In the Pursuit of Effective Affective Computing: The Relationship Between Features and Registration

S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan

***Abstract*—For facial expression recognition systems to be applicable in the real world, they need to be able to detect and track a previously unseen person's face and its facial movements accurately in realistic environments. A highly plausible solution involves performing a "dense" form of alignment, where 60–70 fiducial facial points are tracked with high accuracy. The problem is that, in practice, this type of dense alignment had so far been impossible to achieve in a generic sense, mainly due to poor reliability and robustness. Instead, many expression detection methods have opted for a "coarse" form of face alignment, followed by an application of a biologically inspired appearance descriptor such as the histogram of oriented gradients or Gabor magnitudes. Encouragingly, recent advances to a number of dense alignment algorithms have demonstrated both high reliability and accuracy for unseen subjects [e.g., constrained local models (CLMs)]. This begs the question: Aside from countering against illumination variation, what do these appearance descriptors do that standard pixel representations do not? In this paper, we show that, when close to perfect alignment is obtained, there is no real benefit in employing these different appearance-based representations (under consistent illumination conditions). In fact, when misalignment does occur, we show that these appearance descriptors do work well by encoding robustness to alignment error. For this work, we compared two popular methods for dense alignment—subject-dependent active appearance models versus subject-independent CLMs—on the task of action-unit detection. These comparisons were conducted through a battery of experiments across various publicly available data sets (i.e., CK+, Pain, M3, and GEMEP-FERA). We also report our performance in the recent 2011 Facial Expression Recognition and Analysis Challenge for the subject-independent task.**

***Index Terms*—Active appearance models, automatic facial expression recognition, biologically-inspired appearance descriptors, constrained local models.**

Manuscript received May 15, 2011; revised November 6, 2011 and February 27, 2012; accepted March 6, 2012. Date of publication May 7, 2012; date of current version July 13, 2012. This work was supported in part by the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC) and in part by the National Institute of Mental Health under Grant R01 MH51435. This paper was recommended by Associate Editor M. Pantic.

S. W. Chew and S. Sridharan are with the Speech, Audio, Image and Video Technology Laboratory, Queensland University of Technology, Brisbane, Qld. 4000, Australia (e-mail: sien.chew@qut.edu.au; s.sridharan@qut.edu.au).

P. Lucey and I. Matthews are with Disney Research Pittsburgh, Pittsburgh, PA 15213 USA (e-mail: patrick.lucey@disneyresearch.com).

J. F. Cohn is with the Department of Psychology, University of Pittsburgh/Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15260 USA (e-mail: jeffcohn@cs.cmu.edu).

S. Lucey and J. Saragih are with the Commonwealth Scientific and Industrial Research Organisation, Clayton South, Vic. 3169, Australia (e-mail: simon.lucey@csiro.au; jason.saragih@csiro.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2012.2194485

I. Introduction

RESEARCH INTO affective computing has been very active over the past decade, mainly driven by social, economic, and commercial interests such as marketing, human–computer interaction, health care, security, behavioral science, driver safety, etc. The main goal of this research is to have a computer system being able to automatically detect/infer the emotional state of any person based on various modes (i.e., face, voice, body, and actions) in real time.

The majority of this work has centered on the task of facial expression detection, mostly by way of individual action unit (AU) detection. The predominant approach [1]–[3] to this has been to first locate and track a person's face and facial features, derive a feature representation of the face, and then classify whether a frame contains the AU of interest or not (see Fig. 1). In terms of face alignment [4]–[6], this can be done either coarsely through tracking a couple of key features (i.e., Viola and Jones [7] type approach where the face and eyes are tracked) or highly accurately via a deformable model approach where a dense mesh of 60–70 points on the face is used. The latter is desired due to their accuracy in addition to their ability to infer the 3-D pose parameters (i.e., pitch, yaw, and roll) and viewpoint normalized pixel representations (i.e., synthesize frontal view), which is ideal in situations where there is a lot of head movement, particularly out-of-plane rotations. *Subject-dependent* active appearance models (AAMs) [8], [9] have been widely used in this field [5], [10]–[12] for those reasons, but this approach requires manual labeling of key frames of the training sequence (up to 5% of frames). For applications where the manual labeling of frames is prohibitive (e.g., marketing, security/law enforcement, health care, and HCI), a more generic or *subject-independent* face alignment approach is required. One such approach is the constrained local model (CLM) method [13], [14]. The CLM leverages the generalization capacity of local patch experts and constraints made on the joint deformation, as provided by a point distribution model (PDM). It is similar to AAMs in that it tracks a dense mesh of points on the face that produce both shape and appearance features, but through the utilization of these patches, it has been shown to work well for the subject-independent case (i.e., unseen subjects).

Once the face has been tracked, the normal convention is to apply a bank of filters, followed by a rectification step and, then, a pooling/subsampling strategy.[1] For example, the popular histogram of oriented gradients (HoG) [15]–[18] and Gabor

[1] We have neglected the common step of contrast normalization step here, as we are primarily interested in representations that have invariance to alignment error, not illumination variation.

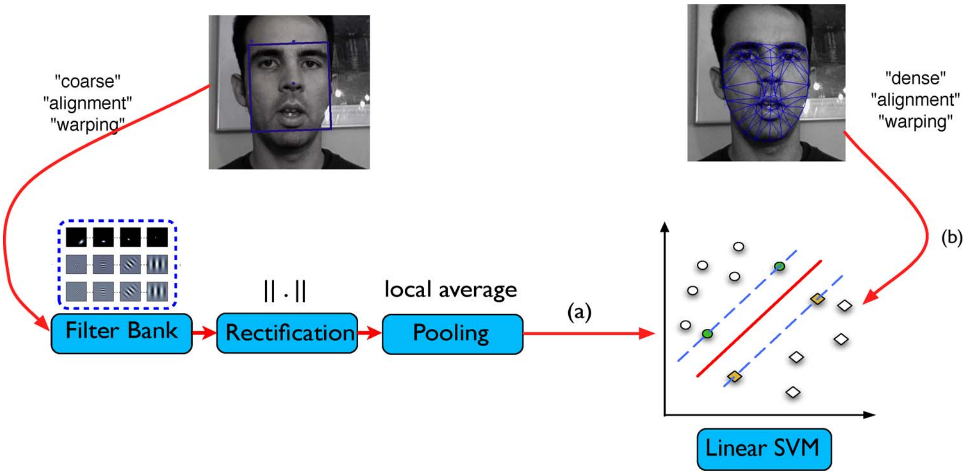


Fig. 1. In this paper, we explore two paradigms for expression detection where a subject's face is (a) coarsely aligned followed by a biologically motivated descriptor (e.g., HoG or Gabor magnitudes) and (b) densely aligned using raw pixels. Both paradigms employ a linear SVM to perform classification.

magnitude [17], [18] descriptors readily fit into this parametric form and have both been successfully applied to expression detection [18] (other exotic variants such as that in [19] are also possible but are outside the scope of this paper). These features have been widely used due to their biological relevance [20], their ability to encode edges and texture, and their invariance to illumination. More recently, Whitehill *et al.* [18] have argued that, for the specific task of expression detection, these features provide a nonlinear classification boundary when using efficient linear classifiers [e.g., linear support vector machine (SVM)]. An inherent problem with this approach, however, is the large memory and computational overheads required for training and testing. Other than cases where there is extreme illumination variation, which is unlikely for the vast majority of applications of this technology at the moment (i.e., the environment should be somewhat constant for health [6], [10] and marketing [21] applications), it begs the question: "*If we have good registration, are appearance descriptors worth the effort?*"

In addition to this central question, this paper specifically addresses other questions which are vital in the quest for effective affective computing.

