Local versus Global Segmentation for Facial Expression Recognition

Jacob Whitehill
Department of Computer Science
University of the Western Cape, South Africa
whitehill@cs.stanford.edu

Christian W. Omlin
Department of Mathematics & Computing Science
University of the South Pacific, Fiji
omlin_c@usp.ac.fj

Abstract

We examined the open issue of whether FACS action units (AUs) can be recognized more accurately by classifying local regions around the eyes, brows, and mouth compared to analyzing the face as a whole. Our empirical results showed that, contrary to our intuition, local expression analysis showed no consistent improvement in recognition accuracy. Moreover, global analysis outperformed *local analysis on certain AUs of the eye and brow regions. We attributed this unexpected result partly to high correlations between different AUs in the Cohn-Kanade expression database. This underlines the importance of establishing a large, publicly available AU database with singly-occurring AUs to facilitate future research.*

1. Introduction

Automatic facial expression recognition has applications to human-computer interaction, interactive computer games, and psychological research. It is also a crucial component of any computer system designed to recognize a signed language in real time. As part of a larger project on the integration of signed with spoken communication, we are studying machine-learning algorithms for the recognition of facial expressions. The well-known Facial Action Coding System (FACS) by Ekman and Friesen [3] provides the framework.

FACS defines expressions in terms of the presence or absence of 44 elementary muscle movements, called *action units* (AUs). Although these muscles are situated within local regions of the face (see Figure 1), AUs can sometimes also impact on facial regions outside of their origin. For example, AU 6 (cheek raise), though triggered by a muscle circling the eye, can also accentuate the nasolabial furrow around the mouth [3]. It is thus unclear whether local classification would improve recognition accuracy through reduced noise, or decrease accuracy due to the loss of relevant, global appearance information.

2. Problem Statement

In this paper, we investigate the open issue of whether local analysis of the face and classification of AUs results in higher AU recognition accuracy than global classification, i.e., from the whole face. Our metric for comparison is the percentage of face pictures that are classified correctly for the presence or absence of a particular AU. We restrict our study to images without partial occlusion or out-of-plane rotation of the head. Support vector machines with linear kernels are used for classification.

3. Related Work

Techniques for expression recognition can be divided roughly into two classes: *appearance-based methods*, and *feature-point methods*. Appearance-based techniques analyze the individual pixels of the face, possibly after performing a dimension reduction or frequency filtering. Feature-point methods instead track the locations of key points of the face (e.g., corners of the eyes) and classify expressions based on the spatial relationship between these points. Best results for each approach are similar:

Tian, et al [8] developed a feature point-based, multi-state model of 7 upper- and 11 lower-face AUs. Using neural networks as classifiers, they achieved over 95% accuracy in each group of AUs. Donato, et al [2] compared a variety of appearance-based methods and achieved 96% accuracy on 12 AUs, both with Gabor filters and independent component analysis. Bartlett, et al [1], in more recent work, used Gabor filters, support vector machines, and hidden Markov models to detect AUs 1, 2, and 4 with up to 90% accuracy.

3.1. Local versus Global Segmentation

Little work has been done comparing the advantage of local versus global regions for facial expression recognition. Most such work has conducted this comparison for prototypical expressions, and no study, to our knowledge, has assessed the comparative performance for FACS AUs.

Lisetti and Rumelhart developed neural networks to classify faces as either smiling or neutral [5]. They compared two networks: one which was trained and tested on the whole face, and one which was applied only to the lower half of the face (containing the mouth). For their application, local analysis of the lower face-half outperformed the global, whole-face analysis.

Padgett and Cottrell compared global to local face analysis in the recognition of six prototypical emotions. In particular, they compared principle component analysis (PCA) on the whole-face (*eigenfaces*) to PCA on localized windows around the eyes and mouth (*eigenfeatures*). The projections onto the eigenvectors from each analysis were submitted to neural networks for expression classification. In their experiments, the localized recognition clearly outperformed global recognition. Padgett and Cottrell attribute these results both to an increased signal-to-noise ratio and to quicker network generalization due to fewer input parameters [7].

