NOVEL REPRESENTATION FOR DRIVER EMOTION RECOGNITION IN MOTOR VEHICLE VIDEOS

Rajkumar Theagarajan, Bir Bhanu*, Albert Cruz†, Belinda Le*, Asongu Tambo**

*Center for Research in Intelligent Systems, University of California, Riverside, CA 92521, USA
† Computer Perception Lab, California State University, Bakersfield, CA 93311, USA

ABSTRACT

A novel feature representation of human facial expressions for emotion recognition is developed. The representation leveraged the background texture removal ability of Anisotropic Inhibited Gabor Filtering (AIGF) with the compact representation of spatiotemporal local binary patterns. The emotion recognition system incorporated face detection and registration followed by the proposed feature representation: Local Anisotropic Inhibited Binary Patterns in Three Orthogonal Planes (LAIBP-TOP) and classification. The system is evaluated on videos from Motor Trend Magazine's Best Driver Car of the Year 2014-2016. The results showed improved performance compared to other state-of-the-art feature representations.

Index Terms— Facial expression, emotion recognition, feature extraction, background texture, anisotropic Gabor filter.

1. INTRODUCTION

Facial expressions are crucial to non-verbal communication of emotion. Automatic facial emotion recognition software has applications in lie detection, human behavior analysis, medical applications, and human-computer interfaces. We develop a system to detect stress and inattention of a motor vehicle operator from a single camera. Previous work in observation of motor vehicle operators employed multiple cameras for 3-D reconstruction [1], but multi-camera systems may introduce too much complexity and too many constraints in the design of a system. Another possible solution is gaze, but as of yet there is no consensus on how to detect inattention from gaze [2]. The goal of our work is a system that can extrapolate high stress and inattention from valence and arousal measurements on a low-cost platform so as to prevent motor vehicle accidents.

To this end, we present a novel dynamic local appearance feature that can compactly describe the spatiotemporal behavior of a local neighborhood in the video. The method is based on Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) [3] and background suppressing Gabor Energy filtering [4], but it is significantly different. We demonstrate that the background suppression concept can be applied to LBP-TOP to improve performance. The system is tested on three data sets collected from the Motor Trend Magazine's Best Driver Car of the Year 2014, 2015 and 2016.

2. RELATED WORK AND CONTRIBUTIONS

The current challenge to dynamic facial emotion recognition is the detection of emotion despite the various extrinsic and intrinsic imaging conditions, and intra-personnel differences in expression. While deep learning has been a growing trend in image processing and computer vision, the effects of transfer learning — using expression data from other datasets [5] — are diminished possibly because of various factors [6]. Thus, hand-crafted features, not learned from the neural networks, are still of great interest to unconstrained facial emotion recognition. This work focuses on the development of a novel hand-crafted feature representation.

Local Binary Pattern (LBP) is the most commonly used appearance-based feature extraction method [7]. LBP is a static texture descriptor and is not suitable for dynamic facial expressions in videos.

A variation of LBP, Volume Local Binary Patterns (VLBP), was developed to capture dynamic textures [8]. VLBP uses 3 parallel planes in the spatiotemporal domain where the center pixel is on the center plane, and it records the dynamic patterns in the neighborhood of each pixel into a $(3n+2)$ dimensional histogram, where n is the number of neighboring pixels.

The high dimensionality of VLBP is 2^{3n+2}, makes it impractical to use due to the rapid increase in dimensionality as the size of the neighborhood increases. An alternate solution to VLBP is the Local Binary Patterns in Three Orthogonal Planes (LBP-TOP). The dimensionality of LBP-TOP $(3*2^n)$ is significantly lower than VLBP. The working of LBP-TOP is described in section 3.

The other major type of appearance feature is the Gabor filter. Traditional Gabor filters are too sensitive in unconstrained settings; it captures all edges within an image, noise included. Cruz *et al.* [4] proposed Anisotropic Inhibited Gabor Filter (AIGF) that is robust to background noise and computationally efficient. Almaev *et al.* [9]

combined the original Gabor filter with LBP-TOP and noticed increased accuracy in classification of facial expressions compared to using only LBP-TOP.

We leveraged the advantages of AIGF and employed LBP-TOP for compact quantification of spatiotemporal information. We named this feature descriptor as Local Anisotropic Inhibited Binary Patterns in Three Orthogonal Planes (LAIBP-TOP).

