Identifying User-Specific Facial Affects from Spontaneous Expressions with Minimal Annotation

Michael Xuelin Huang, Grace Ngai, Kien A. Hua, *Fellow, IEEE,* Stephen C.F. Chan, *Member*, *IEEE* and Hong Va Leong, *Member*, *IEEE Computer Society*

Abstract—This paper presents PADMA (Personalized Affect Detection with Minimal Annotation), a user-dependent approach for identifying affective states from spontaneous facial expressions without the need for expert annotation. The conventional approach relies on the use of key frames in recorded affect sequences and requires an expert observer to identify and annotate the frames. It is susceptible to user variability and accommodating individual differences is difficult. The alternative is a user-dependent approach, but it would be prohibitively expensive to collect and annotate data for each user. PADMA uses a novel Association-based Multiple Instance Learning (AMIL) method, which learns a personal facial affect model through expression frequency analysis, and does not need expert input or frame-based annotation.

PADMA involves a training/calibration phase in which the user watches short video segments and reports the affect that best describes his/her overall feeling throughout the segment. The most indicative facial gestures are identified and extracted from the facial response video, and the association between gesture and affect labels is determined by the distribution of the gesture over all reported affects. Hence both the geometric deformation and distribution of key facial gestures are specially adapted for each user. We show results that demonstrate the feasibility, effectiveness and extensibility of our approach.

Index Terms—Facial affect detection, weakly supervised learning, user-dependent model

1 Introduction

AFFECTIVE computing has attracted a great deal of attention in recent years, and analysis of facial expression is considered to be one of the most effective approaches for automated recognition and interpretation of human affect [1]. However, despite prior successes, applying this work in real-use situations is still difficult because of natural differences among individual users, especially for spontaneous expressions.

Much research in this area focuses on training a generic, or user-independent, facial expression interpretation model that fits the majority of users. The conventional approach uses supervised machine learning [1], which requires a "gold-standard" data set annotated by human experts [2]. The assumption is that when the training dataset is large enough, the machine learning algorithm will be able to recognize and discriminate between different facial expressions. However, individual differences of facial appearance, ethnicity, culture, personality and preference all affect the performance of the user-independent model. The same facial expression might indicate dissimilar affects for different persons. Evidence shows that applying an affect model trained on one dataset to another dataset results in a significant performance drop [1][3][4]. As devices become increasingly mobile and personal, a user-dependent approach seems increasingly reasonable, and would be effective for addressing individual differences. However, the annotation effort required would be too time-consuming and expensive to be feasible with the conventional approach.

Many previous approaches model affects based on simulated, or posed, expressions from actors and actresses in near frontal view [1]. However, there are significant differences between posed and spontaneous, naturally experienced expressions, as has been reported in previous work [5][6][7]. Other real-use issues such as out-of-plane head rotation [8] and illumination variations also make the recognition of affects from spontaneous expressions more challenging. We therefore see two major challenges to the application of facial affect identification in real use. The first is accommodating user differences, especially for spontaneous expressions. The second is collecting and annotating enough data.

Related research efforts have focused on three main approaches. User-specificity through transfer learning [9][10] assumes that the distribution or characteristics of the expressions and/or affects in the target dataset reflect those from the generic dataset This assumption may not

- *Michael Xuelin Huang, Grace Ngai and Stephen C.F. Chan are with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. E-mail: {csxhuang, csgngai, csschan}@comp.polyu.edu.hk.*
- *Kien A. Hua is with the School of Electrical Engineering and Computer Science, Univerity of Central Florida, Orlando, FL 32816. E-mail:kienhua@eecs.ucf.edu.*

be valid due to individual differences. Active learning selects a relatively small portion of discriminant samples for annotation so as to reduce human annotation effort [11]. Bootstrapping has the same objective, for instance, by manually labeling a few essential frames, such as the apex, and extrapolating to the rest by identifying similar frames [2][12]. However, human expertise is still required to identify the needed key frames.

Multiple instance learning (MIL) models the data from coarse-grained bag-level (segment-level) annotation. MIL usually assumes that a bag (segment) is positive if it contains at least one positive instance [13]. This is valid for many binary classification problems; however, in multi-class facial affect modeling, it is not uncommon to have instances with different (affect) labels manifesting within a 1-2 minute segment (bag), or even for complex mixed feelings to occur [14]. Some common affects, such as "neutral", may also occur frequently in a bag that is labeled as something other than neutral.

We propose an approach for Personal Affect Detection with Minimal Annotation (PADMA) that uses a novel association-based multiple instance learning (AMIL) approach. In contrast to conventional MIL methods, AMIL assumes that if an instance occurs frequently in bag(s) labeled with one particular class, but not in others, the instance has a strong association with that label.

PADMA relies on facial features similar to action units (AUs), which describe the visual effects from facial muscle movement defined in the Facial Action Coding System (FACS) [15]. Similar expressions are clustered and key facial gestures extracted. AMIL is then used to correlate key facial gestures (defined as a short sequence of facial behavior over multiple consecutive frames [16]) with user-reported affects to obtain the fine-grained affect labels based on the distribution of the facial gestures. PADMA therefore adaptively extracts and annotates key facial gestures for a user, according to his/her actual response. Our challenge comes from identifying detailed facial gestures and their implications, given only rough overall self-reported information.

The contributions of this paper are as follows. We (1) propose a novel adaptive clustering approach to encode facial response data from individual users; (2) devise a novel AMIL method that automatically identifies key facial gestures from spontaneous expressions and associates them with human affects, given only segment-level labels; (3) show the effectiveness of our method in modeling user-dependent, spontaneous facial affects and demonstrate its superiority compared to its user-independent counterparts. Our approach has two advantages: (1) it does not require expert annotation, and (2) it automatically accommodates user differences.

2 Related Work

A variety of supervised machine learning algorithms have been applied in the research of facial affect recognition. McDuff et al. [17] compared the performance of generative and discriminative classifiers on assigning valence labels to facial action sequences. Littlewort et al. [3] evaluated AdaBoost and Support Vector Machines (SVMs) on recognition of basic emotions. They also used SVMs to recognize AUs and expressions of posed and spontaneous pain [6]. Hoque et al. [5] explored detections of frustration and delight by applying SVMs, Hidden Markov Models and Hidden-state Conditional Random Fields. El Kaliouby et al. [18] inferred the cognitive mental states using dynamic Bayesian networks (DBN). Li et al [19] applied DBN to estimate the intensity of AUs. A comprehensive investigation of spontaneous facial expression recognition can be found in Zeng et al [1] and recent challenges like AVEC [20] and EmotiW [21]. With few exceptions, most of the previous efforts are based on supervised learning, which requires intensive manual labeling of the facial data.

2.1 A Personal Model for Affect Detection

Much current research in affective computing focus on model generalization for new users [22]. However, user-independent models have difficulty accommodating individual differences. Littlewort et al. [3] reported an accuracy drop from 95% to 60% when a model trained on one dataset is tested on another. Michel et al. [4] carried out similar experiments and the accuracy drops from 87.5% to 60.7%. Findings from the first facial expression recognition and analysis (FERA) challenge [8] also show that the user-dependent model generally outperforms the user-independent model.

There have also been efforts in combining generic and user-dependent target data into the same model. Valstar et al. [8] shows that high performance could be achieved for emotion recognition when the priori training data for the target user is available. However, well-labeled user-dependent data is expensive to obtain.

More recently, transfer learning has been advocated for building personalized models with limited target data. Chen et al. [9] used inductive and transductive transfer learning to build person-specific models, with the inductive transfer learning approach, trained on generic data and a small set of labeled target user data, outperforming the user-independent model and its transductive counterpart. In contrast, Chu et al. [10] demonstrated the superiority of the transductive learning approach, which re-weighs the generic training samples most relevant to the test subject. Generally, the effectiveness of transfer learning relies on the distribution similarity between the training and test data, and may fail when the target user behaves differently from the generic data, *i.e.* when the target and the generic data have different distributions.

There are at least two definitions of "user-dependent model". The "quasi-person-specific model" [10] is evaluated based on its performance on *one* specific subject from its training set [8]. The second definition, taken in our work, trains and tests a model for one specific user. This approach often suffers from overfitting, since it is difficult to obtain large amounts of data for any particular user [10]. However, given sufficient data, it should outperform the quasi-person-specific model as it is tailored specifically to accommodate the characteristics of

a single user [9]. Our objective is therefore to investigate methods of obtaining and utilizing such data in a feasible and effective manner.

2.2 Reducing Human Effort of Data Labeling

To reduce the annotation effort, previous work has investigated various degrees of supervision for facial affect modeling.

Unsupervised learning uses clustering to identify similar facial expressions or facial gestures. De la Torre et al. [16] proposed a geometric-invariant clustering technique that segments a specific user's facial behavior into facial gestures. Zhou et al. [23] used Aligned Cluster Analysis to detect facial events from video across multiple individuals. Both approaches identify similar expressions across different users. These methods can successfully discover similar facial events/gestures, but they do not aim to correlate the gesture with the affect, or address the differences in exhibited expression and felt affect across individuals.

