Automatic Recognition of Lower Facial Action Units

Isabel Gonzalez igonzale@etro.vub.ac.be
Hichem Sahli hichem.sahli@etro.vub.ac.be
Werner Verhelst wverhels@etro.vub.ac.be

Joint Research Group on Audio Visual Signal Processing (AVSP),
Vrije Universiteit Brussel, Department ETRO,
Pleinlaan 2, 1050 Brussels

ABSTRACT
The face is an important source of information in multi-modal communication. Facial expressions are generated by contractions of facial muscles, which lead to subtle changes in the area of the eyelids, eye brows, nose, lips and skin texture, often revealed by wrinkles and bulges. To measure these subtle changes, Ekman et al. [5] developed the Facial Action Coding System (FACS). FACS is a human-observer-based system designed to detect subtle changes in facial features, and describes facial expressions by action units (AUs). We present a technique to automatically recognize lower facial Action Units, independently from one another. Even though we do not explicitly take into account AU combinations, thereby making the classification process harder, an average F_1 score of 94.83% is achieved.

Author Keywords
Facial Action Units, AdaBoost, OVL, SVM.

INTRODUCTION
The Facial Action Coding System (FACS) [5] is a human-observer-based system designed to detect subtle changes in facial features, and describes facial expressions by 44 anatomically based Action Units (AUs). AUs can occur individually or in combinations. When AUs occur in combination they may be nonadditive, in which case the combination changes the appearance of the constituents.
FACS coding is very labor-intensive; automating this process has attracted a lot of attention since the early 1990s. Automatic facial analysis requires extraction of features from static images or image sequences, and classification into AUs, or AU combinations. Basically, there are two main approaches for feature extraction: geometry-, and appearance-based. Geometry-based feature extraction consists of detecting and tracking facial feature points. Appearance-based features concern motion and texture changes. Methods to extract these features are Gabor Wavelets Analysis, optical flow analysis, PCA, ICA.

It has been shown that the combination of geometry- and appearance-based features give better recognition results, [4, 10, 13].

State of the art classification methods are Support Vector Machines (SVM) [1, 10], Neural Networks, Hidden Markov Models, and Discriminant Analysis.

FEATURE EXTRACTION
Our features are geometry, and consist of a shape model of 83 facial feature points [6], see Figure 1. Once we have the shape model for all video frames, we apply an affine transformation by warping each image to a common view.
Our initial set of features is the set of all facial feature points $p_i = (x_i, y_i)$, with $i = 1...83$. Next, we calculate for all points p_i, the displacements of the facial feature points from the current frame t relative to those in a neutral frame 0:

$$f(p_i)_t = || p_{i,t} - p_{i,0} ||$$

The last set of features consists of several features calculated from the facial feature points. The equations are constructed from the linguistic description of the FACS manual, and from state of the art work [8, 13]. Table 1 shows these equations; Table 2 presents the description of auxiliary points derived from the shape model. $d(p_i, p_j)$ denotes the Euclidean distance between points p_i and p_j , while $dv(p_i,p_j)$ denotes the vertical distance between the