1) What advantages do these appearance descriptors (i.e., HoG and Gabor magnitudes[2]) have over pixel-based representations for the task of AU detection? When there is close to perfect alignment, is there any benefit in employing these? Does this vary between posed and spontaneous expressions? How does this change when there is poor alignment?
2) What is the difference between subject-dependent (AAM) and subject-independent (CLM) face alignment algorithms in terms of alignment accuracy and AU detection performance?

To quantify the effects ranging across different environments, this paper presents results for a battery of experiments across various data sets which include the posed data set provided by the extended Cohn–Kanade data set (CK+) [12], spontaneous expressions using the UNBC-McMaster shoulder pain archive [6], and the M3 data set [17], in addition to the recent GEMEP-FERA challenge [22]. We also report AU detection results obtained by the CLM in the recent 2011 Facial Expression Recognition and Analysis (FERA2011) Challenge [22] (see Section VIII).

II. Subject-Dependent versus Subject-Independent Deformable Image Alignment Algorithms

As expressions can be subtle, alignment using a deformable model is desired so that the correspondence between various facial features and muscles contracting and controlling the face can be maintained, enhancing the ability of the classifier to detect the facial expression correctly. In addition to this, where there is quite a lot of head movement, particularly out-of-plane rotation, these models can be used to gain the 3-D pose parameters (i.e., pitch, yaw, and roll) [23] and to synthesize a uniform frontal view. In this paper, we will be comparing the *subject-dependent* AAM [8], [9] versus the *subject-independent* CLM [14].

Subject-dependent AAMs are tuned specifically to the subject, camera conditions, and illumination of the target image sequence to be tracked [5], [10], [12] and are able to exhibit "humanlike" accuracy. This tuning is accomplished through the judicious hand labeling of key frames in the target image sequence, where up to 5% of images in a given training sequence need to be manually labeled. In applications in the fields of behavioral science and others where time can be taken to gain an accurate and objective measure, this is a viable solution.

For commercial applications where no enrollment of the subject is possible (e.g., marketing, security/law enforcement, health care, and HCI), a generic or subject-independent face alignment approach is required. Recently, Saragih *et al.* [14] proposed several adaptations to the CLM method, which leverages the generalization capacity of local patch experts and the constraint over joint deformation provided by a PDM. It is similar to AAMs in that it tracks a dense mesh of points on the face that produce both shape and appearance features, but through the utilization of these patches, it generalizes well

[2] In this paper, we selected only HoG and Gabor magnitudes (and not their other variants) for in-depth investigations since these have been heavily employed in the recent expression recognition literature.

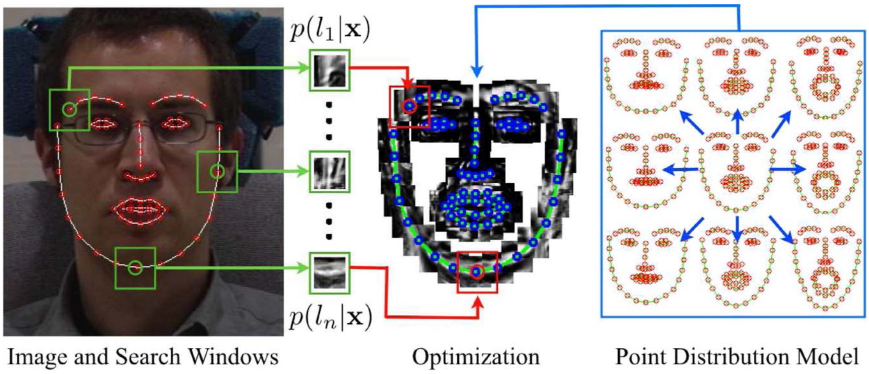


Fig. 2. Illustration of CLM fitting and its two components: (i) Exhaustive local search for feature locations to yield the response maps and (ii) an optimization strategy to maximize the responses of the PDM constrained landmarks.

for the subject-independent case, where the AAM does not. A description of both alignment algorithms is given in the following sections.

A. AAMs

AAMs have been shown to be a good method of aligning a predefined linear shape model that also has linear appearance variation to a previously unseen source image containing the object of interest. In general, AAMs fit their shape and appearance components through a gradient-descent search, although other optimization methods have been employed with similar results [8].

The shape $\mathbf{s}$ of an AAM [8] is described by a 2-D triangulated mesh. In particular, the coordinates of the mesh vertices define the shape $s = [x_1, y_1, x_2, y_2, \ldots, x_n, y_n]$, where n is the number of vertices. These vertex locations correspond to a source appearance image, from which the shape was aligned. Since AAMs allow linear shape variation, the shape $\mathbf{s}$ can be expressed as a base shape $\mathbf{s}_0$ plus a linear combination of m shape vectors $\mathbf{s}_i$

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

where the coefficients $\mathbf{p} = (p_1, \ldots, p_m)^T$ are the shape parameters. These shape parameters can typically be divided into rigid similarity parameters $\mathbf{p}_s$ and nonrigid 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 nonrigid geometric variations associated with determining the object shape (e.g., mouth opening, eyes shutting, etc). Procrustes alignment [8] is employed to estimate the base shape $\mathbf{s}_0$.

Key frames within each video sequence are manually labeled, while the remaining frames are automatically aligned using a gradient descent AAM fitting algorithm described in [9].

B. CLMs

Similarly to the AAM, we want to find the shape $\mathbf{s}$ as in (1), also known as the PDM. CLMs [13] refer to a host of algorithms which utilize an ensemble of local detectors to determine $\mathbf{s}$. All of these methods have the following two goals: 1) Perform an exhaustive local search for each PDM landmark around their current estimate using some kind of feature detector, and 2) optimize the PDM parameters such that the detection responses over all of its landmarks are jointly maximized. Fig. 2 illustrates the components of the CLM fitting.

The particular instance of CLM used in this work is that proposed in [14]. 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 (i.e., $p(l_i|\mathbf{x})$ in Fig. 2). 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. Second, 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 [14] uses an optimization strategy-coined subspace-constrained mean shifts. By assuming that 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 (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. An example of the CLM tracking an unseen face is given in Fig. 3(a).

III. Appearance-Based Features

Under the assumption that there will always be some degree of registration error in a target face image, it is useful to explore features that give invariance to registration. Holistic

