Integrating Geometric and Textural Features for Facial Emotion Classification using SVM Frameworks

Samyak Datta[1], Debashis Sen[2], and R. Balasubramanian[1]

[1] Department of Computer Science and Engineering,
Indian Institute of Technology, Roorkee
[2] Department of Electronics and Electrical Communication Engineering,
Indian Institute of Technology, Kharagpur

Abstract. In this paper, we present a fast facial emotion classification system that relies on the concatenation of geometric and texture-based features. For classification, we propose to leverage the binary classification capabilities of a Support Vector Machine classifier to a hierarchical graph-based architecture that allows multi-class classification. We evaluate our classification results by calculating the emotion-wise classification accuracies and execution time of the hierarchical SVM classifier. A comparison between the overall accuracies of geometric, texture-based and concatenated features clearly indicates the performance enhancement achieved with concatenated features. Our experiments also demonstrate the effectiveness of our approach for developing efficient and robust real-time facial expression recognition frameworks.

Keywords: emotion classification, geometric features, textural features, local binary patterns, DAGSVMs.

1 Introduction

This paper attempts to address the problem of enabling computers to recognize emotions from facial expressions in a fast and efficient manner. Emotion recognition is a challenging problem due to the high degree of variability in the emotions expressed through human faces. Extracting a sub-set of facial features that best captures this variation has been a long-standing problem in the Computer-Vision community. A basic expression recognition framework is expected to involve modules for detecting faces, deciding on an appropriate subset of features to best represent the face which involves a trade-off between accuracy of representation and fast computation and finally, classification of the feature vector into a particular emotion category.

In this paper, we propose a framework for performing fast emotion classification from facial expressions. Two types of features are extracted for each facial image frame: geometric and texture-based. Angles formed by different facial landmark points have been selected as geometric features which is a novel and

speed optimized technique as compared to other expression recognition methods. Spatially enhanced, uniform pattern local binary pattern (LBP) histograms have been used as texture-based features. The hybrid feature vector for classification is then constructed by concatenating both the types of features. The concatenated features are able to capture both types of facial changes - high-level contortions of facial geometry (geometric features) and low-level changes in the face texture (texture-based features). In comparison with previous methods, our approach of using concatenated features results in enhanced performance. The classification module is based on Support Vector Machines. One of the novelties of the work lies in the use of hierarchical SVM architectures to leverage the binary classification of SVMs to multi-class classification problems. The use of hierachical SVMs results in much faster execution times than the traditional SVM-based multi-class classification approaches (such as one-vs-one SVMs) making the system suitable for real-time applications.

The remainder of the paper is structured as follows. Section 2 talks about the current state-of-the-art in the field. Section 3 discusses the proposed architecture of the work in detail where both the feature extraction and classification phases of the emotion recognition system are explained. Section 4 consists of a discussion regarding the results obtained as a consequence of this work and finally and in Section 5, we conclude by discussing the relevance of our work in enhancing the state-of-the-art.

2 Related Work

The state-of-the-art in emotion classification can be broadly divided into two categories: (a) geometric feature based or (b) texture feature based.

Pantic and Rothkrantz [1] used a rule-based classifier on frontal-facial points to achieve an accuracy of 86% on 25 subjects from the MMI database. Similar attempts by were made by Pantic and Patras in 2005 [2] where they tracked a set of 20 fiducial points and obtained an overall recognition rate of 90% on the CK-database. Cohen et. al. [3], 2003 extracted a vector of motion units using the PBVD tracker by measuring the displacement of facial points. More recently, Anwar et al. [7] in their 2014 paper use a set of 8 fiducial points to achieve the state-of-the-art classification rates.

The texture-based methods involve techniques such as Local Binary Patterns (L.B.P.) or applying some image filters to either the entire facial image (global) or some parts (local). Zhang et al. [6] use LBP along with Local Fisher Discriminant Analysis (LFDA) to achieve an overall recognition rate of 90.7%. Although Gabor filters are known to provide a very low error rate, but it is computationally expensive to convolve a face image with a set of Gabor filters to extract features at different scales and orientations.

