Chapter 2

Human Automotive Interaction: Affect Recognition for Motor Trend Magazine's Best Driver Car of the Year

Albert C. Cruz, Bir Bhanu and Belinda T. Le

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65635

Abstract

Observation analysis of vehicle operators has the potential to address the growing trend of motor vehicle accidents. Methods are needed to automatically detect heavy cognitive load and distraction to warn drivers in poor psychophysiological state. Existing methods to monitor a driver have included prediction from steering behavior, smart phone warning systems, gaze detection, and electroencephalogram. We build upon these approaches by detecting cues that indicate inattention and stress from video. The system is tested and developed on data from Motor Trend Magazine's Best Driver Car of the Year 2014 and 2015. It was found that face detection and facial feature encoding posed the most difficult challenges to automatic facial emotion recognition in practice. The chapter focuses on two important parts of the facial emotion recognition pipeline: (1) face detection and (2) facial appearance features. We propose a face detector that unifies state-of-the-art approaches and provides quality control for face detection results, called reference-based face detection. We also propose a novel method for facial feature extraction that compactly encodes the spatiotemporal behavior of the face and removes background texture, called local anisotropic-inhibited binary patterns in three orthogonal planes. Real-world results show promise for the automatic observation of driver inattention and stress.

Keywords: facial emotion recognition, local appearance features, face detection

1. Introduction

In this chapter, we focus on the development of a system to track cognitive distraction and stress from facial expressions. The ultimate goal of our work is to create an early warning system to alert a driver when he/she is stressed or inattentive. This advanced facial emotion recognition technology has the potential to evolve into a human automotive interface that grants

nonverbal understanding to smart cars. Motor Trend Magazine's The Enthusiast Network has collected data of a driver operating a motor vehicle on the Mazda Speedway race track for the Best Driver Car of the Year 2014 and 2015 [1]. A GoPro camera was mounted on the windshield facing the driver so that gestures and expressions could be captured naturalistically during operation of the vehicle. Attention and valence were annotated by experts according to the Fontaine/PAD model [2]. The initial goal of both tests was to detect the stress and attention of the driver as metrics for ranking cars, automatically with computer algorithms. However, affective analysis of a driver is a great challenge due to a myriad of intrinsic and extrinsic imaging conditions, extreme gaze, pose, and occlusion from gestures. In 2014, two institutions were invited to apply automatic algorithms to the task but failed. It proved too difficult to detect face region of interest (ROI) with standard algorithms [3] and it was difficult to find a facial feature-encoding scheme that gave satisfactory results. Quantification of emotion was instead carried out manually by a human expert due to these problems. In this chapter, we discuss groundbreaking findings from analysis of the Motor Trend data and share promising, novel methods for overcoming the technical challenges posed by the data.

According to the U.S. Centers for Disease Control (CDC), motor vehicle accidents (MVA) are a leading cause of injury and death in the U.S. Prevention strategies are being implemented to prevent deaths, injuries, and save medical costs. Despite this, the U.S. Department of Transportation reported that MVA increased in 2012 after 6 years of consecutive years of declining fatalities. Video-based technologies to monitor the emotion and attention of automobile drivers have the potential to curb this growing trend. Existing methods to prevent MVA include smart phone collision detection from video [4], intelligent cruise control systems [5], and gaze detection [6]. The missing link in all these prevention strategies is the holistic monitoring of the driver from video—the key participant in MVA, and the detection of cues indicating inattention and stress. The introduction of intelligent transportation systems and automotive augmented reality will exacerbate the growing problem of MVA. While one would expect autonomous/self-driving cars to decrease MVA from inattention, intelligent transportation systems will return control of the vehicle to the driver in emergency situations. This handoff can only occur safely if the vehicle operator is sufficiently attentive, though his/her attention may be elsewhere from complacency due to the auto piloting system. Augmented reality systems seek to enhance the driving experience with heads-up displays and/or head-mounted displays that can distract the vehicle operator [7]. In short, driver inattention will continue to be a significant issue with cars into the future.

2. Related work

The field of affect analysis dates back to 1872 when Charles Darwin studied the relationship between apparent expression and underlying emotional state in the book, "The Expression of the Emotions in Man and Animals [8]." Communication between humans is a complex process beyond the delivery of semantic understanding. During conversation, we communicate nonverbally with gestures, pose, and expressions. One of the first works in automatic affect analysis by computers dates to 1975 [9]. Since this seminal work, emotion recognition

has found many applications in medicine [10–12], observation analysis (marketing) [13], and deception detection [14–16].

Systems to monitor the emotion and attention of vehicle operators date as far back to a 1962 patent that used steering wheel corrections as a predictor of attention and mental state [17]. Currently, there is much interest in the observation analysis of driver cognitive load, attention, and/or stress from video or biometric signals. While gaze has become a popular method for measuring attention of a driver, there is no consensus on how gaze should be monitored. Wang et al. [18] found that a driver's horizontal gaze dispersion was the most significant indicator of concentration under heavy cognitive load. Mert et al. [19] studied gaze during the handoff between manual vehicle control and autonomous piloting systems. It was found that if a driver was out of the loop it took more time to recover control of the vehicle, increasing the risk of MVA. However, a drawback to both of these methods is that it may not be possible to obtain an accurate measurement of driver gaze from video. A collaboration between AUDI AG, Volkswagen, and UC San Diego developed a video-based system for the detection of attention [20, 21]. This system focused on extracting head position and rotation using an array of cameras. We build upon state-of-the-art with an improved system that detects attention from only a single front-facing camera. In the following, we discuss the two most significant challenges to the system: face detection and facial feature encoding.

