GREEDY SEARCH FOR DESCRIPTIVE SPATIAL FACE FEATURES

Caner Gacav[1], *Burak Benligiray*[1], *Cihan Topal*[1,2]

[1]Visea Innovative Information Technologies, Eskişehir, Turkey
[2]Anadolu University, Department of Electrical and Electronics Engineering, Eskişehir, Turkey

ABSTRACT

Facial expression recognition methods use a combination of geometric and appearance-based features. Spatial features are derived from displacements of facial landmarks, and carry geometric information. These features are either selected based on prior knowledge, or dimension-reduced from a large pool. In this study, we produce a large number of potential spatial features using two combinations of facial landmarks. Among these, we search for a descriptive subset of features using sequential forward selection. The chosen feature subset is used to classify facial expressions in the extended Cohn-Kanade dataset (CK+), and delivered 88.7% recognition accuracy without using any appearance-based features.

Index Terms— facial expression recognition, spatial features, sequential forward selection

1. INTRODUCTION

Facial expressions are important cues that support verbal communication. Analyzing individuals' psychological states and emotions by their facial expressions has become widespread in human behavior analysis and human–computer interaction studies [1, 2]. Automated computer vision methods that gather facial expression data allow these studies to be conducted more effortlessly [3, 4]. As the technology advances, vision systems will be able to sense subtle emotions and sentiments that humans cannot.

Geometric and appearance-based features are commonly used in facial expression recognition. In this study, we focus on spatial features, which are a type of geometric feature. Spatial features are calculated using the displacements of a combination of facial landmarks. Due to the high number of combinations, there are many potential spatial features, of which some are more descriptive. To provide the classifier with a descriptive subset of features with little redundancy, selection can be made based on prior knowledge [5, 6]. For example, Facial Action Coding System (FACS) defines a set of Action Units that produce expressions [1]. Another approach is to explicitly apply dimension reduction [7, 8] or let the classifier handle the selection [9]. As an alternative to these methods, we use a feature selection algorithm to form a descriptive feature subset.

Two combinations of 68 facial landmarks result in 2278 landmark pairs. We handle horizontal and vertical distance variations between these landmark pairs as separate features, thus work with 4556 potential features. Forward sequential feature selection reduces the number of features to 7. The resulting subset of features is used for classification in the extended Cohn-Kanade dataset (CK+) [10]. By using the selected spatial features, 88.7% recognition accuracy is obtained. This result surpasses other methods using only geometric features, and can be improved by utilizing appearance-based features.

2. RELATED WORK

Feature extraction and classification for facial expression recognition is a well-established problem in computer vision [11, 12]. Huang et al. use local binary patterns as appearance based features [7]. They build a canonical subspace of the subsequent frames, and model the lower-dimensional feature space using discriminative canonical correlation analysis.

Lucey et al. detect facial landmarks using active appearance models [10]. These landmarks are used to calculate similarity-normalized shape (SPTS) and canonical appearance (CAPP) features. Suk and Prabhakaran locate facial landmarks using an active shape model, and use displacements between landmarks located from neutral and expressive faces as features [6].

Chen et al. use appearance-based and geometric features [9]. Appearance-based features are represented by histogram of gradients (HOG) from three orthogonal planes. Geometric features are divided into two categories, namely rigid and non-rigid changes. Multiple kernel learning is used to find an optimal combination of these features. Turan and Lam extract features from the eye and mouth regions using local phase quantization and pyramid of HOG descriptors [13]. Features are fused using canonical correlation analysis and classified with SVM.