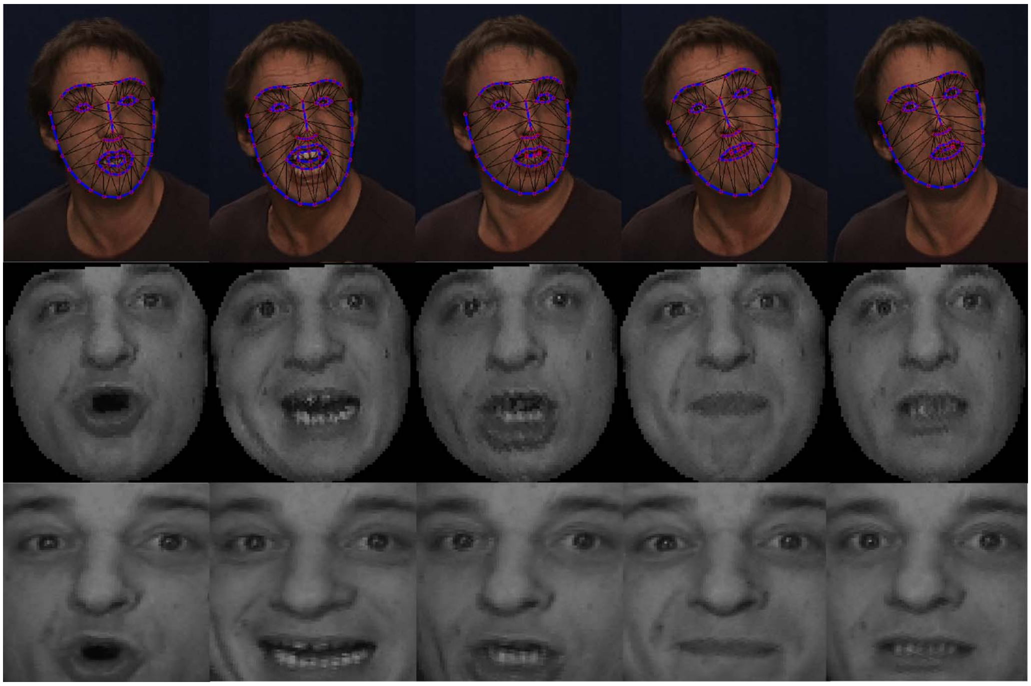


Fig. 3. Example of the CLM tracking a sequence from the GEMEP-FERA data set over time. (a) Once the face is tracked, we extract (b) the canonical normalized appearance features and (c) the similarity normalized appearance features.

invariant features are difficult to derive as one rarely has prior knowledge of how the image geometrically deforms holistically. Instead, it is simpler to adopt a strategy where a single complex holistic deformation in an image, such as those found in facial expressions, can always be broken down into multiple simple deformations (e.g., optical flow, where a single complex deformation can be defined as multiple (one for each pixel) locally constrained translations). Representing an image as a "supervector" of concatenated local region features that are invariant to simple deformations (e.g., translation), an argument can then be made that this supervector will exhibit invariance to more complex holistic registration errors.

Many different techniques for describing local image regions have been proposed in the literature. The simplest feature is a vector of raw pixel-intensity values. However, if an unknown error in registration occurs, there is an inherent variability associated with the *true* (i.e., correctly registered) local image appearance. Due to this variability, an argument can be made that these local pixel appearances are more aptly described by a distribution rather than a static observation point. In addition to the pixel-based representations which we derive from our deformable face alignment algorithm, we investigate two popular methods in vision for obtaining distribution features that exhibit good local spatial invariance: 1) HoG and 2) Gabor magnitudes (GAB).

A. Pixel-Based Representations

Once we have tracked the subject's face by estimating the shape and appearance parameters, we can use this information to derive the following features.

1) **SPTS**: The similarity normalized shape $\mathbf{s}_n$ refers to the 68 vertex points in s_n for both the x- and y-coordinates, resulting in a raw 136-D 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 $\mathbf{s}_n$ can be obtained by synthesizing a shape instance of $\mathbf{s}$, using (1), that ignores the similarity parameters $\mathbf{p}$.
2) **SAPP**: The similarity normalized appearance features $\mathbf{a}_n$ refer to where all the rigid geometric variation (translation, rotation, and scale) has been removed. It achieves this by using $\mathbf{s}_n$ calculated earlier and warps the pixels in the source image with respect to the required translation, rotation, and scale. This is the type of approach that is employed by most researchers [1], [18], 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, and also, some nonfacial information is included due to occlusion. Furthermore, the shape transformation is inherently unknown since substantial variation exists between the shapes of different faces, which therefore makes the z-component of facial points difficult to estimate. An example of this is shown in Fig. 3(c).
3) **CAPP**: The canonical normalized appearance $\mathbf{a}_0$ refers to where all the nonrigid shape variation has been normalized with respect to the base shape $\mathbf{s}_0$. This is accomplished by applying a piecewise affine warp on each triangle patch appearance in the source image so that it aligns with the base face shape. It was shown in [24] that, by removing the rigid shape variation, poor performance was gained. Examples of the CAPP features are shown in Fig. 3(b).

In this paper, we are interested in analyzing the change in performance of the different appearance features (SAPP and

CAPP) between subject-dependent and subject-independent alignment algorithms, as well as across different feature representations (next sections).

B. HoG

HoG [16] is a close relation of the descriptor in Lowe's seminal SIFT approach [15] to code visual appearance. Briefly, the HoG method tiles the input image with a dense grid of cells, with each cell containing a local histogram over orientation bins. At each pixel, the image gradient vector is calculated and converted to an angle, voting into the corresponding orientation bin with a vote weighted by the gradient magnitude. The orientation bins were evenly spaced over 0°–180° (unsigned gradient). Histograms were obtained at different discrete scales using a Gaussian gradient function (in x- and y-) with the variance parameter σ^2 defining the scale. These scale specific histograms are all concatenated into a single feature vector. Shift invariance is naturally encoded in this type of feature through the size of the cell from which the histograms are derived. The larger the cell size, the greater the shift invariance. In this paper, we used a cell size of 12 × 12 over three frequencies and four rotations.

C. Gabor Magnitudes

A 2-D Gabor function is a complex exponential modulated by a Gaussian envelope

$$g_{\omega,\theta}(x,y) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{x'^2+y'^2}{2\sigma^2} + j\omega x'\right\} \tag{2}$$

where $x' = x\cos(\theta) + y\sin(\theta)$, $y' = -x\sin(\theta) + y\cos(\theta)$, x and y denote the pixel positions, ω represents the frequency of the Gabor wavelet, θ represents the orientation of the Gabor wavelet, and σ denotes the standard deviation of the Gaussian function (please refer to [25] on strategies for spacing the filters in the 2-D spatial frequency domain for a fixed number of scales and orientations). These filters are in quadrature where the real part of the filter is even symmetric and the imaginary part of the filter is odd symmetric. When convolved with an input image, the scalar magnitude value of the resultant complex response can be interpreted as the correlation matrix (i.e., distribution) of the local region (defined by σ) for the image components resonating with the central frequency (defined by $\mathbf{k}$) in the direction of θ. Like HoG features, the magnitude values for each orientation and central frequency are concatenated into a vector. In this paper, we use eight different orientations and eight different frequencies and employ AdaBoost to select the top 8% of the most discriminant features as a subset of the entire Gabor feature space for training and testing. Please note that these optimal parameters had been selected for both Gabor magnitudes and HoG during preliminary experiments.

IV. Geometric Invariance via Descriptors

A laundry list of features/descriptors has now been proposed for in computer vision literature for a myriad of matching/classification tasks including expression classification. Biologically inspired descriptors such as HoG and Gabor magnitudes have proved successful in recent state-of-the-art expression detection algorithms [17], [18]. As pointed out by Lecun *et al.* [26], these biologically inspired features/descriptors all share a common parametric form. This parametric form has it roots in the seminal work of Hubel and Wiesel [20] involving the study of the mamallian primary visual cortex (i.e., V1). Typically, an input image is passed through a bank of filters, followed by a rectification step, contrast normalization, and, then, a pooling/subsampling strategy. For example, the very popular HoG [15]–[18] and Gabor magnitude descriptor [17], [18] readily fit into this parametric form. Recently, variants of this parametric have been explored [19], [27], [28], with impressive performance being obtained for a number of vision classification tasks. For example, Serre *et al.* [19] demonstrated that the tuning properties of a majority of cortical cells in the visual cortex could be captured by selecting the parameter values that correspond to a host of different visual stimuli.