2.1. Related work in face detection

Detection of ROI is the first step of pattern recognition. In face detection, a rectangular bounding box must be computed that contains the face of an individual in the video frame. Despite significant advances to the state-of-the-art, detection of face in unconstrained facial emotion recognition scenarios is a challenging task. Occlusion, pose, and facial dynamics reduce the effectiveness of face ROI detectors. Imprecise face detection causes spurious, unrepresentative features during classification. This is a major challenge to practical applications of facial expression analysis. In Motor Trend Magazine's Best Driver Car of the Year 2014 and 2015, emotion was a metric for rating cars. In 2014, two institutions were invited to apply automatic algorithms to the task but all algorithms failed to sufficiently detect face ROI. Quantification of emotion was carried out manually by a human expert due to this problem [22].

Over the past 5 years, face detection has been carried out with the Viola and Jones algorithm (VJ) [10, 23–27]. Since the release of VJ, there have been numerous advances to face detection. Dollár et al. [28] proposed a nonrigid transformation of a model representing the face that is iteratively refined using different regressors at each iteration. Sanchez-Lozano et al. [29] proposed a novel discriminative parameterized appearance model (PAM) with an efficient regression algorithm. In discriminative PAMs, a machine- learning algorithm detects a face by fitting a model representing the object. Cootes et al. [30] proposed fitting a PAM using random forest regression voting. De Torre and Nguyen [23] proposed a novel generative PAM with a kernel-based PCA. A generative PAM models parameters such as pose and expression, whereas a discriminative PAM computes the model directly.

While the field of pattern recognition has historically been about features, ROI extraction is arguably the most important part of the entire pipeline. The adage, "garbage-in garbage-out" applies. In the AV+EC 2015 grand challenge, the Viola and Jones face detector [3] has a 6.5% detection rate and Google Picasa has a 0.07% detection rate. How does one infer the missing 93.95% of face ROIs? Among the "successfully" extracted faces, what is their quality? If one were to fill in the missing values with poor ROIs the extracted features would be erroneous and lead to a poor decision model. To address this, we propose a system that unifies current approaches and provides quality control of extraction results, called *reference-based face detection*. The method consists of two phases: (1) In training, a generic face is computed that is centered in the image. This image is used as a reference to quantify the quality of detection results in the next step. (2) In testing, multiple candidate face ROIs are detected, and the candidate ROI that best matches the reference face in the least squared sense is selected for further processing. Three different methodologies for finding the face ROIs are considered: a boosted cascade of Haar-like features, discriminative parameterized appearances, and a parts-based deformable models. These three major types of face detectors perform well in exclusive situations. Therefore, better performance can be achieved by unifying these three methods to generate multiple candidate face ROIs and quantifiably determine which candidate is the best ROI.

2.2. Related work in facial appearance features

Local binary patterns (LBP) are one of the most commonly used facial appearance features. They were originally proposed by Ojala et al. [31] as static feature descriptors that capture texture features within a single frame. LBP encode microtextures by comparing the current pixel to neighboring pixels. Differences are recorded at the bit level, e.g., if the top pixel is greater than the middle pixel a specific bit is set. Identical microtextures will take on the same integer value. There have been many improvements and variations of LBP over the years as the problems within computer vision became more complex. Independent frame-by-frame analysis is no longer sufficient for analysis of continuous videos.

A variation of LBP that was developed to address the need of a dynamic texture descriptor was volume local binary patterns (VLBP) [32]. VLBP are an extension of LBP into the spatiotemporal domain. VLBP capture dynamic texture by using three parallel frames centered on the current pixel. The need for a dynamic texture descriptor with a lower dimensionality than VLBP inspired the development of local binary patterns in three orthogonal planes (LBP-TOP) [32]. The dimensionality of LBP-TOP is significantly less than VLBP and is computationally less costly than VLBP.

LBP were not always the most popular local appearance feature. Some of the first, most significant works in facial expression analysis by computers used Gabor filters [33]. Gabor filters have historical significance, and they continue to be used in many approaches [34]. Nascent convolutional neural network approaches eventually learn structures similar to a Gabor filter [35]. The Gabor filters are bioinspired and were developed to mimic the V1 cortex of the human visual system. The V1 cortex responds to the gradient images of different orientation and magnitude. It is essentially an appearance-based feature descriptor that

http://dx.doi.org/10.5772/65635

captures all edge information within an image. However, state-of-the-art feature descriptors are known for their compactness and ability to generalize over external and intrinsic factors. The original Gabor filter does not have the ability to generalize in unconstrained settings because it captures all edges within an image, noise included. Furthermore, the Gabor filter is not computationally efficient. The filter produces a response for each filter within its bank. The Gabor filter has been developed into the anisotropic inhibited Gabor filter (AIGF) to model the human visual system's nonclassical receptive field [36]. AIGF generalizes better than the original Gabor filter because of its ability to suppress background noise. A combined Gabor filter with LBP-TOP has been shown to improve accuracy in the classification of facial expressions [37].

A thorough search of literature found no work, which has combined the anisotropic-inhibited Gabor filter and LBP-TOP and this is one of the foci of this chapter. This novel method that compactly encodes the spatiotemporal behavior of a face also removes background texture. It is called *local anisotropic-inhibited binary patterns in three orthogonal planes (LAIBP-TOP).* This feature vector works by first removing all background noise that is captured by the Gabor filter. Only the important edges of the Gabor filter are retained which are then encoded on the *X*, *Y*, and *T* orthogonal planes. The response is succinctly represented as spatiotemporal binary patterns. This feature vector provides a better representation for facial expressions as it is a dynamic texture descriptor and has a smaller feature vector size.

