21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Facial Emotion Recognition in Continuous Video

Albert Cruz, Bir Bhanu and Ninad Thakoor
Center for Research in Intelligent Systems, University of California, Riverside, CA, USA
{*acruz, bhanu, nthakoor*}*@ee.ucr.edu*

Abstract

Facial emotion recognition–the detection of emotion states from video of facial expressions–has applications in video games, medicine, and affective computing. While there have been many advances, an approach has yet to be revealed that performs well on the non-trivial Audio/Visual Emotion Challenge 2011 data set. A majority of approaches still employ single frame classification, or temporally aggregate features. We assert that in unconstrained emotion video, a better classification strategy should model the change in features, versus simply combining them. We compute a derivative of features with histogram differencing and derivative of Gaussians and model the changes with a hidden Markov model. We are the first to incorporate temporal information in terms of derivatives. The efficacy of the approach is tested on the non-trivial AVEC2011 data set and increases classification rates on the data by as much as 13%.

1. Introduction

Facial emotion recognition has applications in medicine (treatment of Asperger's), video games (Xbox Kinect), human-computer interaction (intelligent tutoring systems) and affective computing (embodied agents). A recognition system must detect a subject's underlying emotional state from video of apparent facial expressions.

State of the art approaches use the recently available Facial Emotion Recognition and Analysis Challange 2011 (FERA2011) and Audio/Visual Emotion Challenge 2011 (AVEC2011) data sets [4, 5], where state-of-the-art approaches perform poorly and the videos of expressions are spontaneous and natural. These data sets fall into two categories: (1) videos are pre-segmented so that each video captures only one emotion (as is the case in CK+, MMI-DB and FERA2011); or (2) videos are continuous, spontaneous and unsegmented. This second case is unique to AVEC2011. Videos are interviews, where an embodied agent, the Sensitive Artificial Listener causes emotional reactions. An example is available on the internet [2].

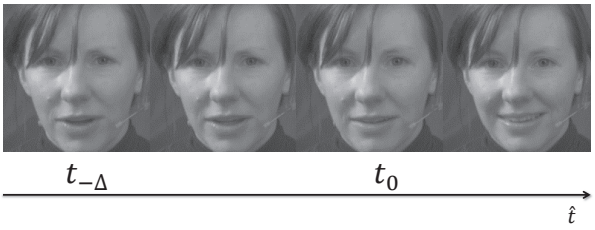

Figure 1. Ambiguity in transitions.

1.1 Motivation

Because the subject can transition freely between emotions within the same video, paradigms from the previous challenges do not hold. In most approaches, features are extracted from the current time t_0, and classification is carried out using this single time point. In Schuller et al. [4], facial features (Uniform Local Binary Patterns and pose tracked with Active Appearance Models) and audio features are fused and classified with a SVM, on a per-frame basis. In Fig. 1, at t_0, the emotion of the subject is ambiguous; either the subject is concerned or happy, and classification using this frame alone would be unsatisfactory. Aggregating appearance features about t_0 would capture previously arched eyebrows, and the smile in the following frame. These are typical expressions associated with either emotion. We propose classifying emotion considering the change in facial features between t_0 and $t_{-\Delta}$, or the temporal derivative. Comparing t_0 and $t_{-\Delta}$, the eyebrows lack an arch, and the lip corners have returned from a downward curl. It can be stated with confidence that she is no longer concerned.

This information needs to be combined with the feature information. Typically this is accomplished with fusion, another paradigm. In these approaches, a fusion approach aggregates results with an averaging process

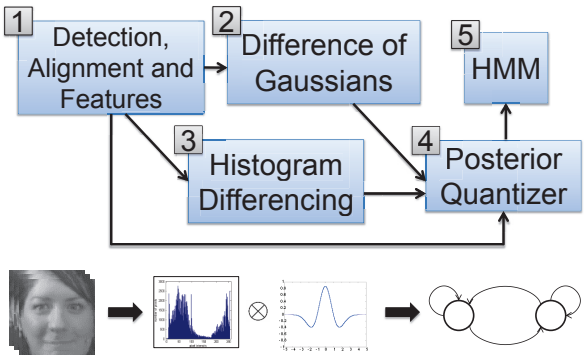

Figure 2. System overview.

to reduce the impact of frames that are labeled differently from their temporal neighbors. Valstar et al. [5] uses a voting approach. Glodek et al. [1] employs feature fusion. The features are not heterogeneous; they perform well in different, exclusive situations. Feature derivatives are useful when emotion is weak or transitioning, while the original features have enough information when emotion is strong. The classifier must be aware of when certain features are performing well, and when some are not–it must be aware of co-occurrences–so it must hold a state.