The performance of human vision is obviously far superior to that of current computer vision systems at the moment, so there is a valid argument to be had in emulating biological processes. However, this explanation is largely unsatisfying from an engineering perspective for understanding why these features are useful. As Berg and Malik [29] elegantly point out, one useful consequence of treating the positive and negative components of oriented edge responses separately (or rectifying them) is that information about zero crossings is not lost under blurring. Instead of blurring the signal response around a zero crossing to zero, the positive and negative responses are both blurred over the area, retaining the information that there was a zero crossing but allowing uncertainty as to its position. This nonlinear process enables an encoding that is able to handle much greater tolerance than traditional pixel representations to geometric misalignment. Lecun *et al.* [26] refer to this blurring process more generically as "pooling," and pooling operations other than blurring (e.g., taking the maximum of a local spatial cell) have been explored in [19], [27], and [28].

V. Experimental Setup and Data Sets

A. Experimental Setup

In this paper, all experiments were for the task of subject-independent AU detection. These experiments were to do the following: 1) Investigate the role of biologically inspired features (Gabor and HoG features) across various levels of registration accuracy and compare them to pixel representations, and 2) compare the registration accuracy between subject-dependent (AAM) and subject-independent face registration algorithms and their subsequent performance for AU detection. To facilitate this, we conducted these experiments across four common facial expression data sets (see Section V-B).

Once we tracked the face and extracted representative facial features (see Sections II and III), the classification of these AUs was performed via a linear SVM. SVMs are an effective method for AU classification, and they are used in many facial expression systems [18], [30]–[32]. In this paper, we used a one-versus-all linear two-class SVM (i.e., AU of interest versus non-AU of interest) in all experiments. For the training of the SVMs, all frames which were manually labeled by expert FACS

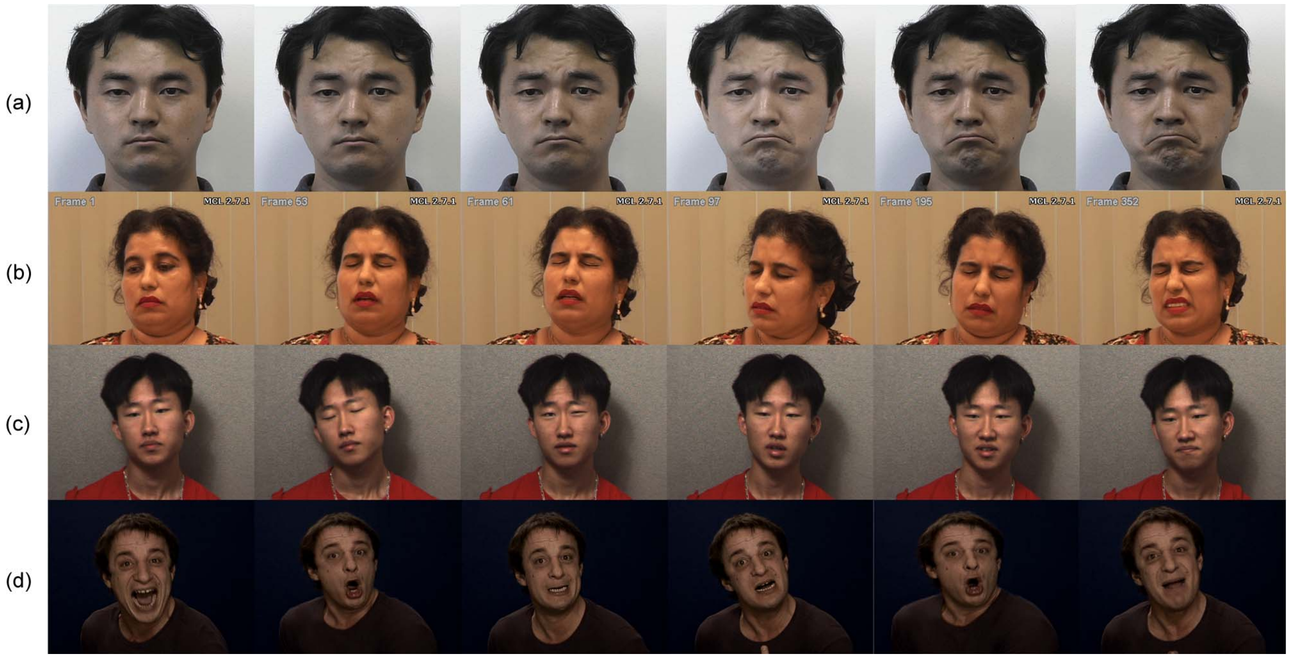


Fig. 4. Examples of the four data sets that we used in this paper. (a) CK+. (b) UNBC-McMaster shoulder pain archive. (c) M3 database. (d) GEMEP-FERA data set.

coders to contain the AU of interest were used as positive examples, regardless whether it occurred with other AUs or alone. The frames that did not have the AU of interest in them were used as negative examples.[3]

Training and testing were conducted using a leave-one-subject-out strategy, so as to maximize the amount of training and testing data. It is also worth noting that all AAM experiments were subject-dependent (i.e., approximately 5% of all images in a given training sequence were used to train an AAM to track that sequence). The CLM used in these experiments had NOT seen any of the images in any of the data sets (i.e., completely generic).

In order to predict whether a video frame contained an AU, the output score from the SVM was used. As there are many more frames with no behavior of interest than frames containing a behavior of interest, the overall agreement between correctly classified frames can skew the results somewhat. As such, we used the receiver-operator characteristic (ROC) curve, which is a more reliable performance measure. This curve is obtained by plotting the hit rate (true positives) against the false alarm rate (false positives) as the decision threshold varies. From the ROC curve, we used the area under the ROC curve (A'), to assess the performance. The A' metric ranges from 0.5 (pure chance) to 1 (ideal classification). An upper bound on the uncertainty of the A' statistic was obtained using the formula $s = \sqrt{A'(100 - A')/\min\{n_p, n_n\}}$, where n_p and n_n are the number of positive and negative examples, respectively [18], [33]. We chose this approach over the F1 metric as the latter relates only to the maximum F1 score on the precision-recall curve which does not give an indication of the generalized performance for different thresholds.[4]

B. Data Sets

The Extended Cohn–Kanade Database: In this paper, we used the extended Cohn–Kanade (CK+) database [12], which contains 593 sequences from 123 subjects. The image sequences vary in duration (i.e., 10 to 60 frames) and incorporate the onset (which is also the neutral frame) to peak formation of the facial expressions [see Fig. 4(a)]. For the 593 posed sequences, full FACS coding of the peak frames is provided. Approximately 15% of the sequences were comparison coded by a second certified FACS coder. Interobserver agreement was quantified with coefficient kappa, which is the proportion of agreement above what would be expected to occur by chance [34]. The mean kappas for interobserver agreement were 0.82 for AUs coded at apex and 0.75 for frame-by-frame coding. An inventory of the AUs that we used in this experiment is given in Table II.

UNBC-McMaster Shoulder Pain Archive: The UNBC-McMaster Shoulder Pain Expression Archive contains video of the faces of adult subjects (129 subjects—63 males and 66 females) with rotator cuff and other shoulder injuries. In the portion released by Lucey *et al.* [6], 200 video sequences spanning 25 subjects were recorded of their faces while they moved their affected (these subjects had various shoulder injuries) and unaffected shoulders. In this data set, considerable head movement occurs during the sequence, and the video sequences have various durations, with sequences lasting from 90 to 700 frames. Within these sequences, the patient may display various expressions multiple times (of which all AUs had been fully

[3] There is the following exception: If similar AUs occurred, these were not used in the negative example pool. Please refer to Table I for our AU pooling strategy. Please also note that this intuitive strategy has not been empirically proved to provide optimal AU detection performance.

