Facial Expression Recognition by Sparse Reconstruction with Robust Features

André Mourão, Pedro Borges, Nuno Correia, and João Magalhães

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

Abstract. Facial expression analysis relies on the accurate detection of a few subtle face traces. According to specialists [3], facial expressions can be decomposed into a set of small Action Units (AU) corresponding to different face regions. In this paper, we propose to detect facial expressions with sparse reconstruction methods. Inspired by sparse regularization and sparse over-complete dictionaries, we aim at finding the minimal set of face atoms that can represent a given expression. l_1 based reconstruction computes the deviation from the average face as an additive model of facial expression atoms and classify unknown expressions accordingly. We compared the proposed approach to existing methods on the well-known Cohn-Kanade (CK+) dataset [6]. Results indicate that sparse reconstruction with l_1 penalty outperforms SVM and *k*-NN baselines with the tested features. The best accuracy (97%) was obtained using sparse reconstruction in an unsupervised setting.

1 Introduction

Facial expressions are commonly represented by the Emotion Facial Action Coding System (EMFACS) proposed by Eckman et al. [3]. This system identifies seven basic facial expressions: *happiness*, *sadness*, *surprise*, *fear*, *anger*, *disgust, contempt* and a state of no expression, *neutral*. These facial expressions are representations of a person's emotional state. Amongst other definitions, EMFACS also constructs a set of rules relating a facial expression to particular face muscle actions (the Action Units, AUs).

Traditional facial expression detection, involves an initial feature extraction step followed by a classifier. Previous approaches for representing facial features have exploited global contours [7] and small binary patterns [10]. Both approaches do not explicitly consider the AUs positions. In contrast, in EMFACS, a facial expression is represented by the articulation of the various AUs.

In this article, we cast facial expression analysis as a signal reconstruction problem of different face components. We decompose the face into regions where most salient AUs are more active. In these regions, we apply a set of Local Gabor filters to detect orientations. With this approach, we bring together into a single method, the advantages of explicit AU analysis and contour-based analysis methods.

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 107–115, 2013.

In the next section, we discuss previous work. Section 3 describes the local frequency analysis of face regions. Section 4 presents the recovery method to detect the facial expression. The evaluation process is discussed in section 5.

2 Related Work

Facial expression representation deals with face features that create and distinguish between facial expressions. In this paper, we have used Emotional Facial Action Coding System (EMFACS) [4] based on the Facial Action Coding System (FACS) [3]. FACS primary goal was "to develop a comprehensive system which could distinguish all possible visually distinguishable facial movements" [3]. FACS is an index of Action Units (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 AU into seven universally recognizable expressions: *happiness*, *sadness*, *surprise*, *fear*, *anger*, *disgust* and *contempt*. We have chosen EMFACS because it is widely recognized and there are facial expression datasets available to the scientific community made according to the EMFACS methodology, such as the CK+ dataset [6].

For facial feature extraction, we applied banks of Gabor wavelets with multiple scales and orientations. Gabor wavelets are widely adopted for facial expression recognition [2, 5, 14] and we have combined them with hard partitioning of the face area into multiple rectangular areas.

Facial expression classification has been tested with multiple features and classifiers in the literature. Zhao et Pietikäinen [16] proposed LBP and Support Vector Machines (SVM) achieving an accuracy of 96.26% on 10-fold validation on the CK+ dataset. Asthana et al. [1] tested various Active Appearance Models (AAM) fitting techniques with SVM. They achieved their best accuracy 95.88% on the CK+ dataset with Iterative Error Bound Minimization Methods scheme, but AAM techniques require manual annotation of the eye position for calibration. A thorough review of facial expression recognition techniques can be found in Valstar et al. [11].

Sparse representation was also tested for face recognition. Wright et al. [13] proposed sparse representation with the facial recognition training data as overcomplete dictionary of face pixels (without any transformation). 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. Other authors have proposed similar approaches for facial expression recognition [15] or face recognition robust to facial expression variations [9]. We propose facial expression recognition as a regression problem of AU regions, using sparse reconstruction with localized analysis of AU regions.

