] S. Ullman. *High-level vision: object recognition and visual cognition*. MIT Press, Cambridge, MA, 1996.

[19] E. Wachsmuth, M. Oram, and D. I. Perrett. Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cereb. Cortex*, 4(5):509522, 1994.

[20] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2008.

[21] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.