SPARSE RECONSTRUCTION OF FACIAL EXPRESSIONS WITH LOCALIZED GABOR MOMENTS

André Mourão, Pedro Borges, Nuno Correia, João Magalhães

Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,
Quinta da Torre, 2829-516 Caparica, Portugal
a.mourao@campus.fct.unl.pt, p.borges@campus.fct.unl.pt, nmc@fct.unl.pt, jm.magalhaes@fct.unl.pt

ABSTRACT

Facial expression recognition depends on the detection of a few subtle facial feature traces. EMFACS (Emotion Facial Action Coding System) is a taxonomy of face muscle movements and positions called Action Units (AU) [1]. AUs can be combined to describe complex facial expressions. We propose to (1) deconstruct facial expressions into face regions, grouping AUs by their proximity and contour direction; (2) recognize facial expressions by combining sparse reconstruction methods with face regions. We aim at finding a minimal set of AU to represent a given expression and apply l_1 reconstruction to compute the deviation from the average face as an additive model of facial micro-expressions (the AUs). We compared our proposal to existing methods on the CK+ [2] and JAFFE datasets [3]. Our experiments indicate that sparse reconstruction with l_1 penalty outperforms SVM and k-NN baselines. On the CK+ dataset, the best accuracy (89.8%) was obtained using sparse reconstruction.

Index Terms— facial expression recognition, sparse reconstruction, Gabor wavelets

1. INTRODUCTION

Ekman et al. [1] proposed the Emotion Facial Action Coding System (EMFACS). This system identifies seven basic facial expressions: *anger, contempt*, *disgust*, *fear, happiness*, *neutral, sadness* and *surprise*. In addition to this list of expressions, EMFACS also defines a set of rules relating a facial expression to a set of AUs.

Traditional facial expression recognition systems rely on an initial feature extraction step followed by a classification algorithm. Previous approaches for representing facial features have exploited global contours [3] and small binary patterns [4]. Both approaches do not explicitly consider the AUs positions. In contrast, EMFACS represents a facial expression as the articulation of the various face AUs. Most previous approaches exploring AUs also rely on a manual fiducial points alignment step [5].

In this article, we formulate facial expression recognition as a signal reconstruction problem of different face components. We divide the face into regions where the most salient AUs are more active using Global Gabor filters and *k*-means clustering. In these regions, we apply a set of Local Gabor filters to detect the face contours. With this approach, our features combine the advantages of explicit AU analysis and contour-based analysis methods. The classifier determines the facial expression by computing the lowest reconstruction error with a dictionary of (labeled) facial expression Local Gabor Moments.

In the next section, we discuss related work. Section 3 describes face region computation. Section 4 presents the facial expression recognition algorithm. The evaluation process is discussed in section 5.

2. RELATED WORK

Facial expression representation tackles the decomposition of an expression into its fundamental elements. In this paper we use the Facial Action Coding System (FACS) [1] to identify facial expressions. The FACS primary goal was "to develop a comprehensive system which could distinguish all possible visually distinguishable facial movements" [1]. EMFACS is an index of AU. An AU is an individual action that humans are able to distinguish, that can be performed by one or more muscles of the face. EMFACS combines AUs into seven universally recognizable expressions: *anger, contempt*, *disgust*, *fear, happiness*, *sadness* and *surprise*. We chose FACS because it is widely used and there are facial expression datasets labeled according to the EMFACS methodology [6], such as the Extended Cohn-Kanade (CK+) [2] and JAFFE [7].

There are two main approaches for facial feature extraction: geometry-based and appearance-based. Geometry-based methods rely on the shape and position of facial components like nose, mouth, eyes and brows to create feature vectors, [5]. Appearance based methods apply contour analysis filters like Gabor wavelets or Local Binary Patterns (LBP) to extract contour features from the face image. Banks of Gabor filters are widely adopted for facial expression recognition [8]–[10].