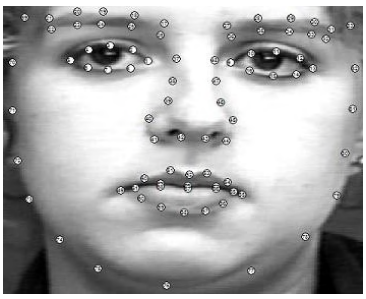


Figure 1. Shape Model, © J.F. Cohn

Id	Equation	Ref
1	$(dv(52,15)_t + dv(52,7)_t)/2 - (dv(52,15)_0 + dv(52,7)_0)/2$	[13]
2	$d(52,O)_t - d(52,O)_0$	[13]
3	$d(52,58)_t - d(52,58)_0$	[13]
4	$d(44,41)_t - d(44,41)_0$	(*)
5	$(d(A,49)_t + d(A,55)_t) - (d(A,49)_0 + d(A,55)_0)$	[8]
6	$\| d(55,9)_t - d(55,9)_0 \|$	[13]
7	$\| d(49,1)_t - d(49,1)_0 \|$	[13]
8	$d(55,49)_t - d(55,49)_0$	[13]
9	$(dv(15,29)_t + dv(7,19)_t)/2 - (dv(15,29)_0 + dv(7,19)_0)/2$	[13]
10	$(d(9,27)_t + d(1,17)_t)/2 - (d(9,27)_0 + d(1,17)_0)/2$	[13]
11	$d(J,J')_t - d(J,J')_0$	[13]
12	$(d(15,11)_t + d(7,3)_t)/2 - (d(15,11)_0 + d(7,3)_0)/2$	(*)
13	$d(63,52)_t - d(63,52)_0$	(*)
14	$d(62,51)_t - d(62,51)_0$	(*)
15	$d(64,53)_t - d(64,53)_0$	(*)
16	$d(58,67)_t - d(58,67)_0$	(*)
17	$d(57,66)_t - d(57,66)_0$	(*)
18	$d(59,68)_t - d(59,68)_0$	(*)
19	$d(B,B')_t - d(B,B')_0$	(*)
20	$d(58,52)_t - d(55,49)_t$	(*)

Table 1. Equations formulated according to AU Description in FACS manual [5] by I. Gonzalez and P. Fan, ETRO, VUB (*); by Zhang [13], and el Kaliouby [8].

two points:

$$dv\,(p_i, p_j)_t = |y_{i;t} - y_{j;t}|$$

Equations (1) to (19) are estimates on the current frame *t* relative to the neutral frame *0*, and compute the following: (1) the average vertical distance between the upper lip center and the eyes; (2) the distance of the upper lip center to the center point of the nose; (3) the mouth height in the center of the upper and lower lip; (4) the width between the outer nostril wings; (5) the width of the mouth, from left corner to center and from right corner to center; (6) and (7), the distance from the mouth corner to the inner eye corner, right and left respectively; (8) the mouth width, from left to right corner; (9) the average vertical distance between lower eye lid center and upper brow center; (10) the average distance between inner eye corner and inner brow; (11) the distance between the inner brows; (12) the parts of the lips, in all points; and (19) the mean height of the mouth. Finally, (20) computes the difference between the height and the width of the mouth. All feature sets, i.e. the facial feature points, the displacements, the other geometric features of Table 1, and a combination thereof, are evaluated in the experiments.

CLASSIFICATION

We compared classification by AdaBoost and SVM. AdaBoost (Adaptive Boosting) is a machine learning algorithm [7] that constructs a strong classifier as a linear combination of weak classifiers. AdaBoost was also used as a feature selection technique. For the feature selection, we used decision stumps as weak classifiers. The feature that minimizes the classification error on the weighted samples is chosen in each iteration.

A Support Vector Machine (SVM) is a supervised method used for classification and regression, and has a well-founded mathematical theory [12]. Through training, a SVM builds a hyper plane, or set of hyper planes, in a high or infinite dimensional feature space which separates two or more classes. The maximum-margin hyper plane is the hyper plane that maximizes the distance from it to the nearest data point on each side. The larger the margin is, the lower the generalization error of the classifier.

EXPERIMENTS

Our system was trained on the Cohn and Kanade's DFAT-504 dataset [9]. This database consists of 486 sequences of facial displays that are produced by 98 University students from 18 to 30 years old, of which 65% is female. All sequences are annotated by certified FACS coders. We selected all sequences where at least one of the analyzed AUs is present. This resulted in 364 sequences from 94 subjects (66% female). Training and evaluation is done on the last frame of each sequence, which contains the apex of the expression.

Our final goal is to develop a system that detects behavioral patterns in communication between two persons, more specifically between mother and infant. We therefore selected a set of AUs from state of the art work in mother-infant communication, e.g. [11]. As lower facial AUs do not involve lots of wrinkles and furrows, which need appearance-based features, we started with the analysis of the lower AUs. Table 3 shows the AUs we analyze.

Point	Description
A	Midpoint between two mouth corners [8]
O	Center point of nose [13]
J/J'	Right/Left Inner Brow point [13]
B/B'	Mean inner upper/lower lip point (*)

Table 2. Auxiliary points used in Table 1.