[4] We, however, did use the F1 metric in reporting our AU performance in the FERA2011 challenge.

TABLE I

IN SELECTING NEGATIVE TRAINING EXAMPLES, THE AUs WHICH ARE IN CLOSE PROXIMITY WITH ONE ANOTHER HAD BEEN CATEGORIZED INTO THE FOLLOWING POOLS AS THE DIFFERENCES BETWEEN THEM ARE QUITE SUBTLE (E.G., FOR AU1, WE DID NOT INCLUDE ANY FRAMES WHICH HAD AU2 NOR AU4 IN THE NEGATIVE POOL)

AU Pool	Facial Region	AUs Involved
1	*upper face*	1−2−4
2	*middle face*	6−7−9−10
3	*lower face*	12−15−17−18−25−26

TABLE II

AU INVENTORY OF THE NUMBER OF INSTANCES OF AUs THAT WERE PRESENT IN THE VARIOUS DATA SETS USED IN THIS PAPER

AU	CK+	Pain	M3	GEMEP
1	173	–	16319	1600
2	116	–	13722	1631
4	191	1074	2204	1356
6	122	5557	7980	1808
7	119	3366	7980	2123
9	–	423	–	–
10	–	525	5471	2034
12	111	6887	28017	2725
15	89	–	4232	1026
17	196	–	8383	822
18	–	–	–	419
25	287	2407	28865	812
26	48	2093	19782	499
43	–	2434	–	–

FACS coded). In total, there were over 48 000 frames used. An inventory of the AUs used for these experiments is also given in Table II. An example of the data set is given in Fig. 4(b). For full details of this freely available data set, please see [6].

M3 Database: The spontaneous M3 [17] facial expression database was recorded from a hundred participants of a false-opinion paradigm. This paradigm proved effective at evoking a plethora of emotion-related facial expressions, where subjects first fill out a questionnaire regarding their opinions about a social or political issue and then attempted to deceive an experienced interviewer for monetary gains. For each subject, approximately 2 min of video was recorded, which was coded by two separate certified FACS coders. As these expressions are spontaneous, a mixture of AUs tend to overlap with one another (i.e., co-occurrence), at varying levels of intensity. Other facets of the spontaneous nature include speech-related mouth movements and out-of-plane head rotations. An example of the data set is given in Fig. 4(c). Ground-truth FACS coding was provided by expert coders. Data from 28 of the subjects were available for our experiments. In particular, we divided this data set into 17 subjects for training (97 000 frames) and 11 subjects for testing (67 000 frames). For the other three data sets, we conducted our experiments using a leave-one-subject-out strategy for training and testing. An inventory of the AUs used in this experiment is given in Table II.

1) GEMEP-FERA Database: The GEMEP-FERA database [22] contains audiovisual recordings of ten actors expressing a total of 15 emotions together with a variety of AUs which had been FACS coded. In all of these recordings, the actors were instructed to utter meaningless phrases (such as the sustained vowel "aaa") with the aid of a professional director. The key difference between this data set and the CK+ and UNBC-McMaster shoulder pain archive is that expressions had been displayed in the presence of speech, which generated a substantial amount of rigid head and body motion. An example of the data set is given in Fig. 4(d). In our experiments, we focused on the following AUs: $\{1\ \ 2\ \ 4\ \ 6\ \ 7\ \ 10\ \ 12\ \ 17\ \ 18\ \ 25\ \ 26\}$. The number of these AUs are given in Table II.

VI. EXPERIMENT I: THE ROLE OF FEATURES

We had two main interests here: 1) comparing different AAM pixel representations across noise levels (SAPP versus CAPP) and 2) comparing these pixel representations against biologically inspired features (i.e., HoG and Gabor magnitudes). To facilitate these goals, we added various amounts of geometric noise to the test images. To do this, the similarity normalized base template had an interocular distance of 50 pixels. For a fair comparison, we took into account differing face scales between testing images. This is done by first removing the similarity transform between the estimated shape and the base template shape and then computing the root-mean-squared pixel-error (RMS-PE) between the 68 points. We obtained the poor initial alignment by synthetically adding affine noise to the ground-truth coordinates of the face. We then perturbed these points with a vector generated from white Gaussian noise. The magnitude of this perturbation was controlled to give a desired RMS-PE from the ground-truth coordinates (which were the AAM-tracked landmarks). During learning, the initially misaligned images were defined to have between 5–30 RMS-PE. This range of perturbation was chosen as it approximately reflects the range of alignment error that can be experienced using subject-independent face alignment algorithms. Examples of the poor tracking are given in Fig. 5. In our experiments, all training images were clean (i.e., zero noise), and they were tested across different noise levels (i.e., 5–30 RMS-PE). After all images were registered, they were downsampled to 48×48 pixels. As can be seen in Fig. 5(c), when the amount of noise is increased, the piecewise affine warp which synthesizes the CAPP image causes significant deformation to the face [observe the lip area in Fig. 5(c)] which is a much noisier representation than the SAPP image. As this is the case, all HoG and Gabor features were calculated on the SAPP pixel representations. The results for these experiments are given in Fig. 6.

As can be seen, there is a gradual drop-off in performance as the amount of noise is increased across all the data sets (a–d). The first thing to note is the performance of both the AAM representations $(\text{SAPP} = \text{black},\ \text{CAPP} = \text{blue})$. As observed from the two pixel representations, the performance is very similar, so there is little difference between these two at the zero-noise condition, but it was observed that the performance of CAPP appeared to deteriorate more rapidly than that of