One of the earliest works [3] in facial expression recognition applied Gabor wavelets and Linear Discriminant Analysis on the JAFFE dataset . Shan et al. [4] popularized Local Binary Pattern features for facial expression

recognition. With an SVM classifier they achieved of 88.4% precision on the CK+ dataset. The combination of Gabor and SVM achieved a precision of 86.9%. Shan et al. concluded that the extraction of LBP features is faster than Gabor wavelets and more resilient to face images with low resolution. Besides Gabor wavelets, Local Binary Patterns and Aspect Appearance Models (requiring proper key-point registration) many different classifiers are proposed in the literature. Littlewort et al. proposed Gabor analysis and Support Vector Machines (SVM) [8]. Tian et al. were among the first to model the temporal dynamics of facial expressions [11]. Moriyama et al. addressed the same task with Hidden Markov models [12]. A thorough review can be found in Tian et al. [13]. More recently, sparse representation for face recognition has been proposed by Wright et al. [14]. A dictionary of face pixels (without any transformation) is used to reconstruct a subject's face. They suggest that as long as the feature space is large enough to represent the original space a regression approach with proper regularization is adequate for face recognition. This triggered a series of other works following similar methods. Zhang et al. [15] applied compressive sensing and sparse representation to create a robust facial expression recognition algorithm. Other authors have pursued similar approaches for facial expression recognition [16] or face recognition robust to facial expressions variations [17]. In contrast to previous approaches, we propose to recognize facial expression as a regression problem of AU regions, more specifically, sparse reconstruction with localized analysis of AU regions.

3. LOCALIZED GABOR-FILTER MOMENTS

Action Units (AU) were identified as the muscular basis actions that produce a facial expression. They have been studied for their ability to associate a facial expression to well identified face key-points (basis muscles). These facial muscle actions are perceived as the intended expression by the visual cortex of the human brain. Since Gabor filters can model base perception functions of the human visual system, they have been widely used in facial expression recognition. In this section, we propose to merge these two ideas and integrate the localized analysis of Action Units and the frequency decomposition provided by Gabor filters.

3.1 Gabor filter-bank

Combinations of Gabor filters, are widely applied in the facial expression recognition literature [8][9] because of their natural ability to detect facial contours (i.e., eyes, nose, mouth, brows and wrinkles), and filter out most of the existing noise [10].

To extract information concerning the face contours and expression traces, several Gabor filters at different orientations and scales will capture the different details of a facial expression. This allows building a dictionary of facial traits and the corresponding intensity. Thus, a Gabor filter is computed as the convolution $f_{\theta,m}(x,y) = \iint I(x_1,y_1) * g_{m\theta}(x-x_1,\, y-y_1)\, dx_1 dy_1$ where I is the face image and $g_{m\theta}$ is the Gabor filter with scale $m \in \{0,1,2,3\}$ and orientation $\theta \in \{0°, 30°\ 60°, 90°, 120°, 150°\}$. Dictionaries with this configuration have been found to work well on a number of domains, namely, facial expressions recognition [10] and image retrieval [18]. This filter is applied to the face image to detect facial traces with a given orientation and scale.

3.2 Robust features

Since we aim at inferring a facial expression automatically and without any human intervention, we cannot rely on approaches that manually register the position of facial key-points on an image. In this section, we detail how to improve feature robustness to poor alignment and subject variations.

3.2.1 Localized Gabor-filter moments

Instead of tracking each AU point, we propose a localized analysis of face regions grouping nearby AU. By inspecting the FACS data, we decomposed the face image, according to the observed per-expression variations. This way, each face region groups a set of AU, and a local analysis of each region allows a specific assessment of the face traits in a particular direction. This renders a greater sense of locality to the Gabor filters output.

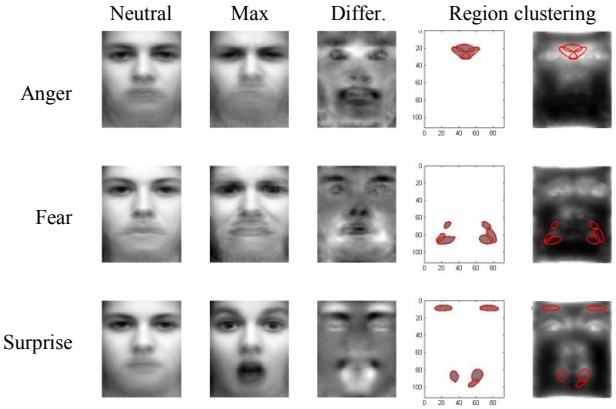

Figure 1. Creation of the expression regions. The represented expressions are *anger*, *fear* and *surprise*.