In our previous work, we used the variations in Euclidean distances between landmark pairs as spatial features [14], which gives slightly worse results than handling horizontal and vertical distances independently [10]. Leave-one-subject-out was used instead of 10-fold cross-validation, producing

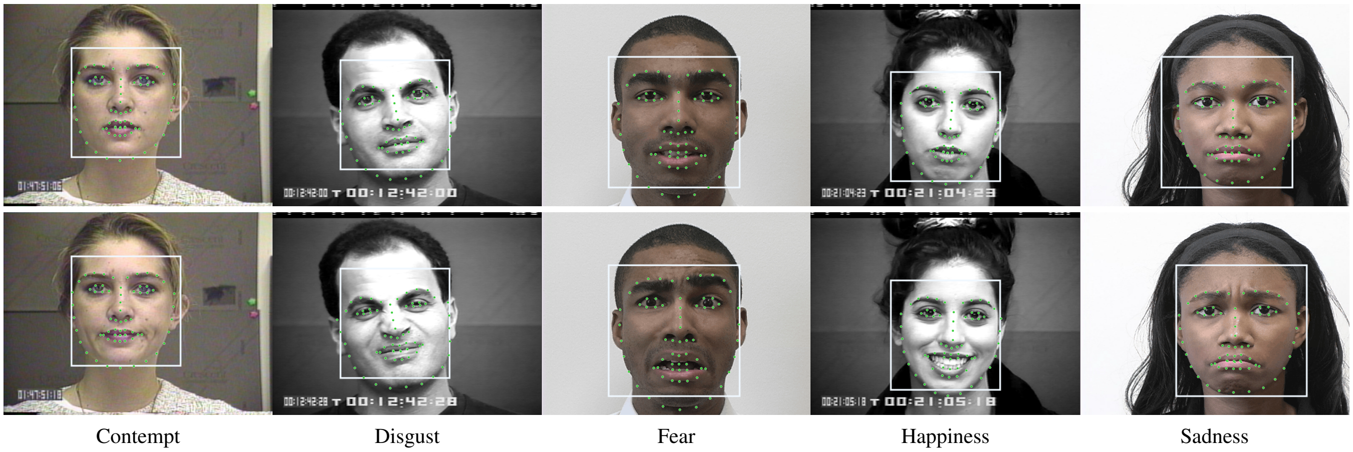


Fig. 1: Rectangles enclose the face regions, green markings indicate the facial landmarks. Landmarks move distinctively with different facial expressions.

optimistic results due to CK+ dataset providing small number of examples for some classes.

Deep learning methods have grown to be an important part of literature for all computer vision problems. However, the modest sizes of current datasets may be limiting their prevalence in facial expression recognition [10, 15]. Deep learners optimize feature design, feature selection and classification steps jointly. Liu et al. design a deep belief net that trains for these steps iteratively [16].

3. PROPOSED METHOD

We start by detecting the face region with the Viola and Jones algorithm [17]. Then, 68 facial landmarks are localized using Kazemi and Sullivan's method [18]. See Fig. 1 for face detections and located landmarks on neutral and expressive face images taken from the CK+ dataset. The distances between landmark pairs change as the subject expresses an emotion. We use the horizontal and vertical variations in these distances as features. A descriptive subset is chosen among these features using forward sequential selection, and used for classification by an SVM.

3.1. Extracting Spatial Features

In the feature extraction step, the horizontal and vertical distances between all landmark pairs are calculated. Using the relative displacements of landmarks provides robustness against translations between the neutral and the expressive face. By taking the difference between distance vectors obtained from a neutral and an expressive face, the displacement caused by the facial expression is captured [10]. This approach can also be interpreted as an implicit calibration using the neutral state of the face.

The CK+ dataset provides a set of consecutive frames where the subject gradually displays the intended expression. While extracting features, we only use the first and the last frames, which are fully neutral and fully expressive (see Fig. 1). For both images, 68 facial landmarks are located, which form 2278 different pairs. Since we handle the horizontal and vertical distances between the landmarks independently, a distance vector with the size of 4556 is obtained from each image in the pair. The difference of these two distance vectors results in a feature vector of the same size for each example. This large feature vector includes non-descriptive and redundant elements.

3.2. Sequential Forward Selection of Features

In Section 3.1, we extracted a large feature vector, composed of non-descriptive and redundant features, along with useful ones. To form a descriptive subset, we use sequential forward selection (SFS). This is a greedy search algorithm that iteratively selects the feature that improves the recognition accuracy the most. A feature's usefulness is defined by the improvement it provides to recognition accuracy when used with the previously selected features.