3 A Robust Dictionary for Facial Expression Analysis

Action Units (AU) were identified as the muscular basis actions underlying to every facial expression. They have been studied for its 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 analysis as the state-of-the-art. 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

Using ideas already deployed in visual analysis of facial expressions [5] [14], we implemented a dictionary of Gabor filters. Combinations of these filters, are widely applied in the facial expression analysis literature 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 [2].

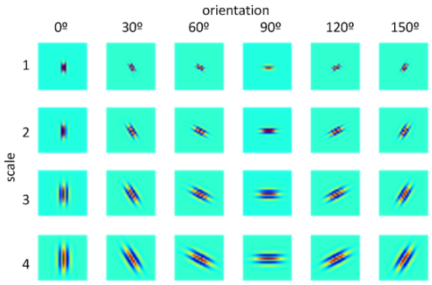

Fig. 1. Gabor filter-bank

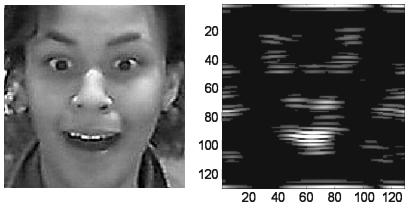

Fig. 2. Facial expression example and its filter output (scale 3 and orientation 0°)

Gabor filters are edge detector filters composed by a two-dimensional wave sign weighted by an exponential decay that can be applied at a given orientation and scale, see Manjunath and Ma [8] for a detailed discussion. 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 us to build a dictionary of facial traits and the corresponding intensity. Thus, a Gabor filter is computed as

$$f_{\theta,m}(x,y) = \iint I(x_1,y_1) * g_{m\theta}(x-x_1,\, y-y_1)\, dx_1 dy_1, \tag{1}$$

where I is the face image and $g_{m\theta}$ is the Gabor filter with scale m and orientation θ. Figure 1 presents the dictionary of Gabor filters at multiple scales and orientations. Dictionaries with this configuration have been found to work well on a number of domains, namely, facial expressions analysis [2] and image retrieval [8].

This filter is applied to the face image to detect facial traces with a given orientation and scale. This corresponds to the convolution of the Gabor filter with the face

image to produce the filter output, as illustrated in Figure 2. The detected face contours with the orientation of the filter are represented in white – it is clear that the mouth area is highly expressive on the 0° (horizontal) orientation.

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. Thus, in this section, we detail how to make Gabor wavelets robust to small alignment variations and to subject variations.

Localized Gabor-Filter Moments

A facial expression is represented by the position of the various face components - different expression will make specific AU (mouth wide open: AU26 and arched eyebrows: AU1+AU2+AU5 equal *Surprise*). To classify an expression, it is necessary to estimate the state of each the AU and compare them to the existing facial expression models.

Instead of tracking each AU point, we propose a localized analysis of face regions grouping nearby AU. Examining the FACS data and the contour representation provided by the dictionary of Gabor filters, we followed a hard-partitioning of the face image where we observed the largest variations per expression (for example, eyes/brows area and mouth area). This way, each face region groups a set of AUs, and

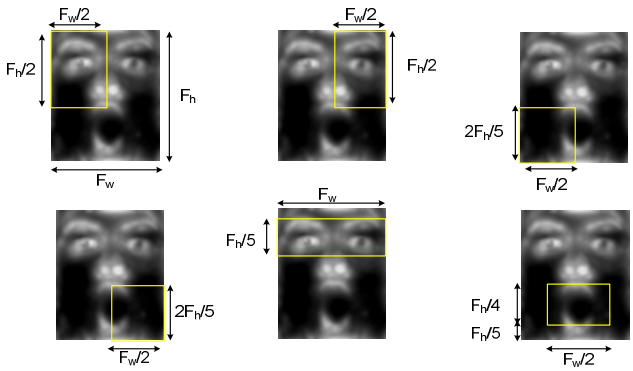

Fig. 3. AU regions highlighted on a *surprise* Gabor face. Each region groups a set of AUs and the Local Gabor filters analyze the face traits direction in each region. F_h is the face image height and F_w is the face image width.