3. Technical approach

Automatic facial emotion recognition by computers has four steps: (1) region-of-interest (ROI) extraction, also known as face detection, (2) registration, colloquially known as alignment, (3) feature extraction, and (4) classification/regression of emotion. This chapter will focus on two important parts of the facial emotion recognition pipeline: face region-of-interest extraction and facial appearance features.

3.1. Reference-based face detection

Reference-based face detection consists of two phases: (1) In the training phase, a reference face is computed with avatar reference image. This face represents a well-extracted face and quantifies the quality of detection results in the next step. (2) In testing, multiple candidate face ROIs are detected, and the candidate ROI that best matches the reference face in the least squared sense is selected for further processing. Three different methodologies for finding the face ROI are combined: a boosted cascade of Haar-like features (Viola and Jones (VJ) [3], a discriminative parameterized appearance model (SIFT landmark points matched with iterative least squares), and a parts-based deformable model. VJ was selected because of its ubiquitous use in the field of face analysis. Discriminative parameterized appearance models were recently deployed in commercial software [38]. Parts-based deformable models showed promise for face ROI extraction in the wild [39]. Despite the success of currently used methods, there is still much room for improvement. In the Motor Trend data, there are segments

of video where one extractor will succeed when others fail. Therefore, better performance can be achieved by unifying these three methods to generate multiple candidate face ROIs and quantitatively determine which candidate is the best ROI. Note that Refs. [38, 39] use VJ for an initial bounding box so running more than one face detector is not excessive for state-of-the-art approaches.

3.1.1. Reference-based face detection in training

The avatar reference image concept generates a reference image of an expressionless face. It was previously used for registration [40] and learning [41]. A proof of optimality of the avatar image concept is given in the previous work [42]. Let I be an image in the training data D. To estimate the avatar reference image $R_{ARI}(x)$, take the mean across all face images:

$$R_{ARI}(x, y) = \frac{1}{N_D} \sum_{i \in D} I_i(x, y) \tag{1}$$

where N_D is the number of training images; (x, y) is a pixel location; and I_i is the i-th image in the dataset D. The process iterates by rewarping D to R_{ARI} to create a more refined estimate of the reference face. The procedure is described as follows: (1) compute reference using Eq. (1) from all training ROIs D, (2) warp all D to the reference, and (3) recompute Eq. (1) using the warped images from the previous step. Steps (2) and (3) are iterated for three times which was empirically selected in Ref. [40]. Results of the reference face at different iterations are shown in **Figure 1**. SIFT-Flow warps the images in step (2) and the reader is referred to [43] for a full description of SIFT-Flow. In short, a dense, per-pixel SIFT feature warp is computed with loopy belief propagation. After this point, a R_{ARI} represents a well-extracted reference face.

Figure 1. Iterative refinement of the avatar reference face. It represents a well-extracted face.

3.1.2. Reference-based face detection in testing

To robustly detect a face, three different pipelines simultaneously extract the ROI. We fuse a discriminative parameterized appearance model, a part-based deformable model, and the

http://dx.doi.org/10.5772/65635

Viola and Jones framework. In Viola and Jones (VJ), detection of the face is carried out with a boosted cascade of Haar-like features. Because of the near-standard use of VJ, we omit an in-depth explanation of the method. The reader is referred to [3] for the details of the algorithm.

3.1.2.1. Discriminative parameterized appearance model

Consider a sparse appearance model of the face. The face detection problem can be framed as an optimization problem that fits the landmark points representing the face. A face is successfully detected when the gradient descent in the fitness space of the optimization problem is complete. Traversing the fitness space can be viewed as a supervised learning problem [38], rather than carrying out a gradient descent with Gauss-Newton algorithm [44]. In the training phase the following equation is minimized:

$$\min_w \left\| s(p + w(p)) - s(p^*) \right\| \tag{2}$$

where s is a function that computes SIFT features; w is a flow vector to be optimized; p^* is manually labeled landmark points; and the vector p has horizontal and vertical components $\mathbf{p} = (x, y)$. Computing the Hessian of the model is computationally undesirable, and supervised learning of the descent from p^* avoids computing this directly. In testing, face alignment is carried out with linear least squares.

3.1.2.2. Parts-based deformable models

Parts-based deformable models represent a face as a collection of landmark points similar to PAMs. The difference is that the most likely locations of the parts are calculated with a probabilistic framework. The landmark points are represented as a mixture of trees of landmark points on the face [39]. Let Φ be the set of landmark points on the face. A facial configuration L is modeled as $L = \{p_i : i \in \Phi\}$. Alignment of the landmark points is achieved by maximizing the posterior likelihood of appearance and shape. The objective function is formulated as follows:

$$\epsilon(I, L, j) = \sum_i u_{ij}\, \mathbf{s}(p_i) + \sum_{(i,k)} \left(b_1(i, j, k)\tilde{x}^2 + b_2(i, j, k)\tilde{x} + b_3(i, j, k)\tilde{y}^2 + b_4(i, j, k)\tilde{y} \right) \tag{3}$$

where ϵ is the objective function to be minimized; I is the video frame; j is the mixture index; k is the landmark point indexes; u_{ij} is the template of mixture j at point i; s is an appearance feature; b_1, b_2, b_3, and b_4 are the spring rest and rigidity parameters of the model's shape. $\tilde{x}$ and $\tilde{y}$ are the displacement in horizontal and vertical directions from i and k:

$$\tilde{x} = x_i - x_k \tag{4}$$

$$\tilde{y} = y_i - y_k \tag{5}$$

Inference is carried out by maximizing the following:

$$\max_j \left(\max_L (\epsilon(I, L, j)) \right) \tag{6}$$

which enumerates over all mixtures and configurations. The maximum likelihood of the model which best fits the parameters is computed with the Chow-Liu algorithm [45].