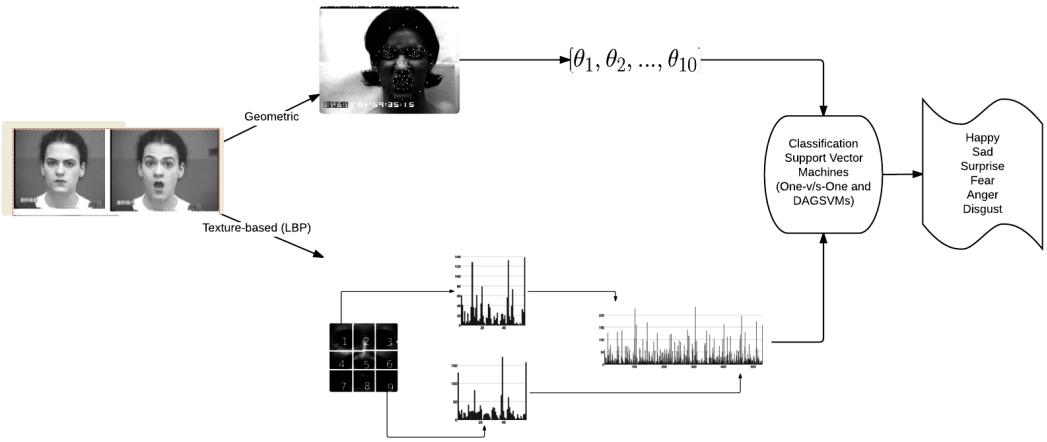

Fig. 1. A flowchart depicting the proposed architecture of our hybrid feature based facial emotion classification system.

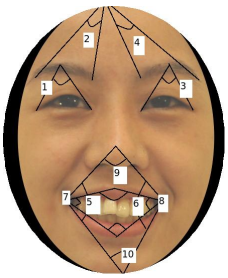

Fig. 2. A frontal face image depicting the facial angles (numbered from 1 to 10) used as geometric features. These angles are computed from the lines joining the 17 key facial feature points as detected by A.S.M. algorithm.

3 Proposed Architecture

In the proposed architecture, as shown in figure 1, both geometric and texture-based features have been used for classifying emotions. Two particular frames of interest from the extended Cohn-Kannade (CK+) database - neutral and peak expression have been selected for each subject. The calculation of geometric features involves using both the frames whereas texture-based features only make use of the peak-expression image frame.

3.1 Geometric features

Facial angles subtended by lines joining some key facial feature points have been selected as candidates for geometric features as shown in figure 2. Solely relying on facial angles as geometric features allows facial images of different sizes and

orientations to be treated in a similar manner and removes the need for intra-face normalization of features.

The face detector proposed by Viola and Jones based on Haar cascades [11] has been applied to both the frames followed by the Active Shapes Model (ASM) algorithm [9] to locate the 17 key facial points. Subsequently, 10 facial angles have been computed for each frame as shown in figure 2. The facial angles have been computed from the landmark points using the following set of basic co-ordinate geometry formulas.

Let $A(x_1, y_1).O(x_0, y_0)$ and $B(x_2, y_2)$ be three points in the two-dimensional Euclidean space. We define 2 vectors $\boldsymbol{OA} = (x_1 - x_0, y_1 - y_0)$ and $\boldsymbol{OB} = (x_2 - x_0, y_2 - y_0)$. The angle between $\boldsymbol{OA}$ and $\boldsymbol{OB}$ is given by:

$$\theta = cos^{-1} \frac{\boldsymbol{OA} \cdot \boldsymbol{OB}}{|\boldsymbol{OA}||\boldsymbol{OB}|} \tag{1}$$

The difference in the values (in degrees) in the corresponding facial angles between the neutral and peak expressions serves as the 10 geometric features for the face which help capture the high-level distortions in the facial geometry across emotion classes.