Other work has focused on reducing the human labor. Zhang et al. [11] uses an interactive technique that initializes the affect labels with Bayesian networks and then uses mutual information to select informative data for human correction. The goal is to label only the most optimal data. Zhu et al. [12] used dynamic cascades to identify frames that are proximal to the apex between onset and offset to increase the amount of training data for AU detection. De la Torre et al. [2] labels only the apex of the AUs and automatically predicts the corresponding onset and offset. However, the above approaches all require human expertise to locate and label some essential data, such as the apex frames, which is time-consuming and expensive, and probably infeasible for user-dependent modeling.

Weakly supervised learning based on coarse-grained segment-level annotation, rather than the fine-grained frame-level annotation, has been attracting recent attention. Xu et al. [24] divided a facial sequence into 20 representative sub-motions based on optical flow, and applied "bag of motion words" to recognize basic emotion in the facial sequences. Their work targets sub-motions at the facial gesture level, which last around 100 frames (*i.e.* 4 seconds). Ashraf et al. [25] use clustering to recognize expressions of pain on the Shoulder Pain UNBC-McMaster dataset (UNBC) [26]. Their method aims at deciding a segment-level result for individual video sequences, whose length ranges from 48 to 518 frames. However, there has not been much work on multi-class classification at the segment level for spontaneous facial affect recognition, which is a more practical way to obtain data in real use situations.

Multiple instance learning (MIL) has recently received much attention as an effective approach for weakly supervised object detection and segmentation [27][28][29]. Some standard supervised learning methods have been adapted for MIL, such as MilBoosting [30] and MI-Forest [31]. Sikka et al. [32] introduced "concept segments" (*i.e.* facial gestures) for pain expression recognition using MILBoost. They aggregated over the frames in a gesture by max-pooling and estimated segment-level label from the gesture probability. However, this method considers only a subset of instances in the positive bag, and ignores a potentially large number of ambiguous instances [27], which might contribute to the classification if properly explored. Ruiz et al. [33] uses a regularization term to discard non-informative features and multi-concept MIL to allow different contributions from concepts (expressions). Their method maps the instances to a bag-level vector for the bag-classifier. Similarly, Chen et al.'s [34] MILES transformed the bag instances to a vector via an instance similarity measure, turning MIL into a standard supervised learning problem. However, the mapping may be biased by the representativeness and the number of instances in the bag, and information may be lost during the frame-to-segment aggregation.

The majority of the prior work on MIL target the bag-level binary classification problem [13]. However, facial affect recognition is normally a multi-class problem. AMIL analyzes the distribution of expressions across bags and deduce the bag label based on expressions that strongly correlated with few or even a single bag class. This allows AMIL to support multi-class learning and to output a per-frame fine-grained result, which can also facilitate bag-level recognition [25]. In this sense, AMIL resembles work from Xiao et al [27], which uses instance similarity to maximize information utilization.

3 System Overview

Fig. 1 illustrates the PADMA process. Similar to De la Torre et al. [16], we consider facial expression and affective state to correspond at the level of temporal facial gestures. That is, a change in the affective state results in a change in the user's facial expression, which is captured as a sequence of expression labels.

We start with an affect elicitation process to obtain samples of spontaneous facial affect from the user. Following Gross et al. [14], our affect elicitation uses video clips selected to arouse specific affects. The clips are chosen to arouse only one user affect at a time. To verify that the affects were elicited, users are asked to select an affect (including "Neutral" and "None of the Above") that best represents their overall feeling after watching each video clip. In this work, we focus on 5 basic affective states and two higher-level mental states (Fig. 1, Row 1).

We capture the user's facial response during affect elicitation (Fig. 1, Row 2). Each frame in the user response video is processed to extract the facial features, which are then combined into a feature vector. This produces a sequence of facial feature vectors (Fig. 1, Row 3). Clustering is then applied to group together feature vectors similar to each other. For each cluster, a distinct cluster label, or *expression label*, is introduced and used in place of the feature vectors to characterize each of the frames in this cluster (Fig. 1, Row 4). This frame-labeling procedure, similar to vector quantization, encodes the large number of possible feature vectors as a relatively small set of labels to simplify and facilitate facial expression analysis in the subsequent steps.

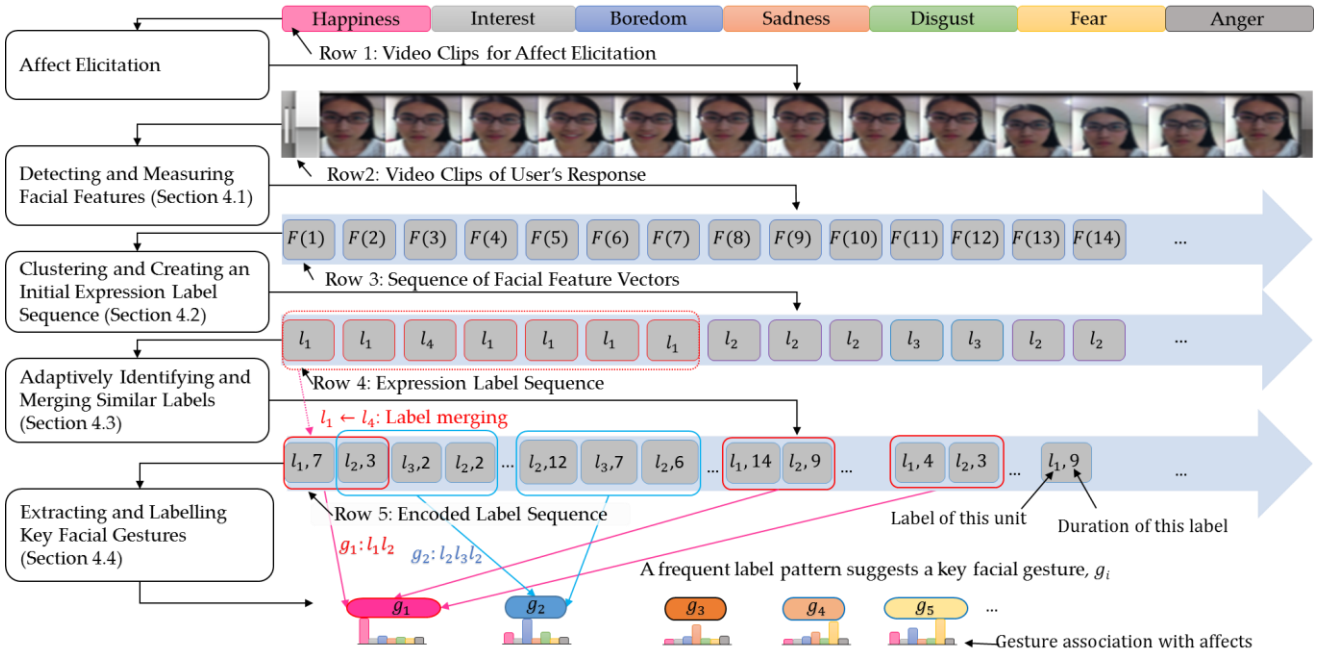

Fig. 1. Personal Affect Detection with Minimal Annotation. Feature vectors (row 3) are extracted from the response video (row 2) to the stimulus (row 1). These vectors are clustered to create initial expression labels, which are then used to label the corresponding frames to create expression sequences (row 4). An adaptive merging process combines similar clusters (row 5) and key facial gestures are extracted. Association-based Multiple Instance Learning (AMIL) is then used to determine the relation between key facial gestures and affects depending on the occurrences of the gestures in segments and the corresponding self-reported affect labels.

A facial expression, with its onset, apex and offset, can therefore be represented as a sequence of expression labels. Run-length encoding and cluster merging are then applied to identify frequently-occurring expression label sequences, or *key facial gestures*, from the response sequence (Fig. 1, Row 5). Given the user's self-reported affects from the elicitation process, we infer the affective state that is expressed by each key facial gesture by analyzing the distribution of the key facial gestures across the entire sequence of the response video and the correlation between the key facial gestures and the affects (Fig. 1, Row 6). Once the correlation is identified, the affective state of a user at any given point in time can be identified by looking for key facial gestures that occur around that time period.

4 METHODOLOGY

4.1 Detecting and Measuring Facial Gestures

The affect elicitation process constructs a *response video* that contains the user's facial expressions for a given set of affects, such that these expressions may be detected and measured automatically.

Previous work in psychology and computer vision has proven the value of using AUs-based analysis for interpreting and analyzing facial expressions [3][6][18] [17]. Facial AUs are descriptors of facial movements, which constitute the essential representation of a facial expression. Indeed, it is possible to describe all facial expressions as combinations of different AUs.

We follow an approach from previous work [16][23] to extract facial features referring to AUs. We apply Constrained Local Models (CLM) [35] to track 66 facial landmarks from the response video. This model is trained on the CMU Multi-PIE Face database [36], which contains over 750,000 images from 337 people. However, due to the nature of the training data, this model fails to track some of the mouth movements, such as mouth corner depression. To improve the tracking accuracy, we also apply the Supervised Descent Method [37] to validate and optimize the 2D landmark locations. During the CLM optimization procedure, the 2D and 3D landmarks and other global and local parameters are adjusted iteratively until the face fitting regression model converges. Removing the rigid transformation from the acquired 3D shape compensates for the influence of out-of-plane rotation and produces the aligned 3D landmarks.