Before starting SFS, we randomly segment CK+ dataset as the training and test set. At the start of the k^{th} iteration of the algorithm, $k-1$ features are already selected. Features that are not among the selected are candidates. To test a candidate, it is grouped with the selected features, and the resulting vector is $L2$ normalized. The normalized feature vectors from the training set are used to train a multiclass SVM. This classifier uses the normalized vectors from the test set for recognition. The candidate whose addition improves the recognition accuracy the most is selected and the algorithm moves on to the next iteration. The algorithm stops when none of the candidates can improve recognition accuracy.

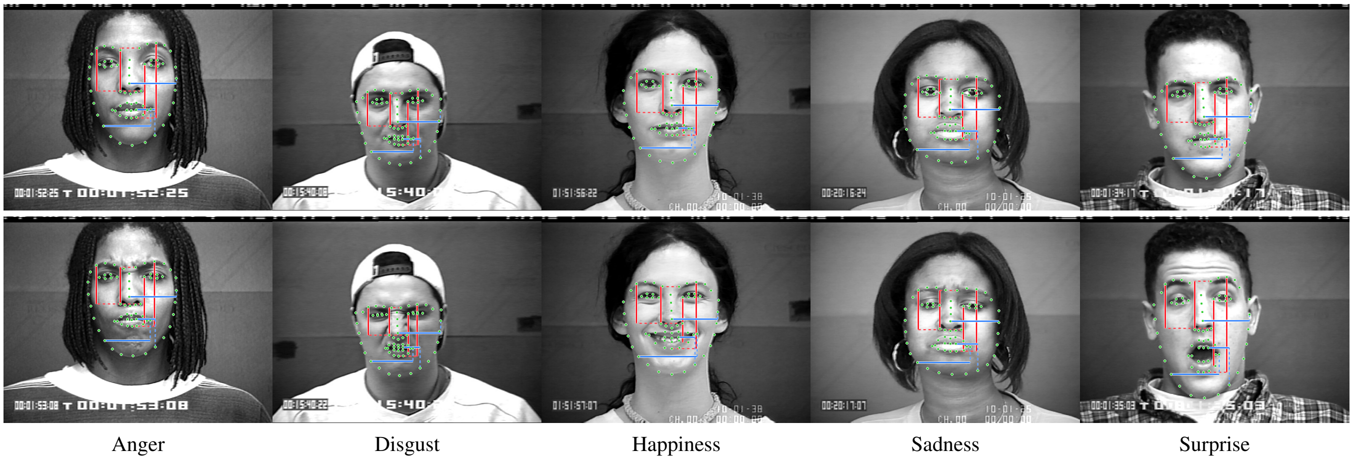


Fig. 2: The blue bars show the horizontal and the red bars show the vertical distances between two facial landmarks. Each bar starts at the first facial landmark, and its end is connected to the second facial landmark with a dashed line. The changes in the lengths of these bars are the features used for classification. Note that for each expression, at least one of the bars shorten or elongate distinctly.

Table 1: Number of examples for each facial expression in the CK+ dataset.

Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise	**Total**
45	19	59	25	69	28	82	**327**

4. EXPERIMENTAL RESULTS

We conducted our experiments on the extended Cohn-Kanade dataset [10]. The CK+ dataset contains images of faces with seven different facial expressions. These expressions are labeled as anger, contempt, disgust, fear, happiness, sadness and surprise. The dataset is composed of 327 image sequences gathered from 123 subjects. These image sets are consecutive frames that start at a neutral expression and end when the subject is expressing the respective emotion intensely. Number of examples for each expression is given in Table 1[1]. Since gathering definitive examples from classes such as contempt and fear is more difficult, examples from these classes are lower in number.

4.1. Feature Selection

To apply SFS as described in Section 3.2, the dataset is segmented into the training and test sets with 0.6 and 0.4 ratios, while protecting the class frequencies of the original dataset. The features chosen by SFS are illustrated in Fig. 2. The distances used as features are plotted as horizontal or vertical bars. Lengths of these bars change distinctly with the respective expression.