3.2 Texture-Based features

For texture-based feature extraction, local binary pattern histograms have been selected due to their simplicity, intuitiveness and computational efficiency. The $LBP_{8,1}$ operator has been used in our experiments which essentially computes the LBP code for each pixel (x_c, y_c) using the expression:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c).2^n \tag{2}$$

where i_k is the gray-scale intensity value of the pixel with co-ordinates (x_k, y_k) and $s(x)$ is 1 if $x \geq 0$ and 0 otherwise.

A variant of the traditional LBP operator that has been used as part of this work is the uniform pattern LBP operator denoted by $LBP^{u2}_{8,1}$. Uniform patterns are those binary patterns that have at most 2 bit transitions when the 8-bit LBP code is interpreted as a circular string. For example, 11000010 is not a uniform pattern whereas 11110000 is (3 and 2 bit transitions respectively). Using only uniform patterns (all non-uniform pattern LBP codes are assigned to a single bin) brings down the number of histogram bins from 256 to merely 59. After computing the $LBP^{u2}_{8,1}$ codes for each pixel, a histogram of the LBP values is constructed.

For computing texture-based features, all peak expression facial images are aligned and cropped to a uniform spatial resolution of 120×120 pixels and are then divided into 9 equal sized blocks of 40×40 pixels each. The local $LBP^{u2}_{8,1}$ histogram is computed for each sub-image and then concatenated into a global spatially-enhanced uniform pattern LBP histogram for the facial image. The feature vector thus formed is of size 59 x 9 = 531.

3.3 Concatenated features

The graph in Figure 3 shows the variation in values of the 10 geometric features for each of the 6 emotion classes. From a visual inspection, it is evident that the facial feature angles do a good job in differentiating between classes such as "Happy", "Surprise" and "Fear". However, "Disgust" and "Sad" are difficult to differentiate due to very low (almost non-existent) inter-class variance. The inability to completely capture variations between certain emotion classes arises due to the fact that these features only capture the high-level distortions in the facial geometry. Examples of such distortions would include the opening/closing of the mouth and widening of the eyes or curvature of lips. Hence, simply using geometric features is not sufficient to train a facial expression classifier with good discriminative powers.

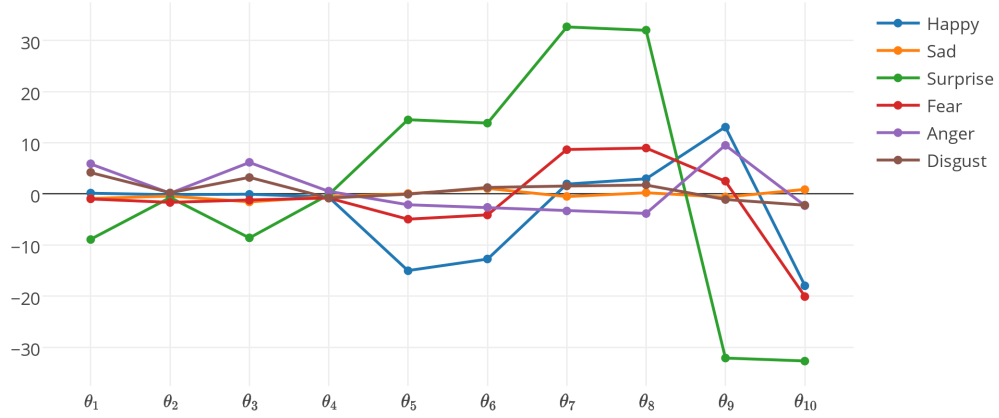

Fig. 3. Variation of geometric features across the 6 emotion classes

On the other hand, the L.B.P. histograms contain information about the distribution of micro-patterns such as edges, flat areas and wrinkles which represent some of the finer details of the face. To illustrate with an example, the histogram in Figure 4 shows a comparative analysis between the spatially concatenated LBP histograms of the "disgust" and "sad" emotion classes. As noted earlier, geometric features were not able to capture the inter-class variations between these two classes. However, through a simple visual inspection, we can

see that there are significant differences in the values of the histogram bins (e.g. peaks near the bin numbers 60, 80 and 120) between the 2 classes. These differences help impart discriminative powers to the classifiers trained using these features.

