

# Person-Independent Facial Expression Detection using Constrained Local Models

Sien. W. Chew, Patrick Lucey, Simon Lucey, Jason Saragih, Jeffrey F. Cohn and Sridha Sridharan

**Abstract**—In automatic facial expression detection, very accurate registration is desired which can be achieved via a deformable model approach where a dense mesh of 60-70 points on the face is used, such as an active appearance model (AAM). However, for applications where manually labeling frames is prohibitive, AAMs do not work well as they do not generalize well to unseen subjects. As such, a more coarse approach is taken for person-independent facial expression detection, where just a couple of key features (such as face and eyes) are tracked using a Viola-Jones type approach. The tracked image is normally post-processed to encode for shift and illumination invariance using a linear bank of filters. Recently, it was shown that this preprocessing step is of no benefit when close to ideal registration has been obtained. In this paper, we present a system based on the Constrained Local Model (CLM) method which is a generic or person-independent face alignment algorithm which gains high accuracy. We show these results against the LBP feature extraction on the CK+ and GEMEP-FERA datasets.

## I. INTRODUCTION

The field of affective computing has matured to the stage where fully automatic systems exist for a variety of tasks [1], [2], [3]. The approach commonly adopted in developing these systems is to first track the face and facial features, derive some feature representation from the face and then based on these features do some classification of facial expressions. If there is poor registration, this error would propagate through subsequent processing stages (i.e. feature extraction and classification), which ultimately dictates the detection performance. For example, when considering facial actions units AU1 and AU2 (i.e. eyebrows), the misalignment of a couple of pixels may affect performance. This is further emphasized when one wishes to detect subtle or low-intensity (e.g. ‘a’ and ‘b’) AUs.

To accommodate this, it is preferred that very accurate registration is obtained which can be achieved via a deformable model approach where a dense mesh of 60-70 points on the face is used. Such an approach is desired due to this accuracy in addition to their ability to infer the 3D pose parameters (i.e. pitch, yaw and roll) and features (i.e. synthesize frontal view), which is ideal in situations where there is a lot of head

S.W. Chew and S. Sridharan are with the Speech, Audio, Image and Video Technology Laboratory at Queensland University of Technology, Brisbane, Australia. Email: {sien.chew@qut.edu.au, s.sridharan@qut.edu.au}

P. Lucey is with Disney Research Pittsburgh and J.F. Cohn is with Department of Psychology, University of Pittsburgh/Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15260. Email: {patrick.lucey@disneyresearch.com, jeffcohn@cs.cmu.edu}

S. Lucey and J. Saragih are with the Commonwealth Science and Industrial Research Organisation (CSIRO), Australia; Email: {simon.lucey@csiro.au, jason.saragih@csiro.au }

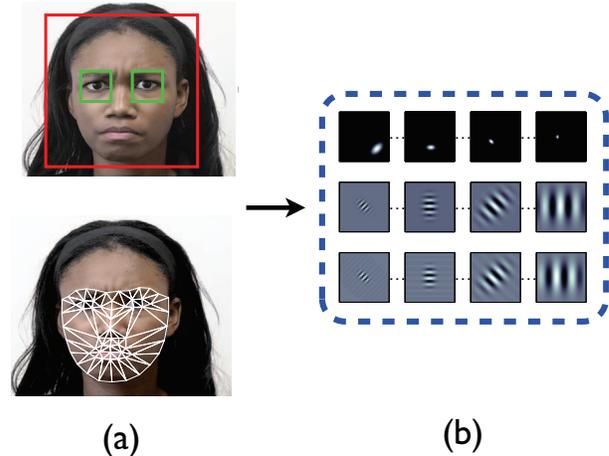


Fig. 1. (a) For person-independent facial expression detection, most systems register (a) the face and facial features coarsely (i.e. track the face and eyes) as deformable models such as AAMs do not generalize well to unseen subjects. (b) After tracking, an image of the face is obtained and this image is normally post-processed to encode against shift invariance. In this paper, we use Constrained Local Models (CLM) developed by Saragih et al. [4], which is a person-independent deformable model which can obtain very accurate dense models. We show by using this highly accurate approach, post-processing the image to gain shift-invariance is of little benefit.