Finally, Littlewort, et al compared whole-face, upper-half, and lower-half face segmentations for the recognition of prototypical facial expressions. In their experiments, the whole-face analysis clearly outperformed the other two segmentation strategies by several percentage points [6].

4. Experiment

We assessed the comparative performance of the local and global segmentations in the task of FACS AU recognition for the following AUs: 1, 2, and 4 (brow AUs); 5, 6, and 7 (eye AUs); and 15, 17, 20, 25, and 27 (mouth AUs). We denote this combined set of AUs as $\mathscr{A}$.

Each AU number, along with a sample image and the number of samples in our test database (described later), is shown in Figure 1.

As data set we used the Cohn-Kanade AU-Coded Facial Expression Database [4]. This database contains images of individual human subjects performing a variety of facial expressions. In the public version of this database, 97 different human subjects, ranging from ages to 18 to 30, performed six prototypical expressions: anger, disgust, fear, joy, sadness, and surprise. For each subject and expression, the database contains a sequence of face images beginning with the "neutral" expression (containing no AUs) and ending with the target expression. Certified FACS coders mapped each image sequence in the database to the set of AUs that were exhibited in that sequence.

Our experiments required the positions of the eyes and mouth in each image. We used a subset of the Cohn-Kanade Database containing 580 images from 76 human subjects. From each image sequence of each subject, we used the first two images, which contained the "neutral" expression, and the last two images, in which the target expression was most pronounced. We classified only those AUs for which at least 40 positively labelled images were available in our data subset.

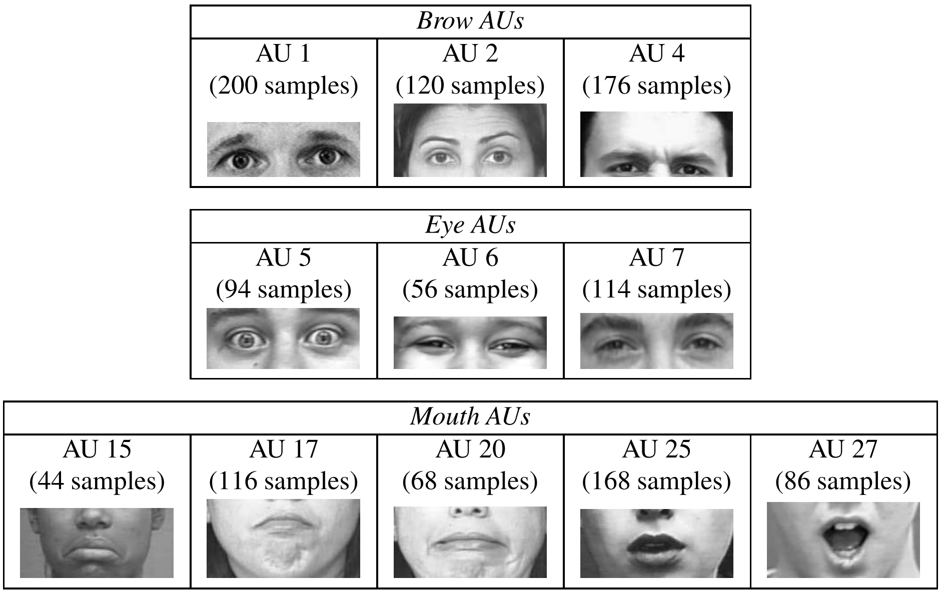
AU recognition was performed in four stages: image processing, image segmentation, feature extraction, and classification. We describe each stage below.

4.0.1. Image Processing. Prior to segmenting the local and global regions, all images were rotated and scaled such that the coordinates of the eyes and mouth were constant over all images. The face width was set to 64 pixels; the inter-ocular distance was set to 24 pixels; and the y-distance between the eyes and mouth was 26 pixels.

4.0.2. Image Segmentation. For the local expression analysis, images were segmented by cropping square regions around the center of the eyes brows, and mouth. The center of the brows was estimated by shifting the center of the eyes up by one-fourth the inter-eye width. In all cases, the width of each square was 24 pixels.