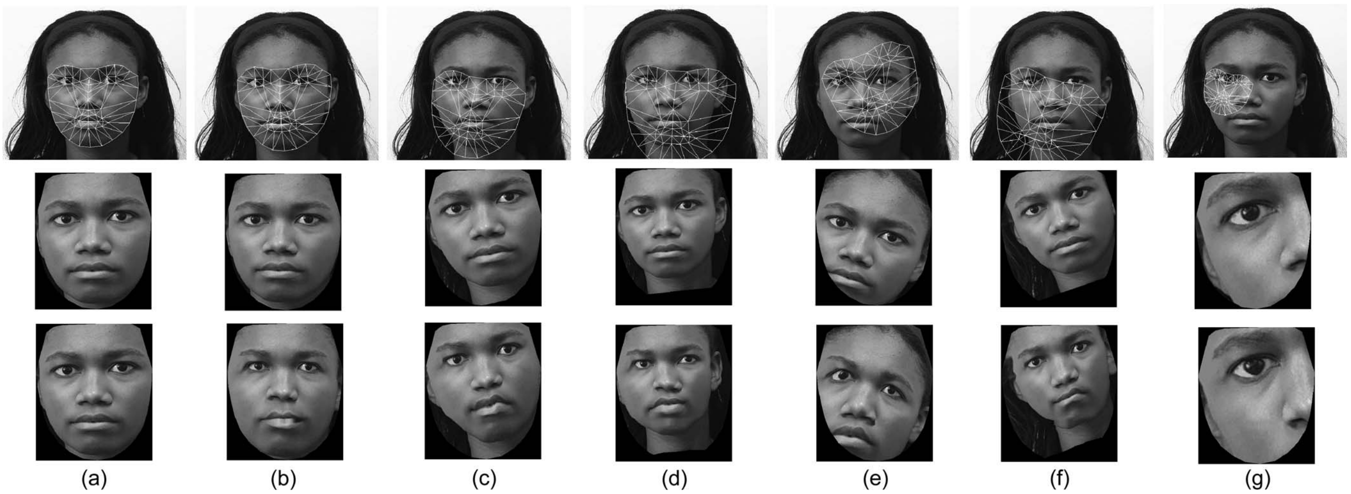


Fig. 5. In our experiments, we compared the (row 2) SAPP and (row 3) CAPP features from the AAM across various geometric noise levels, which is symptomatic of poor registration in subject-independent algorithms: (a) Ideal tracking, (b) 5 RMS-PE, (c) 10 RMS-PE, (d) 15 RMS-PE, (e) 20 RMS-PE, (f) 25 RMS-PE, and (g) 30 RMS-PE. From this, it can be seen when the amount of noise is increased. The piecewise affine warp which synthesizes the CAPP image causes significant deformation to the face which is a much noisier representation than the SAPP image.

SAPP on the CK+ and Pain experiments. What is interesting though is that, when the amount of noise was increased, the biologically inspired features outperformed the pixel representations (particularly for the CK+ [see Fig. 6(a)] and M3 data sets [see Fig. 6(c)]). This supports the reason for the combination of coarse registration and a biologically inspired descriptor is widely used in the literature [17]. As the amount of head motion present in the majority of applications that this system is applied on can be considered to be limited, having a coarse registration (noise from 0–15 RMS-PE) would produce only slight degradation in performance when these features are employed.

VII. Experiment II: AAM Versus CLM

A. Comparing Alignment Accuracy

In comparing the alignment accuracy of both the AAM and CLM to manually landmarked images, we first normalized all tracked AAM and CLM points and manual landmarks for similarity to a common mesh size and rotation, with an interocular distance of 50 pixels and aligned to the center of the eye coordinates. For the CK+ database, we compared against 393 manually landmarked images; for the UNBC-McMaster Pain database, we compared against 2584 manually landmarked images; for M3, we compared against 1990 images; and, for the GEMEP, we compared against 963 manually landmarked images. The alignment curves are given in Fig. 7. As can be seen for the CK+ data set [see Fig. 7(a)], nearly all of the AAM landmarks are within 2-pixel root-mean-square (rms) error of the manual landmarks, which is negligible when one considers that this is based on a distance of 50 pixels between the center of the eyes. The CLM is within 5 pixels which is also very accurate.

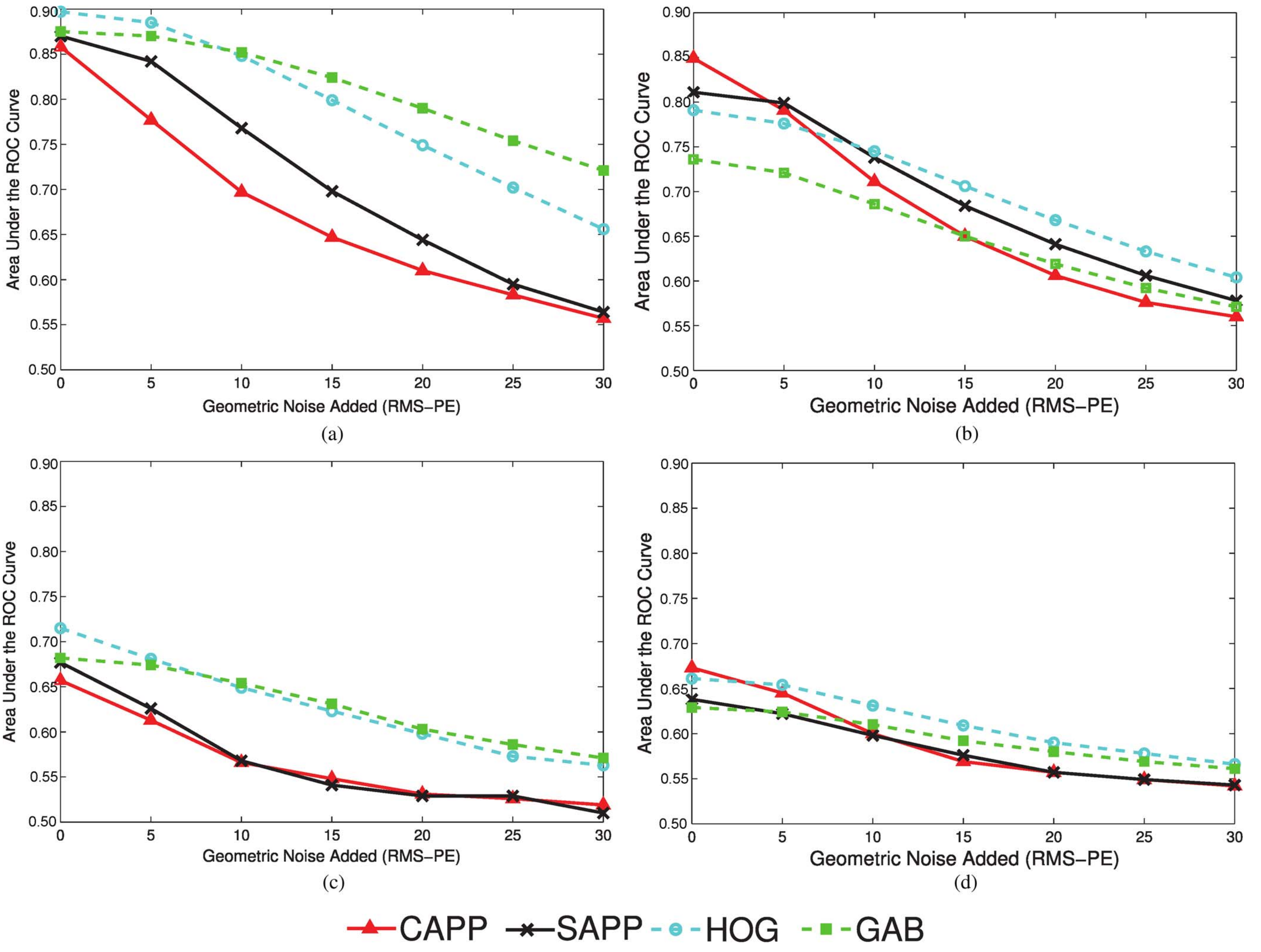
For the more visually complex data sets such as the UNBC-McMaster Pain Archive [see Fig. 7(b)], the AAM performed very well while the CLM performance was not as good which highlights the benefit of a subject-dependent approach. However, as the majority of images were tracked within 10 pixels, which is a reasonable result, considering the relatively significant quantities of head motion in the data set, similar findings can be found for both the M3 [see Fig. 7(c)] and GEMEP-FERA [see Fig. 7(d)] data sets. Using the results in the previous section, it can be seen that, even though there was a drop-off in performance across these noise levels, the discrepancy between the different pixel representations and HoG and Gabor features would be minimal.

B. Comparing AU Detection Performance

The experiments in Section VI were conducted using subject-dependent AAMs which provide ideal face registration. These subject-dependent AAMs were tuned specifically to a particular subject to counter for high appearance variabilities such as illumination, pose, and camera conditions. The drawback of this is that alignment accuracy deteriorates once the target population is large, and having to learn specific models for each available subject becomes infeasible. On the other hand, subject-independent CLMs are well suited to handle the problem of subject dependence as they are able to generalize well to unseen subjects. The tradeoff, however, is that CLMs exhibit a deterioration in alignment accuracy as compared to AAMs. The experiments in this section evaluate the role of features in CLM-derived pixel representations and determine if the implementation of such features could improve AU detection performances to ideal AAM levels.