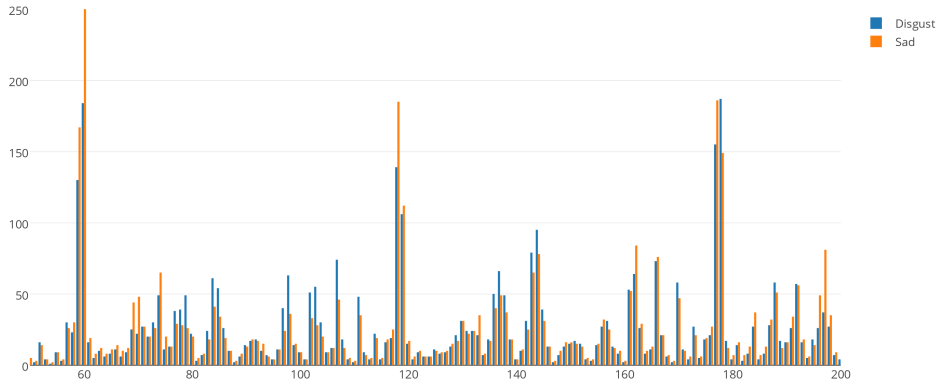

Fig. 4. Variation of LBP histogram features between "disgust" and "sad"

Further, dividing the facial image into sub-images and spatially concatenating their corresponding LBP histograms preserves the spatial texture-information while representing the global description of the face. The 531 uniform pattern $LBP^{u2}_{8,1}$ histogram features are concatenated with the 10 geometric facial feature angle features to form a combined, hybrid 541-dimensional feature vector. These hybrid vectors are used as a representative of the face for classification purposes.

3.4 Classification

Support Vector Machines (SVMs) are primarily binary classifiers that find the best separating hyperplane by maximizing the margins from the support vectors. Support vectors are defined as the data points that lie closest to the decision boundary. Some common SVM architectures such as one-vs-one, one-vs-rest and directed acyclic graph SVMs (DAGSVMs) [8] have been proposed to leverage the binary classification capabilities of an SVM classifier for multi-class classification problems.

In the one-vs-one scheme for an $n-$ class classification problem, $\binom{n}{2}$ binary SVM classifiers are trained corresponding to each pair of classes. The test point is put across all $\binom{n}{2}$ SVMs and the winning class is decided on the bass of

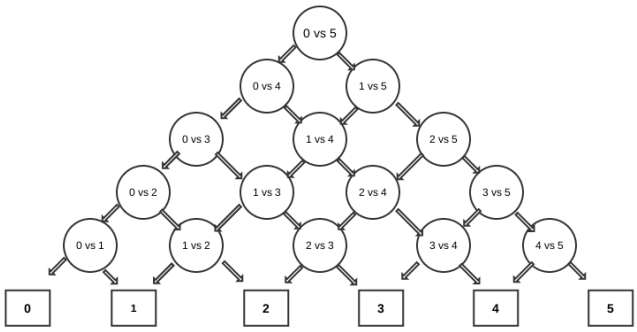

Fig. 5. Decision diagram for a 6-class DAGSVM.

a majority vote. On the other hand, in a hierarchical SVM architecture such as DAGSVM, $\binom{n}{2}$ SVMs are trained, but the test point only has to go across $(n-1)$ SVMs by traversing the directed graph as shown in figure 5. To evaluate the DAGSVM for a test point x, starting at the root node, a binary SVM is evaluated. Depending on the classification at this stage, the node is exited either via the left or the right edge. Then, the value of the binary SVM corresponding to the next node is evaluated until we reach one of the leaf nodes. The final classification is the class associated with the leaf node.

The statistical classification algorithm of SVMs is compared with an instance-based learning technique, the k-nearest neighbors (k-NN) classifier. The results are summarized in Tables 1 and 2. The motivation behind selection of the two classifiers lies in the fact that while SVMs construct an explicit model from the training data, instance-based techniques refrain from such generalizations. k-NNclassifies data points by comparing new problem instances with those seen in training.