1.1.1. Contributions. Our contributions to state-of-the-art facial emotion recognition are: (1) to the best of our knowledge, we are the first to use feature derivatives, and (2) we propose a novel algorithm that estimates the derivative of features with respect to time and fuses the information from features and feature derivatives with a hidden Markov model for facial emotion recognition.

2. Technical Approach

The proposed system overview is shown in 2: (1) Face ROI is detected with a boosted cascade of Haar-like features, aligned with SIFT Flow to a reference image, and Local Phase Quantization (LPQ) texture features are extracted. The derivative of features are estimated with two methods: (2) with a fine spatial granularity with Difference of Gaussians (DoG) and (3) with a course spatial granularity with histogram differencing (HD) of LPQ histograms. (4) A support vector machine (SVM) outputs posterior probabilities for emotion labels from each of the three feature vectors and yhe posterior probabilities are quantized into a single observation vector. (5) A hidden Markov model computes the optimal emotion labels, taking of advantage of the co-occurances between $x(t)$ and $x'(t)$. where T is the length of the video. $\mathbf{U}$ has m^i observable symbols.

2.1. Detection, Alignment and Features

Faces are extracted with a boosted cascade of Haar-like features. After extraction, faces are aligned with SIFT Flow to the Avatar Reference Image [6]. The parameters of this algorithm are the number of iterations I. After alignment, Local Phase Quantization (LPQ) histograms are extracted [3] in each of $n \times n$ local regions and the histograms are concatenated to form the feature vector (this process is also called cells, or gridding).

2.2. Modeling Temporal Changes

The derivative of features $x'(t)$ is approximated by two methods: convolution with a DoG filter and difference of feature histograms. DoG has a fine spatial granularity, in that it captures local changes happening at the pixel. Histogram differencing has a course spatial granularity, in that captures global changes happening between the histograms of each cell.

2.1.1. Local Derivatives with DoG. A DoG filter is employed as opposed to a finite difference because the finite difference is sensitive to noise. The i-th feature $\langle x(t) \rangle_i$ is convolved with the DoG filter to approximate the gradient of $x(t)$ with the following equation:

$$\langle x'_{\mathrm{DoG}}(t) \rangle_i \approx \langle x(t) \rangle_i \otimes h(t) \tag{1}$$

where $h(t) \sim N(0, \sigma_1) - N(0, \sigma_2)$ and $\sigma_1 = 4\sigma_2$. The effect of Eq. 1 is a 1-D temporal gradient of $\langle x(t) \rangle_i$ that has been low-pass filtered to remove noise. $h(t)$ is discretized to $2l$, where $3\sigma_2 = l$, retaining approx. 99% of the energy of the larger Gaussian.

2.1.2. Global Derivatives with HD. Let the feature vector $x(t)$ be composed of a set of n^2 histograms $\{H_1(t), H_2(t), ..., H_{n^2}(t)\}$. The histogram difference is computed with the l_1 metric, for each histogram, between $t - \delta$ and $t + \delta$. This is similar to shot transition detection for key frames, except the histogram of features is used, as opposed to color histograms. A new feature vector $x'_{\mathrm{HD}}(t) \epsilon \Re^{1 \times n^2}$ is generated where:

$$\langle x'_{\mathrm{HD}}(t) \rangle_i = \| H_i(t - \delta) - H_i(t + \delta) \|_1 \tag{2}$$

where $H_i(t)$ is the ith histogram at time t, and δ is a spacing parameter.