For global analysis, the face square region was cropped at a width of 64 pixels around (x_c, y_c), where x_c is the x-coordinate of the midpoint between the eyes, and y_c is the y-coordinate of the midpoint between the eyes and mouth. See Figure 2 for an illustration of image segmentation.

4.0.3. Feature Extraction. Each segmented image was converted into a Gabor representation using a bank of 40 Gabor filters. Five spatial frequencies (spaced in half-octaves) and eight orientations (spaced at $\pi/8$) were used. Feature vectors were calculated as the com-



Brow AUs		
AU 1 (200 samples)	AU 2 (120 samples)	AU 4 (176 samples)

Eye AUs		
AU 5 (94 samples)	AU 6 (56 samples)	AU 7 (114 samples)

Mouth AUs				
AU 15 (44 samples)	AU 17 (116 samples)	AU 20 (68 samples)	AU 25 (168 samples)	AU 27 (86 samples)

Pictures courtesy of Carnegie Mellon University Automated Face Analysis Group, `http://www-2.cs.cmu.edu/afs/cs/project/face/www/facs.htm`.

Figure 1. Classified AUs and Prevalence in Dataset

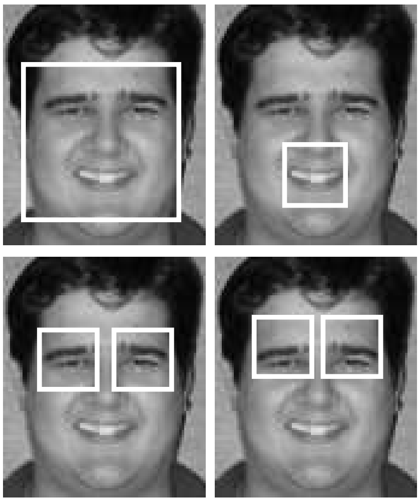


Figure 2. The global segmentation (top-left); and the local segmentations of the mouth (top-right), eye (bottom-left), and brow regions (bottom-right).

plex magnitude of the Gabor jets, and vectors were then subsampled by a factor of 16 and normalized to unit length as in [2].

4.0.4. Classification. We trained and tested all classifiers on the subset of the Cohn-Kanade database described in Section 4. Each trained classifier detected the presence or absence of one AU, regardless of whether it occurred in combination. We did not attempt to account for non-additive AU combinations.

Ten-fold cross-validation was employed to test the generalization performance. Each fold contained all the images of a particular group of subjects. None of the validation folds contained the same human subject. We calculated mean accuracies over the ten test folds. When comparing recognition accuracy between two facial segmentations, we performed matched-pairs t-tests in order to assess the statistical significance of any difference in mean performance.

5. Results

Recognition accuracies (%) are displayed in Table 1 for both the local and global segmentations; the particular local segmentation depended on the region in which the AU is centered. Whenever a statistically significant difference was identified (for 95% confidence, the p value of the t-test must be less than 0.05), the superior segmentation is listed. When no statistically

Table 1. Cross-validation recognition accuracies for all AUs.

AU Recognition Accuracy			
	Segmentation		
AU #	Local	Global	Best
Brow AUs			
1	77.99	89.67	Global
2	88.29	94.23	=
4	86.65	89.73	=
Eye AUs			
5	94.08	92.60	=
6	87.80	94.44	Global
7	93.86	92.43	=
Mouth AUs			
15	94.96	95.07	=
17	90.67	91.47	=
20	96.51	95.16	=
25	96.49	95.08	=
27	98.16	99.42	Global
Avg	**91.41**	**93.57**	

significant difference was present, an = sign is listed. In some cases (e.g., AU 2), the mean accuracies between segmentations may differ by several percentage points and yet not be statistically significant.

To summarize the results, the local segmentation failed to achieve any consistent and statistically significant advantage over the global segmentation in terms of recognition accuracy. More surprising is that, for for AUs 1 and 6, the global strategy achieved higher accuracy. Overall, the average accuracy over all AUs was higher for the global segmentation.