movement, especially out-of-plane head rotations. *Person-dependent* active appearance models (AAMs) [5], [6] have been widely used in this field [7], [8], [9], [3] for those reasons but this approach requires manual labeling of key frames of the target sequence (up to 5% of frames). For applications where manually labeling frames is prohibitive (e.g. marketing, security/law enforcement, health-care and HCI), a more generic or *person-independent* face alignment approach is required. As AAMs do not generalize well to unseen subjects, a more coarse approach is taken for person-independent facial expression detection, where just a couple of key features (such as face and eyes) are tracked using a Viola-Jones type [10] approach. After tracking, an image of the face is obtained and this image is normally post-processed to encode for shift and illumination invariance. The normal convention is to apply a bank of linear filters on the extracted face image (i.e. Gabor [11], HOG [12], Box [10], SIFT [13], LBP [14]). These features have been widely used due to their biological relevance, their ability to encode edges and texture, and their invariance to illumination. An inherent problem with this method is the large memory and computational overheads required for training and testing these filter banks (e.g. using Gabor filters). An

example of this is shown in Figure 1.

In a recent paper [15], it was shown that when there is close to ideal registration, post-processing the tracked image is of little benefit in consistent illumination conditions (which is indicative of the conditions expected in most of the above applications). However as noted earlier, for person-independent facial expression detection, getting close to perfect registration is very difficult to obtain. Recently, Saragih et al. [4] developed a generic or person-independent face alignment algorithm which leverages the generalization capacity of local patch experts which has shown face alignment accuracy close to person-dependent AAMs.

In this paper, we make the following contributions:

- We present a person-independent facial expression detection system using the CLM to track the face and features, which can generate a synthesized canonical appearance view (Section II).
- We show via a slue of experiments on the CK+ and GEMEP-FERA dataset, that there is little benefit in applying the LBP features compared to the raw pixels, when there is close to ideal registration (Sections V & VI).
- Based on these evaluations, we describe our approach to the FERA challenge as well as report our results (Section VI).

## II. CONSTRAINED LOCAL MODELS (CLM)

Constrained Local Models (CLM) was devised by Saragih et al. [4], with the goal of finding the shape  $\mathbf{s}$ , which is described by a 2D triangulated mesh. In particular, the coordinates of the mesh vertices define the shape  $\mathbf{s} = [x_1, y_1, x_2, y_2, \dots, x_n, y_n]$ , where  $n$  is the number of vertices. These vertex locations correspond to a source appearance image, from which the shape was aligned. The shape  $\mathbf{s}$  can be expressed as a base shape  $\mathbf{s}_0$  plus a linear combination of  $m$  shape vectors  $\mathbf{s}_i$ :

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^m p_i \mathbf{s}_i \quad (1)$$

where the coefficients  $\mathbf{p} = (p_1, \dots, p_m)^T$  are the shape parameters. These shape parameters can typically be divided into rigid similarity parameters  $\mathbf{p}_s$  and non-rigid object deformation parameters  $\mathbf{p}_o$ , such that  $\mathbf{p}^T = [\mathbf{p}_s^T, \mathbf{p}_o^T]$ . Similarity parameters are associated with a geometric similarity transform (i.e. translation, rotation and scale). The object-specific parameters, are the residual parameters representing non-rigid geometric variations associated with the determining object shape (e.g., mouth opening, eyes shutting, etc.). Procrustes alignment [5] is employed to estimate the base shape  $\mathbf{s}_0$ .