The direction and intensity of the facial movements can be calculated from the normalized distances and angles between the corresponding facial landmarks. This generates facial features that are similar to Motion-Units [38], which describe facial movement like Ekman's AUs; but are numeric and directional in nature, unlike AUs which are classified into discrete intensity levels.

Fig. 2 shows the facial landmarks used in our work. The wired face shows the 3D facial landmarks from the tracking result, and the facial image shows the locations and indices (i.e. numbers in the bracket) of the corresponding 2D facial landmarks. Table 1 presents the descriptions and measurements of the 20 facial features calculated from the aligned 3D landmarks.

Fig. 3 shows sample frames from our experiment data. Both head pose movement and lighting condition (e.g. dissimilar illumination and camera exposure, etc.) pose significant challenges for the appearance-based features,

especially with elderly people with natural wrinkles. Hence, to ensure robustness in real-use situations with various environmental variations, we focus on geometric facial features, which avoids the noise from the textural/appearance channel.

Fig. 2 shows the facial landmarks used in our work. The wired face shows the 3D facial landmarks from the tracking result, and the facial image shows the locations and indices (i.e. numbers in the bracket) of the corresponding 2D facial landmarks. Table 1 presents the descriptions and measurements of the 20 facial features calculated from the aligned 3D landmarks.

Fig. 3 shows sample frames from our experiment data. Both head pose movement and lighting condition (e.g. dissimilar illumination and camera exposure, etc.) pose significant challenges for the appearance-based features, especially with elderly people with natural wrinkles. Hence, to ensure robustness in real-use situations with various environmental variations, we focus on geometric facial features, which avoids the noise from the textural/appearance channel.

4.2 Clustering and Creating an Initial Expression Label Sequence

After the facial features are detected and measured, a user's facial response can be represented as a sequence of facial feature vectors, $F(j) = (F_i(j), i = 1, \dots, 20)$, where each $F_i(j)$ is the measurement of feature i for a given frame j. This gives a quantified representation of the user's facial expressions in the response video, which is highly dimensional and difficult to manage. Dimensionality reduction is therefore used to render the changing user facial expressions more manageable.

We normalize each facial feature measurement of a user to a range between 0 and 1, then apply K-means clustering [39] to cluster together similar facial feature vectors. This allows us to identify *expression labels* for distinct facial expressions, which will essentially function as a low dimensional representation of the facial expression in the user's response.

Since the purpose of the expression labels is to represent different expressions, we need to avoid clustering together markedly different expressions. The cluster number K is therefore chosen to be large (500 in our experiments). We perform a preliminary clustering on a random 10% subset of data to seed the initial locations for the cluster centroids. To compensate for the randomness in the clustering process, we repeat the entire

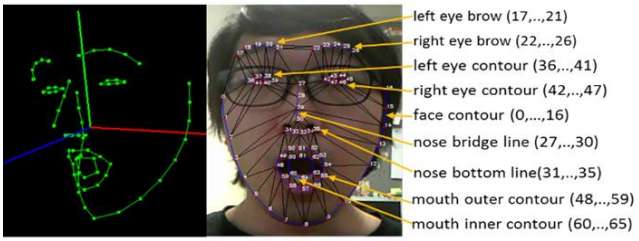

Fig. 2. Facial landmarks tracked by CLM. The wired face (left) presents the tracked 3D facial landmarks. The facial image (right) shows the locations and indices (i.e. numbers in the bracket) of the corresponding 2D facial landmarks.

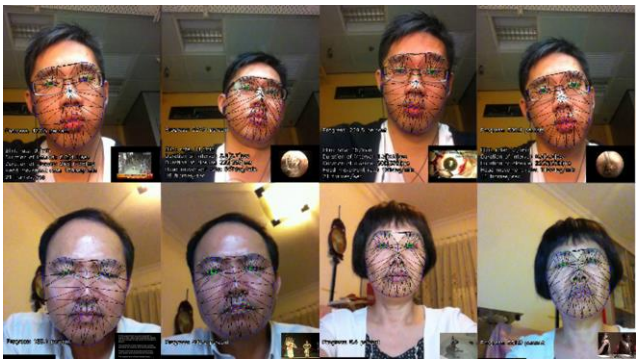

Fig. 3. Example frames from our experiment data. The face tracking model is able to correctly locate the landmarks and gives precise geometric features, regardless of the lighting and facial appearance conditions.

clustering process 10 times, and choose the result that gives us the most compact clusters, or the lowest intra-cluster distance, averaged over all clusters.

The centroids of the resulting clusters are then assigned unique IDs, which function as labels for the feature vectors. Each feature vector, $F(j), j \in [1, n]$, is then

TABLE 1. FACIAL FEATURES USED IN PADMA.

Feature	Implication	Measurement
$f_{1,2,3,4}$	Inner and outer brow movement	Distance between eye brow corner and corresponding eye corners (left & right)
$f_{5,6}$	Eye brow movement	Distance between the eye center and the corresponding brow center
$f_{7,8}$	Eye lid movement	Sum distance between corresponding landmarks on the upper and lower lid
f_9	Upper lip movement	Distance between landmark 33 and 51
$f_{10,11}$	Lip corner puller	Distance between the mouth corner and the corresponding eye outer center
f_{12}	Eye brow gatherer	Distance between inner eye brow corners
f_{13}	Lower lip depressor	Distance between landmarks 8 and 57
f_{14}	Lip pucker	Perimeter of the mouth outer contour
f_{15}	Lip stretcher	Distance between the mouth corners
f_{16}	Lip thickness variation	Sum distance between corresponding points on the outer and inner contours
f_{17}	Lip tightener	Sum distance of corresponding points on the upper and lower mouth outer contour
f_{18}	Lip parted	Sum distance of corresponding points on the upper and lower mouth inner contour
f_{19}	Lip depressor	Angle between mouth corners and lip upper center
f_{20}	Cheek raiser	Angle between nose wing and nose center

replaced with the label for its corresponding cluster. This gives us a sequence of *expression labels* $L(j)$, where n denotes the number of frames in a user's response video.

4.3 Adaptively Identifying and Merging Similar Labels

Theoretically, given the sequence of expression labels $L(j)$, identifying facial gestures should simply be a matter of looking for frequently occurring subsequences in $L(j)$. In practice, however, it is a challenge to decide on the number of clusters K used in the clustering process. A large K leads to redundant expression labels, where similar facial expressions are assigned to different clusters. This manifests as temporal jittering in $L(j)$, when the facial gesture sequence "bounces" back and forth between two labels over a short duration (Fig. 4). A small K, on the other hand, may assign markedly distinct expressions to the same cluster, which may result in inadequate expression labels and the loss of potential indicative expressions.

PADMA adaptively learns the proper number of clusters in a manner similar to G-means [40]. The underlying assumption is that changes in human facial expressions are usually continuous and progressive, and do not exhibit back-and-forth changes as would be suggested by temporal jittering. Since the jitter is caused when similar facial expressions are split into different clusters as a result of an over-large K, we merge similar clusters by minimizing temporal jittering in $L(j)$, subject to their distribution in the response video.

Table 2 illustrates the process for identifying jitters and merging clusters from the sequences of expression labels $L(j)$. Run-length encoding is used to decompose $L(j)$ into $S(u)$ and $D(u)$, where $S(u)$ is the encoded sequence of expression labels and $D(u)$ the frame duration of the labels. For example, $L(j) = \{l_a, l_a, l_a, l_a, l_b, l_b, l_c, l_c, l_c, l_c, l_c\}$ will be decomposed to $S(u) = \{l_a, l_b, l_c\}$ and $D(u) = \{4,2,5\}$ (Table 2, Step a).

We assume that a facial expression normally lasts for at least T_t frames, which we define as the duration threshold for expression label transition. We then identify a jitter by looking for all instances of u where the following conditions are fulfilled:

$$l_p = S(u-1) = S(u+1);\ l_q = S(u);\ p \neq q; \qquad (1)$$
$$D(u-1) > T_t,;\ D(u) < T_t;\ D(u+1) > T_t$$

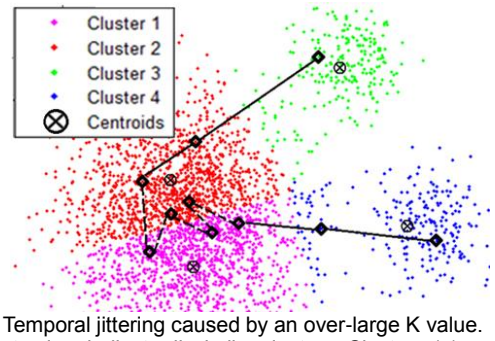

Fig. 4. Temporal jittering caused by an over-large K value. Different colors indicate dissimilar clusters. Clusters 1 (purple) and 2 (red) are similar clusters that should be merged. The black lines denote an expression sequence. The dotted lines indicate jittering between cluster 1 and 2.