6. Analysis

We view two factors as possibly responsible for the statistically indistinguishable, and sometimes even significantly superior performance of the global segmentation relative to the local strategy. The first is that certain AUs may affect regions of the face outside of the AUs' muscular origin (see Section 1), and therefore the global segmentation may profit from this non-local appearance information. The second is that, due to the high degree of AU correlation in the Cohn-Kanade database, one AU in one face region may be predictive of another AU elsewhere in the face.

6.1. Inter-AU Correlation

Some AUs are easier to detect to classify than others, both by humans and, as witnessed by the results of Table 1, by computerized classification. Suppose now that AU i were more difficult to classify than AU j: If it were known that AU i were highly correlated with another AU j (e.g., $\rho_{ij} = 1$), then a classifier for AU i could attempt to classify instead AU j, and then output the same result for AU i. Note that the *global* segmentation could benefit from this correlation even if AUs i and j occur in different parts of the face. A *local* segmentation strategy, on the other hand, would be unable to observe AU j's appearance changes on the face and thus would not profit from this correlation.

This hypothesis is supported by the matrix of inter-AU correlations over our data subset given in Table 2. Correlation coefficients over the entire Cohn-Kanade database are similar. We considered the correlation between AUs i and j to be high if $|\rho_{ij}| \geq 0.60$; the corresponding entries are shown in bold. Notice how AUs in one region of the face may be highly correlated with AUs in a different region. In particular, AU 1 is highly correlated with AU 25, and AU 2 is highly correlated with both AU 25 and AU 27.

6.1.1. Experiment. In order to test whether inter-AU correlation was responsible for the superior performance of the global classifier for certain AUs, we performed the following experiment: To the feature vectors of both the global and local segmentations, we appended the classification label $\mathrm{au}_i \in \{0,1\}$ of every AU in $\mathscr{A}$ *except* the one to be classified. For instance, for a classifier for AU 1, we augmented the standard Gabor feature vector $\mathscr{F}_n$ of each classified image n to be:

$$\mathscr{F}_n \,.\, (\mathrm{au}_2, \mathrm{au}_4, \mathrm{au}_5, \mathrm{au}_6, \mathrm{au}_7, \mathrm{au}_{15}, \mathrm{au}_{17}, \mathrm{au}_{20}, \mathrm{au}_{25}, \mathrm{au}_{27})$$

where the dot . represents vector concatenation, and au_i is the true classification label for AU i in image n. Each feature vector was thus given perfect knowledge of the presence or absence of every other AU in $\mathscr{A}$. If the correlation was truly responsible for the global segmentation strategy's superior classification performance, then the local classifier should perform at least as well as the global classifier when using the modified feature vectors.

In our experiment, we modified the feature vectors for AUs 1, 2 6, and 27 - all the AUs for which the global segmentation had shown superior performance [1]. Classification results are displayed in Table 3.

[1] In the case of AU 2, the difference in mean accuracy was statistically insignificant.

Table 2. Inter-AU correlation matrix for our subset of the Cohn-Kanade database. Entries ρ_{ij} where $|\rho_{ij}| \geq 0.60$ (other than self-correlation) are marked in bold.

	Brow AUs			Eye AUs			Mouth AUs				
AU #	1	2	4	5	6	7	15	17	20	25	27
1	1.00	**0.69**	0.26	0.59	-0.08	-0.07	0.39	0.15	0.38	**0.69**	0.58
2	**0.69**	1.00	-0.18	**0.76**	-0.13	-0.23	0.02	-0.08	0.01	**0.65**	**0.83**
4	0.26	-0.18	1.00	-0.13	0.46	**0.73**	0.26	**0.61**	0.45	0.17	-0.23
5	0.59	**0.76**	-0.13	1.00	-0.08	-0.15	-0.09	-0.15	0.05	**0.65**	**0.76**
6	-0.08	-0.13	0.46	-0.08	1.00	0.53	-0.09	0.26	0.18	0.11	-0.13
7	-0.07	-0.23	**0.73**	-0.15	0.53	1.00	-0.08	0.43	0.31	0.14	-0.21
15	0.39	0.02	0.26	-0.09	-0.09	-0.08	1.00	0.54	-0.10	-0.15	-0.08
17	0.15	-0.08	**0.61**	-0.15	0.26	0.43	0.54	1.00	-0.12	-0.26	-0.21
20	0.38	0.01	0.45	0.05	0.18	0.31	-0.10	-0.12	1.00	0.54	-0.12
25	**0.69**	**0.65**	0.17	**0.65**	0.11	0.14	-0.15	-0.26	0.54	1.00	**0.65**
27	0.58	**0.83**	-0.23	**0.76**	-0.13	-0.21	-0.08	-0.21	-0.12	**0.65**	1.00