3.1.2.3. Least square selection

We compare the results of all three pipelines to check if a face has been properly detected. The problem is posed where we must quantify the accuracy of each extraction pipeline. We minimize the candidate face ROI I_k to the reference of a face R_{ARI} in the least squared sense:

$$\min_k \sqrt{\sum p \left(I_k(x,y) - R_{ARI}(x,y)\right)^2} \tag{7}$$

where I_k is a candidate face ROI from one of the face extraction pipelines k. It is possible that Eq. (7) failed to generate a candidate face. There are two causes for this: (A) there are no candidate face ROIs generated, or (B) the selected face is a false alarm, e.g., it is not a face, or the bounding box is poorly centered. To prevent (B), the face selected in Eq. (7) must have a distance to the reference of no greater than parameter T, which is empirically selected in training. If the detector fails because of (A) or the threshold is less than T, the last extracted face should be used for processing further in the recognition pipeline. Note when comparing this proposed method to other detectors in **Table 1** we count (A) and (B) as a failure of the method.

%	Viola and Jones (VJ)	Constrained local models (CLM)	Supervised descent method (SDM)	Proposed face detector
True positive rate	60.27 ± 10.53	68.36 ± 9.80	<u>81.37 ± 17.60</u>	**86.29 ± 8.90**
F1-score	74.52 ± 19.67	80.81 ± 7.17	<u>89.47 ± 11.22</u>	**92.43 ± 5.07**

Viola and Jones is the worst performer with the highest variance. Constrained Local Models and Supervised Descent Method are acceptable but have a high variance. The proposed method is the best performer. Higher is better for both metrics. Bold: Best performer. Underline: Second best performer.

Table 1. Face detection rates for the Motor Trend Magazine's Best Driver Car of the Year.

3.2. Local anisotropic inhibited binary patterns in three orthogonal planes

3.2.1. Gabor filter

A Gabor filter is a bandpass filter that is used for edge detection at a specific orientation and scale. Images are typically filtered by many Gabor filters at different parameters, called a bank. It is modulated by a sine and a cosine. When it is modulated by a sine, the Gabor filter finds symmetric edges. When it is modulated by a cosine, the Gabor filter finds antisymmetric edges. According to Grigorescu et al. [36], a Gabor filter at a specific orientation and magnitude is:

$$g(x, y; \gamma, \theta, \lambda, \sigma, \phi) = \exp\left(\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\cos\left(\frac{2\pi x'}{\gamma} + \phi\right) \tag{8}$$

http://dx.doi.org/10.5772/65635

where γ is the spatial aspect ratio that effects the eccentricity of the filter; θ is the angle parameter that tunes the orientation; and λ is the wavelength parameter that tunes the filter to a specific spatial frequency, or magnitude. In pattern recognition this is also referred to a scale. σ is the variance of the distribution. It determines the size of the filter. ϕ is the phase offset that is taken at 0 and π. x' and y' are defined as follows:

$$x' = x\cos\theta + y\sin\theta \tag{9}$$

$$y' = -x\sin\theta + y\cos\theta \tag{10}$$

The Gabor filter can be used as local appearance filter by tuning the filter to a local neighborhood while still varying the orientation: $\sigma/\lambda = 0.56$ and varying θ. For the rest of the chapter, $g(x, y; \theta, \phi)$ represents g with $\gamma = 0.5$, $\lambda = 7.14$, and $\sigma = 3$, and with varying θ and ϕ. Given an image I, the Gabor energy filter is given by:

$$E(x, y; \theta) = \sqrt{\Big((I * g)(x, y; \theta, 0)\Big)^2 + \Big((I * g)(x, y; \theta, \pi)\Big)^2} \tag{11}$$

which corresponds to the magnitude of filtering the image at the phase values of 0 and π.

3.2.2. Anisotropic-inhibited Gabor filter

The original formulation of the Gabor energy filter does not generalize well. The Gabor energy filter captures all edges and magnitudes within the image, including the edges due to noisy background texture. For example, MPEG block encoding artifacts that present as a grid-like repeating pattern. In the field of facial expression recognition, face morphology causes creases along the face that are not a part of the background texture thus a better contour map can be extracted by removing the background texture of the face. In order to eliminate the background texture detected by the Gabor filter, we build upon the Anisotropic Gabor energy filter. To suppress the background texture, we take a weighted Gabor filter:

$$\tilde{g}(x, y; \theta) = (E * w)(x, y) \tag{12}$$

where the weighted function w is:

$$w(x, y) = \frac{1}{\|DoG(x, y)\|} h(DoG(x, y)) \tag{13}$$

where $h(x) = H(x) * x$, where $H(x)$ is the Heaviside step function; $DoG(.)$ is the difference of Gaussians:

$$DoG(x, y; \theta) = \frac{1}{2\pi K^2 \sigma^2} e^{\frac{x^2+y^2}{2K^2\sigma^2}} - \frac{1}{2\pi \sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{14}$$

w resembles a ring. Eq. (12) retrieves the background texture of (x, y) without the texture of (x, y) itself by weighting E by the ring-like filter w. The resulting anisotropic-inhibited Gabor filter is described as follows:

$$\hat{g}(x, y; \theta) = h(E(x, y; \theta) - \alpha \times \tilde{g}(x, y; \theta)) \tag{15}$$

where α is a parameter that affects how much of the background texture is removed. α ranges from 0 to 1, where 0 indicates no background texture removal and 1 indicates complete background texture removal. The first term of Eq. (15) defines the original Gabor energy filter that captures all edges including background edges. The second term subtracts the weighted

Gabor filter with a specified alpha, depending on how much background suppression is needed. We follow [46] where a value of $\alpha = 1$ was empirically selected.