Our objective is to identify the face regions where the dominant changes occur in each facial expression. The facial expression sequences used to create the features are from the CK+ dataset [2]. We extracted the first (neutral expression) and last (peak expression) images from the dataset. Initially, the frame with the peak facial expression was taken from the sequence (Figure 1, Column 2) and passed through the Gabor filters. To identify the regions with the dominant changes, we subtracted the Gabor filters output of the neutral image (Figure 1, Column 1). The process was repeated for all expression images and an

average Gabor filter output difference was taken per expression (Figure 1, Column 3).

A k-means clustering is then applied to the energy difference image. The number of clusters k is estimated using the techniques described in [19]. The detected clusters are present in Figure 1, Column 3 and 4. Lighter colored regions represent bigger facial expression changes. The dictionary of Gabor filters is applied to the entire face image and the mean and variance of the filters output energy are computed for the clusters regions.

The output of each filter is represented by its mean and variance to improve the features robustness to variations in the face position and alignment. Since there are six regions and twenty-four filters (four scales, six orientations), the dimensionality of the robust representation is 288. Thus, a facial expression j is represented by the vector $f_j = (f_{j_1}, \ldots, f_{j_{288}})$ where $f_{j_i} = avg(Image_{ROI})$ and $f_{j_{i+1}} = std\_dev(Image_{ROI})$. These features provide a robust measure of the face contours surrounding a group of AU key-points. The localized measure of simple statistics (mean and variance) reduce the effect of poor face alignment.

3.2.2 Intra-subject normalization

To increase the relation between Gabor image and facial expressions, a proper normalization must be performed. When a facial expression occurs, the different muscles must act accordingly and position themselves at some distance from its neutral position. We argue that facial expressions are best represented as the difference between the subject neutral expression and the current expression. Thus, the neutral face feature vector is subtracted from the facial expression feature vectors. The sparse reconstruction algorithm uses the resulting vector.

In some situations, it might be easy to obtain the individual's neutral face, while in other situations the individual's average face might be easier to obtain. We compared these two scenarios and a third one where the global average face is the normalizing vector.

4. SPARSE RECONSTRUCTION OF FACIAL EXPRESSIONS

Let us consider a set of k training face images, where each image i contains a facial expression label $z_i \in \{happy, sad, surprise, fear, anger, disgust, contempt\}$. We also define

$$\mathbf{D} = [f_1 \cdots f_k \ ],$$

as the dictionary of Localized Gabor Moments (of dimension m) of all k training examples, where $f_j = (f_{j,1}, \ldots, f_{j,m})$.

One can reconstruct an unseen face image feature vector y_i, as a linear combination of a set of several facial expressions, i.e., the columns of the dictionary $\mathbf{D}$. The reconstruction algorithm gets more support data by reconstructing a facial expression from several images belonging to all expressions. This intuition relies on the fact that micro-expressions are present in all expressions, making it easier to use support data from a different facial expressions that share a common micro-expression in some particular AU. This helps the reconstruction algorithm in minimizing the global representation error. Figure 2 illustrates the reconstruction coefficients of an example image depicting a *surprise* facial expression. The figure illustrates the contribution of each dictionary component to the sparse reconstruction of the image test face. The different colors indicate the label of the dictionary components. The image is classified with the facial expression whose dictionary components minimize the reconstruction error.

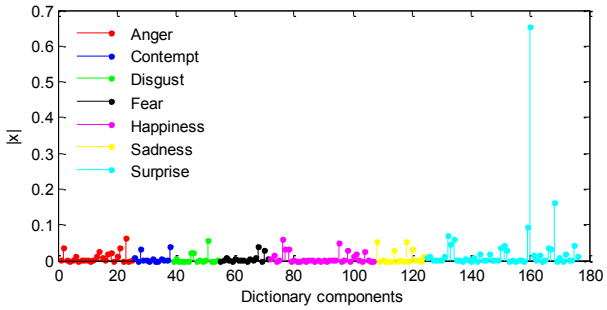

Figure 2. Reconstruction coefficients of an example image labeled with the *surprise* facial expression.