TABLE 2. IDENTIFYING JITTERS AND MERGING EXPRESSION LABELS

Input: expression label sequence of user's response L	
Output: expression label sequence with merged labels S	
a	$(S, D) = \text{runLengthEncoding}(L)$
	do
b	$J = \text{countJitter}(S, D)$ via Equation (1)
c	$\mu_i, \sigma_i = \text{calculateInterClusterDistance}(\text{c}_i, \text{c}_{j\neq i})$
d	$\lambda = \mu_J + \xi \cdot \sigma_J$
	foreach *pair* (l_p, l_q) **do**
e	**If** $j_{p,q} > \lambda$ **then**
f	$\tau_{p,q} = \min(\tau_p, \tau_q)$
	If $d_{pq} < \tau_{p,q}$ **then**
g	$(S, D) = \text{updateSequences}(S, D, l_p \leftarrow l_q)$
	While *number of clusters being successfully merged* > 0

For each jitter between two expression labels, l_p and l_q, we increment the corresponding entry j_{pq} in the jitter frequency matrix J (Table 2, Step b).

Simultaneously, we calculate the *cluster distance* d_{ij} for each cluster c_i. d_{ij} is defined as the Euclidean distance between the centroids of clusters c_i and c_j. μ_i and σ_i are then the mean and standard deviation of the distances between c_i and all other clusters (Table 2, Step c).

We define the *jitter frequency threshold* $\lambda = \mu_J + \xi \cdot \sigma_J$, where μ_J is the mean and $\sigma_J$is the standard deviation of all the nonzero data in J, and ξ is a parameter that models the probability of jitter between two expression labels (Table 2, Step d).

If j_{pq} is larger than λ, then the clusters corresponding to l_p and $l_q$are potential candidates for merging. For each pair of such clusters, we calculate $\tau_{p,q} = \min(\tau_p, \tau_q)$, $\tau_i = \mu_i - \sigma_i/2$. (Table 2, Step f) If $d_{pq} < \tau_{pq}$, c_p and c_q will be merged. (Table 2, Step g)

The algorithm iterates until no more labels are merged.

4.4 Extracting and Labeling Key Facial Gestures

After similar labels in the expression sequence have been merged, frequently-occurring subsequences are identified from the entire expression sequence. We perform a data stream mining, which is similar to "frequent sequence mining" [41] using Apriori [42] to identify the most significant sequences and accelerate the searching. The identified sequences are then regarded as key facial gestures. The distribution of each key facial gesture can be analyzed to infer its association with each affect.

Our measure is inspired by the tf-idf [43] measure used in information retrieval. We recast our problem as that of retrieving the most appropriate user affect, given a "query" of a facial gesture $g_i, i \in [1, m]$, where m is the number of facial gestures. We define v_i, the *response clip-set* for affect a_i, as the set of response video segments that were reported by the user as exhibiting affect a_i -- that is, a_i is user-reported to be the main affect experienced when viewing the corresponding elicitation clip. Therefore, we expect the facial gestures in v_i to exhibit mainly affect a_i and neutral, with a few other affects also included.

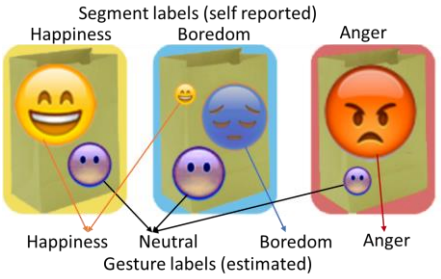

Fig. 5. Analyzing gesture distribution. Different bag colors represent different self-reported labels for three response clip-sets. Emoticons represent facial gesture sequences. The size of the emoticon is proportional to the frequency of occurrence of the gesture. A gesture (e.g. purple) that commonly occurs across different bags is identified as neutral, while gestures that occur primarily in a particular clip-set are considered indicative of the affect associated with the bag.

Fig. 5 illustrates with an example. Three response clip-sets are shown, corresponding to the affects "happy", "sad" and "angry", respectively. Facial gestures that occur almost exclusively in one response clip-set are identified as exhibiting that particular affect. On the other hand, facial gestures that occur regularly across multiple response clip-sets most likely are not representative of any particular affect, hence labeled as neutral. Therefore, our goal is to identify key facial gestures that commonly occur in v_i, but not in response clip-sets for other affects.

We define $f(g_j, v_i)$ as the frequency of occurrence of the key facial gesture g_j in response clip v_i. The *inverse affect frequency* (IAF) of a gesture quantifies the indicative value of a gesture by measuring its "rarity", on the basis that very common gestures have little indicative value:

$$\mathrm{IAF}(g_j, V) = log \frac{1 + |V|}{|\{v \in V : f(g_j, v) > 0\}|} \quad (2)$$

V is the set of all response clip-sets; $|V|$ denotes the number of different self-reported affects, and $|\{v \in V : f(g_j, v) > 0\}|$ represents the number of response clips that contain g_j. Since g_j represents an existing key facial gesture in the response clips, the denominator is always nonzero. We set the numerator to be $1 + |V|$ to ensure the resulting IAF is larger than zero, so that it will not eliminate the contribution of other factors after the multiply operation.

The *response frequency* (RF), on the other hand, measures the prevalence of a facial gesture over the duration of v_i. Given the set of all key facial gestures G:

$$\mathrm{RF}(g_j, v_i) = \frac{f(g_j, v_i)}{max\{f(g, v_i) : g \in G\}} \quad (3)$$

$max\{f(g, v_i) : g \in G\}$ denotes the maximum frequency of any gesture occurring in v_i, and normalizes bias towards longer response clips. The RFIAF value is calculated as:

$$\mathrm{RFIAF}(g_j, v_i, V) = \mathrm{RF}(g_j, v_i) * \mathrm{IAF}(g_j, V) \quad (4)$$

After obtaining the RFIAF values between each gesture and affect, the possible facial gestures are extracted with multi-scale moving windows [32]. Denoting G_w as the set of gestures occurring in the windows that span over the w-th element in the run-length encoded sequence, we define the association between affect a_i and the gestures in G_w as follows:

$$R(\mathrm{G}_w, a_i) = \sum_{g_j \in \mathrm{G}_w} \mathrm{RFIAF}(g_j, v_i, V) \quad (5)$$

4.5 Calculating Segment- and Frame- Level Affects

Identifying the key facial gestures and associating them with the corresponding affect gives us a *description* of how a person expresses a particular affect. Given this information, identifying the affect is then a matter of looking for key facial gestures.

Using the same multi-scale moving windows over each segment, we calculate the segment-level affect label a according to the $R(\mathrm{G}_w, a_i)$ values across all windows:

$$a = \mathrm{argmax}_{a_i \in A} \sum_{w \in W} R(\mathrm{G}_w, a_i) \quad (6)$$

where $A = \{a_1, \ldots, a_{|V|}\}$ is the set of affects and W denotes the elements in the run-length encoded sequence.

Similarly, we can also estimate the frame-level label from the gesture-level estimation. For the k-th frame, the affect label is estimated by:

$$a^{(k)} = \mathrm{argmax}_{a_i \in A} R(\mathrm{G}_{\Phi(k)}, a_i) \quad (7)$$

Φ denotes the frame mapping from the original video sequence to the run-length encoded sequence.

5 Experimental Validation

The contribution of our approach is a novel, weakly supervised method that uses Association-based Multiple Instance Learning (AMIL) to identify human affects from video data in real-use scenarios for user-dependent affect modeling. It does not require expert annotation, nor does it require much human work for labeling. We shall validate its correctness and effectiveness in two aspects:

Contribution of our novel AMIL approach. AMIL differs from other MIL approaches by using an information retrieval-inspired approach that uses the distribution of a pattern across *all* bags for labeling. We evaluate the impact of this assumption against that of other MIL models, both at the segment (bag) level and at the frame level. For this purpose, we will reconstruct two high-performing MIL methods as representatives of current state-of-the-art [27], and compare the performance of our approach with theirs on a publicly-available dataset as well as our own dataset.

Overall performance of the PADMA method. We argue that a weakly-supervised user-dependent model would be more appropriate in real-use contexts with spontaneous expressions. We will therefore evaluate PADMA against user-independent approaches. For a better understanding of the role of training data and user effort, as well as the advantages and disadvantages of each approach, we will also explore issues such as learning speed, training set size, and the nature of the problem.

Following previous approaches [44], we use the weighted average precision, recall or F-measure (F1) as an evaluation metric. The performance for a particular affect $\overline{P_c}$ is the weighted average performance of that affect over all the subjects: $\overline{P_c} = \sum_{s=1}^{N_s} w_{cs} * p_{cs}$, where $w_{cs} = \frac{N_{cs}}{\sum_{i=1}^{N_s} N_{ci}}$. Here s denotes the index of the subject and c denotes the index of the class (affect). p_{cs} therefore is the recognition performance on affect c for subject s, and N_s the number of subjects. N_{cs} denotes the number of instances in the

ground truth data for subject s that are labeled with affect c. The overall performance $\bar{P}$ can similarly be represented by $\bar{P} = \sum_{c=1}^{N_c} \bar{P}_c / N_c$, where N_c is the number of affects.