To obtain an image that contains only the strongest edges and corresponding orientations, we take the edges with the strongest magnitude across N different orientations:

$$AIGF(x, y) = \max_{\theta} \hat{g}(x, y; \theta) \tag{16}$$

The resulting output of Anisotropic Inhibited Gabor Filter is an image that is $M \times N$. Results are given in **Figure 2**.

Figure 2. (a) Original frame, (b) result of Gabor energy filter (Eq. (15) with $\alpha = 0$), and (c) result of Anisotropic Gabor Energy Filtering.

We build upon the work in Ref. [46], but the proposed approach is significantly different. The anisotropic Gabor energy filter (AIGF) further computes the orientations corresponding to the maximum edges as follows:

$$\Theta(x, y) = \operatorname{argmax}_{\theta} \tilde{g}(x, y; \theta) \tag{17}$$

A soft histogram is computed from Θ with votes weighted by the maximal edge response $AIGF$. For the proposed approach, we use $AIGF$ and do not compute a soft histogram.

3.2.3. Local binary patterns

Local binary patterns (LBP) encode local appearance as a microtexture code. The code is a function of comparison to the intensity values of neighboring pixels. Some formulations are invariant to rotation and monotonic grayscale transformations [31]. At present LBP and its many variations are one of the most widely used feature descriptors for facial expression recognition. LBP result in a texture descriptor with dimensionality of 2^n where n is a parameter that controls the number of pixel neighbours. The LBP code of a pixel at (x, y) is given as follows:

$$LBP(x, y) = \sum_{\{u,v\} \in N_{x,y}^{LBP}} \operatorname{sign}\big(I(u, v) - I(x, y)\big) \times 2^q \tag{18}$$

where (u, v) iterates over points in the neighborhood of $N_{x,y}^{LBP}$; sign(.) is the sign of the expression; q is a counter starting from 0 that increments on each iteration; and $N_{x,y}^{LBP}$ is the neighborhood of

points about (x, y) (see **Figure 3A**). 2^q encodes the result of the intensity difference in a specific bit. A histogram is taken for further compactness and tolerance of registration errors. Each pixel in I is encoded with an LBP code from Eq. (18) then an n-level histogram is extracted from LBP. Typically, the image is segmented into nonoverlapping regions and a histogram is extracted from each region [47]. While powerful and effective for static images, LBP lacks the ability to capture temporal changes in continuous video data.

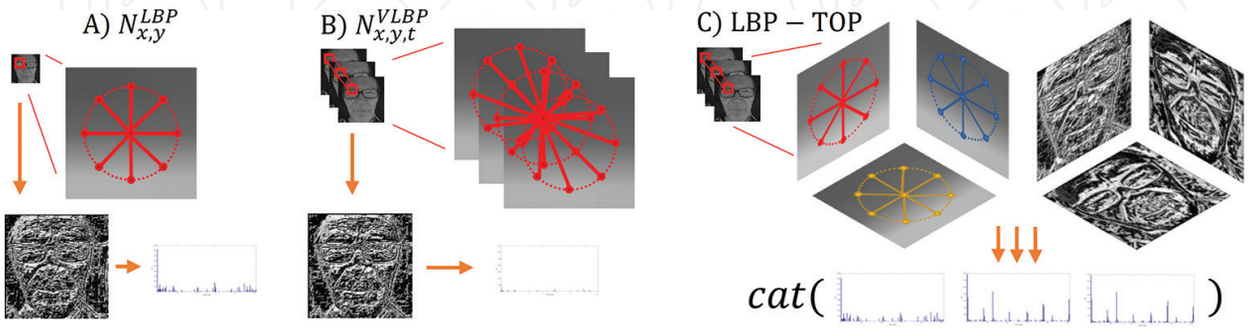

Figure 3. (A) In LBP, microtexture is encoded in the XY-plane. (B) In VLBP, this is extended to the spatiotemporal domain by including neighbors in the three planes parallel to the current frame. (C) In LBP-TOP, local binary patterns are separately extracted in three orthogonal planes and the resultant histograms are concatenated. This greatly reduces feature vector size over treating the volume as a 3D microtexture.

3.2.4. Volumetric local binary patterns

Volume local binary patterns (VLBP) and local binary patterns in three orthogonal planes (LBP-TOP) are variations of LBP that were developed to capture dynamic textures for video data. In VLBP, the circle of neighboring points in LBP is scaled up to a cylinder. VLBP computes code values as a function of three parallel planes centered at $\{x, y, t\}$. That is, the middle plane contains the center pixel. VLBP coding is obtained by the following equation:

$$VLBP(x, y, t) = \sum_{k\in\{-L,0,L\}} \sum_{\{u,v\}\in N_{x,y,t}^{VLBP}} \text{sign}\big(I(u, v, k) - I(x, y, t)\big) \times 2^q \tag{19}$$

where k iterates over three time points: t, $t-L$, and $t+L$. $N_{x,y,t}^{VLBP}$ is the set of spatiotemporal neighbours of $\{x, y, t\}$ (see **Figure 3B**). A large set of $N_{x,y,t}^{VLBP}$ results in a large feature vector while a small $N_{x,y,t}^{VLBP}$ results in a small feature vector. As with LBP, a histogram is taken for further compactness. The maximum grey-level from Eq. (19) is $2^{(3n+2)}$, thus VLBP are more computationally expensive to calculate and require larger feature vector.