In light of the related work, the key contributions of this paper are:

- Develop a novel appearance-based feature descriptor (LAIBP-TOP) that combines anisotropic filtering and local binary patterns.
- Evaluate the system on an unconstrained and continuous video dataset from Motor Trend's Best Driver Car of the year for 2014 – 2016 [17].
- Analyze and compare the results with state-of-the-art features, including Convolutional Neural Networks.

3. TECHNICAL APPROACH

This section elaborates on the computation of the dynamic appearance-based feature vector. Fig.1 shows the proposed system for dynamic facial expression recognition. We experimented with two different algorithms [10] and [11] for detecting a face in a video frame. Next, the detected faces are registered using similarity transform and Avatar Image Transformation [12].

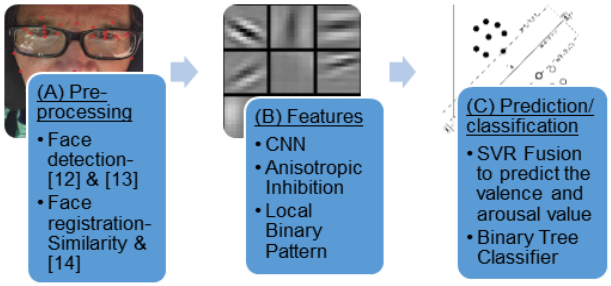

Fig. 1: The system for classification of human emotions.

A. Background Texture Removal - AIGF

The Gabor filter is used for edge detection at a specific orientation and scale. The Gabor filter [9] at a specific orientation and magnitude is expressed as:

$$g(x, y; \gamma, \theta, \lambda, \sigma, \varphi) = e^{\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}} \cos\left(2\pi\frac{x}{\lambda}+\varphi\right) \quad (1)$$

where x and y are the pixel location, γ is the spatial aspect ratio, θ is the angle, λ is the wavelength, σ^2 is the variance and φ is the phase offset taken at 0 and π. x' and y' are defined as: $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$. For the rest of this paper, $g(x, y; \theta, \varphi)$ is shorthand for $g(x, y; \gamma, \theta, \lambda, \sigma, \varphi)$. The image $I(x, y)$ is filtered by $g(x, y; \theta, 0)$ and $g(x, y; \theta, \pi)$ and the magnitude sum of both is taken as the result; this is called the Gabor energy filter:

$$E(x,y;\theta) = \sqrt{\left((I*g)(x,y;\theta,0)\right)^2 + \left((I*g)(x,y;\theta,\pi)\right)^2} \quad (2)$$

Regular Gabor energy filter captures all edges and magnitudes, including edges from noise due to background texture. This background information is not important for the classification of facial expressions and emotions. AIGF utilizes a weighted Gabor filter given by:

$t(x, y: \theta) = (E * w)(x, y)$, where, the weighted function w is:

$$w(x, y) = \frac{1}{\|DoG(x,y)\|} h(DoG(x,y)) \quad (3)$$

where $DoG(x, y)$ is a difference of Gaussians and $h(z) = H(z) * z$, $H(z)$ is the Heaviside step function. The resulting AIGF is described as:

$$g'(x,y;\theta) = h\left(E(x,y;\theta) - \alpha t(x,y;\theta)\right) \quad (4)$$

α is a constant ranging from 0 to 1, where $\alpha = 0$ indicates no background texture removal and $\alpha = 1$ indicates complete background texture removal.

Next, we perform non-maximal suppression to the output of the AIGF, this results in an image that contains only strong edges of the face important for the determination of expression and emotion. Fig. 2 shows an example of the driver's facial expression along with various outputs.

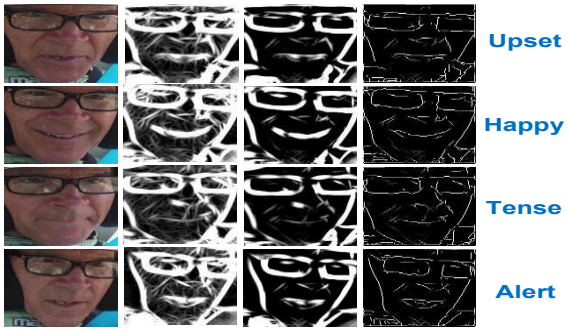

Fig. 2: (a) ROI, (b) Output from original Gabor filter α=0, (c) Output with α=1, (d) After non-maximal suppression.

When interpreting human emotions, we are concerned only with the following facial features as indicators of emotion: eyebrows, eyes, nose, mouth and crease of the cheeks. All other regions are considered as background, and we hypothesize that by eliminating this background noise from the original Gabor filter will leave us with only the strong edges from the points of interest.