The CLM uses a host of algorithms which utilize an ensemble of local detectors to determine  $\mathbf{s}$ . All of these methods have the following two goals: (i) perform an exhaustive local search for each PDM landmark around their current estimate using some kind of feature detector, and (ii) optimize the PDM parameters such that the detection responses over all

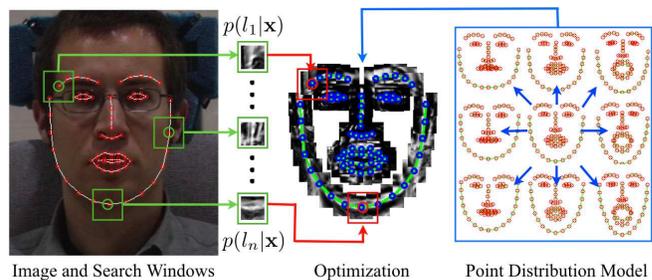


Fig. 2. CLM fitting is performed by an exhaustive search in a patch through feature detectors, and the feature responses jointly maximized through constrained-mean shifts. This figure was taken from [4].

of its landmarks are jointly maximized. Figure 2 illustrates the components of the CLM fitting.

The particular instance of CLM used in this work is that proposed in [4]. The method uses linear SVMs over power normalized image patches to discriminate aligned from misaligned mesh vertex coordinates. Composing the SVM classification score with a Sigmoid function generates a likelihood map over the vertices within a local search region around its current estimate. This allows a Bayesian treatment of the alignment problem. The advantage of using the linear SVM over more sophisticated classifiers is twofold. First, it allows rapid computations of the mesh vertices' probability maps using efficient normalized cross correlation. Secondly, the linear model's limited capacity results in better generalization to unseen subject identities.

Once likelihood maps for each mesh vertex have been computed, the CLM variant in [4] uses an optimization strategy coined subspace constrained mean-shifts. By assuming the vertex likelihoods are conditionally independent given the shape, optimization proceeds by alternating two steps: 1. compute a single mean-shift update for each vertex independently of all others, and 2. project the mean-shifted vertex coordinates onto the subspace of the shape model in Equation 1. By virtue of its interpretation as an instance of the EM algorithm, this simple two step procedure is provably convergent. To encourage convergence to the global optimum in cases with gross initial misalignment, this optimization strategy is applied on a pyramid of smoothed versions of the likelihood maps, which is similar to the heuristic often used in AAM alignment but with the difference that smoothing is applied directly to the objective rather than indirectly through the image. For full details please see [4]<sup>1</sup>.

## III. APPEARANCE-BASED FEATURES

Various techniques have been proposed in literature to extract key attributes of an image that are useful in classifying facial expressions. The most basic feature available is basically a vector of raw appearance pixels from the facial region. The major drawback in such a simplistic approach is an inherent variability associated with the correctly registered local image appearance when errors in registration are

<sup>1</sup>For demonstration of the CLM in action and details of getting access to the CLM API, please visit [www.jsaragih.com](http://www.jsaragih.com)

present. This variability, when described as a distribution instead of a static observation point, could be used to normalize for either rigid or non-rigid geometric variations.

### A. Pixel-Based Representations

Once the subject's face have been CLM-tracked by estimating the shape and appearance parameters, the following features can be derived:

- 1) **SPTS**: The similarity normalized shape,  $s_n$ , refers to the 68 vertex points in  $s_n$  for both the  $x$ - and  $y$ -coordinates, resulting in a raw 136 dimensional feature vector. These points are the vertex locations after all the rigid geometric variation (translation, rotation and scale), relative to the base shape, has been removed. The similarity normalized shape  $s_n$  can be obtained by synthesizing a shape instance of  $s$ , using Equation 1, that ignores the similarity parameters  $\mathbf{p}$ .
- 2) **SAPP**: The similarity normalized shape,  $\mathbf{a}_n$ , refers to where all the rigid geometric variation (translation, rotation and scale) has been removed. It achieves this by using  $s_n$  calculated above and warps the pixels in the source image with respect to the required translation, rotation and scale. This is the type of approach is employed by most researchers [1], [2] as only coarse registration is required (i.e. just face and eye locations). When out-of-plane head movement is experienced some of the face is partially occluded which can affect performance, also some non-facial information is included due to occlusion.
- 3) **CAPP**: The canonical normalized appearance  $\mathbf{a}_0$  refers to where all the non-rigid shape variation has been normalized with respect to the base shape  $s_0$ . This is accomplished by applying a piece-wise affine warp on each triangle patch appearance in the source image so that it aligns with the base face shape. In [8], it was shown by removing the rigid shape variation, poor performance was gained.