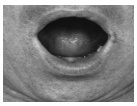
AU	Action	Description	
12	Oblique	Lip Corner Puller	
20	Horizontal	Lip Stretcher	
23	Orbital	Lip Tightener	
25	Up/ Down	Lips Part	
27	Up/ Down	Mouth Stretch	

Table 3. Action Units [5] with their description.

AU	F_1	Acc	Hit / FA
12	95.37 95.37	97.25 97.25	92.79/0.79 92.79/0.79
20	87.30 87.88	95.60 95.60	80.88/1.01 85.29/2.03
23	92.68 92.50	98.35 98.35	90.48/0.62 88.10/0.31
25	98.28 98.80	97.25 98.08	97.94/5.48 98.97/5.48
27	97.99 98.01	99.18 99.18	97.33/0.35 98.67/0.69

Table 5. Results for SVM, features selected by AdaBoost (first row), or OVL (second row). Results given by F_1 score, accuracy (Acc), Hit and False Alarm Rate (FA).

A binary classifier is constructed for each AU. Positive examples for a given AU are those samples that contain the specific AU, regardless of the occurrence of other AUs. Negative examples are the samples not containing that AU. Generalization to new subjects was tested using leave-one-subject-out cross validation, in which all sequences of the test subject are excluded from training. Performance was measured by the F_1 score. It considers both the precision p and the recall r to compute the score:

$$p = tp/(tp + fp)$$

$$r = tp/(tp + fn)$$

with tp the number of items correctly labeled as belonging to the positive class (true positives); fp the items incorrectly labeled as belonging to the positive class (false positives); and fn the items labeled as belonging to the negative class but which belong to the positive class (false negatives). The F_1 score can be interpreted as a weighted average of the precision and recall:

$$F_1 = 2*(p*r)/(p+r)$$

We experimented with different feature sets, and classifiers SVM, and AdaBoost. We combined feature selection with SVM classification. Table 4 gives an overview of the AUs with the number of positive and negative samples.

AU	12	20	23	25	27
P	111	68	42	291	75
N	253	296	322	73	289

Table 4. Number of positive (P) and negative (P) samples for each AU.

SVM Versus AdaBoost

We first started with two state of the art classification techniques in AU recognition. SVMs were trained and evaluated using LibSVM [2], while AdaBoost is constructed using OpenCV. For SVM, we tested kernels such as Linear, and Radial Basis Functions (RBF). We experimented with different variants of boosting algorithms, such as Discrete, Real, Gentle AdaBoost, and LogitBoost. The results show that SVM outperforms AdaBoost for almost every AU. SVM achieves an average F_1 score of 92.77%, while AdaBoost has an average of 90.73%. We tried all feature sets, and the best results were always the geometric features of Table 1, or a combination of those, and the displacements of facial feature points.

Feature Selection and SVM Classification

The features used in these experiments consist of estimations on the entire shape model. Feature selection is performed by AdaBoost, or according to the overlapping coefficient (OVL). AdaBoost is a well-known technique for feature selection, and is used in state of the art AU recognition. Because we wanted to know how discriminative the features for each AU are, and how well AdaBoost was able to select these discriminative features, we also tried out selecting features according to the OVL coefficient. OVL [3] is defined to be the area intersected by graphing two probability density functions, in our case two normal distributions. So, if $f_1(x)$ and $f_2(x)$ are two probability density functions defined on the n-dimensional real numbers R^n, then the OVL is defined as:

$$OVL = \int_{R^n} min(f_1(x), f_2(x))dx$$

In our case, $f_1(x)$ and $f_2(x)$ are the distributions of positive, respectively negative samples for each feature. The lower the OVL coefficient, the more a feature is discriminative. We evaluate a SVM for each AU with the first k features selected by AdaBoost, or according to the OVL coefficient. We chose the number of features for which the SVM has the best performance. Table 5 gives an overview of the best results. OVL slightly outperforms AdaBoost. OVL achieves on average a F_1 score of 94.51%, whereas AdaBoost has an average of 94.32%. Taking the best technique for each AU, we achieve an average F_1 score of

AU\Method	All feats	Low feats
12	95.37	95.41
20	87.88	88.89
23	92.68	93.83
25	98.80	98.63
27	98.01	97.37

Table 6. F_1 score for classification with facial features of the entire shape model (all feats) versus the lower facial features (Low Feats).