Further, a comparison, in terms of both classification accuracy and execution times between two of the multi-class SVM frameworks: one-vs-one and DAGSVMs has been presented as part of this work. Since using DAGSVMs involves putting the test example through a lesser number $(n-1)$ of SVMs than one-vs-one SVMs ($\binom{n}{2}$), a drastic reduction in the execution time is expected.

4 Results and Discussions

The Cohn-Kannade extended (CK+) dataset consists of 593 image sequences (frames of a video shot) from 123 subjects, out of which 327 sequences are labelled as belonging to one of the 7 emotion categories: happy, sad, surprise, fear, anger, disgust and contempt. Labelled facial expression images (neutral and peak frames) for the six basic emotions - happy, sad, surprise, fear, anger and disgust have been used in the tests. All the results are 10-fold cross validated.

The classification accuracies for the various SVM architectures for geometric, texture-based and hybrid features are summarized in Table 1. For benchmarking

Table 1. A comparison of classification accuracies of different SVM architectures for different feature extraction techniques.

Architecture	Classification Accuracy (%)		
	Geometric	Texture (LBP)	Geometric + Texture
One-vs-One	78.15	88.52	91.85
DAGSVMs	76.67	86	89.26

Table 2. A comparison of classification accuracies of the k-NN algorithm for different values of k and feature extraction techniques.

k	Classification Accuracy (%)		
	Geometric	Texture (LBP)	Geometric + Texture
3	73.528	64.7	69.93
5	75	67.65	68.29
7	75.29	66.01	67
9	76.76	63.4	64.37

purposes, the classification accuracies for SVM-based classifers have been compared with those of the k-Nearest neighbor algorithm for different values of k. The results are reported in Table 2.

It is clear that irrespective of the SVM classifier architecture used, there is a significant enhancement in the overall classification accuracies with our approach of using hybrid features in place of simply using geometric or LBP features. For example, a one-vs-one SVM classifier gives an overall recognition rate of 78.15% and 88.52% with only geometric and LBP-based features respectively, which increases to 91.85% when using a hybrid feature set.

In the case of k-NN classifiers, there is a sharp decrease in the classification accuracies as we move from geometric to texture (LBP)-based or hybrid features to due to the increase in the dimensionality of feature vectors. The geometric feature vector has 10 attributes which increases to 531 and 541 in the case of texture and hybrid features respectively. This demonstrates the inability of the k-NN classifier to work well in high dimensional feature spaces. However, when the recognition rates of texture-based and hybrid features are compared, both of which are high dimensional, we see the hybrid feature vectors outperforming again as evident in Table 2 .

The confusion matrices for the 6-class classification problem using our approach of hybrid-features is shown in Table 3 for both one-vs-one and directed acyclic graph SVMs (DAGSVMs). In both cases, it is evident that “happy” and “surprise” are easiest whereas “sad” and “fear” are the most difficult to classify.

A comparison of the execution times of one-vs-one and DAGSVMs averaged over 30 test samples and 10-folds in figure 6 clearly shows the gain in computational efficiency in terms of time while using DAGSVMs. The reported execution

Table 3. Confusion matrix for SVM classification using hybrid features.

Predicted class

Actual Class	Happy	Sad	Surprise	Fear	Anger	Disgust
Happy	**98.53**	0	0	0	0	1.47
Sad	0	**58.82**	0	0	35.29	5.88
Surprise	0	0	**95.83**	1.43	0	2.86
Fear	6.25	6.25	6.25	**75**	6.25	0
Anger	0	5.71	0	0	**91.43**	2.86
Disgust	0	3.28	0	0	1.64	**95.08**

(a) One-vs-One SVMs

Predicted class

Actual Class	Happy	Sad	Surprise	Fear	Anger	Disgust
Happy	**100**	0	0	0	0	0
Sad	0	**50**	15	0	35	0
Surprise	1.43	0	**94.29**	1.43	0	2.86
Fear	10.53	0	5.26	**78.94**	5.26	0
Anger	0	7.89	0	0	**86.84**	5.26
Disgust	0	1.72	1.72	0	6.89	**89.65**

(b) DAGSVMs

times are on an Intel Core(TM) i3-2330M CPU with a clock speed of 2.20GHz. Since the overall classification accuracy for DAGSVMs (89.26%) is not significantly less than one-vs-one SVM classifiers (91.85%), the reduction in time may render DAGSVM as a suitable candidate for real-time emotion classification problems.