### B. Local Binary Pattern Operators

Local binary pattern operators (LBP) were derived from a general definition of texture in a local neighborhood. It gained popularity as an effective and yet computationally simple texture descriptor which exhibited invariance to illumination. Since its initial conception, the LBP operator had expanded into a vast family of LBP-based feature detectors [16], [17].

1) *LBP Neighbourhood Grids*: In the seminal work of Ojala et al. [14], the first incarnation of the LBP operator utilized a  $3 \times 3$  local rectangular grid consisting of 8 neighbours, and relied on the gray value of the centre pixel as a reference threshold. They demonstrated how an effective local representation of texture  $T$  could be afforded, by thresholding all neighbourhood pixel values  $g_P$  into a binary number through the gray value of the centre pixel  $g_c$ ,

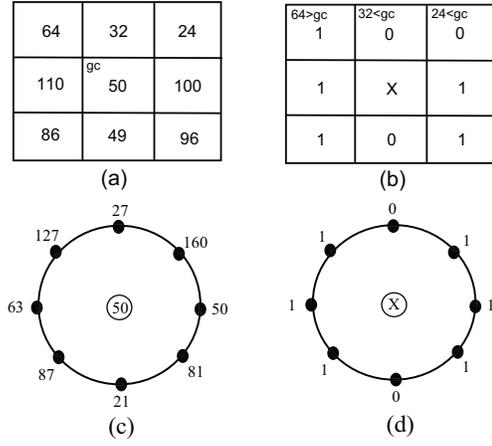


Fig. 3. (a) local binary pattern operator with a  $3 \times 3$  rectangular grid, (b) corresponding thresholded values  $T = 10011011_2 = 155_{10}$ , (c) local binary pattern operator with an 8-neighbourhood circular grid, (d) corresponding thresholded values  $T = 10111011_2 = 187_{10}$ .

$$T = \sum_{P=0}^{P-1} t(g_P - g_c) 2^P \quad (2)$$

$$t(\cdot) = \begin{cases} 1 & \text{if } g_P \geq g_c \\ 0 & \text{otherwise.} \end{cases}$$

where the size of the neighbourhood is described by radius  $R$  having  $P$  equally spaced pixels. Circular grids provide a better fit to unorthogonal objects, such as faces. In the same degree, a local texture model could be described by a joint distribution of pixel value differences in the circular grid without loss of information (Figure 3).

2) *Analyzing the Spatial-Shift Mechanism in LBP*: A recent paper [15] suggested that two popular appearance-based features (Gabor magnitudes and Histograms of Oriented Gradients (HOG)) functioned by encoding spatial-shift invariance into input images that were poorly aligned. When good alignment was ensured, a marked reduction in the benefits these features introduced was observed. In this subsection, we analyze whether the concept of spatial-shift encoding plays an active role in the LBP operator.

Briefly, Gabor magnitude responses are obtained by convolving an input image with a bank of two-dimensional spatial bandpass filters. It is common practice to design filterbanks with evenly-spaced spatial frequencies and orientations,

$$g(x, y) = K \exp \left\{ -\pi \left[ a^2 (x - x_0)_r^2 + b^2 (y - y_0)_r^2 \right] \right\} \exp \left\{ j 2\pi (\mu_0 x + \nu_0 y) + P \right\}. \quad (3)$$

Sine frequencies  $\mu_0$  and  $\nu_0$  of Equation 3 control the degree of pixel sampling in the  $x$ -/ $y$ -dimensions of an input image. The weight applied onto local cells of the image is determined by constant  $K$ . Local cells are selected through modulation of the sine carrier  $\exp \{ j 2\pi (\mu_0 x + \nu_0 y) \}$  with the Gaussian envelope  $\exp \{ -\pi [ a^2 (x - x_0)_r^2 + b^2 (y - y_0)_r^2 ] \}$ .

Envelope size, shape and position are controlled through the following variables –  $a$  and  $b$  (elongation factor),  $x_0$  and  $y_0$  (rotation). A filterbank consisting of multiple orientations and spatial frequencies is commonly employed to generate Gabor magnitude features (e.g.  $8 \times 8$  filterbank in [11] achieved excellent AU detection performance).

