FACIAL EMOTION RECOGNITION WITH ANISOTROPIC INHIBITED GABOR ENERGY HISTOGRAMS

Albert Cruz, Bir Bhanu and Ninad S. Thakoor

Center for Research in Intelligent Systems, University of California, Riverside
Riverside, CA 92521-0425, USA

ABSTRACT

State-of-the-art approaches have yet to deliver a feature representation for facial emotion recognition that can be applied to non-trivial unconstrained, continuous video data sets. Initially, research advanced with the use of Gabor energy filters. However, in recent work more attention has been given to other features. Gabor energy filters lack generalization needed in unconstrained situations. Additionally, they result in an undesirably high feature vector dimensionality. Non-trivial data sets have millions of samples; feature vectors must be as low dimensional as possible. We propose a novel texture feature based on Gabor energy filters that offers generalization with a background texture suppression component and is as compact as possible due to a maximal response representation and local histograms. We improve performance on the non-trivial Audio/Visual Emotion Challenge 2012 grandchallenge data set.

Index Terms— Feature extraction, image texture analysis, facial emotion recognition, anisotropic inhibition, Gabor energy filter

1. INTRODUCTION

Expression/emotion recognition and analysis has numerous applications such as human-computer interaction, video games, medicine, and lie detection [1, 2]. A particular interest is given to facial expression and emotion detection because facial expressions are critically imporant in non-verbal communication [3, 4]. In video-based facial emotion recognition, face video of a human is captured. Computer algorithms must detect their facial expressions and infer their underlying emotional state. Gabor energy filters were critically important in early automatic facial emotion recognition approaches. Among the first papers utilizing these features for facial emotion/expression was Lyon and Akamatsu [5].

1.1. Motivation and Related Work

In recent emotion recognition grand challenges, Gabor filters features have fallen out of favor. Approaches using appearance features have given more attention to Local Phase Quantization (LPQ) [6] features, Local Binary Patterns (LBP) features [7] and their derivatives. Some of the reasons for this are:

(1) *Lack of generalization*. Gabor energy filters do not generalize well in unconstrained settings. State-of-the-art local appearance features have computational steps in an effort to be more generalizable. Uniform LBP micro-textures are rotation invariant, and robust to monotonic grayscale transformations. LPQ features are robust to blur. A Gabor energy filter simply captures edge magnitudes/orientations.

(2) *Compounds dimensionality problem*. Gabor energy filters produce a response for each filter in its filter bank. For example, a Gabor energy response at 8 orientations, 4 scales, and a square image of 150×150 results in a dimensionality of 7.2×10^5. This is very large compared to Uniform LBP's dimensionality of 5900.

The Gabor filter is an important feature in all aspects of image processing and computer vision, and, though it suffers from the drawbacks listed above, we assert that it can still be effectively applied to facial emotion with the proposed modifications in this paper. Gabor energy filtering is an important feature in all aspects of computer vision, and, though it suffers from the drawbacks listed above, we assert that it can still be effectively applied to facial emotion. A table of related work is given in Table 1.

1.2. Contribution

We develop a novel procedure that improves the Gabor energy filter. It exploits anisotropic texture inhibition for facial emotion recognition for the first time. It generalizes well because of its ability to supress background texture. It has a low feature vector dimensionality because of local histograms. We demonstrate its efficacy on the non-trivial AVEC2012 grandchallenge data set [12].

2. TECHNICAL APPROACH

The system overview is shown in Figure 1: For each frame, (1) face region-of-interest is detect with a cascade of Haar-like features [13] and are aligned with Avatar Image Alignment [9]; (2) aligned face ROI is filtered a bank of Gabor filters and their Gabor energy is computed; (3) anisotropic texture inhibition retains edges that are not consistent with background texture; a compact representation is achieved by first (4) obtaining the maximal response from each filter orientation and (5) computing a local histogram in different parts

Table 1. Related Work and Their Local Appearance Feature Approach

Feature	Reference	Notes
DCT	Ma and Khorasani [8]	Discrete cosine transform used as feature for classification.
LBP	Yang and Bhanu [9]	Uniform Local Binary Patterns. Top approach in FERA2011 grand-challenge [4] for the frame-level discrete emotion challenge.
FPLBP and TPLBP	Wolf *et al.* [10]	Four-Patch and Three-Patch Local Binary Patterns. Binary patterns of intensity differences, higher order than LBP. Highly ranked approach for Labeled Faces in the Wild [11] grand-challenge.