5.1 Experiment Setup

Our evaluation requires a dataset of spontaneous facial responses from multiple subjects, labeled with facial affect labels at both coarse and fine-grained levels, with sufficient samples for each individual.

There are a number of existing datasets from previous work. Chu et al. [10] and Valstar et al. [8] were tested on GEMEP-FERA [8], which consists of posed (simulated) expressions from 7 actors. Chu et al.'s work [10] was tested on Extended Cohn-Kanade (CK+) [44] and RU-FACS [7]. Although these two datasets contain a good number of subjects, the data for any one subject is limited: around 100 frames for CK+ and 2.5 minutes for RU-FACS. DISFA [45] is annotated with AUs rather than facial affects, and the individual data is limited, around 4 minutes for each subject. Likewise, BP4D-spontanous [46] provides only short segments and limited individual data. MAHNOB-HCI [47] and DEAP [48] have sufficient individual data, but they do not provide frame-level facial affect annotation.

Most prior MIL research [32][33][34] was evaluated on the UNBC dataset [26], which contains segment and frame-level annotation from multiple subjects. Since this satisfies our requirements, we will use the UNBC dataset as a basis for comparison with state-of-the-art methods.

The UNBC dataset contains 200 segments from 25 subjects with shoulder pain. Subjects performed active and passive arm movement with their affected and unaffected limbs. Expert coders gave Observer Pain Intensity (OPI) rating for each segment, ranging from 0 (no pain) to 5 (strong pain). We follow previous work [25][32][33] and define the segment-level label according to OPI, *i.e.* OPI≥3 is labeled as "pain" and OPI=0 as "no pain", and intermediate intensities of 1 and 2 are omitted. Selecting subjects who have more than one video segment in the dataset gives us 147 segments and 23 subjects. For the frame-level label, we follow previous work and determine the label according to the Parkachin and Solomon pain intensity (PSPI) [49], where PSPI>0 is labeled as pain, and PSPI=0 is marked as no pain [25].

The UNBC data is essentially a binary pain/no pain classification problem, which is arguably simpler than the multi-class affect classification problem. We therefore construct our own dataset of spontaneous facial affect responses. We shall refer to this dataset as Mobile Spontaneous Affect Response Video (or MSARV for short). The dataset consists of 11 Asian test subjects (5 female, aged 21-56, mean 32.4, standard deviation 11.8). Most are university students and staff.

MSARV contains segments of spontaneous facial affects, captured on a mobile device. Elicitation videos are presented to the subject, and the front camera of the mobile device is used to capture the facial response. This produces a response video at 480x640 resolution and 30 frames per second. In total, the dataset contains 817,080 frames. The resulting head poses exhibited in the dataset, as estimated by the face tracker, are: pitch: mean 5.6°, sd 6.3°; yaw: mean -1.3°, sd 2.5°; roll: mean 3.6°, sd 3.5°.

Video segments used for emotion elicitation include amusing scenes from the comedy "Gags", talks on popular technology from "Engadget", academic lectures on advanced topics, sad scenes from "Grey's Anatomy" and "Les Misérables", eye and ear surgeries, trailers from horror and ghost movies, and video clips depicting abuse of pregnant woman, children and elderly people.

We assembled different segments into two elicitation videos, each approximately 40 minutes and containing 25 short segments. The content of each video is selected to elicit the following affects: happiness (1'30"x3), interest (1'8"x6), boredom (1'15"x5), sadness (2'19"x3), disgust (2'13"x2), fear (2'14"x3) and anger (1'32"x3). (The numbers in the brackets indicate the average length and the number of segments.) These are the same affects that are covered in most of the publicly-available datasets [8][44]. Some of the affects (e.g. happiness) are more easily aroused than others (e.g. sadness) [1], which accounts for the difference in the length and number of the elicitation videos. Segments are kept between 1-2 min to avoid habituation to the stimuli while being long enough to arouse an affect [14].

The experiment was performed in a private area (a research lab). Each subject was randomly assigned to watch one of the elicitation videos. Subjects were instructed to behave naturally and knew in advance that their expressions were being recorded. None of the subjects reported feeling inhibited with their emotions during the experiments.

Even though the elicitation videos were carefully chosen to elicit a particular affect, it does not mean that the viewers will necessarily feel that affect when viewing it. Therefore, after each video segment, subjects are asked to select an affect label (happiness, interest, boredom, sadness, disgust, fear, anger, or none/neutral) that best describes their *overall* feeling while watching the video segment. This is the limit of human annotation required in our approach. It takes only 1~2 seconds for each video segment and no particular expertise. They are used as segment-level ground truth to evaluate our weakly supervised learning approach.

To obtain the gold-standard ground truth labels for frame-level evaluation, we followed a cued-recall procedure [50], which requires the subject to recall the felt affects from memory by the provision of visual information. This retrospective affect-judgment has been validated [50], proved to be consistent with external observations, and successfully used in previous work engagement detection. The response video was synchronized with the corresponding elicitation video. The subject then watched the videos, together with two observers. Every 4 seconds, the subject and the observers were asked to label the response expression in the current frame with one of the affect labels in MSARV. If the subjects' self-evaluation is consistent with the observers' evaluation, the corresponding affect label is accepted as

TABLE 3. MSARV VIDEO DETAILS: ELICITATION VIDEOS, USER RESPONSES AND MANUALLY ANNOTATED GROUND TRUTH LABELS.

Affect	Length of elicitation video for each affect (sec)	Number of elicitation video segments for each affect	Number of response videos self-reported to be exhibiting each affect	Percentage of frames exhibiting each affect in the ground truth
N	0	0	1	3%
H	270	3	32	9%
I	408	6	87	27%
B	375	5	82	29%
S	417	3	13	6%
D	266	2	20	10%
F	402	3	16	7%
A	279	3	24	9%

N: neutral, H: happiness, I: interest, B: boredom, S: sadness, D: disgust, F: fear, A: anger

ground truth. In cases of disagreement, the subject and the observers discussed until a mutual agreement was achieved. In addition to the 7 affects that we focus on in this work, facial expressions with no particular affect-related indication are marked as "neutral".

Table 3 summarizes the details for the MSARV dataset. We show the length and number of the elicitation videos, the user response and the ground truth labeling, with respect to each affect.

5.2 Evaluation at the Segment Level

Our first evaluation compares the recognition result at the segment level on both the UNBC and MSARV datasets.

The user-independent model is evaluated using leave-one-subject-out cross-validation. The overall result is the average performance, weighted by the amount of testing data for each individual.

To evaluate user-dependent learning, we performed leave-one-segment-out cross-validation. Each segment is used for testing in turn, with the remaining segments from the same individual used for training. Similarly, averaging over the test iterations gives us the overall result. We exclude subjects whose segments exhibit only one label type (i.e. only "pain" or only "no pain") on UNBC, which yielded 22 subjects with 145 segments.

Our state-of-the-art "competitors" are based on two weakly-supervised SVM/MIL-based models from previous work [28][32][33]. The first model, which we refer to as vMIL (for *vector*-based MIL), uses max-pooling [51], which has been shown to be effective for feature aggregation [32], to extract segment-level features. Each segment is represented by one feature vector, and the individual feature values are chosen as the value with maximum deviation from the mean for that feature. The second model, referred to as sMIL (for *subset*-based MIL), uses the subset representation method, which represents each segment with cluster centroids from K-Means clustering. We empirically choose K=20 in our experiment. To determine the segment recognition result for sMIL for pain detection on UNBC, we follow previous work [25] and rely on a frame threshold determined by the equal error rate (EER). The classifier for both models is the support vector machine (SVMs) [52], which generally performs well on pattern recognition applications, including state-of-the-art affect detection [44]. Our particular SVMs are implemented by the sequential minimal optimization algorithm [53], using polynomial kernels and parameters determined by grid search. Finally, for affect classification on MSARV, sMIL uses majority voting based on the results of frames in the subset to determine the segment-level result.

TABLE 4. PERFORMANCE AND COMPARISON TO STATE-OF-THE-ART MIL METHODS ON USER-INDEPENDENT LEARNING, UNBC DATASET (PERFORMANCE METRIC: ACCURACY AT EQUAL ERROR RATE)

MILES [34]	MILIS [54]	MilBoosting [30]	MI-Forest [31]	MS-MIL[32]	RMC-MIL[33]	AMIL (ours)
78.2	76.9	76.9	75.8	83.7	85.7	84.4

Table 4 shows that AMIL outperforms MS-MIL [32] and is comparable with RMC-MIL, the current highest-performing approach [33], for user-independent learning. This shows that our information retrieval-based assumption is effective at modeling facial affects.

Table 5 presents user-dependent and user-independent segment-level recognition results on UNBC and MSARV.