B. Local Binary Pattern - Three Orthogonal Planes

The output of Equation (4) has a feature vector length of L_xL_y, where L_x and L_y are the row and column dimensions of the image, respectively. To further reduce the feature vector length and to measure spatiotemporal motion, we build upon LBP-TOP. LBP-type features apply a local neighborhood operation comparing the current pixel's values to the values of its neighbors:

$$q(x,y,\theta) = \{H(g'(x,y,\theta) - g'(u,v,\theta)) \,|\, (u,v) = N_{xy}^{8}(z)\} \quad (5)$$

$$q'(x,y,\theta) = \sum_{i=0}^{7} 2^i q_i(x,y,\theta) \quad (6)$$

where H is the step function used in equation (3), $N_{xy}^8(z)$ is z-th, 8-connected neighbor of *(x, y)*, *i* is the index of iteration over the 8-connected neighborhood of *(x, y)*. To quantify 3-D motion, Equation (6) is carried out in the XY, XZ and YZ, where the XY plane contributes to the spatial information, and the XT and YT frames contribute to the temporal information. Finally, to quantify spatiotemporal motion a histogram of each plane is taken. The feature vector length is reduced to $3*2^8$

4. EXPERIMENTAL RESULTS

4.1 Motor Trend Dataset

The videos in the dataset were privately recorded in collaboration with Motor Trends Magazine and are not yet available for public disclosure. The dataset for 2014-2016 [17] contains videos that are each approximately 1 minute and 45 seconds long. The videos consist of a driver racing different sports cars. The camera used to record the driver is a GoPro that takes 24 frames a second. These videos are unconstrained, continuous and provide real world data of complex emotions exhibited while test driving cars on a race track. The emotions that the driver displayed in our dataset are classified as stressed **(ST)**, tense **(TE)**, upset **(UP)**, alert **(AL)**, excited **(EX)** and happy **(HA)**. A greater variety of emotions were investigated and these emotions were the only emotions with a significant representation in the data.

4.2 Metrics and ground-truth

The metrics used for classifying the emotions are Arousal and Valence as shown in Fig. 4. Arousal is the measure of how energetic one feels and valence is the measure of how pleasant a feeling is. The range of emotion intensity varies from -1 to +1. The ground-truth labelling for the valence and arousal values was done by 3 experts for all the videos in the 2014 dataset and 7 experts for the 2015 and 2016 datasets. The final ground-truth for each video was obtained by taking the average of each individual ground-truths. Both the video and audio were analyzed while labeling the ground-truth.

4.3 Regression experiment results

• **2014 dataset-** We detected the face using [10] and used similarity transform for registration. 21,093 out of 25,270 faces were detected correctly. We used Support Vector Regression (SVR) with 9 fold leave-one-video-out cross-validation approach and a volume of 8 frames as input for the proposed approach and state-of-the-art features. The output of SVR is the predicted arousal and valence values.

Table I. shows the average and standard deviation of the correlation and RMS error between the predicted values and ground-truth values (2014 dataset) for the proposed approach and compared with the state-of-the-art features. The scale for correlation is -1 to +1, where +1 indicates that the ground-truth and predicted values have a perfect linear relationship and -1 means the opposite. The RMS error indicates the distance between the ground-truth and predicted value and its scale is from 0 to 1, where 0 indicates that there is no error. If the correlation is +1 and the RMS error is 0, it indicates that the ground-truth and predicted values are the same.

Table I: Comparison of Correlation and RMS

Features	Arousal		Valence	
	Correlation Avg/Std	RMS Avg/Std	Correlation Avg/Std	RMS Avg/Std
LBP [9]	0.103 ± 0.075	0.253 ±0.114	0.006 ±0.153	0.502 ±0.206
VLBP [9]	0.381 ± 0.102	0.243 ±0.148	0.306 ±0.147	**0.130 ±0.091**
LBP-TOP [9]	0.282 ±0.112	0.162 ±0.082	0.370 ±0.252	0.213 ±0.134
Gabor filter [14]	0.057 ± 0.136	0.193 ±0.126	0.130 ±0.151	0.394 ±0.181
LGBP-TOP [9]	0.077 ± 0.097	0.656 ±0.192	0.280 ±0.139	0.721 ±0.182
Proposed	**0.580 ±0.101**	**0.147 ± 0.064**	**0.525 ± 0.127**	0.134 ±0.042

Fig.3 shows the plot between the predicted and ground-truth values for arousal and valence for our proposed approach in Table I. From Table I it is observed that our approach outperforms the existing state-of-the-art. During cross-validation the predicted values of LBP, Gabor filter and LGBP-TOP had negative correlation with the ground-truth. The main reason for this is the presence of background texture noise, which is removed in our proposed approach and hence corroborates for the improvement in performance.