Once it is understood how one Gabor filter essentially selects and weights different local cells, it may be readily established that multiple local cells of weighted pixels are formed by a bank of Gabor filters. Likewise, HOG divides an image into a grid of cells, and a local histogram is computed using a weighted vote of image gradients from each pixel. When one perceives the mechanics underlying these two algorithms as a *fusion* of multiple local cell outputs, it may be appreciated as an alignment of spatial-shifts.

3) *Spatial Shift Coding in LBP?*: Firstly, the same method of tiling an input image with a grid of cells is common to both LBP and HOG. In Figure, 3, it was noted that the gray-intensity values of every pixel in  $g_c$  and  $g_P$  needed to be taken into account for the calculation of local texture  $T$ . Likewise, global texture is represented as a combination of all local textures. LBP operators may thus be perceived as spatial-shift encoders, in a similar respect to Gabor filters and HOG descriptors.

#### IV. CLASSIFICATION USING SUPPORT VECTOR MACHINES

Support vector machines (SVM) have been recognized as an effective algorithm in numerous pattern recognition and facial expression recognition applications [1], [11], [18], [2]. This type of supervised binary classifier attempts to maximize the Euclidean distance between support vectors by locating the optimal position of the hyperplane,

$$\arg \min_w \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} + \lambda \sum_i \delta_i \quad (4)$$

$$\text{subject to } y_i \mathbf{w}^T x_i \geq 1 - \delta_i; \quad \forall_i$$

where  $\mathbf{x} \in \mathbb{R}^{m \times n}$  is a matrix containing positive and negative training examples,  $y_i \in \{\pm 1\}$  is a vector containing labels corresponding to  $\mathbf{x}$ ,  $\lambda$  is the penalty term,  $\xi_i$  represent slack variables, and support vector weights  $w$  are solved by minimising Equation 4. In this paper, linear kernel SVMs [19] were used in all experiments.

#### V. CLM EXPERIMENTS ON THE EXTENDED COHN-KANADE (CK+) DATASET

We conducted AU and emotion detection experiments on the Extended Cohn-Kanade (CK+) [3] dataset using CLM-tracked CAPP and SAPP features. In each experiment, a leave-one-out strategy was employed for training and testing.

The dataset consists of 123 subjects displaying a range of posed emotions in a mostly frontal view angle. Head movement was minimal. AU detection results in Table I show that there is little difference in classification rates between CAPP and SAPP features when rigid head motion was

TABLE I  
AU CLASSIFICATION RATES FOR THE CK+ DATASET (F1-SCORE).  $N$  REPRESENTS THE NUMBER OF POSITIVE EXAMPLES AVAILABLE, AND  $\mu$  REPRESENTS THE WEIGHTED MEAN.

AU	$N$	PIX-CAPP	PIX-SAPP	LBP-CAPP	LBP-SAPP
1	173	0.75	0.72	0.74	0.58
2	116	0.73	0.75	0.74	0.71
4	191	0.73	0.67	0.71	0.60
6	122	0.70	0.66	0.66	0.57
7	119	0.56	0.59	0.52	0.60
12	111	0.78	0.76	0.76	0.77
15	89	0.75	0.48	0.59	0.40
17	196	0.77	0.59	0.72	0.60
25	287	0.85	0.81	0.83	0.75
26	48	0.26	0.27	0.22	0.27
$\mu$	–	0.73	0.67	0.70	0.62

TABLE II  
CONFUSION MATRIX FOR EMOTION CLASSIFICATION ON THE CK+ DATASET.

	Anger	Cont	Disg	Fear	Happy	Sad	Surp
Anger	<b>70.1</b>	4.5	9.0	5.5	1.0	7.9	2.0
Cont	3.3	<b>52.4</b>	3.6	11.3	11.3	6.4	11.7
Disg	2.8	0.4	<b>92.5</b>	0.4	1.7	1.3	0.9
Fear	5.6	2.1	1.4	<b>72.1</b>	11.0	1.8	6.0
Happy	0.8	0.7	1.5	1.9	<b>94.2</b>	0.1	0.9
Sad	17.4	7.9	13.1	7.3	1.3	<b>45.9</b>	7.1
Surp	0.8	2.1	0.5	1.3	1.1	0.7	<b>93.6</b>

kept minimal. The confusion matrix for emotion classification rates is shown in Table II.