Similar to the synthetic experiments in Section VI, PIX (CAPP) was compared against HoG and GAB across all four data sets, but this time, only clean test images were utilized. Experimental results illustrated in Fig. 8 once again suggest that little benefit could be obtained from utilizing HoG and GAB on CLM-derived pixel representations at 0 RMS-PE. Although certain appearance descriptors may be relatively "cheap" to compute (e.g., local binary pattern operators), these analyses provide additional insights into their fundamental operations and could serve as a platform to inspire future methods in either appearance descriptors or dense facial alignment methods.

For AU detection (see Fig. 9), the performance of the CLM was very similar (within 3%) to the AAM in all but the



(a)

(b)

(c)

(d)

Fig. 6. Plots showing the average AU detection performance across all noise levels for the pixel (PIX), HoG, and Gabor features (GAB) for the following data sets: (a) CK+, (b) the UNBC-McMaster Pain Archive, (c) M3, and (d) GEMEP-FERA. Detection performance was evaluated using the weighted mean A′ proportional to the number of positive examples (i.e., the higher the number of positive examples, the higher the weighting).

UNBC-McMaster Pain Archive. In this archive, there was substantially more rigid head motion as compared to the other three data sets, which indicates that the CLM does provide similar face alignment performances to the AAM when out-of-plane rigid head motion is kept minimal, but it is still outperformed by the AAM in aligning the face from a synthesized frontal view.

VIII. FERA2011 Challenge Results

Motivated by these findings, we decided to participate in the recent FERA2011 challenge [22] as an opportunity to evaluate our CLM system. In this challenge, participants were assigned the task of automatically recognizing a total of 12 AUs (AUs 1 2 4 6 7 10 12 15 17 18 25 26) from the GEMEP-FERA testing partition. Here, half of the test subjects were the same as the subjects in the training partition. Our CLM (CAPP) AU detector was trained on using examples from the GEMEP-FERA training partition and the CK+ data set based on a leave-one-subject-out strategy. AU detection thresholds were obtained where the maximum F1 scores occurred in the respective precision-recall curves. The AU classification rates achieved in the challenge are presented in Table III.

In comparison to the baseline system [22] (which employed local binary pattern operators in combination with an RBF kernel SVM), our CLM system achieved much better results. The poor detection of AUs 6 and 25, however, was somewhat puzzling. The poor detection of AU 25 could be attributed to poor tracking of the mouth region, particularly on the lips (since subjects were constantly speaking). Registration error of key points at the mouth region could thus propagate through to the cheek region (AU 6) and subsequently cause incorrect pixel warping due to the poorly tracked mouth region, hence explaining the poor detection of AU 6.

IX. Conclusion

The field of affective computing is fast maturing (the recent FERA2011 challenge is indicative of this), and fully automatic facial expression recognition systems will be a reality soon. However, a major hurdle in the pursuit of achieving this is the accuracy and robustness of a generic (i.e., subject-independent) face and facial feature tracker where a large number (e.g., 60 or 70) of fiducial points on the face image are accurately detected. This paper illustrated that, with the advent of subject-independent face alignment methods such as CLMs, this goal

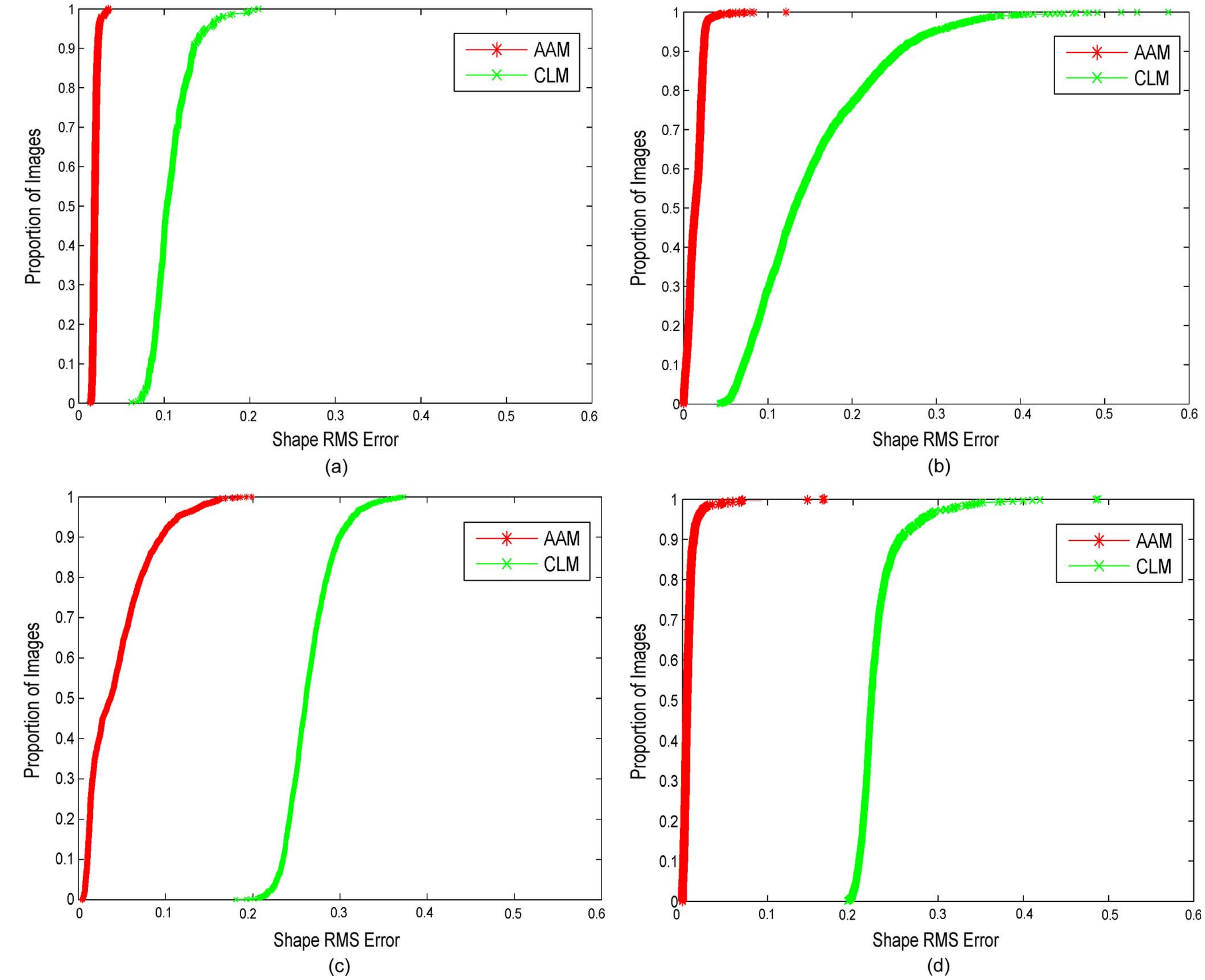


Fig. 7. Fitting curves comparing the registration accuracy between the AAM and CLM across the (a) CK+, (b) UNBC-McMaster Pain Archive, (c) M3, and (d) GEMEP-FERA data sets. Shape rms error is presented as a ratio with respect to interocular distance (50 pixels).