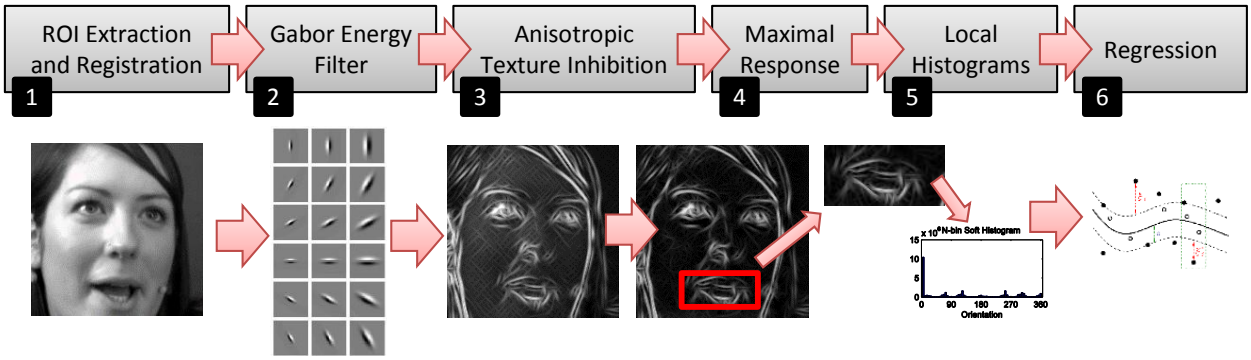

Fig. 1. System overview.

of the face; finally, (6) ϵ Support Vector Regression estimates emotion labels.

2.1. Gabor Energy Filter

Let f be a face image. Contours of the face are measured with a Gabor filter:

$$g_{\theta,\phi}(x,y) = e^{-\frac{\tilde{x}^2+\gamma^2\tilde{y}^2}{2\sigma^2}} \cos\left(2\pi\frac{\tilde{x}}{\lambda}+\phi\right) \quad (1)$$

where $\{x,y\}$ is the pixel location $\tilde{x} = x\cos\theta + y\sin\theta$, $\tilde{y} = -x\sin\theta + y\cos\theta$, γ is the spatial aspect ratio, σ is the standard deviation of the Gaussian, λ is the wavelength, θ is the angle parameter where $\theta \in [0,\pi)$ and ϕ is the phase offset taken to be 0 and π. The Gabor energy of f is:

$$E_\theta(x,y) = (f * g_{\theta,0})^2(x,y) + (f * g_{\theta,\pi})^2(x,y) \quad (2)$$

2.2. Generalization Step

We want to extract the edges of the face belonging to strong contours, e.g. the outline of the mouth. However, not all contours on the face are significant to facial expressions, e.g. contours appearing in hair texture. Grigorescu *et al.* [14] demonstrated that anisotropic texture inhibition reduced the effect of background texture. Background texture that is not consistent with the object edges is quantified as:

$$t_\theta(x,y) = (E_\theta * w)(x,y) \quad (3)$$

where w is a weight function:

$$w(x,y) = \frac{1}{\|g(\mathrm{DoG})\|_1} g(\mathrm{DoG}(x,y)) \quad (4)$$

Table 2. Breakdown of Videos Used in Testing

Fold	Videos Used
1	4, 5, 7, 11, 12, 17, 18, 20, 25, 29, 32
2	1, 2, 7, 9, 11, 18, 21, 23, 24, 30, 31
3	6, 12, 13, 14, 18, 20, 22, 26, 28, 29, 30

where DoG is a Difference of Gaussians and $g(z) = H(z) * z$, where H is a Heaviside step function. We remove background texture from the Gabor energy of f by subtracting the estimated background texture:

$$\tilde{b}_\theta(x,y) = g(E_\theta(x,y) - \alpha t_\theta(x,y)) \quad (5)$$

where α is a weight that effects how much background texture is removed. It is taken to be 1 for full suppression [14]. When $\alpha = 0$, there is no texture suppression and it is the Gabor energy. An example of the effect of α and the result of anisotropic inhibition is given in Figure 2. Note that the Gabor energy succeeds in detecting the important contours of the faces (the jaw line, eyebrows, mouth, eyes, etc.). However, it also detects contours resulting from texture of the face. These contours are removed in the anisotropic inhibited Gabor energy; it can be observed that there are no patterns detected on the forehead and cheeks.