94.55%. In general, the most discriminative features are also selected by AdaBoost, though not in the order given by the OVL coefficient. AUs that have no real discriminative features, such as AU20 (with lowest OVL of 0.51) benefit the most of doing feature selection according to the OVL coefficient. AUs with low OVL coefficients, such as AU25 and AU27, achieve the best results.

The Effect of Upper Facial Features

In the previous experiments, upper facial features were always included in the feature selection with the best results, so we wanted to know why, and what the effect was of excluding the upper facial features. For this experiment, we restart AdaBoost with only the features of the lower face. The OVL coefficient is independently calculated for each feature; we take only the lower facial features. On average, we get a slightly better performance using only these features, i.e. 94.83%. Using AdaBoost as feature selection technique slightly outperforms OVL, i.e. on average, we obtain 94.59% for AdaBoost and 94.16% for OVL. Removing upper facial features from the set, makes AdaBoost select different (lower facial) features in another order than when all the features are present. Table 6 shows the F_1 score of the best results using all features versus features from the lower part of the face only. For all AUs, when all features are present, features from the eyes and brows are included in the selection process. For AU12, AU20, and AU23, these features are not highly discriminative. AdaBoost does a better selection when these features are not present. For AU25 and AU27, these features are moderately discriminative, thereby a slightly worse performance is obtained when removing them. The moderate discriminativity of these features could be due to the high percentage of sequences that only include AU25 or AU25+AU27, without action in the upper facial region.

CONCLUSION AND FURTHER RESEARCH

We achieve an average F_1 score of 94.83%. Performing feature selection before classification improves the results. AdaBoost also selects the most discriminative features, though not in the order given by the OVL coefficient. The feature selection always includes upper facial features. Using only lower facial features gives on average a better performance. In our further research, we will continue to look for the best feature set for each AU, combining geometry- and appearance-based features. In future work, we will extend the AU list to the entire face, investigate the dynamics of the AUs, and analyze the results and possible differences when the technique is applied on posed and spontaneous expressions.

REFERENCES

1. Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I., and Movellan, J. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In IEEE International Conference on Systems, Man & Cybernetics (2004), 592-597.
2. Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
3. Clemons, T.E. and Bradley Jr., E.L. A nonparametric measure of the overlapping coefficient. Computational Statistics &. Data Analysis. 34, 1 (2000), 51-61.
4. Cohn, J.F., Reed, L., Moriyama, T., Xiao, J., Schmidt, K., and Ambadar, Z. Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. In Proc. FG 2004, IEEE Computer Society (2004), 129-135.
5. Ekman, P., Friesen, W., and Hager, J.C. The Facial Action Coding System. Second Edition. Research Nexus eBook (2002).
6. Hou, Y., Sahli, H., Ravyse, I., Zhang, Y., and Zhao, R. Robust Shape-Based Head Tracking. In Proc. ACIVS 2007, Springer-Verlag (2007), 340-351.
7. Freund, Y., and Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55, 1 (1997), 119-139.
8. el Kaliouby, R. Mind-Reading Machines: Automated Inference of Complex Mental States. PhD thesis, University of Cambridge, Computer Laboratory (2005).
9. Kanade, T., Cohn, J.F., and Tian, Y. Comprehensive database for facial expression analysis. In Proc. FGR 2000, IEEE Computer Society (2000), 46-53.
10. Mahoor, M.H., Cadavid, S., Messinger, D.S., and Cohn, J.F. A framework for automated measurement of the intensity of non-posed facial action units. In Proc. CVPR4HB 2009, 12-18.
11. Messinger, D., Mahoor, M., Chow, S., and Cohn, J.F. Automated Measurement of Facial Expression in Infant-Mother Interaction: A Pilot Study, Infancy, 14, 3 (2009), 285-305.
12. Vapnik, V.N. The nature of statistical learning theory. Second Edition, Springer-Verlag (1999).
13. Zhang, Y.M. and Ji, Q.A. Active and dynamic information fusion for facial expression understanding from image sequences. PAMI, 27, 5 (2005), 699-714.