For the UNBC dataset, our reconstructed models, vMIL and sMIL, achieve 83.7% and 85.7% accuracy respectively. This is comparable to reported performance from similar approaches in literature (vMIL and MS-MIL [32] both achieve 83.7%; sMIL and RMC-MIL [33] both achieve 85.7%), and suggests that our reconstructed models are state-of-the-art.

Using AMIL for feature aggregation in user-independent learning achieves performance close to the best result (sMIL: 85.7% vs 84.4% -- a difference of 2 segments). When used for user-dependent learning in PADMA, AMIL outperforms the other feature aggregation approaches by 5% (81.6% vs 76.6%). This is close to the best performing overall model (user-independent sMIL: 81.6% vs 85.7% -- 7 segments). Hence, PADMA achieves performances that are generally comparable to state-of-the-art on UNBC.

Unexpectedly, user-dependent learning on UNBC does not perform as well as user-independent learning. Inspecting the data suggests two possible reasons. First, even though UNBC contains a good number of subjects, there is limited data *per subject* (mean: 1525 frames, sd:

TABLE 5. RESULT COMPARISON ON UNBC AND MSARV (PERFORMANCE METRIC: ACCURACY AND F-MEASURE)

Dataset	UNBC			MSARV		
Method	vMIL	sMIL	AMIL	vMIL	sMIL	AMIL
User-dependent	76.2(0.74)	76.6(0.76)	**81.6**(0.81)	30.5(0.33)	53.5(0.55)	**72.0**(0.71)
User-independent	83.7(0.82)	**85.7**(0.86)	84.4(0.84)	16.4(0.18)	11.6(0.13)	**32.4**(0.33)

Numbers in and out of the bracket denote the F-measure and accuracy at equal error rate, respectively.

712 frames), which makes it difficult for the user-dependent model to generalize. Secondly, it appears that the expression of pain may be somewhat more universal, and thus easier to generalize across different users, than other high-level mental states such as interest. This is supported by the fact that pain is usually measured according to the PSPI score, which only considers a small subset of facial action units.

In contrast to UNBC, MSARV is relatively richer, with more data per subject and multiple affect labels.

Performance results on MSARV are promising. Table 5 shows that the user-dependent models significantly outperform their user-independent counterparts across the board. Using AMIL for feature extraction, PADMA achieves the highest performance with 72.0% accuracy and 0.71 F1 – 18% higher than the next best-performing model (sMIL: 53.5% – a difference of 51 segments), and twice as accurate as the best user-independent model (user-independent AMIL: 32.4% -- 109 segments). This suggests that, in certain contexts, user-dependent models significantly outperforms user-independent learning, and the AMIL assumptions provide the best performance.

Data analysis suggests that the performance difference between UNBC and MSARV are mainly due to the difference in their affect attributes. Affects on MSARV include both basic emotions and mental states such as interest and boredom, which commonly occur in daily human-computer interactions. Although it is reported that basic emotions are universal across cultures [1], in real use, it appears that the *manifestation* of these affects as spontaneous expressions still differ across subjects. It also appears that high-level mental states are manifested differently between people, and may be more challenging to recognize [1]. For instance, some subjects react to boredom by looking away, while others frown or change postures. Modeling this individuality in a user-independent manner would be difficult.

Table 6 gives the confusion matrix and the performance metrics of PADMA on UNBC. Both precision and recall are high, which demonstrates that AMIL is effective for binary classification.

Table 7 shows the same measurements on the multi-class MSARV data. In general, the majority of the segments are recognized correctly. PADMA performs worst on sadness (F1: 0.50), fear (F1: 0.62) and neutral. This is consistent with previous findings [44], which states that these emotions are naturally more difficult to recognize. The problem is exacerbated in MSARV, since it contains only spontaneous expressions, and most of the time, sadness and fear were not elicited to a high degree.

TABLE 6. CONFUSION MATRIX AND PERFORMANCE OF PADMA FOR SEGMENT-LEVEL USER-DEPENDENT LEARNING ON UNBC. ROWS: ANNOTATED (TRUTH) CLASS; COLUMNS: RECOGNIZED CLASS. F-MEASURE OVER ALL SUBJECTS: 0.81.

	Pain	No Pain	Precision	Recall	F1
Pain	**34**	21	0.85	0.62	0.72
No Pain	6	**86**	0.80	0.93	0.86

TABLE 7. CONFUSION MATRIX AND PERFORMANCE OF PADMA FOR SEGMENT-LEVEL USER-DEPENDENT LEARNING ON MSARV. ROWS: ANNOTATED (TRUTH) AFFECT; COLUMNS: RECOGNIZED AFFECT. F-MEASURE OVER ALL AFFECTS FOR ALL SUBJECTS IS 0.72.

	N	H	I	B	S	D	F	A	Precision	Recall	F1
N	0	0	1	0	0	0	0	0	NA	0.00	0.00
H	0	**27**	5	0	0	0	0	0	0.69	0.84	0.76
I	0	6	**58**	16	4	1	1	1	0.70	0.67	0.68
B	0	4	9	**66**	2	0	1	0	0.76	0.80	0.78
S	0	0	4	1	**7**	0	1	0	0.47	0.54	0.50
D	0	1	1	1	0	**16**	0	1	0.80	0.80	0.80
F	0	1	0	1	1	3	**9**	1	0.69	0.56	0.62
A	0	0	5	2	1	0	1	**15**	0.83	0.63	0.71

N: neutral, H: happiness, I: interest, B: boredom, S: sadness, D: disgust, F: fear, A: anger

Post-experiment interviews can help us to understand this phenomenon. For sadness, the subjects noted that they were less likely to feel sad without knowing the context of the video segment. Therefore, if they had previously watched the movie that contains the elicitation video segment, (re)watching the short segment would cause them to recall the movie, and a stronger feeling of sadness is successfully induced. However, if they had not previously watched the movie, they were less likely to feel that emotion. The result is that for some subjects (3 out of 11), the affect of sadness was never successfully aroused during the elicitation process. Fear proved to be another emotion that was hard to elicit. The subjects noted that even though the movie segments might be scary, the fact that they knew that they were in an experiment was counterproductive to eliciting fear.

Surprisingly, the "neutral" affect was rarely reported in our experiments. Post-experiment interviews suggest that this is because "interest" and "boredom" were available as options, and users who did not feel that any of the basic emotions applied to them tended to choose one of those two affects instead.

5.3 Evaluation at the Frame Level

In addition to the segment-level, the frame-level label is also of interest to us, as it is useful for precise understanding of the temporal affect changes within a segment. For example, it can shed light on the exact moment a patient feels pain, or the moment that a user becomes interested in the stimulus. Frame-level performance can also be considered as an approximation of the gesture-level accuracy.

The frame-level ground truth on UNBC is obtained through the PSPI score, while MSARV provides the frame-level observations.

Table 8 presents the frame-level performance on UNBC and MSARV. Unsurprisingly, the frame-level performance is similar to the segment-level performance. User-independent learning with AMIL outperforms user-dependent on UNBC, while user-dependent learning

TABLE 8. FRAME-LEVEL RECOGNITION PERFORMANCE OF AMIL ON UNBC AND MSARV.

Dataset	UNBC	MSARV
User-dependent	58.5(0.62)	**59.2(0.59)**
User-independent	**71.3(0.69)**	25.7(0.25)

performs better on MSARV. This may also be a result of insufficient individual training data for user-dependent learning on UNBC. More interestingly, performance at the frame-level is lower than at segment-level in general (71.3% vs. 84.4% on UNBC; 59.2% vs. 72.0% on MSARV).

Fig. 6 presents the overall F-measure across all affects for each subject on MSARV at the both segment-level and frame-level. For all but one subject, segment-level recognition achieves a higher accuracy. This makes sense as it is challenging to recognize the fine-grained result from the coarse-grained data labels. However, given that we achieve good results on segment-level recognition, this shows that it is possible to extract and label users' key facial gestures with only a very small amount of annotation. This also demonstrates that if one wants to obtain the affect implication behind facial gestures, deducing them from the segment-level label would be an effective approach.

The performance difference for different users is also due to a data sparsity problem, as not all affects were successfully aroused in some of the subjects. This suggests that a longer affect elicitation process, or a dynamic affect elicitation process that selects video clips to show the user based on the affects that have already been successfully elicited, might be more effective.

Fig. 7 shows three examples of key facial gestures identified by PADMA. Each image represents one expression label contained in the gesture. This shows that our approach can successfully identify both dynamic and static indicative gestures. For example, gesture (a) represents an expression transition from onset to apex of happiness. Gesture (b) is associated with boredom by our user, and indicates a fast transition from onset to apex and then offset. Gesture (c) is a static gesture that was associated with sadness. It contains only one expression label, indicating the subject kept her apex expression for a long duration. Sequences (b) and (c) give a sense of the challenge posed by the MSARV data: spontaneous expressions often do not exhibit exaggerated facial muscle movement, which suggests that facial affect detection in

Fig. 7. Examples of identified key facial gesture sequences in MSARV. (a) happy, (b) bored, (c) sad. The resulting gestures may contain different number of expression labels. Gesture (a) presents a transition from neutral to apex; gesture (c) shows a long-lasting sad expression. Note the subtle difference between (b) and (c).

real-use situations may be a much more complex task than the posed expression alternative.