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, and increase the robustness to small pans and rotations in the face image. In Figure 3,

we show an average face (created from the data from CK+ dataset) with the different rectangular areas highlighted.

The dictionary of Gabor filters is applied to each one of these regions to obtain the face regions contours. To improve the features robustness, each the output of each filter is represented by its mean and variance. These features are of particular interest because they are highly robust to poor facial alignment. Since there are six regions and twenty-four filters (four scales: m, six orientations: θ), 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 $f_{j_{i+1}} = std\_dev(Image_{ROI})$, with $\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$, $m \in \{1, 2, 3, 4\}$ and $Image_{ROI}$ is the feature vector from one of the ROI (from Figure 3) of the image.

Normalization: Deviation from Neutral

To increase the relation between Gabor filters output 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 neutral expression and the current expression. Thus, we subtract the features of a given expression features from the neutral face features and represent a facial expression as this normalized vector.

In some situations, it might be easy to obtain the individual's neutral face, while in others 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 variable. The CK+ dataset allows for both approaches as it contains the neutral face for every expression.

4 Sparse Reconstruction with Robust Features

Let us consider a set of k training face images, where each image j contains a facial expression label $l_j \in \{happy, sad, surprise, fear, anger, disgust, contempt\}$. We also define **D** as the dictionary of Localized Gabor Moments (of dimension m) of all k training examples:

$$\mathrm{D} = [f_1^T \cdots f_m^T \; f_{m+1}^T \cdots f_{2m}^T \; f_{2m+1}^T \quad \cdots \quad f_{km}^T]. \tag{2}$$

One can reconstruct an unseen face image feature vector y_i, as a linear combination of a set of several micro-expressions, i.e., the columns of the dictionary **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.

Thus, 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 to a few dimensions. This is cast as the quadratic optimization problem:

$$x_i = \arg\min_{x_i} \|y_i - \mathbf{D} \cdot x_i\| \qquad \text{subject to } \|x_i\|_0 < \varepsilon \tag{3}$$

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 in a few micro-expressions of the dictionary. We implemented the FISTA optimization algorithm to handle the l_1 constrained minimization.

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 of the vector x_i is given by the facial expression that most contributed to the minimization of representation error:

$$l_i = \arg\min_{j} \left\| y_i - \mathbf{D} \cdot r_j \cdot x_i \right\| \tag{4}$$

where r_j is an indicator matrix containing all elements equal to zero except for the elements corresponding the facial expression j. This allows reconstructing the y_i image with a dictionary

$$\mathrm{D} = \left[[0 \cdots 0] \ [f_{m+1}^T \cdots f_{2m}^T] \cdots \ [0 \ \cdots \ 0] \right] \tag{5}$$

containing the columns corresponding to the j^{th} facial expression and the columns corresponding to the other facial expressions set to zero.

5 Evaluation

5.1 Experimental Setup

To assess the facial expression detection performance, we followed a standard pattern recognition experiment setup. The dataset was split into a training set (70%) and a test set (30%). 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).

Dataset. The dataset chosen for facial expression detection was the CK+ dataset [6]. It is a comprehensive set of sequences of labeled face images. It contains images with various facial expressions: *anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* and a *neutral* face for each sequence. 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 [12]; (ii) faces are aligned by detecting the best eye pair; and cropped (iii) to ensure that the images are correctly aligned for facial expression recognition.

5.2 Results and Discussion

We conducted two experiment to assess the proposed methods: first we examined the Localized Gabor Moments (LGM), comparing it to the average of all Gabor filters (GM), and the full set of grayscale face pixels similarly to [13] (results in Table 1). Second, we evaluated influence of the different feature normalizations: non-normalized features; normalized with own neutral facial expression; normalized with own average facial expression; and normalized with average of all observed faces (results in Table 2).