#### VI. CLM EXPERIMENTS ON THE GEMEP-FERA DATASET

The GEMEP-FERA dataset [20] contains video recordings of 10 actors displaying various AUs and emotions. A meaningless phrase was uttered by the actors while standing up, which resulted in substantial rigid head and body motion. Attempting AU and emotion detection in the presence of speech made the tasks particularly challenging.

##### A. AU Sub-Challenge

The AU training partition was supplemented with training instances from the CK+ dataset, and evaluated in a leave-one-out configuration. In addition to this, a comparison was made between CLM-tracked pixel representations (PIX-CAPP and PIX-SAPP) and uniform LBP features ( $R = 1, N = 8$ ). This was done to test the spatial-shift hypothesis of the LBP operator (Section III-B.3). Classification results shown in Table III suggested little benefit could be obtained from using LBP features. As slight improvements in classification was observed from PIX-CAPP over PIX-SAPP pixel representations, we selected PIX-CAPP to train all our AU models for

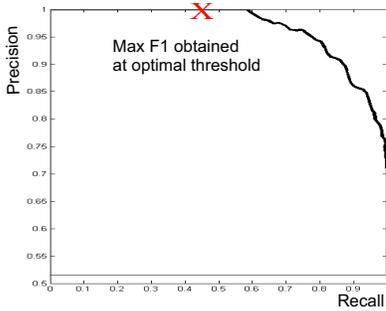


Fig. 4. Graphic describing how thresholds were obtained for the AU detector. Training instances from the GEMEP-FERA training partition were supplemented with positive and negative AU instances from the CK+ dataset. All training instances were power-normalized, and evaluated in a leave-one-out configuration. The detection threshold for each AU was obtained from the threshold where the maximum F1-score occurred in a precision-recall curve.

TABLE III

FERA-GEMEP DATASET AU TRAINING PARTITION (F1-SCORES).  $N$  REPRESENTS THE NUMBER OF POSITIVE EXAMPLES AVAILABLE, AND  $\mu$  REPRESENTS THE WEIGHTED MEAN.

AU	$N$	PIX-CAPP	PIX-SAPP	LBP-CAPP	LBP-SAPP
1	1600	0.62	0.55	0.55	0.49
2	1631	0.55	0.47	0.47	0.47
4	1356	0.41	0.41	0.41	0.41
6	1808	0.69	0.65	0.63	0.63
7	2123	0.63	0.67	0.60	0.59
10	2034	0.60	0.60	0.56	0.56
12	2725	0.74	0.74	0.72	0.70
15	1026	0.34	0.43	0.33	0.37
17	822	0.30	0.27	0.29	0.27
18	419	0.33	0.26	0.23	0.17
25	812	0.31	0.30	0.30	0.30
26	499	0.20	0.22	0.20	0.20
$\mu$	—	0.56	0.54	0.52	0.51

use in the testing partition; with the exception of AUs 7, 15 and 26 in which PIX-SAPP was used instead.

The detection thresholds for the test partition were selected from classification scores obtained from the training set, based on the maximum F1-score in a precision-recall curve (Figure 4). Classification rates for the AU test partition are presented in Table IV.

### B. Emotion Sub-Challenge

A five-way forced choice strategy was employed to determine the emotion class of every frame in a test sequence. The emotion class which obtained the highest score in a winning frame gets a single vote, and the class with the highest number of votes in a sequence ( $\hat{S}_1$ ) wins the sequence. An additional constraint was placed on the majority voting scheme. In order to win a sequence,  $\hat{S}_1$  must obtain at least  $0.1N$  votes more than the emotion with the next highest

TABLE IV

FERA-GEMEP DATASET AU TESTING PARTITION (F1-SCORES). BASELINE SCORES [20] ARE INCLUDED IN BRACKETS.  $\mu$  REPRESENTS THE MEAN.