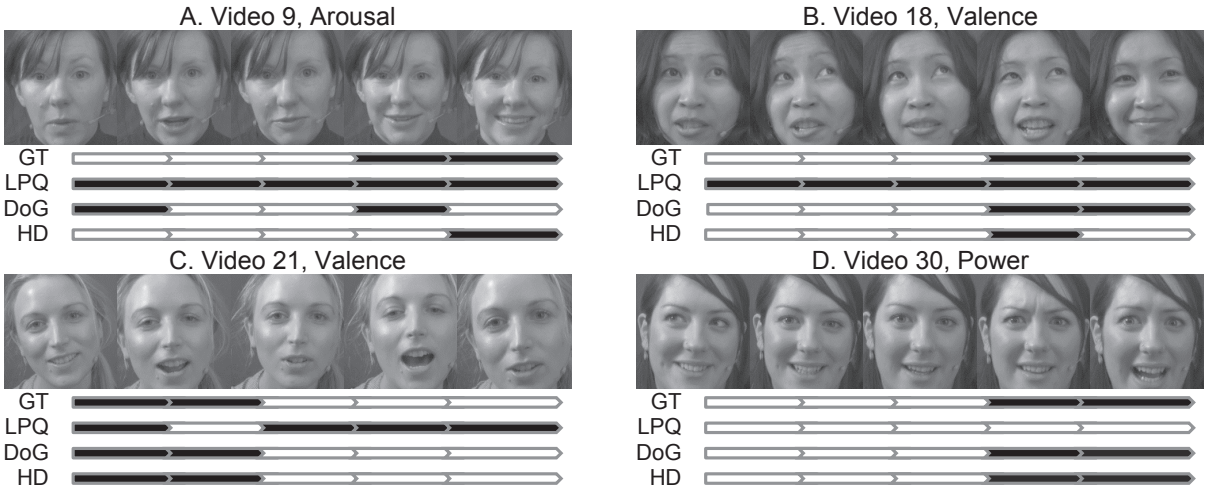

Figure 3. Examples of a co-occurrences.

2.1.3. Observation Quantization. A linear SVM is trained to output posterior probabilities. Let $\tilde{w}_i(t)$ be the estimated label at time t from a matcher using the i-th feature set. We hypothesize that a co-occurrence exists, where the feature derivatives perform better when emotion is weak or transitioning, e.g. $\tilde{w}_{DoG}(t)$ would properly classify t_0 in Fig. 1, and $\tilde{w}_{LPQ}(t)$ would not. The output of the SVM must be fused in such a way as to capture the combination of outputs of the SVM. First, the posterior probabilities across all videos for each matcher are quantized into m bins with k-means clustering. Let $v_i(t)$ be the set of membership of $\tilde{w}_i(t)$ at time t, ranging from 0 to $m-1$. Second, the quantized probabilities $v_i(t)$ of each matcher are combined into a single observation matrix. Let $u(t)$ be the combined, quantized observation at time t:

$$u(t) = \sum_{i=1}^{n} m^{(i-1)} v_i(t) \tag{3}$$

where n is the number of different matchers. Let $\mathbf{U}$ be the observation sequence defined as:

$$\mathbf{U} = \{u(t) : 0 < t \leq T\} \tag{4}$$

2.1.3. Hidden Markov Model. Co-occurance aware fusion is realized with a Hidden Markov model (HMM). We formulate out HMM as follows: given the observation sequence $\mathbf{U}$, and the HMM, am optimal corresponding state sequence $\mathbf{Y} = y(0)\, y(1) ... y(T)$ must be chosen. $\mathbf{Y}$ is taken to be the estimated labels; the number of states of the model are equal to the number of classes p. The state transition probability distribution matrix A and observation probability distribution matrix B are estimated from training data. We assign labels with:

$$\mathbf{Y} = \text{argmax}_{y(0)...y(T)} p(y(0) ... y(T), ... \mathbf{U} \,|\lambda(A, B)) \tag{5}$$

where λ is the model. Eq. 5 is solved with dynamic programming, with the Viterbi algorithm. Because the joint probabilities of the matchers are estimated, the model can fuse information from each matcher in a more meaningful way, as opposed to simply aggregating the labels.

	Act.	Pow.	Unpred.	Val.
LPQ+DOG	0.373	-0.274	-0.345	-0.407
LPQ+HD	0.732	-0.373	-0.543	-0.224

Table 1. Q-statistics for each feature set.

3. Experimental Results and Discussion

The training and development sets of the AVEC2011 dataset [4] consists 63 videos of 13 different individuals, where frontal face videos are taken during an interview when a subject is engaged in conversation by an embodied agent. We divide the set of videos into four roughly equal, mutually exclusive sets for cross-validation (one set is used to testing; the rest for training). The frames are originally $\sim$ 50fps resulting in 868999 frames. We subsample by a factor of 16. AVEC2011 quantizes emotion along four dimensions: arousal, power, valence and expectancy. An emotion consists of a real value along each of these four emotions. AVEC2011 quantizes each dimension into bins of high and low, therefore $p = 2$.