Table 1. Accuracy results for the rectangular features

	SR	**SVM**	***k*-NN: k = 1**	***k*-NN: k = 3**
LGM	**0.97**	**0.95**	**0.79**	**0.78**
GP	0.82	0.29	0.76	0.68
Pixels	0.89	0.29	0.76	0.70

Table 2. Accuracy results for the rectangular features

	SR	**SVM**	***k*-NN: k = 1**	***k*-NN: k = 3**
No normalization	0.88	**0.91**	0.79	0.68
Individual neutral	**0.97**	0.95	0.79	0.78
Individual's average face	**0.96**	0.88	0.71	0.66
Global average face	**0.87**	0.86	0.71	0.64

The best results were obtained using proposed the sparse reconstruction with Localized Gabor Moments and features with own neutral subtraction (97% accuracy) and average individual's face subtraction (96% accuracy). The SVM came close with 96% accuracy with features with own neutral subtraction, but performed much worse using other types of neutral faces. 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 (only 170 faces for all expressions), which lead to a bias towards the facial expressions with more images (for example *surprise*).

Localized Gabor Moments perform better than the other tested features, as they are more resilient to small changes in the face (pans and rotations). Individual neutral subtraction is better for classification, as the differences between the neutral and the peak expression are only the ones provoked by the expression (little to no noise present), but this neutral face is only available in very specific settings and might not be possible to obtain it in an unsupervised setup. This was fundamental for the sparse reconstruction approach.

Finally, it should be noted that using the individual's average face to normalized new facial expressions works almost as well (1% accuracy difference), and the normalizing vector can be easily captured in a real setting.

6 Conclusions

In this article, we proposed a facial expression detection 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 (removing the need for manual intervention) by examining regions grouping AUs. 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 results (97%) in the CK+ dataset. Our approach performs on par with the state of the art techniques [1, 11, 15, 16] and can be performed in a fully unsupervised setting.

Acknowledgements. This work has been partially funded by the Portuguese National Foundation under the projects UTA-Est/MAI/0010/2009, PTDC/EIA-EIA/105305/2008 and PEst-OE/EEI/UI0527/2011.

References

1. Asthana, A., et al.: Evaluating AAM fitting methods for facial expression recognition. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–8. IEEE (2009)
2. Dahmane, M., Meunier, J.: Continuous emotion recognition using Gabor energy filters. In: Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction, pp. 351–358 (2011)
3. Ekman, P., et al.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
4. Ekman, P.: Facial expression and emotion. The American Psychologist 48(4), 384–392 (1993)
5. Littlewort, G., Fasel, I.: Fully automatic coding of basic expressions from video. INC MPLab Technical Report 6 (2002)
6. Lucey, P., et al.: The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101. IEEE (2010)
7. Lyons, M., et al.: Coding facial expressions with Gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205. IEEE Comput. Soc. (1998)
8. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(8), 837–842 (1996)
9. Nagesh, P.: A compressive sensing approach for expression-invariant face recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1518–1525. IEEE (2009)
10. Shan, C., et al.: Robust facial expression recognition using local binary patterns. In: IEEE International Conference on Image Processing 2005, pp. II–370. IEEE (2005)
11. Valstar, M.F., et al.: Meta-Analysis of the First Facial Expression Recognition Challenge. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society 42 (2012)

12. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision 57(2), 137–154 (2004)
13. Wright, J., et al.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 210–227 (2009)
14. Yeasin, M., et al.: Recognition of facial expressions and measurement of levels of interest from video. IEEE Transactions on Multimedia 8(3), 500–508 (2006)
15. Zhang, S., et al.: Robust facial expression recognition via compressive sensing. Sensors (Basel, Switzerland) 12(3), 3747–3761 (2012)
16. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 915–928 (2007)