• **2014, 2015 and 2016 datasets-** The integrated dataset consists of 308,202 frames and 226,630 faces were correctly detected. Facial detection was done using Constrained Local Models [11] and registration using Avatar Image method [12]. We used SVR with leave one year out cross validation and a volume of 24 frames as input to our proposed approach and state-of-the-art features. Table II shows the correlation and standard deviation of our results and comparison with the state-of-the-art features.

It is observed from Table II that the LBP-TOP improves the performance. CNN pre-trained on the CASIA dataset [13] has the third best performance and falls below expectations, corroborating the findings in [5] and the motivation for this work. The proposed method performs better than all other unimodal feature representations, and even greater performance can be gained when fusing CNN with the proposed approach.

4.4 Classification experiment results on 2014 dataset

In our work we employed the valence and arousal domain of the Fontaine emotional model [15]. We used the predicted values from the SVR as input to the C4.5 binary tree [16] to classify the emotions. Table III shows the confusion matrix of the classified emotions for the 2014 dataset.

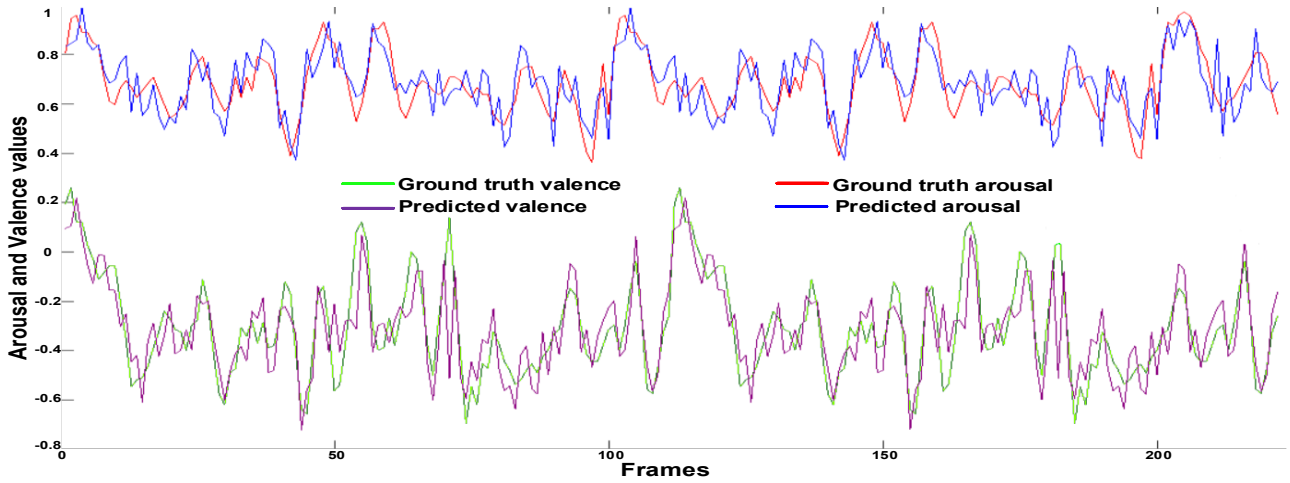

Fig. 3: Predicted and ground-truth values for Arousal and Valence (2014 dataset).

Table II: Comparison of Correlation with state-of-the-art (2014, 2015 and 2016 dataset integrated together).

	LBP [9]	VLBP [9]	LBP -TOP [9]	Gabor [14]	AI Gabor [4]	LGBP -TOP [9]	CNN [5]	Proposed	Fusion
Val.	0.255 ± 0.203	0.312 ± 0.156	0.341 ± 0.123	0.393 ± 0.243	0.491 ± 0.081	0.504 ± 0.196	0.563 ± 0.121	0.598 ± 0.135	0.633 ± 0.188
Ar.	0.111 ± 0.095	0.291 ± 0.128	0.321 ± 0.207	0.302 ± 0.214	0.336 ± 0.214	0.370 ± 0.313	0.441 ± 0.181	0.494 ± 0.142	0.527 ± 0.201

Table III: Confusion matrix for classification of emotion (2014 dataset). Stressed (ST), Tense (TE), Upset (UP), Alert (AL), Excited (EX), Happy (HA)

Class	**ST**	**TE**	**UP**	**AL**	**EX**	**HA**
ST	43	7	4	5	0	0
TE	3	36	6	2	4	1
UP	2	4	35	3	2	0
AL	4	5	2	18	2	1
EX	1	2	0	0	11	3
HA	0	0	1	0	2	13

Fig. 4 shows the Valence vs. Arousal plane with the data points corresponding to Table III overlaid on it. The predicted valence and arousal values are converted from Cartesian to polar coordinates to fit in the pie chart.