Note that in Figure 2-(E), contours from the teeth are detected in Gabor energy. This is an exemplar case where a small patch of texture is detected and inhibited with the proposed algorithm.

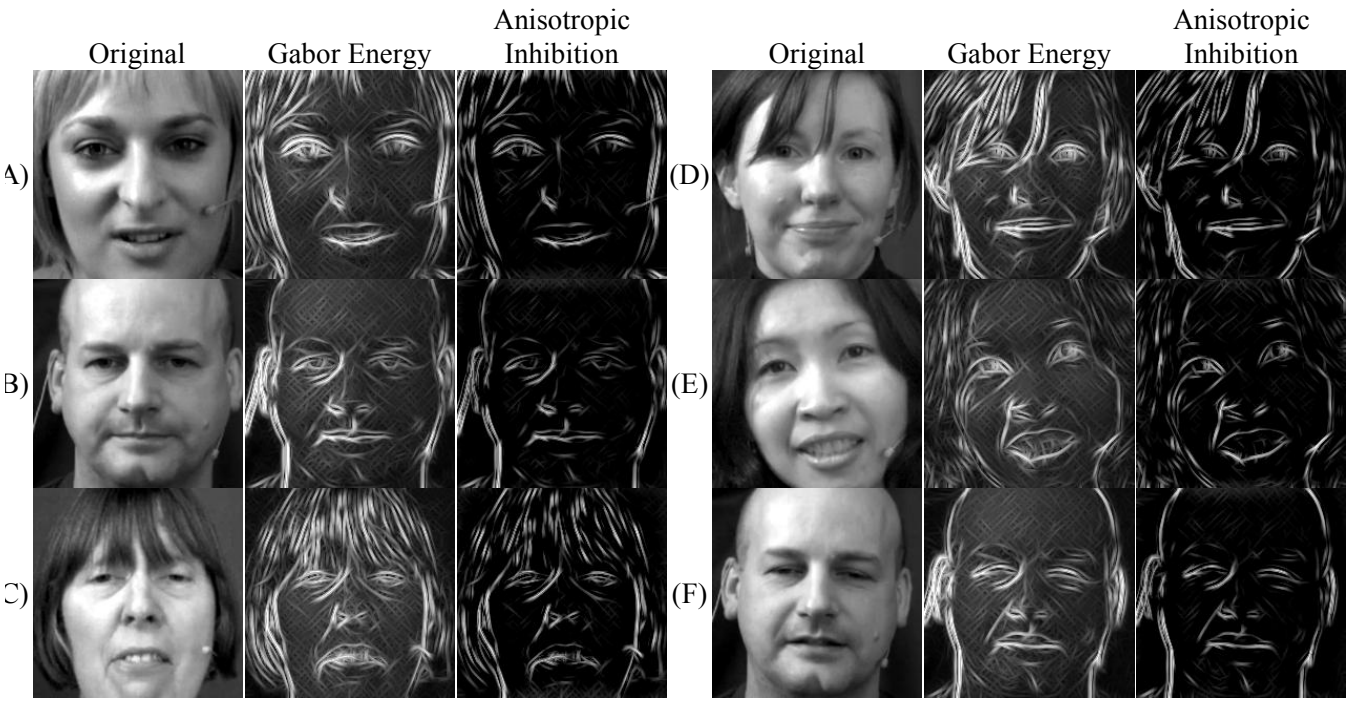

Fig. 2. Maximal representation of anisotropic inhibition versus Gabor energy, $b(x, y)$ for varying α, for 64 orientations and $\sigma = 4$.