AU	Person-Independent	Person-Specific	Overall
1	<b>0.78</b> (0.63)	<b>0.53</b> (0.36)	<b>0.72</b> (0.57)
2	<b>0.72</b> (0.68)	<b>0.67</b> (0.40)	<b>0.71</b> (0.59)
4	<b>0.43</b> (0.13)	<b>0.64</b> (0.30)	<b>0.52</b> (0.19)
6	<b>0.66</b> (0.85)	<b>0.40</b> (0.26)	<b>0.60</b> (0.46)
7	<b>0.55</b> (0.49)	<b>0.64</b> (0.48)	<b>0.59</b> (0.49)
10	<b>0.47</b> (0.45)	<b>0.55</b> (0.53)	<b>0.50</b> (0.48)
12	<b>0.78</b> (0.77)	<b>0.75</b> (0.69)	<b>0.77</b> (0.74)
15	<b>0.16</b> (0.08)	<b>0.16</b> (0.20)	<b>0.16</b> (0.13)
17	<b>0.47</b> (0.38)	<b>0.30</b> (0.35)	<b>0.41</b> (0.37)
18	<b>0.45</b> (0.13)	<b>0.42</b> (0.24)	<b>0.44</b> (0.18)
25	<b>0.31</b> (0.80)	<b>0.22</b> (0.81)	<b>0.27</b> (0.80)
26	<b>0.54</b> (0.37)	<b>0.27</b> (0.47)	<b>0.43</b> (0.42)
$\mu$	<b>0.53</b> (0.45)	<b>0.46</b> (0.42)	<b>0.51</b> (0.45)

TABLE V

CLASSIFICATION RATE FOR EMOTION DETECTION: TESTING PARTITION. BASELINE SCORES [20] ARE INCLUDED IN BRACKETS.  $\mu$  REPRESENTS THE MEAN.

Emotion	Person-Independent	Person-Specific	Overall
Anger	<b>0.43</b> (0.86)	<b>0.08</b> (0.92)	<b>0.26</b> (0.89)
Fear	<b>0.20</b> (0.07)	<b>0.70</b> (0.40)	<b>0.40</b> (0.20)
Joy	<b>0.75</b> (0.70)	<b>0.09</b> (0.73)	<b>0.52</b> (0.71)
Relief	<b>0.88</b> (0.31)	<b>0.90</b> (0.70)	<b>0.88</b> (0.46)
Sadness	<b>0.87</b> (0.27)	<b>1.00</b> (0.90)	<b>0.92</b> (0.52)
$\mu$	<b>0.62</b> (0.44)	<b>0.55</b> (0.73)	<b>0.60</b> (0.56)

number of votes ( $\hat{S}_2$ ); such that  $\hat{S}_1 - \hat{S}_2 \geq 0.1N$ , where  $N$  represents the number of frames in a test sequence. If  $\hat{S}_1 - \hat{S}_2 < 0.1N$ , then the emotion with the highest mean score is chosen as the sequence winner. Classification rates on the test partition for emotion detection are presented in Table V.

### C. Discussion of Test Partition Results

Our key focus of this work was on the person-independent portion of the challenge. For both the person-independent AU and emotion tasks, our CLM approach achieves much better results over the baseline system [20]. This can be attributed to the fact that better registration can be gained.

Person-specific classifier models were not trained in our experiments, mainly due to the limited amount of training data available. This explains the poor person-specific AU and emotion scores obtained. Understanding this, adapting models to specific subjects using current methods may still be considered as a challenge. Due to the recent Brisbane floods, the tasks of person-specific AU and emotion detection have not been fully investigated.

Detection on AU 25 was notably poor compared to the

baseline system [20]. This was mainly due to effects made by mouth movements, as a result of interference from speech. It was interesting to note that a further 5% improvement in the overall mean AU score could be achieved if AU 25 was excluded from the calculation.

Normally, anger can be characterized by the lowering of the brow (i.e. AU 4), but this was observed not to be solely the case for anger expressed during speech. We noted that substantial amounts of head movement acted as a good indicator for ‘angry speech’. One explanation for the poor detection of anger by our CLM system was that these rigid movements (i.e. translation, rotation, and scale) were effectively normalized in both SAPP and CAPP features.