Although these features are chosen automatically, they are justifiable when inspected individually. It can be said that the movements of the mouth and eyebrows are particularly effective in recognizing facial expressions. Three of the features describe the movements of the eyebrows, while four describe movements of the mouth. The bottommost blue bar only describes the widening of the mouth, as its other end is anchored to a stationary part of the jaw (see Fig. 2, Happiness). The leftmost red bar is another feature that describes a single factor, the vertical eyebrow movement (see Fig. 2, Disgust and Surprise).

An unexpected feature is the horizontal distance variation between the ear and the nose. SFS chooses features even when they provide a very marginal increase in accuracy. Assuming this was the case, we tried eliminating this feature, which resulted in a 3% decrease in accuracy. Considering that additional features improve classification accuracy with diminishing returns, this difference is actually significant. We speculate that the role of this feature may be to sense head pose variation. Since the distance from a person's nose to ear cannot change, the feature extracted from this landmark pair will be non-zero only when the head pose changes. Head movement is limited to intense expressions such as anger and surprise, which may be the reason that this feature is descriptive.

[1] The Contempt example from Subject-129 is mislabeled as Surprised in the dataset. We used the corrected label.

Table 2: Confusion matrix of classification using the selected spatial features in the CK+ dataset.

	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Anger	0.78	0.04	0.09	0	0	0.09	0
Contempt	0.21	0.64	0	0.05	0.05	0.05	0
Disgust	0.05	0	0.93	0.02	0	0	0
Fear	0	0	0	0.80	0.04	0	0.16
Happiness	0	0	0	0.01	0.99	0	0
Sadness	0.11	0.11	0.03	0.03	0	0.64	0.08
Surprise	0	0	0	0	0	0	1

4.2. Classification Accuracy

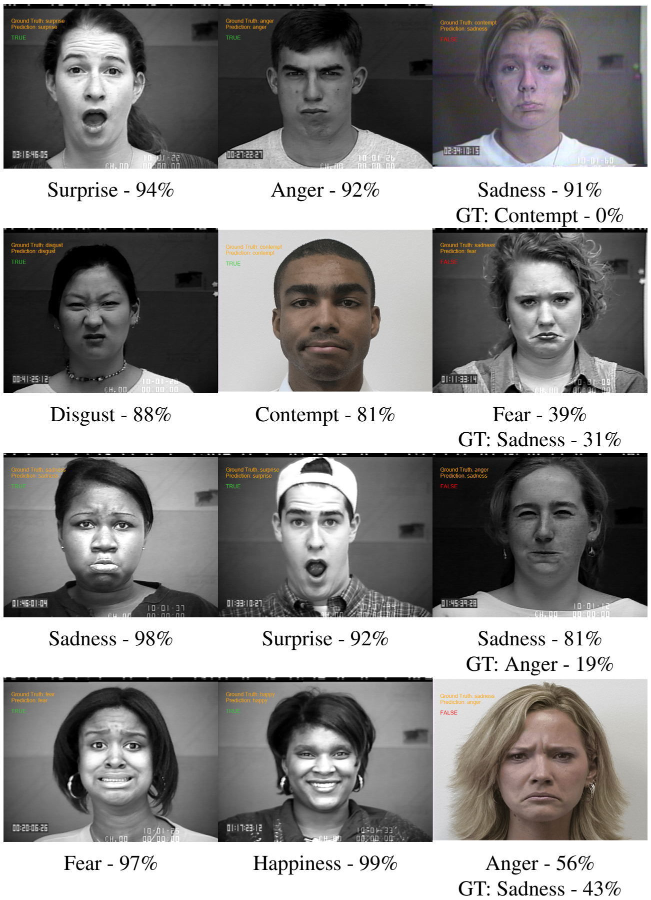
We tested the selected features for classification using 10-fold cross-validation. The feature vector is $L2$ normalized and a one-against-one multiclass SVM with RBF kernel is used as the classifier. See Table 2 for the confusion matrix. Accuracy is 88.7%, and the mean of accuracies for individual classes is 82.4%. See Fig. 3 for examples of classifications and respective posterior probabilities.