2.3. Computational Efficiency Step

Equation 5 has retrieved the contours of f and removed the background texture. However, after this step, there is a response for each value of θ that was used. A method is needed to reduce the feature vector dimensionality. A representation of $\tilde{b}_\theta(x, y)$ is created that retains the maximal response for each pixel:

$$b(x, y) = \max\left\{\tilde{b}_\theta(x, y) \,|\theta = \theta_1, ..., \theta_N\right\} \quad (6)$$

Separately, an orientation map is constructed that contains the dominant orientation for each pixel:

$$\Theta(x, y) = \text{argmax}_\theta\left\{E_\theta(x, y) \,|\theta = \theta_1, ..., \theta_N\right\} \quad (7)$$

Equation 6 measures the strength of the contour and Equation 7 measures the orientation of the contour. A histogram is computed from b and Θ. This response is broken into M nonoverlapping, equally sized regions to account for face morphology. Note that this process is similar to local histograms in Local Binary Patterns [15]. For a given region A:

$$h_A(\theta_i) = \sum_{\forall(x,y)\in A|\Theta(x,y)=\theta_i} b(x, y) \quad (8)$$

Unlike Local Binary Patterns, a soft histogram is created for the edge orientations. With the proposed approach, instead of accumulating a vote for a specific orientation, the value of the magnitude at that pixel is used. This is realized by b; a conventional method would have 1 in place of b in a standard histogram for equation in Equation 8. Finally, the M histograms are concatenated and form the feature vector for f, see Figure 3.

3. EXPERIMENTS

The Audio/Visual Emotion Challenge 2012 (AVEC2012) data contains 32 continuous, frontal face videos of an interview subject being emotionally engaged. It is more challenging than previous data sets because the data is not acted or sponsored by the interviewer. The subject is speaking with an interviewer. As a result, emotions are subtle and more difficult to detect than in other data sets. The development subset contains 1 million frames. We present results on the development partition because it is the only set where ground truth labels are public. Results are generated with a 3-fold cross validation, where each fold was randomly generated. The specific folds used in these experiments are given in Table 2. The reader is referred to Schuller et al. [12] for a more in depth description of the data. The classes are described in terms of the Fontaine emotion model [16]. For this model, Ekman emotions [17] are projected into a four dimensional, Euclidean space. Emotion can be positive or negative valued along these dimensions. The dimensions are: arousal, whether or not an individual is positively or negatively interested in the situation; valence, the individuals over-

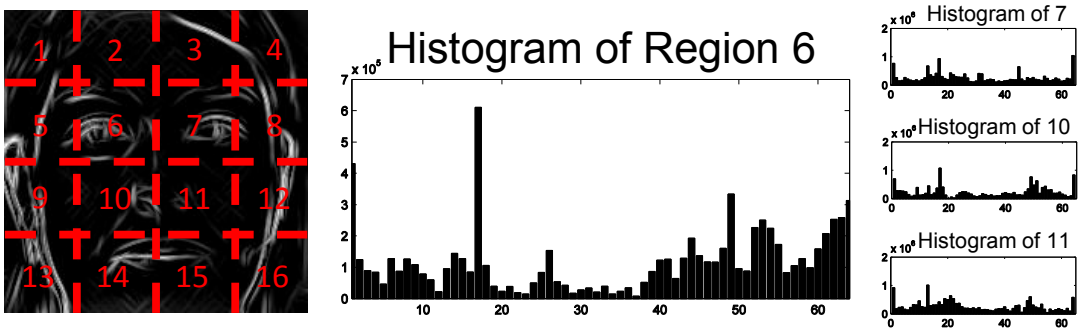

Fig. 3. An example of computing histograms in $M = 4$ local regions. The histograms from all regions are concatenated and taken to be the feature vector.

Table 3. Test Results on AVEC2012 Development Set Frame-Level Sub-Challenge

Feature	Arousal	Expectancy	Valence	Power	Average
Proposed	.4173 ± .0354	.1425 ± .0511	.3470 ± .0621	.1236 ± .0334	.2576
FPLBP	.4252 ± .0373	.1079 ± .0507	.2908 ± .0656	.0934 ± .0325	.2293
LBP	.4345 ± .0387	.0721 ± .0299	.2571 ± .0640	.0883 ± .0319	.2130
DCT	.0338 ± .0154	.0775 ± .0240	.0762 ± .0239	.0626 ± .0346	.0625
TPLBP	.0237 ± .0335	.0467 ± .0251	.0857 ± .0295	.0385 ± .0280	.0487
ACM	.0584 ± .0280	.0225 ± .0014	.0712 ± .0499	.0120 ± .0067	.0410

all feeling of himself/herself or the situation; power, whether or not the individual feels in control; and unpredictability, whether or not the individual is familiar with the situation. This formulates the problem as four regression problems, one for each emotion. A ground truth signal is given for each video. The correlation between the predicted emotion value and the ground truth is computed, and this value is averaged across all videos to give the performance. While there are many metrics that could be used to grade performance in regression, average correlation with the ground truth is the single official metric.