(%)		Arousal	Expectancy	Power	Valence	Average
LPQ [6]		65.0	56.4	56.7	62.1	60.1
DoG (l)	3	63.1	56.3	56.0	58.3	58.4
	5	64.3	56.5	56.1	58.2	58.8
	7	64.2	55.5	56.3	57.4	58.4
HD (δ)	3	64.5	57.1	60.8	67.2	62.4
	5	64.3	58.3	61.4	60.1	61.0
	7	65.0	58.7	60.9	60.0	61.2
Proposed		64.8	69.5	65.8	74.2	67.7
A priori		51.1	42.6	55.0	59.2	-
[4]		60.2	56.0	58.3	63.6	59.5
[1]		58.2	53.7	53.7	53.2	54.7

Table 2. Results on AVEC2011 data.

For Avatar Image Registration $I = 3$. For LPQ features, $n = 8$ with a dimensionality of 16384. δ, the histogram differencing spacing, and l, the half-width of the DoG filter, are selected experimentally from $\{3, 5, 7\}$ to give the best classification rate, see Tab. 2 ($l = 5$ and $\delta = 3$). For the quantizer, $m = 2$. The proposed approach uses the parameters that give the best classification rate.

In Sec. 1, we hypothesized that a cooccurance existed where LPQ and derivative of LPQ features perform well in different, exclusive situations and that the HMM could take advantage this co-occurrence. In Fig. 3 we provide examples of co-occurrences, where LPQ features perform poorly, but feature derivatives were able to properly classify emotion during the transition. In the figure, "GT" stands for ground truth, for each feature set, the class with the favored posterior is shown in white for low, and black for high. We verify that such a co-occurrence has an impact on classification with a Q-statistic:

$$Q_{ij} = \frac{(n_{00}n_{11} - n_{01}n_{10})}{(n_{00}n_{11} + n_{01}n_{10})} \quad (6)$$

where Q_{ij} is the Q-statistic metric that measures how disparate the performance is between i and j; n_{00}, the number of samples where i and j misclassified the same sample; n_{11}, the number of samples where i and j correctly classified the same sample; n_{01}, the number where i misclassified and j correctly classified; n_{10}, the number where j misclassified and i correctly classified. Note that $Q_{ij} \epsilon [-1, 1]$. Two disparate feature sets that have coocurances where one performed better than the other should have a low Q_{ij}. The Q-statistics are given in Tab. 1. The Q-statistics are low in most cases, verifying the hypothesis of cases of disparate performance.

In Tab. 2, *a priori* indicates the *a priori* rate of the classes. The baseline approach [4] and Glodek et al. [1] are included for comparison (a ranking approach from AVEC2011). The proposed method improves state of the art by 7.6% over the top performer in the FERA2011 emotion challenge [6].

4. Conclusion

In this paper we verified that, for facial emotion recognition where a subject can freely express emotions, a derivative of features was more suitable than using the features themselves. The derivative was estimated with histogram differencing and DoG features. A HMM fused the output of SVM matchers. Co-occurrences were demonstrated to exist between feature derivatives and features, and their impact on classification was positively demonstrated. The proposed approach increased classification results on the non-trivial AVEC2011 data set.

Acknowledgements. Support for this work was provided for in part by NSF grants 0727129 and NSF IGERT: Video Bioinformatics Grant DGE 0903667. The contents and information do not reflect the position or policy of the U.S. Government.

References

[1] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker. Multiple classifier systems for the classification of audio-visual emotional states. In *Proc. Affective Computing and Intelligent Interaction Workshop on AVEC*, 2011.

[2] G. McKeown. Chatting with a virtual agent: The semaine project character spike. Website, February 2011. http://www.youtube.com/watch?v=6KZc6e_EuCg.

[3] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Proc. Image and Signal Processing*, 2008.

[4] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011: The first international audio/visual emotion challenge. In *Proc. Affective Computing and Intelligent Interaction Workshop on AVEC*, 2011.

[5] M. F. Valstar, B. Jiang, M. Méhu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Proc. IEEE AFGR Workshop on Facial Expression Recognition and Analysis Challenge*, 2011.

[6] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In *Proc. IEEE AFGR Workshop on Facial Expression Recognition and Analysis Challenge*, 2011.