3.2.5. Local binary patterns in three orthogonal planes

LBP-TOP was developed as an alternative to VLBP. VLBP and LBP-TOP differ in two ways. First, LBP-TOP uses three orthogonal planes that intersect at the center pixel. Second, VLBP considers the cooccurrences of all neighboring points from three parallel frames, which make for a larger feature vector. LBP-TOP only considers features from each separate plane and then concatenates them together, making the feature vector much shorter when compared to VLBP for large values of n. LBP-TOP performs LBP on the three orthogonal planes corresponding to the XY, XT, and YT axes (see **Figure 3C**). The XY plane contributes the spatial

information and the XT and YT frames contribute the temporal information. These planes intersect at the center pixel. Whereas in Eq. (19), VLBP captures a truly three- dimensional microtexture, LBP-TOP computes LBP codes separately on each plane. The resulting feature vector dimensionality of LBP-TOP is 3×2^n.

3.2.6. Local anisotropic inhibited Gabor patterns in three orthogonal planes

In the proposed method, the computational efficiency of LBP-TOP is applied to images filtered with the anisotropic-inhibited Gabor filter. The suppression of background texture provides an image that only contains the edges separate from the background texture. These edges are the significant boundaries of facial features that are useful when determining expression and emotion. Local anisotropic binary patterns' (LAIBP) code values are computed as follows:

$$LAIBP(x, y) = \sum_{\{u,v\} \in N_{x,y}^{LBP}} \text{sign}\big(AIGF(u, v) - AIGF\big(x, y\big)\big) \times 2^q \tag{20}$$

where $g(u, v)$ is the maximal edge magnitude from Eq. (16). LAIBP-TOP features are extracted in a similar fashion to LBP-TOP: Compute $LAIBP$ codes from Eq. (20) in XY, XT, and YT planes and concatenate the resultant histograms. A comparison of AIGF, LBP, and the proposed method, LAIBP, are given in **Figure 4**. The proposed method (LAIBP-TOP) is significantly different from LBP-TOP because we introduce background texture removal from Eq. (16).

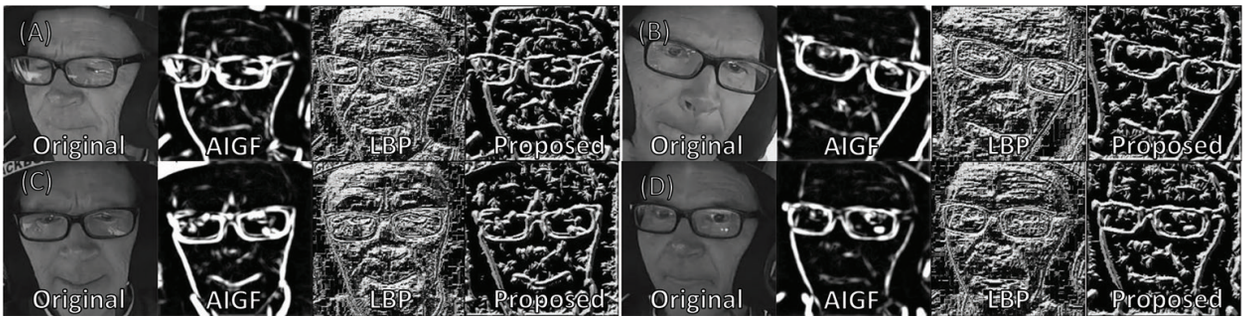

Figure 4. From left to right: The original frame, anisotropic inhibited Gabor filter (AIGF), local binary patterns (LBP), and the proposed method local anisotropic inhibited binary patterns (LAIBP). Note that the proposed method has more continuous lines compared to AIGF. LBP is susceptible to JPEG compression artifacts.

4. Experimental results

4.1. Datasets

Data in this work have been provided by Motor Trend Magazine from their Best Driver Car of the Year 2014 and 2015. They consist of frontal face video of a test driver as he drives one of 10 automobiles around a racetrack. Parts of the video will be released publicly on YouTube at a later date. The videos are 1080p HD quality captured with a Go Pro Hero 4 and range from 231 to 720 seconds in length. The camera is mounted on the windshield of the car facing the driver's face. The dataset was labeled with the Fontaine emotional model [2] rather than facial action units or emotional categories to quantize emotion. Emotions such as happiness,

http://dx.doi.org/10.5772/65635

sadness, etc. occupy a space in a two-dimensional Euclidean space defined by valence and arousal. The objective of the dataset is to detect the valence and arousal of an individual on a per-frame basis. Valence, also known as evaluation-pleasantness, describes positivity or negativity of the person's feelings or feelings of situation, e.g., happiness versus sadness. Arousal, also known as activation-arousal, describes a person's interest in the situation, e.g., eagerness versus anxiety.

4.2. Metrics

For face detection results, we use true positive rate and F_1 score. F_1 score is given by:

$$2 \times \frac{(\text{Precision})(\text{Recall})}{(\text{Precision}) + (\text{Recall})} \tag{21}$$

For both metrics, higher is better. For full recognition results, we use root mean squared (RMS) error and correlation. The correlation coefficient is given by:

$$\frac{E\left[\left(\boldsymbol{y}_d - \mu_{y_d}\right)\left(\boldsymbol{y} - \mu_y\right)\right]}{\sigma_{y_d}\sigma_y} \tag{22}$$

where $E[.]$ is the expected operation; $\boldsymbol{y}_d$ is the vector of ground-truth labels for a video; $\boldsymbol{y}$ is the vector of predicted labels for a video; μ_{y_d} and μ_y are the mean of ground-truth and prediction, respectively; and σ_{y_d} and σ_y are the standard deviation of ground-truth and prediction, respectively.

4.3. Results comparing different face detectors

Face detection results are given in **Table 1**. In general, VJ is the worst performer with the highest variance. Though CLM and SDM have acceptable detection rates, they too have a high variance and some videos are a total failure with no face extraction. The proposed algorithm improves detection rates on both datasets and reduces variance.

4.4. Results comparing different facial appearance features

For the full recognition pipeline: The landmarks for the inner corner of the eyes and the tip of the nose are used as control points for a course registration. These points are the least effected by face morphology. An ϵ-SVR is used for prediction of valence and arousal values [48].