Avatar Image Registration is run for three iterations, which is empirically selected from previous work [18]. The following texture features are compared: (1) Anisotropic texture inhibition (ACM) [14], where the entire image (contour map) is taken to be the feature, (2) the proposed anisotropic inhibited Gabor energy histograms, (3) Uniform Local Binary Patterns (LBP) [7], (4) Three-Patch (TPLBP) and Four-Patch Local Binary Patterns (FPLBP) [10] and (5) Discrete Cosine Transform [8]. For anisotropic contour maps: $N = 64$, values of θ were selected such hat $\theta_{N+1} = \pi$, $\sigma = 4$ and $\alpha = 1$, which were chosen empirically. Additionally: $\gamma = .5$ and $\lambda = \sigma/.56$ [14]. All local histogram computing methods are calculated in neighborhoods of 8×8. For LBP, patterns are from an 8-member neighborhood with a radius of 1. TPLBP and FPLBP parameters are the same parameters as in Wolf *et al.* [10].

The results are given in Table 3. There is a clear dichotomy of feature performance. Either the feature is uncorrelated, or it shows promise. While the performance is relatively low, this is a non-trivial, difficult data set. Note that the LBP method is the exact method that ranked highest in the FERA2011 grand challenge [9]. The proposed method, LBP and FPLBP are the top performers. The proposed method does much better than other methods in the categories of expectancy, valence and power. Note that simply using the anisotropic inhibited contour map (ACM)the proposed method without the generalization stepis the worst performer. This is because the entire image is taken to be the feature. This has an extreme sensitivity to alignment, and demonstrates that local histograms in the computational efficiency step also grant some generalization capability. This would be similar to using the LBP image as a feature as opposed to histograms of LBP images.

4. CONCLUSION

In this paper we demonstrated that a maximal Gabor energy representation captures all the contours in an image. It includes contours from background texture that may not be important in facial emotion recognition. We proposed anisotropic inhibition for suppressing the background texture of a face image. It was also demonstrated that this anisotropic inhibition did poorly by itself and required generalization. We proposed maximal representation and local histograms to generalize anisotropic inhibition. We improved correlation on the Audio/Visual Emotion Challenge 2012, see Table 3.

Acknowledgements. This work was supported in part by NSF grants 0727129 and 0903667. The contents and information do not reflect the position or policy of the U.S. Government.

5. REFERENCES

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE CVPR*, 2011.

[2] R. Adolphs, D. Tranel, H. Damasio, and A. Damasio, "Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala," *Letters to Nature,*, vol. 372, pp. 669 – 672, 1994.

[3] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. SMC B*, vol. 42, no. 4, pp. 996 – 979, 2012.

[4] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. R. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. SMC B*, vol. 42, no. 4, pp. 966 – 979, 2012.

[5] M. Lyons, S. Akamatsu, and M. Kamachi J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. IEEE AFGR*, 1998.

[6] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*. Springer Berlin / Heidelberg, 2008.

[7] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. PAMI*, vol. 29, no. 6, pp. 915 – 928, 2007.

[8] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Trans. SMC B*, vol. 34, no. 3, pp. 1588 – 1595, 2004.

[9] S. Yang and B. Bhanu, "Understanding discrete facial expressions in video using an emotion avatar image," *IEEE Trans. SMC B*, vol. 42, no. 4, pp. 980 – 992, 2012.

[10] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. PAMI*, vol. 33, no. 10, pp. 1978 – 1990, 2011.

[11] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[12] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012 the continuous audio/visual emotion challenge," in *Proc. ACM ICMI*, 2012.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE CVPR*, 2001.

[14] C. Grigorescu, N. Petkov, and M. A. Westenberg, "Contour detection based on nonclassical receptive field inhibition," *IEEE Trans. IP*, vol. 12, no. 7, pp. 729 – 739, 2003.

[15] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803– 816, 2009.

[16] Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[17] P. Ekman, *The Handbook of Cognition and Emotion*, chapter Basic emotions, pp. 45 – 60, John Wiley & Sons, New York, NY, 1999.

[18] A. Cruz, B. Bhanu, and S. Yang, "A psychologically-inspired match-score fusion model for video-based facial expression recognition," in *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011.