Table 3. Recognition accuracies with SVMs for the local and global segmentations, using both the standard and augmented feature vectors (with inter-AU correlation information).

	Feature Vector			
AU #	Local	Augmented Local	Global	Augmented Global
1	77.99	86.15	89.67	89.93
2	88.29	92.66	94.23	94.36
6	87.80	87.47	94.44	94.62
27	98.16	98.16	99.42	99.42

The local segmentations for AUs 1 and 2 benefited significantly ($p << 0.01$) from the appended correlation information. The corresponding global segmentations, on the other hand, were nearly equal despite the added correlation data. Moreover, after augmenting both the local and global feature vectors for AUs 1 and 2, no statistically significant difference between the local and global segmentations remained ($p > 0.05$). We thus conclude that inter-AU correlation was responsible for the higher recognition accuracy of the global segmentation for these AUs.

For AUs 6 and 27, however, the augmentation of the actual AU labels did not appreciably benefit the local segmentations. Correspondingly, the increase in recognition accuracy achieved by the global segmentation remained statistically significant ($p < 0.01$ for AU 6 and $p < 0.02$ for AU 27). These results are consistent with the relatively low correlations of AU 6 with other AUs in $\mathscr{A}$, as shown in Table 2. It is conceivable, however, that the global classifier for AUs 6 and 27 did profit from correlations with AUs not in $\mathscr{A}$, and that this contributed to the global classifier's higher performance.

7. Summary and Conclusions

We have investigated empirically the effect of a global versus local face segmentation on the task of FACS AU recognition. The local segmentation strategy failed to demonstrate a consistent advantage for any AU over both classifiers. Some AUs were even classified more accurately with both classifiers when the entire face was analyzed.

We also studied the effect of inter-AU correlation within facial images on the recognition accuracies of various AUs. Our results show that this correlation effect can impact recognition rates significantly for some AUs. Such correlation effects may be of little consequence when recognizing prototypical expressions, in which high AU correlation is natural. They are of considerable importance, however, when analyzing single AUs, as recognition rates will appear misleadingly high. We would thus like to underline the importance of establishing a large, publicly available AU database with singly-occurring AUs to facilitate future research.

Acknowledgements

We thank Dr. (name removed) and Dr. (name removed) of the (their institution) for their helpful feedback. This research was funded in part by the (agency name omitted to preserve anonymity).

References

[1] M. Bartlett, G. Littlewort, B. Braathen, T. Sejnowski, and J. Movellan. A prototype for automatic recognition of spontaneous facial actions. In S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems, Vol 15. MIT Press*, 2003.

[2] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[3] P. Ekman and W. Friesen. *Manual for the Facial Action Coding System*. Consulting Psychologists Press, 1977.

[4] T. Kanade, J. Cohn, and Y. L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 46 – 53, March 2000.

[5] C. Lisetti and D. Rumelhart. Facial expression recognition using a neural network. In *Proceedings of the 11 th International FLAIRS Conference*, 1998.

[6] G. Littlewort, I. Fasel, M. S. Bartlett, and J. Movellan. Fully automatic coding of basic expressions from video. Technical Report 03, University of California San Diego MPLab, 2002.

[7] C. Padgett and G. Cottrell. Representing face images for emotion classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, 1997.

[8] Y. L. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.