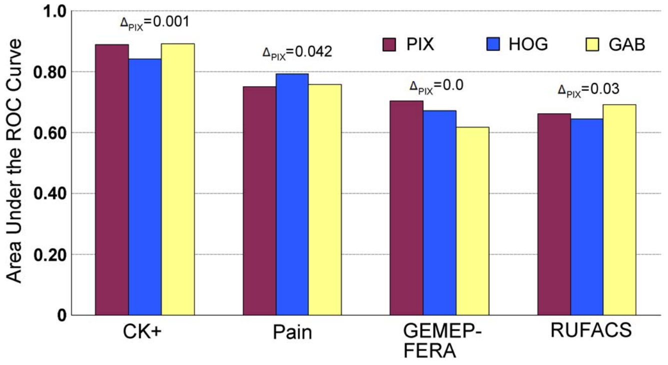


Fig. 8. Comparison of CLM-derived pixel representations (PIX) versus HoG and Gabor magnitudes (GAB) in AU detection on the four facial expression data sets. Detection performance was evaluated using the weighted mean A′ scores averaged across all AUs. Δ_{PIX} represents the amount of benefit gained from using either HoG or GAB features over PIX.

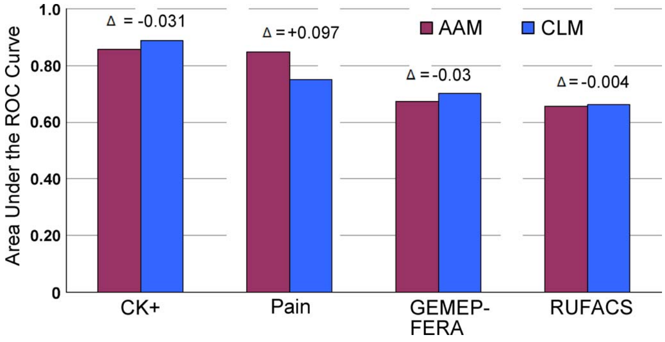


Fig. 9. Comparison of the overall performance between subject-dependent AAMs versus subject-independent CLMs in AU detection performance on the four facial expression data sets. Detection performance was evaluated using the weighted mean A′ scores averaged across all AUs. Δ represents the difference in detection accuracy between the AAM and CLM.

is becoming closer as the accuracy that it can achieve is comparable to that of the subject-dependent AAMs. With this high accuracy achieved, it was also shown that the benefit of employing biologically inspired features over pixels is nullified (given that illumination conditions are known and reasonably consistent) as these features essentially provide shift invariance which is not required when close to ideal registration is achieved. This was demonstrated over four publicly available

TABLE III
GEMEP-FERA DATA SET AU TESTING PARTITION RESULTS (F1 SCORES) ACHIEVED BY THE CLM IN THE FERA2011 CHALLENGE. BASELINE SCORES [22] ARE ALSO SHOWN. μ REPRESENTS THE MEAN

AU	CLM System	Baseline System
1	**0.78**	0.63
2	**0.72**	0.68
4	**0.43**	0.13
6	**0.66**	0.85
7	**0.55**	0.49
10	**0.47**	0.45
12	**0.78**	0.77
15	**0.16**	0.08
17	**0.47**	0.38
18	**0.45**	0.13
25	**0.31**	0.80
26	**0.54**	0.37
μ	**0.53**	0.45

data sets, and motivated by these results, the usefulness of the CLM was also demonstrated in the recent FERA2011 challenge which performed very well in the subject-independent section. This bodes well for the future of this technology as it shows that we are getting closer to the lofty goal of having "effective" affective computing.

Future work will look into the problem of making the classifier invariant in the temporal domain which has the potential to improve AU and expression detection performance. Once these areas can be fully explored and quantified, a better understanding on which approach can be best used for a particular application can be made.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contribution of the National Research Organization and reviewers' comments.

REFERENCES

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[2] T. Wu, M. Bartlett, and J. Movellan, "Facial expression recognition using Gabor motion energy filters," Machine Perception Laboratory, Univ. California, San Diego, CA, Tech. Rep..

[3] A. Ashraf, S. Lucey, and T. Chen, "Re-interpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1335–1341, Jul. 2010.

[4] J. Whitehill and C. Omlin, "Haar features for FACS AU recognition," in *Proc. 7th IEEE Int. Conf. FG*, 2006, pp. 97–101.

[5] 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 *Proc. Int. Conf. Autom. Face Gesture Recog.*, I. Matthews, Ed., 2006, pp. 155–160.

[6] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "PAINFUL DATA: The UNBC-McMaster shoulder pain expression archive database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2011, pp. 57–64.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, 2001, vol. 1, pp. 511–518.

[8] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[9] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, Nov. 2004.

[10] A. Ashraf, S. Lucey, J. Cohn, K. M. Prkachin, and P. Solomon, "The painful face II—Pain expression recognition using active appearance models," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1788–1796, Nov. 2009.

[11] A. Asthana, J. Saragih, M. Wagner, and R. Goecke, "Evaluating AAM fitting methods for facial expression recognition," in *Proc. Int. Conf. Affective Comput. Intell. Interact.*, 2009, pp. 1–8.

[12] 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 *Proc. IEEE Workshop CVPR Human Commun. Behav. Anal.*, 2010, pp. 94–101.

[13] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *Proc. Brit. Mach. Vis. Conf.*, 2006, vol. 3, pp. 929–938.

[14] J. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. ICCV*, 2009, pp. 1034–1041.

[15] D. G. Lowe, "Distinctive image features from scale—Invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.

[17] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, Sep. 2006.

[18] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Towards practical smile detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2106–2111, Nov. 2009.

[19] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[20] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.

[21] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 500–508, Jun. 2006.

[22] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Pro. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2011, pp. 921–926.

[23] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2-D +3-D active appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, pp. 535–542.

[24] A. Ashraf, S. Lucey, J. Cohn, T. Chen, Z. Ambadar, and K. Prkachin, "The painful face: Pain expression recognition using active appearance models," in *Proc. 9th Int. Conf. Multimodal Interfaces*, Nagoya, Japan, 2007, pp. 9–14.

[25] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer. A*, vol. 4, no. 12, pp. 2379–2393, Dec. 1987.

[26] Y. Lecun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. ISCAS*, 2010, pp. 253–256.

[27] N. Pinto and D. Cox, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2011, pp. 8–15.

[28] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox, "A high-throughput screening approach to discovering good forms of biologically inspired visual representation," *PLoS Comput. Biol.*, vol. 5, no. 11, p. e1000579, 2009. DOI: 10.1371/journal.pcbi.1000579.

[29] A. Berg and J. Malik, "Geometric blur for template matching," in *Proc. Comput. Vis. Pattern Recog.*, 2001, pp. 607–614.

[30] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *Proc. Int. Conf. Autom. Face Gesture Recog.*, 2006, pp. 223–228.

[31] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *J. Image Vis. Comput.*, vol. 24, no. 6, pp. 615–625, Jun. 2006.

[32] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Proc. CVPRW*, Jun. 2006, p. 149.

[33] C. Cortes and M. Mohri, "Confidence intervals for the area under the ROC curve," *Proc. Adv. Neural Inf. Process. Syst.*, pp. 305–312, 2004.

[34] J. Fleiss, *Statistical Methods for Rates and Proportions*. New York: Wiley, 1981.

Authors' photographs and biographies not available at the time of publication.