Full regression results and a comparison to other state-of-the-art facial appearance features are given in **Table 2**. Experiments employed a 9-fold, leave-one-video-out cross-validation. For correlation, higher is better; for RMS lower is better. In **Table 2**, the correlation and RMS values for valence and arousal labels by the proposed method performed the best for valence and second best for arousal. Removal of background noise and then implementing LBP-TOP provided better results. RMS values for the proposed method are also the best for arousal and second best for valence. The proposed method has the best average correlation and the lowest average RMS value. Graphs comparing the ground-truth and predicted labels are given in **Figure 5**. It was found that frames with extreme head rotation tended to have lower correlation and higher error due to the difficulty of registering the dataset.

Features	Valence		Arousal		Average	
	Correlation	**RMS**	**Correlation**	**RMS**	**Correlation**	**RMS**
LBP	0.0066	0.5025	0.1032	0.2526	0.0549	0.3776
VLBP	0.3060	**0.1292**	**0.3810**	0.2428	0.3435	0.1860
LBP-TOP	0.3705	0.2134	0.0819	0.1624	0.2262	0.1879
Gaborenergy filter	0.1296	0.3937	0.0569	0.1935	0.0933	0.2936
LGBP-TOP	0.2805	1.1207	0.0787	1.2559	0.1796	1.1883
Proposed	**0.4446**	0.2054	0.2801	**0.1547**	**0.3624**	**0.1801**

Note: The proposed method has better average correlation for valence and arousal. Bold indicates best performing feature.

Table 2. Correlation and RMS for prediction of valence and arousal emotion categories on the Motor Trend Magazine Best Driver's Car of the Year.

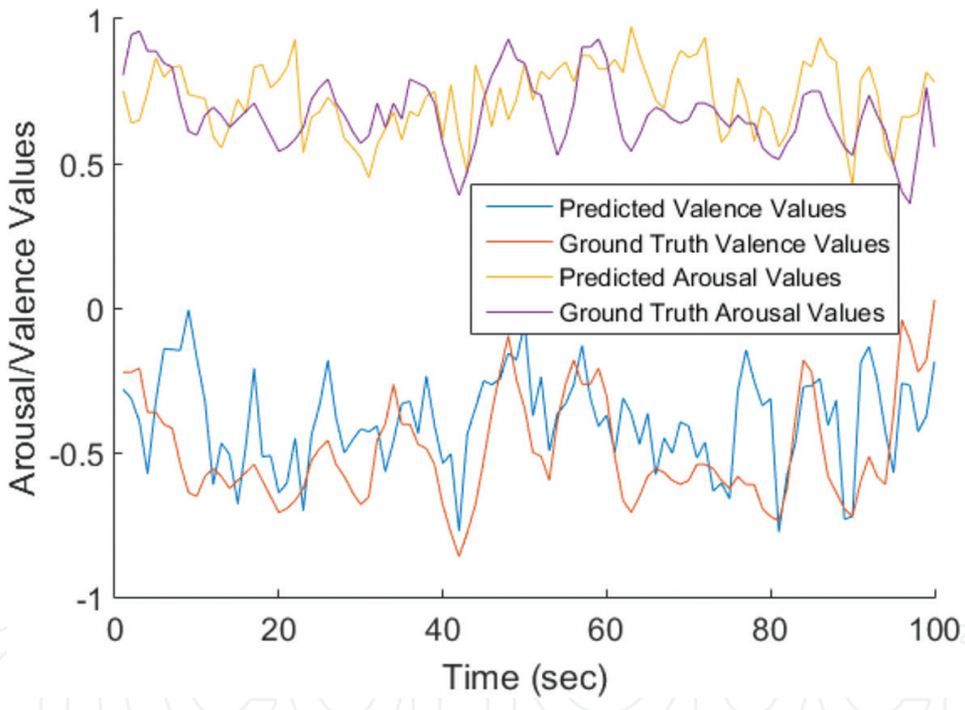

Figure 5. The predicted values are graphed with the values for valence and arousal.

5. Conclusions

In this chapter, we proposed a system to perform facial expression recognition on a brand new dataset. This dataset is unconstrained and unique. We proposed a new feature vector that is robust to background noise and capable of capturing dynamic textures. We also proposed a novel method for fusing the output of many face detectors. Both approaches provided better results than other state-of-the-art methods. In the future work, the face detection scheme will be scaled up to a 3D model to better detect the extreme out of plane head rotations.

Author details

Albert C. Cruz[1]*, Bir Bhanu[2] and Belinda T. Le[2]

*Address all correspondence to: acruz37@csub.edu

1 COMputer Perception LAB (COMPLAB), California State University, Bakersfield, CA, USA

2 Center for Research in Intelligent Systems (CRIS), University of California, Riverside, CA, USA

References

[1] K. Reynolds, "At 2015 Best Driver's Car, What is the Driver Experiencing?," *Motor Trend Magazine*, 2015. [Online]. Available: http://www.motortrend.com/news/the-future-of-testing-measuring-the-driver-as-well-as-the-car/. [Accessed: 26-Apr-2016].

[2] 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.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition. CVPR 2001*, vol. 1, 2001.

[4] J. White, C. Thompson, H. Turner, B. Dougherty, and D. C. Schmidt, "WreckWatch: Automatic traffic accident detection and notification with smartphones," *Mob. Networks Appl.*, vol. 16, no. 3, pp. 285–303, 2011.

[5] S. Echegaray, "The modular design and implementation of an intelligent cruise control system," in *2008 IEEE International Conference on System of Systems Engineering*, 2008, pp. 1–6.