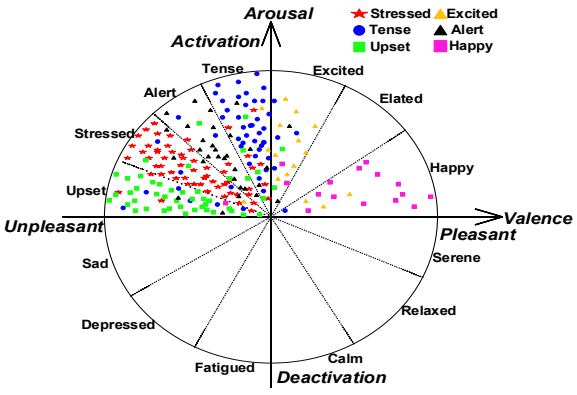

Fig. 4: Valence vs. Arousal plane with classification results.

From Table III, it is observed that our system achieved a correct classification rate of 70.28% with a false positive rate of 18.02%. The reason for high false positive is that the emotions **ST**, **TE**, **AE** and **UP** are not mutually exclusive and have significant overlap and similarly for **HA** and **EX**. The data is unbalanced with more data points for stressed, tense, upset and alert compared to excited and happy because, the driver is nervous as he is test driving a car that he has never driven before at a very high speed.

5. CONCLUSIONS

We investigated the efficacy of systems for the prediction of psychophysiological state of motor vehicle operators from video of a single, front facing camera. To achieve this goal, a novel facial appearance feature representation is presented that builds upon background suppression, Gabor filters, and local binary patterns. The system is tested on real world data from Motor Trend Magazine's Best Driver Car of the Year. Results showed improved performance compared to other state-of-the-art feature representations.

ACKNOLEDGEMENT

This work was supported in part by NSF grant 1330110 and ONR grant N00014-12-1-1026 . The contents of the information do not reflect the position or policy of US Government.

REFERENCES

[1] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 818–830, 2014.

[2] Y. Wang, B. Reimer, J. Dobres, and B. Mehler, "The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 26, pp. 227–237, 2014.

[3] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using volume local binary patterns," *Proc. ECCV 2006 Work. Dyn. Vis.*, vol. 4358, pp. 165–177, 2006.

[4] A. C. Cruz, B. Bhanu, and N. S. Thakoor, "Background suppressing Gabor energy filtering," *Pattern Recognit. Lett.*, vol. 52, pp. 40–47, 2015.

[5] H. W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443–449, 2015.

[6] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 503–510, 2015.

[7] O. T. and P. M. . M. T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24(7):971 - 987, 2002.

[8] G. Zhao and M. Pietikainen, "Dynamic texture recognition using volume ¨local binary patterns," *Dynamical Vision, Springer,* pp. 165–177, 2007.

[9] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," Proceedings - Humaine Association *Conference on Affective Computing and Intelligent Interaction*, ACII 2013, pp. 356–361, 2013

[10] X. Zhu and D. Ramanan "Face detection, pose estimation, and landmark localization in the wild," *IEEE Computer Vision and Pattern Recognition,* pp. 2879-2886, 2012.

[11] S. Cheng, A. Asthana, S. Zafeiriou, J. Shen, and M. Pantic, "Real-time generic face tracking in the wild with CUDA," *ACM Multimedia System*, pp. 148–151, 2014.

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

[13] D. Yi, Z. Lei, S. Liao and S. Z. Li, "Learning Face Representation from Scratch". *arXiv preprint* arXiv:1411.7923. 2014.

[14] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

[15] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, 2007.

[16] Weka V. 3.8.1, The University of Waikato, New Zealand.

[17] http://www.motortrend.com/news/the-future-of-testing-measuring-the-driver-as-well-as-the-car/

[18] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari and M. Mirza, "Combining modality specific deep neural networks for emotion recognition in video.", in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 543-550, 2013.

[19] B. K. Kim, H. Lee, J. Roh and S. Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition", in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 427-434, 2015.

[20] A. Dhall, R. Goecke, J. Joshi, M. Wagner and T. Gedeon, 2013, "Emotion recognition in the wild challenge (EmotiW) challenge and workshop summary," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 371-372, 2013.