For a better understanding, we also investigate the contribution of the PADMA adaptive clustering process, which first chooses a large K and then merges extraneous clusters. We compare our performance against the alternative of using a fixed number of clusters. In our previous experiments, the merging step reduces the number of clusters from 500 to 246~487, depending on the actual facial responses of the subjects. We present experiments with five K values (100, 400, 500, 750, 1000), which lie both within and beyond the range of the final number of clusters achieved through an adaptive K. It can be seen that the adaptive approach outperforms the fixed approach for all affects (Fig. 8). This bears out our hypothesis: if K is too small, some of the key facial gestures will be merged together, and cause many key facial gestures to be labeled as "neutral". In contrast, an over-large K produces different expression labels for similar key facial gestures, which makes their distribution sparser and decreases their RFIAF values. We conclude therefore that a proper way to determine the cluster number is essential for facial affect modeling and the post-clustering merging of redundant expression clusters is an essential step in our algorithm.

5.4 Learning Speed and Amount of Training Data

In this section we further evaluate the performance of our model as a function of user effort. We have shown that user-independent learning outperforms user-dependent learning when there is insufficient training data per user, as in the UNBC dataset. Since time and effort from end-

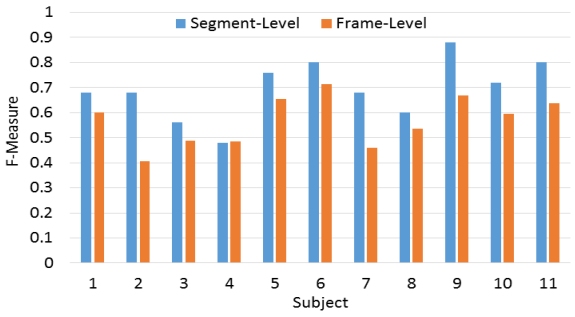

Fig. 6. Per subject user-dependent learning on MSARV. Performance at segment-level and frame-level for all affects.

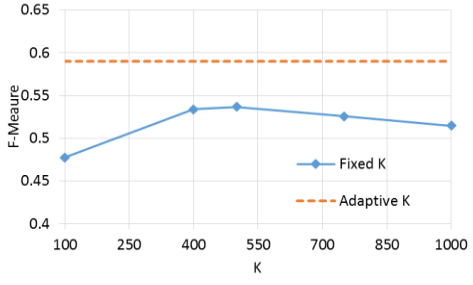

Fig. 8. Comparison between fixed and adaptive K for PADMA. The adaptive clustering approach clearly outperforms using a fixed number of clusters.

users poses a major challenge for user-dependent learning, we wish to understand the impact of training data on system performance.

The evaluation process involves multiple iterations with an incremental training set. User-independent models are evaluated using leave-some-subjects-out cross-validation, and user-dependent models with leave-some-segments-out cross-validation. On each iteration, a subset with a certain number of subjects or segments is selected as the training set using a rolling window, with the test set being the rest of the data. The final performance result for each iteration is averaged over all training subsets.

Fig. 9 shows the impact of training data on performance for the UNBC dataset. The user-dependent model starts off with lower performance compared to the user-independent counterpart, but as the number of segments in the training set increases, the user-dependent model (best F1: 0.81) rapidly approaches the performance of the user-independent model (best F1: 0.84). Given that the segments on UNBC are generally very short (238 frames per segment on average), this means that the user-dependent model can achieve performance comparable to the user-independent model with relatively little effort. However, the learning speed flattens out after 6 segments. This may be due to the fact that UNBC contains on average only 6 segments for any individual subject. Though the user-dependent model does not outperform the user-independent model, the performance difference is small; and the steep upward trend of the learning curve suggests that given enough training data, it is likely that the user-dependent model would outperform the user-independent model.

To validate this hypothesis, we run similar experiments on the MSARV dataset, where more data per subject is available. We use the boredom affect as an example to investigate the amount of additional effort required to extend an existing model to incorporate additional affects. On the user-dependent model, we train on all response video segments self-reported as *not* boredom and increment the amount of training data by adding one boredom-labeled segment (≈1min per segment) at a time, each time testing on the remainder of the boredom-labeled segments. Likewise, for the user-independent model, we train a basic model on a subset of subjects, and perform leave-some-subjects-out cross-validation by incrementing the amount of training data one subject at a time (≈6.8min of new boredom-labeled data per subject), while testing on the rest of the subjects.

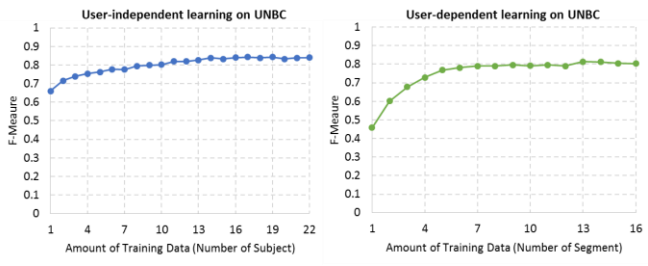

Fig. 9. Learning speed vs amount of training data on UNBC. Comparison between user-independent and user-dependent learning of AMIL. The user-dependent model shows a faster learning speed.

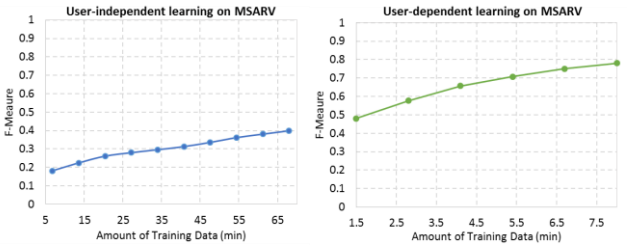

Fig. 10. Learning speed vs amount of training data on MSARV for the "boredom" affect. Comparison between user-independent and user-dependent learning of AMIL. Even with small amounts of individual data, the user-dependent model outperforms the user-independent model in both learning speed and accuracy.

Fig. 10 compares the learning speed between the user-dependent and user-independent models, illustrating performance as a function of time required from subjects. The points on the curve represent the F1 performance for the "boredom" affect for that iteration.

The results are encouraging. The learning speed of the user-dependent model increases much faster than the user-independent model. 5 more training segments (≈6min) increase F1 of the user-dependent model by 0.3. For the user-independent model, however, adding data from 9 more subjects (≈60min) improves F1 by only 0.22. In addition, the user-dependent model, trained on one segment, already outperforms the user-independent model trained on data from 10 different subjects.

We conclude that given sufficient weakly-labeled individual data, user-dependent learning can outperform user-independent learning. Furthermore, to achieve comparable performance, the training set required for the user-dependent technique is relatively small compared to that of the user-independent techniques.

6 Conclusions and Future Work

This paper presents PADMA, a method for building user-dependent models for facial affect recognition, which uses novel algorithms for adaptive clustering and association-based multiple instance learning. We evaluate on two datasets containing expressions of pain and spontaneous facial expressions over 8 affects: neutral, happiness, interest, boredom, sadness, disgust, fear and anger. Experiments demonstrate that PADMA can effectively extract a user's facial gestures and correctly assign the corresponding affect labels without the need for human annotation at the frame level.

To verify the efficacy of our approach, we present evaluations comparing our method with related weakly supervised models on both user-dependent and user-independent learning. Our results conclude that PADMA can successfully model spontaneous facial affects in a practical manner.

Our experiments demonstrate the effectiveness of PADMA on the UNBC and MSARV datasets. It is not difficult to see that this approach can be directly applied

to everyday computing activities. For instance, a user could update the model by self-reporting his/her feelings every time after watching a YouTube video that he/she feels strongly about. This should result in a higher accuracy than by trying to elicit diverse emotions within a short time period. Moreover, continuous data collection in real-use situations will provide more comprehensive expression data and more accurate gesture distribution models, which will further improve the generalizability and accuracy of the model.

We see several possibilities for future work. First, it would be interesting to investigate the possibility of using PADMA as an active learning mechanism. Once the computer captures sufficient expressions or indicative facial gestures occurring with an unknown affect, it could prompt the user for an incremental affect update. Furthermore, if there is sufficient facial data, it would be interesting to use AMIL to recognize the *intensity* given the *type* of the affect, as well as the temporal relationship between gestures.

Second, this study focuses on single affect learning. Our elicitation videos are selected to be short and are intended to arouse one single affect. We foresee, however, that a mixture of affects may occur in real-use scenarios. Apart from introducing a label that combines multiple affects, a potential solution may be just to use all reported affects as valid bag labels when calculating the RFIAF values in the learning phase, and adopt a threshold or use a classifier to determine the occurrence of each affect.

Third, as in previous work [55], we intend to investigate multimodal approaches that supplements our facial features with additional sources of input signals, and to extend our approach to detect the intensity as well as the type of affect. It would also be interesting to extend our dataset to support other higher-order affects, such as "frustration", "stress" and "thinking".