[6] R. C. Coetzer and G. P. Hancke, "Eye detection for a real-time vehicle driver fatigue monitoring system," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2011, pp. 66–71.

[7] J. L. Gabbard, G. M. Fitch, and H. Kim, "Behind the glass: Driver challenges and opportunities for AR automotive applications," *Proc. IEEE*, vol. 102, no. 2, pp. 124–136, 2014.

[8] C. Darwin, "The expression of the emotions in man and animals," *Am. J. Med. Sci.*, vol. 232, no. 4, p. 477, 1872.

[9] F. I. Parke, "A model for human faces that allows speech synchronized animation," *Comput. Graph.*, vol. 1, no. 1, pp. 3–4, 1975.

[10] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Avec '13*, 2013, pp. 21–30.

[11] M. Kächele and M. Schels, "Inferring depression and affect from application dependent meta knowledge," in *ACM Multimedia Workshops*, 2014, pp. 41–48.

[12] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De La Torre, "Detecting depression from facial actions and vocal prosody," in *Proceedings—2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, 2009.

[13] S. Yang and M. Kafai, "Zapping Index: using smile to measure advertisement zapping likelihood," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 432–444, 2014.

[14] S. Demyanov, C. Leckie, and J. Bailey, "Detection of deception in the mafia party game," in *ACM International Conf. Multimedia*, 2015, pp. 335–342.

[15] R. Mihalcea and M. Burzo, "Towards multimodal deception detection – step 1: building a collection of deceptive videos," *ACM Int. Conf. Multimodal Interact.*, pp. 189–192, 2012.

[16] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. Nunamaker, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 36–43, 2005.

[17] P. Fletcher, "Automobile driver attention indicator," US 3227998 A, 1966.

[18] 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, no. PA, pp. 227–237, 2014.

[19] N. Merat, A. H. Jamson, F. C. H. Lai, M. Daly, and O. M. J. Carsten, "Transition to manual: Driver behaviour when resuming control from a highly automated vehicle," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 27, no. PB, pp. 274–282, 2014.

[20] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer, "Looking-in and looking-out vision for Urban Intelligent Assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking," *IEEE Intell. Veh. Symp. Proc.*, no. Iv, pp. 115–120, 2014.

[21] 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.

[22] K. Reynolds, "2014 motor trend's best driver's car: How we test," *Motor Trend Magazine*, 2014.

[23] F. De Torre and M. H. Nguyen, "Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[24] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," *ICMI'12—Proc. ACM Int. Conf. Multimodal Interact.*, no. Section 4, pp. 485–492, 2012.

[25] A. Cruz, B. Bhanu, and N. Thakoor, "Facial emotion recognition in continuous video," *Int. Conf. Pattern Recognit.*, pp. 1880–1883, 2012.

[26] J. R. Williamson, W. Street, T. F. Quatieri, B. S. Helfer, R. Horwitz, and B. Yu, "Vocal biomarkers of depression based on motor incoordination and timing," in *ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 41–47.

[27] G. A. Ramirez, T. Baltrušaitis, and L. P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction Workshops*, 2011, vol. 6975, pp. 396–406.

[28] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 1078–1085.

[29] E. Sanchez-Lozano, F. De la Torre, and D. Gonzalez-Jimenez, "Continuous regression for non-rigid image alignment," in *European Conf. Computer Vision*, 2012, pp. 250–263.

[30] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *European Conf. Computer Vision*, 2012, pp. 278–291.

[31] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[32] 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.

[33] 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, FG 1998*, 1998, pp. 200–205.

[34] F. Ringeval, M. Valstar, E. Marchi, D. Lalanne, and R. Cowie, "The AV + EC 2015 multimodal affect recognition challenge: Bridging across audio, video, and physiological data categories and subject descriptors," in *Proc. ACM Multimedia Workshops*, 2015.

[35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Adv. Neural Inf. Process. Syst. 27 (Proceedings NIPS)*, vol. 27, pp. 1–9, 2014.

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

[37] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," *Proc.—2013 Hum. Assoc. Conf. Affect. Comput. Intell. Interact. ACII 2013*, pp. 356–361, 2013.

[38] X. Xiong and F. De La Torre, "Supervised descent method and its applications to face alignment," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[39] S. Cheng, A. Asthana, S. Zafeiriou, J. Shen, and M. Pantic, "Real-time generic face tracking in the wild with CUDA," *Proc. 5th ACM Multimed. Syst. Conf. - MMSys '14*, no. 1, pp. 148–151, 2014.

[40] 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.

[41] A. C. Cruz, B. Bhanu, and N. Thakoor, "Facial emotion recognition with expression energy," *ACM Int'l. Conf. Multimodal Interact. Work.*, pp. 457–464, 2012.

[42] E. Cambria, G.-B. Huang, L. L. C. Kasun, H. Zhou, C. M. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li, V. C. M. Leung, L. Feng, Y.-S. Ong, M.-H. Lim, A. Akusok, A. Lendasse, F. Corona, R. Nian, Y. Miche, P. Gastaldo, R. Zunino, S. Decherchi, X. Yang, K. Mao, B.-S. Oh, J. Jeon, K.-A. Toh, A. B. J. Teoh, J. Kim, H. Yu, Y. Chen, and J. Liu, "Extreme learning machines [trends & controversies]," *IEEE Intell. Syst.*, vol. 28, no. 6, pp. 30–59, 2013.

[43] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 15–49, 2015.

[44] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.

[45] C. Chow and C. Liu, "Discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, 1968.

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

[47] A. Cruz, B. Bhanu, and N. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Trans. Affect. Comput.*, vol. PP, no. 99, pp. 1–1, 2014.

[48] C.-C. Chang and C.-J. Lin, "LIBSVM," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.