Formally, given the unseen face image feature vector y_i, we wish to minimize the difference between this feature vector and the $\mathbf{D} \cdot x_i$ linear combination, while concentrating the x_i non-null components on a few dimensions. This can be cast as the following optimization problem:

$$x_i = \arg\min_{x_i} \|y_i - \mathbf{D} \cdot x_i\|_2^2 + \lambda \|x_i\|_1$$

λ is the sparsity penalty weight for the reconstruction. The l_1 norm is particularly important, because it aims at maximizing the number of null entries in the x_i vector, thus, it tries to minimize the error by concentrating its representation on a few columns of the dictionary. We implemented the Fast Iterative Thresholding Algorithm (FISTA) optimization algorithm [20] to handle the l_1 regularized minimization problem. FISTA, is an algorithm that solves the above optimization problem by taking an unconstrained approximation of the original problem. FISTA has its roots in the iterative soft thresholding algorithm (ISTA). ISTA iterates by solving the unconstrained problem at a given point using a thresholding function and then taking a gradient step from the local optimal point. FISTA improves the ISTA algorithm by using an improved method to determine the next point at which the minimization step is evaluated that guarantees faster convergence.

To classify a face image with its facial expression, the contribution that each facial expression provides to the minimization of the error is the selected expression. The label expression z_i of the vector y_i is given by the facial

expression that most contributed to the minimization of the representation error, i.e.,

$$z_i = \arg\min_j \left\| y_i - \mathbf{D} \cdot \mathbf{R}_j \cdot x_i \right\|$$

where $\boldsymbol{R}_j$ is an indicator matrix containing all elements set to zero except for the elements corresponding the facial expression j. This allows reconstructing the y_i image with a dictionary containing the columns corresponding to the j^{th} facial expression and the remaining columns set to zero.

5. EVALUATION

5.1 Experimental setup

To assess the facial expression recognition performance, we followed a standard pattern recognition experiment setup. A ten-fold cross-validation procedure was used. Each image is labeled with a facial expression, which is used to measure accuracy. The proposed sparse reconstruction method (SR) is compared to a *k*-NN classifier (with the Euclidean distance) and an SVM classifier (with no kernel). The FISTA algorithm is run with $\lambda = 0.01$ and for a maximum of 1000 iterations (parameters estimated in previous experiments).

Datasets. The datasets chosen for facial expression recognition was the CK+ dataset [2] and the JAFFE dataset [7]. The CK+ is a comprehensive set of sequences of labeled face images. The JAFFE dataset contain the images of 10 Japanese female facial expressions. Both datasets contains images with the facial expressions: *anger*, *disgust*, *fear*, *happiness*, *sadness, surprise* and *neutral*. The CK+ dataset also includes the *contempt* facial expression.

5.2 Results and discussion

Before passing the face images to the facial expression analyzer, the dataset images are pre-processed as follows: (i) a face image dataset is pre-processed to detect every existing faces [21]; (ii) the eyes are detected to align and register the face image; and (iii) images are cropped to remove non-face regions.

We conducted two experiments to assess the proposed methods: first we examined the Localized Gabor Moments (LGM), comparing it to (i) the full set of grayscale face pixels similarly to [14], (ii) the average of all Gabor filters (GM), and (iii) the state-of-the-art Local Binary Patterns (LBP). Table 1 and Table 2 present the results on both datasets. Second, we evaluated influence of the different feature normalizations: non-normalized features; normalized with the subject neutral expression; normalized with the subject average face; and normalized with the global average face (results are in Table 3).

In the CK+ dataset, Table 1 and Table 2, the best results were obtained using sparse reconstruction and Localized Gabor Moments features, normalized with the subject neutral expression (89.8%). Local Binary Patterns were the second type of features. The SVM method came close with 88.9% precision, but performed worse with other types of normalizations. The *k*-NN did not achieve good results for any experiment. We believe that the main reason behind this is the lack of training images, which lead to a bias towards the facial expressions with more images (*surprise*).

Localized Gabor Moments performs better than the other tested features, Table 1. Individual neutral subtraction is better for classification, as the differences between the neutral and the peak expression are the only ones provoked by the expression (little to no noise present). In the JAFFE dataset, Table 3, the sparse reconstruction method achieved the best result in all experiments but one. These results show that the proposed features LGM and SR methods, are a powerful combination to beat other state-of-the-art methods. Both in the CK+ dataset, which is highly heterogeneous, and the JAFFE dataset, which targets a very specific population, the proposed techniques achieved competitive results.