5 Conclusion

In this paper, we have presented a framework for fast emotion classification from facial expression.

Our experimental results show an enhanced performance when using a concatenated feature vector which is a combination of geometrical and texture-based LBP features. We also present a comparative analysis of two major multi-class, SVM-based architectures for classification, namely one-vs-one and DAGSVMs (hierarchical SVMs). Our results indicate that both the systems give almost equal performance. However, using hierarchical multi-class SVM architectures leads to increased efficiency in terms of computation time.

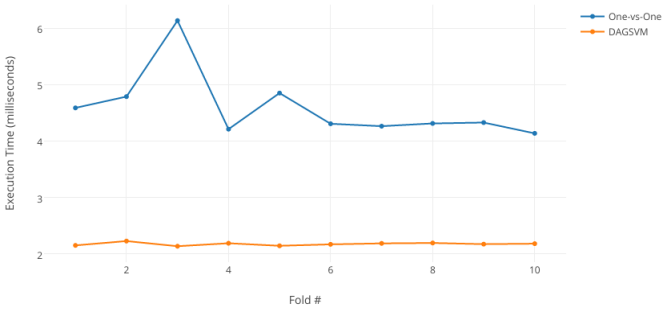

Fig. 6. Comparison between the total execution time of one-vs-one (blue) and DAGSVMs (orange) for the classification of 30 test points.

References

1. M. Pantic and J.M. Rothkrantz, *Facial Action Recognition for Facial Expression Analysis from Static Face Images*, IEEE Trans. Systems, Man arid Cybernetics Part B, vol. 34, no. 3, pp. 1449-1461, 2004.
2. M. Pantic, I. Patras, *Detecting Facial Actions and their Temporal Segments in Nearly Frontal-view Face Image Sequences*, Proc. IEEE conf. Systems, Man and Cybernetics, vol. 4, pp. 3358-3363, Oct 2005.
3. I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang, *Facial Expression Recognition From Video Sequences: Temporal and Static Modeling*, Computer Vision and Image Understanding, vol. 91, pp. 160-187, 2003.
4. W. Zheng, X. Zhou, C. Zou, and L. Zhao, *Facial Expression Recognition Using Kernel Canonical Correlation Analysis (KCCA)*, IEEE Trans. Neural Networks, vol. 17, no. 1, pp. 233238, Jan 2006.
5. I. Kotsia, I. Buciu and I. Pitas, *An Analysis of Facial Expression Recognition under Partial Facial Image Occlusion*, "Image and Vision Computing, vol. 26, no. 7, pp. 1052-1067", Jul 2008.
6. S. Zhang, X. Zhao and B. Lei, *Facial Expression Recognition Based on Local Binary Patterns and Local Fisher Discriminant Analysis*, WSEAS Transactions on Signal Processing, issue 1, vol. 8, pp.21-31, Jan 2012.
7. A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, *Frame-Based Facial Expression Recognition Using Geometrical Features*, Advances in Human-Computer Interaction, vol. 2014, April 2014.
8. J.C. Platt, N. Cristianini and J. S. Taylor, *Large Margin DAGs for Multiclass Classification*, Advances in Neural Information Processing Systems (NIPS), 1999.
9. S. Milborrow and F. Nicolls, *Active Shape Models with SIFT Descriptors and MARS*, International Conference on Computer Vision Theory and Applications (VISAPP), 2014.
10. P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and Matthews, *The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression*, IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), 2010.
11. P. Viola, M. J. Jones, *Robust Real-Time Face Detection*, International Journal of Computer Vision, 2004.