ACKNOWLEDGMENT

The authors wish to thank the experiment subjects for their time and effort. This work was partially supported by grants PolyU 5235/11E and PolyU 5222/13E from the Hong Kong Research Grants Council.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[2] F. De la Torre, T. Simon, Z. Ambadar, and J. F. Cohn, "FAST-FACS: A Computer-Assisted System to Increase Speed and Reliability of Manual FACS Coding," in *Affective Computing and Intelligent Interaction*, 2011, no. 1, pp. 1–10.

[3] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of Facial Expression Extracted Automatically from Video," in *Proc. Comput. Vis. Pattern Recognit. Workshop, 2004. CVPRW '04*, 2004, p. 80.

[4] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proc. 5th Int. Conf. Multimodal interfaces - ICMI '03*, 2003, p. 258.

[5] M. E. Hoque, D. J. McDuff, and R. W. Picard, "Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 323–334, Jul. 2012.

[6] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain," in *Proc. the ninth Int. Conf. Multimodal interfaces - ICMI '07*, 2007, p. 15.

[7] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic Recognition of Facial Actions in Spontaneous Expressions," *J. Multimed.*, vol. 1, no. 6, Sep. 2006.

[8] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-Analysis of the First Facial Expression Recognition Challenge.," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, Jun. 2012.

[9] J. Chen, X. Liu, P. Tu, and A. Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1964–1970, Nov. 2013.

[10] W.-S. Chu, F. De La Torre, and J. F. Cohn, "Selective Transfer Machine for Personalized Facial Action Unit Detection," in *2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3515–3522.

[11] L. Zhang, Y. Tong, and Q. Ji, "Active Image Labeling and Its Application to Facial Action Labeling," in *ECCV '08 Proc. 10th Eur. Conf. on Comput. Vis.*, 2008. pp 706-719.

[12] Y. Zhu, F. De la Torre, Y.-J. Zhang, and J. F. Cohn, "Dynamic Cascades with Bidirectional Bootstrapping for Action Unit Detection in Spontaneous Facial Behavior," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 79–91, Apr. 2011.

[13] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowl. Eng. Rev.*, vol. 25, no. 01, p. 1, Mar. 2010.

[14] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9. pp. 87–108, 1995.

[15] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978.

[16] F. De la Torre, J. Campoy, Z. Ambadar, and J. F. Cohn, "Temporal Segmentation of Facial Behavior," *2007 IEEE 11th Int. Conf. Comput. Vis.*, pp. 1–8, 2007.

[17] D. McDuff, R. El Kaliouby, K. Kassam, and R. Picard, "Affect valence inference from facial action unit spectrograms," *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work.*, pp. 17–24, Jun. 2010.

[18] R. El Kaliouby and P. Robinson, "Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures," in *2004 Proc. Comput. Vis. Pattern Recognit. Workshop*, 2004, pp. 154–154.

[19] Y. Li, S. M. Mavadati, M. H. Mahoor, and Q. Ji, "A unified probabilistic framework for measuring the intensity of spontaneous facial action units," in *2013 10th IEEE Int. Conf. Autom. Face Gesture Recognit., FG*, 2013.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014," in *Proc. the 4th Int. Workshop on Audio/Visual Emotion Challenge - AVEC '14*, 2014, pp. 3–10.

[21] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion Recognition In The Wild Challenge 2014," in *Proc. 16th Int. Conf. on Multimodal Interaction - ICMI '14*, 2014, pp. 461–466.

[22] R. Kaliouby and P. Robinson, "Generalization of a Vision-Based Computational Model of Mind-Reading," pp. 582–589, 2005.

[23] F. Zhou, F. De la Torre, and J. F. Cohn, "Unsupervised discovery of facial events," *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2574–2581, Jun. 2010.

[24] L. Xu and P. Mordohai, "Automatic Facial Expression Recognition using Bags of Motion Words," in *Proc. Brit. Mach. Vision Conf.*, 2010, pp. 13.1–13.13.

[25] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The Painful Face - Pain Expression Recognition Using Active Appearance Models.," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1788–1796, Oct. 2009.

[26] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 57–64.

[27] Y. Xiao, B. Liu, Z. Hao, and L. Cao, "A similarity-based classification framework for multiple-instance learning.," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 500–15, Apr. 2014.

[28] X. He and S. Gould, "An Exemplar-Based CRF for Multi-instance Object Segmentation," in *2014 IEEE Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 296–303.

[29] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL Training for Weakly Supervised Object Localization," in *2014 IEEE Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 2409–2416.

[30] P. Viola, J. C. Platt, and C. Zhang, "Multiple Instance Boosting for Object Detection," in *Adv. Neural Inf. Process. Syst.*, 2006, p. 1417.

[31] C. Leistner, A. Saffari, and H. Bischof, "MIForests: Multiple-instance learning with randomized trees," in *Eur. Conf. Comput. Vis.*, 2010, vol. 6316 LNCS, pp. 29–42.

[32] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain localization using multiple instance learning," in *2013 10th IEEE Int. Conf. Autom. Face Gesture Recognit (FG)*, 2013, pp. 1–8.

[33] A. Ruiz, J. Van de Weijer, and X. Binefa, "Regularized Multi-Concept MIL for weakly-supervised facial behavior categorization," in *Proc. Brit. Mach. Vision Conf.*, 2014.

[34] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 1931–1947, 2006.

[35] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable Model Fitting by Regularized Landmark Mean-Shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Sep. 2010.

[36] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *2008 8th IEEE Int. Conf.on Automatic Face & Gesture Recognition*, 2008, pp. 1–8.

[37] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *2013 IEEE Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[38] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Comput. Vis. Image Underst.*, vol. 91, no. 1–2, pp. 160–187, Jul. 2003.

[39] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[40] G. Hamerly and C. Elkan, "Learning the k in k-means," in *In Adv. Neural Inf. Process. Syst. (NIPS)*, 2003.

[41] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Min. Knowl. Discov.*, vol. 15, no. 1, pp. 55–86, Jan. 2007.

[42] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *Proc. of the 20th Int. Conf. on Very Large Data Bases VLDB '94*, 1994, pp. 487–499.

[43] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.

[44] P. Lucey, J. F. 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 *2010 IEEE Comput. Vis. Pattern Recognit - Workshops*, 2010, no. July, pp. 94–101.

[45] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, pp. 151–160, 2013.

[46] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, 2014.

[47] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, pp. 42–55, 2012.

[48] S. Koelstra, C. Mühl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, pp. 18–31, 2012.

[49] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, pp. 267–274, 2008.

[50] E. L. Rosenberg and P. Ekman, "Coherence between expressive and experiential systems in emotion," *Cogn. Emot.*, vol. 8, no. 3, pp. 201–229, May 1994.

[51] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring bag of words architectures in the facial expression domain," in *Eur. Conf. Comput. Vis., ECCV*, 2012, vol. 7584 LNCS, pp. 250–259.

[52] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[53] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," in *Advances in kernel methods*, 1999, pp. 185 – 208.

[54] Z. Fu, A. Robles-Kelly, and J. Zhou, "MILIS: Multiple instance learning with instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, pp. 958–977, 2011.

[55] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

Michael Xuelin Huang received the BEng and MEng degrees in Automation Science from Beihang University in 2007 and 2010, respectively. Currently, he is working toward the PhD degree in the Department of Computing at the Hong Kong Polytechnic University, Hong Kong SAR, China. His research interests include affective computing, human computer interaction, and pattern recognition.

Grace Ngai received her Ph.D. degree from Johns Hopkins University in Computer Science in 2001. She worked for Weniwen Technologies, a natural language and speech firm in Hong Kong, and joined the Hong Kong Polytechnic University in 2002. She is currently an associate professor at the Department of Computing. Her research interests are in affective computing, human computer interaction, wearable computing, and education.

Kien A. Hua received the B.S. degree in computer science and the M.S. and PhD degrees in electrical engineering, all from the University of Illinois at Urbana-Champaign, in 1982, 1984, and 1987, respectively. He was at IBM from 1987-90. He joined the University of Central Florida in 1990 and is currently a professor in the School of Electrical Engineering and Computer Science. From 2003-05, he served as the Interim Associate Dean for Research of the College of Engineering and Computer Science. He has published widely, including several papers recognized as best/top papers at various international conferences. He has served as a general chair, vice-chair, associate chair, demo chair, and program committee member for numerous conferences. He is a Fellow of the IEEE.

Stephen C.F. Chan received his Ph.D. degree in Electrical Engineering from the University of Rochester in 1987. He had worked for the National Research Council of Canada, and was the Canadian representative for the ISO-10303 STEP standard for the exchange of industrial product data. He is currently an associate professor in the Department of Computing at the Hong Kong Polytechnic University. His research interests are data and text mining, human-computer interaction and service-learning.

Hong Va Leong received his PhD degree from the University of California at Santa Barbara in 1994 and joined the Hong Kong Polytechnic University. His research interests lie in distributed systems, distributed databases, and mobile computing. He has served on the program committee and organizing committee of numerous international conferences as well as chairing some of them. He is a member of the ACM, IEEE Computer Society and IEEE Communications Society.