Finally, we observed that normalization played a crucial role in all methods. It should be noted that using the individual's average face to normalize new face images or the individual's neutral face, are the most robust normalization procedures

Table 1. Best precision results for the different features on the CK+ dataset.

Features	SR	SVM	K-NN
LGM	**89.8%**	**88.9%**	**76.4%**
Pixels	76.4%	80.1%	71.3%
Gabor Pixels	77.3%	78.2%	69.4%
LBP	81.9%	82.4%	55.6%

Table 2. Precision for the LGM features on the CK+ dataset.

Normalization	SR	SVM	K-NN
None	80.6%	82.9%	69.4%
Subject neutral	**89.8%**	**88.9%**	**76.4%**
Subject average	87.0%	83.8%	68.1%
Global average	78.7%	79.2%	60.2%

Table 3. Precision for the LGM features on the JAFFE dataset.

Normalization	SR	SVM	k-NN$_{(k=1)}$
None	89.5%	86.8%	89.5%
Subject neutral	**93.7%**	92.4%	79.1%
Subject average	92.4%	**95.7%**	**89.5%**
Global average	90.4%	82.6%	80.9%

6. CONCLUSIONS

In this article, we propose a facial expression recognition approach based on the sparse reconstruction of a facial expression with robust representations of Localized Gabor-filter Moments. The method relaxed the correct positioning of AUs points by examining regions grouping AUs (removing the need for manual intervention). This creates a

robust representation, unaffected by small variations in face alignment and rotation. Signal reconstruction by sparse approximation with a dictionary of AU regions obtained the best result (89.8%) in the CK+ dataset and was the most consistent method across all experiments. Future work will focus on the study of the facial AU points clustering.

Acknowledgements. This work has been partially funded by the project PTDC/EIA-EIA/111518/2009 funded by the *Fundação para a Ciência e Tecnologia* of Portugal.

7. REFERENCES

[1] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.

[3] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205.

[4] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, 2005, pp. II–370.

[5] W. S. S., G. M. Araujo, E. A. B. da Silva, and S. K. Goldenstein, "Facial fiducial points detection using discriminative filtering on principal components," in *2010 IEEE International Conference on Image Processing*, 2010, pp. 2681–2684.

[6] P. Ekman, "Facial expression and emotion," *Am. Psychol.*, vol. 48, no. 4, pp. 384–392, Apr. 1993.

[7] J. G. Michael J. Lyons, Miyuki Kamachi, "Japanese Female Facial Expressions (JAFFE): Database of digital images." 1997.

[8] G. Littlewort and I. Fasel, "Fully automatic coding of basic expressions from video," *INC MPLab Tech. Rep.*, p. 6, 2002.

[9] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans Multimed.*, vol. 8, no. 3, pp. 500–508, 2006.

[10] M. Dahmane and J. Meunier, "Continuous emotion recognition using Gabor energy filters," in *Affective Computing and Intelligent Interaction*, 2011, pp. 351–358.

[11] Y. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, 2002, pp. 229–234.

[12] T. Moriyama, T. Kanade, J. F. Cohn, Z. Ambadar, and H. Imamura, "Automatic recognition of eye blinking in spontaneously occurring behavior," in *Proceedings. 16th International Conference on Pattern Recognition*, 2002, vol. 4, pp. 78–81.

[13] Y. L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of face recognition*, 1st ed., S. Z. Li and A. K. Jain, Eds. Springer, 2005, pp. 247–275.

[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–27, Feb. 2009.

[15] S. Zhang, X. Zhao, and B. Lei, "Robust facial expression recognition via compressive sensing.," *Sensors (Basel).*, vol. 12, no. 3, pp. 3747–61, Jan. 2012.

[16] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *Face and Gesture 2011*, 2011, pp. 336–342.

[17] P. Nagesh, "A compressive sensing approach for expression-invariant face recognition," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1518–1525.

[18] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, 1996.

[19] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 24, no. 3, pp. 381–396, 2002.

[20] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[21] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.