## VII. CONCLUSIONS AND FUTURE WORK

In this paper we looked at the task of person-independent facial expression detection. A key component of this is to ensure that we register the face accurately, which is difficult to do for unseen subjects. In this paper, the Constrained Local Model (CLM) method was utilized to obtain accurate person-independent tracking and registration. We reported classification rates for both AU and emotion on the GEMEP-FERA dataset, using CLM-derived pixel representations (SAPP and CAPP) with a linear SVM.

Comparisons made between pixel-representations and LBP features demonstrated that when illumination conditions are kept constant, little benefit could be obtained from the features when close to ideal registration had been attained. These findings seemed to suggest that a major function of the LBP operator is to encode shift-invariance.

In future, we hope to extend this work in three aspects. Firstly, AU and emotion detection will be investigated using the temporal nature of the signals. Secondly, emotions expressed simultaneously with speech were observed to co-occur with unique head motion patterns (e.g. rapid nodding of the head with ‘angry speech’). 3D pose parameters (i.e. pitch, yaw, and roll) available from the CLM tracker could identify these patterns, and may improve detection performance when fused with the existing system. Thirdly, the work described in this paper focused on the importance of ensuring accurate alignment, and how spatial invariance may be incorporated once this is achieved. However, classification was performed using only standard SVMs. We hope to apply a novel concept of a “modified correlation filter”, which has shown promising improvements over the SVM in terms of both person-specific and person-independent tasks in preliminary facial expression recognition experiments.

## VIII. ACKNOWLEDGMENTS

This research was supported in part by the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC) and the National Institute of Mental Health grant R01 MH51435.

## REFERENCES

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Fully automatic facial action recognition in spontaneous behavior,” in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 223–228.

[2] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, “Towards practical smile detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2106–2111, 2009.

[3] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *Proceedings of the IEEE Workshop on CVPR for Human Communicative Behavior Analysis*, 2010.

[4] J. Saragih, S. Lucey, and J. Cohn, “Face alignment through subspace constrained mean-shifts,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[5] T. Cootes, G. Edwards, and C. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[6] I. Matthews and S. Baker, “Active appearance models revisited,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[7] S. Lucey, I. Matthews, C. Hu, Z. A. F. de la Torre, and J. Cohn, “Aam derived face representations for robust facial action recognition,” in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, I. Matthews, Ed., 2006, pp. 155–160.

[8] A. Ashraf, S. Lucey, J. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. . Solomon, and B.-J. Theobald, “The painful face: pain expression recognition using active appearance models,” in *Proceedings of the 9th international conference on Multimodal interfaces*. Nagoya, Aichi, Japan: ACM, 2007, pp. 9–14.

[9] A. Asthana, J. Saragih, M. Wagner, and R. Goecke, “Evaluating aam fitting methods for facial expression recognition,” in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009.

[10] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[11] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, “Dynamics of facial expression extracted automatically from video,” *Journal of Image and Vision Computing*, vol. 24, no. 6, pp. 615–625, 2006.

[12] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[13] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[14] T. Ojala, M. Pietikainen, and D. Harwood, “A comparative study of texture measures with classification based on feature distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[15] P. Lucey, S. Lucey, and J. Cohn, “Registration invariant representations for expression detection,” in *International Conference on Digital Image Computing: Techniques and Applications: Techniques and Applications*, 2010, pp. 255–261.

[16] G. Bai, W. Jia, and Y. Jin, “Facial expression recognition based on fusion features of lbp and gabor with lda,” in *2nd International Congress on Image and Signal Processing*, Oct. 2009, pp. 1–5.

[17] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, Jun. 2007, pp. 915–928.

[18] M. Valstar and M. Pantic, “Fully automatic facial action unit detection and temporal analysis,” in *Computer Vision and Pattern Recognition Workshop CVPRW 06*, Jun. 2006, pp. 149–149.

[19] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

[20] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, “The first facial expression recognition and analysis challenge,” in *Proc. IEEE Intl Conf. Automatic Face and Gesture Recognition*, 2011, in print.