SPTS features are the horizontal and vertical displacements of individual facial landmarks after any similarity transformation is rectified [10]. To compute Geometry Features, facial landmarks are organized in a triangle mesh [9]. The edge lengths and angles of the triangles in the mesh change with facial expressions. These changes are used as a type of geometric feature. See Table 3 for a comparison of classification accuracies using different geometric features. The proposed geometric feature outperforms SPTS and Geometry Features for nearly all emotions.

Note that we have deliberately limited our comparison to geometric features. It is common practice in the literature to combine geometric features with appearance-based features to boost performance. The combination of SPTS and CAPP yield 88.4% classification accuracy [10], and the combination of Geometry Features and HOG from three orthogonal planes yield 93.6% classification accuracy [9].

Table 3: Classification accuracies using different geometric features in CK+.

	SPTS [10]	Geometry Features [9]	Proposed
Anger	0.35	**0.89**	0.78
Contempt	0.25	0.39	**0.64**
Disgust	0.68	0.90	**0.93**
Fear	0.22	0.36	**0.80**
Happiness	0.98	**0.99**	**0.99**
Sadness	0.04	**0.64**	**0.64**
Surprise	**1**	0.99	**1**
Total	0.665	0.847	**0.887**



Surprise - 94%

Anger - 92%

Sadness - 91%
GT: Contempt - 0%

Disgust - 88%

Contempt - 81%

Fear - 39%
GT: Sadness - 31%

Sadness - 98%

Surprise - 92%

Sadness - 81%
GT: Anger - 19%

Fear - 97%

Happiness - 99%

Anger - 56%
GT: Sadness - 43%

Fig. 3: The left two columns contain correct classifications, and the rightmost column contains incorrect classifications. Posterior probabilities are provided in percentages. For the incorrectly classified images, ground truth and its posterior probability is indicated in the second line.

5. CONCLUSION

Geometric and appearance-based features tend to capture different representations of facial expressions, hence work well together. Consequently, improvements for either of these feature types will be beneficial for facial expression recognition systems. In this study, we proposed geometric features derived from landmark pairs, including many non-descriptive and redundant ones. Instead of using this feature vector directly or applying a dimension reduction method, we used sequential forward selection to find a descriptive subset.

The selected spatial features yield 88.7% recognition accuracy and surpass other purely geometric features in the literature. To obtain better results, the selection can be done in an extended feature set, including many additional geometric and appearance-based features. A feature selection algorithm that searches for a larger part of the feature subset space is also expected to improve our results.

6. REFERENCES

[1] Paul Ekman and Erika L Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, Oxford University Press, 1997.

[2] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2003.

[3] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 1, pp. 39–58, 2009.

[4] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, H. Rao, J. Kim, L. Lo Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye, "Decoding children's social behavior," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 23, no. 2, pp. 97–115, 2001.

[6] Myunghoon Suk and Balakrishnan Prabhakaran, "Real-time mobile facial expression recognition system–a case study," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014.

[7] Xiaohua Huang, Guoying Zhao, Matti Pietikäinen, and Wenming Zheng, "Robust facial expression recognition using revised canonical correlation," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2014.

[8] Hazim Kemal Ekenel and Bülent Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1377–1388, 2004.

[9] Junkai Chen, Zenghai Chen, Zheru Chi, and Hong Fu, "Dynamic texture and geometry features for facial expression recognition in video," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2015.

[10] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.

[11] Maja Pantic and Leon J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, no. 12, pp. 1424–1445, 2000.

[12] Beat Fasel and Juergen Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[13] Cigdem Turan and Kin-Man Lam, "Region-based feature fusion for facial-expression recognition," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2014.

[14] C. Gacav, B. Benligiray, and C. Topal, "Sequential forward feature selection for facial expression recognition," in *Proc. Signal Processing and Communications Applications Conference (SIU)*, 2016.

[15] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba, "Coding facial expressions with gabor wavelet," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 1998.

[16] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] Paul Viola and Michael J Jones, "Robust real-time face detection," *International Journal of Computer Vision (IJCV)*, vol. 57, no. 2, pp. 137–154, 2004.

[18] Vahid Kazemi and Josephine Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.