1 Machine Learning Methods for Social Signal Processing

Ognjen Rudovic, Mihalis A. Nicolaou and Vladimir Pavlovic

1.1 Introduction

In this chapter we focus on systematization, analysis, and discussion of recent trends in machine learning methods for Social signal processing (SSP)(Pentland 2007). Because social signaling is often of central importance to subconscious decision making that affects everyday tasks (e.g., decisions about risks and rewards, resource utilization, or interpersonal relationships) the need for automated understanding of social signals by computers is a task of paramount importance. Machine learning has played a prominent role in the advancement of SSP over the past decade. This is, in part, due to the exponential increase of data availability that served as a catalyst for the adoption of a new data-driven direction in affective computing. With the difficulty of exact modeling of latent and complex physical processes that underpin social signals, the data has long emerged as the means to circumvent or supplement expert- or physics-based models, such as the deformable musculo-sceletal models of the human body, face or hands and its movement, neuro-dynamical models of cognitive perception, or the models of the human vocal production. This trend parallels the role and success of machine learning in related areas, such as computer vision, c.f., (Poppe 2010, Wright et al. 2010, Grauman & Leibe 2011), or audio, speech and language processing, c.f., (Deng & Li 2013), that serve as the core tools for analytic SSP tasks. Rather than emphasize the exhaustive coverage of the many approaches to data-driven SSP, which can be found in excellent surveys (Vinciarelli et al. 2009, Vinciarelli et al. 2012), we seek to present the methods in the context of current modeling challenges. In particular, we identify and discuss two major modeling directions:

- Simultaneous modeling of social signals and context, and
- Modeling of annotators and the data annotation process.

Context plays a crucial role in understanding the human behavioral signals that can otherwise be easily misinterpreted. For instance, a smile can be a display of politeness, contentedness, joy, irony, empathy or a greeting, depending on the context. Yet, most SSP methods to date focus on the simpler problem of detecting a smile as a prototypical and self-contained signal. To identify the smile as a social signal one must simultaneously know the location of *where* the subject is (outside, at a reception, etc.), *what* his or her current task is, *when* the signal was displayed

(timing), and *who* the expresser is (expresser's identity, age and expressiveness). (Vinciarelli et al. 2009) identify this as the W4 quadruplet (*where, what, when, who*) but quickly point out that comprehensive human behavior understanding requires the W5+ sextuplet (*where, what, when, who, why, how*), where the *why* and *how* factors identify both the stimulus that caused the social signal (e.g., funny video) as well as how the information is passed on (e.g, by means of facial expression intensity). However, most current SSP methods, including the data-driven ones, are not able to provide a satisfactory answer to W4, let alone W5+. Simultaneously answering the W5+ is a key challenge of data driven SSP.

Another key factor in machine learning-based SSP is the *curse* of annotations. Unlike in many *traditional* machine learning settings, social signals are frequently marked by multiple annotators, be those experts or novices, with an unknown *ground truth*. Because of the often subjective interpretation of social signals, annotations reflect both the annotators bias and the potential temporal lag in marking the time-course of the signal. Hence, modeling of the annotators themselves and deriving the *gold* standard, in addition to modeling the expresser and its signal, is another crucial factor for full and robust automated social signal understanding. We therefore analyze recent approaches to the annotation modeling process in this context.

The two modeling challenges are universal across different signal modalities (e.g., visual or auditory). In the rest of this chapter we focus on one signal domain, that of facial signals, that most ubiquitously illustrates the new data-driven modeling directions. Specifically, we consider the problems of Facial Expression Measurements and describe the state-of-the-art in machine learning methods as they relate to modeling of the signal-and-context and the annotators/annotations.

1.2 Facial Expression Analysis

There are two main streams in the current research on automatic analysis of facial expressions. The first considers holistic facial expressions such as facial expressions of six basic emotions (fear, sadness, happiness, anger, disgust, surprise) proposed by Ekman (Ekman et al. 2002), and facial expressions of pain, for instance. The second considers local facial expressions, described with a set of facial muscle actions named Action Units (AUs), as defined in the Facial Action Coding System (FACS) (Ekman et al. 2002). In what follows, we review the existing machine learning approaches for automated classification, temporal segmentation and intensity estimation of facial expressions, and relate these approaches to the W5+ context design.

1.2.1 Classification of Facial Expressions

Different methods have been proposed for classification of facial expressions from image sequences. Depending on how these methods perform classification of facial

expressions they can be divided into frame-based and sequence-based methods. The frame-based methods for classification of facial expressions of six-basic emotion categories (Ekman et al. 2002) typically employ static classifiers such as rule-based classifiers (Pantic & Rothkrantz 2004, Black & Yacoob 1997), Neural Networks (NN) (Padgett & Cottrell 1996, Tian 2004), Support Vector Machine (SVM) (Bartlett et al. 2005, Shan et al. 2009), and Bayesian Networks (BN) (Cohen et al. 2003). SVMs and its probabilistic counterpart, Relevance Vector Machine (RVM), have been used for classification of facial expressions of pain (Lucey et al. 2011, Gholami et al. 2009). For instance, (Lucey et al. 2011) addressed the problem of pain detection by applying SVMs either directly to the image features or by applying a two-step approach, where AUs were first detected using SVMs, the outputs of which were then fused using the Logistic Regression model. Similarly, for the static classification of AUs, where the goal is to assign to each AU a binary label indicating the presence of an AU, the classifiers based on NN (Bazzo & Lamar 2004, Fasel & Luettin 2000), Ensemble Learning techniques (such as AdaBoost (Yang et al. 2009*a*) and GentleBoost (Hamm et al. 2011)), and SVM (Chew et al. 2012, Bartlett et al. 2006, Kapoor et al. n.d.), are commonly employed. These static approaches are deemed context-insensitive as they focus on answering only one context question, i.e., *how*. Recently, (Chu et al. 2013) proposed a transductive learning method, named Selective Transfer Machine (STM), where a SVM classifier for AU detection is personalized by attenuating person-specific biases, thus, simultaneously answering the context questions *who* and *how*. This is accomplished by learning the classifier and reweighting the training samples that are most relevant to the test subject during inference.

The common weakness of the frame-based classification methods is that they ignore dynamics of target facial expressions or AUs. Although some of the frame-based methods use the features extracted from several frames in order to encode dynamics of facial expressions, models for dynamic classification provide a more principled way of doing so. With a few exceptions, most of the dynamic approaches to classification of facial expressions are based on the variants of Dynamic Bayesian Networks (DBN) (e.g., Hidden Markov Models (HMM) and Conditional Random Fields (CRF)). For example, (Otsuka & Ohya 1997, Shang & Chan 2009) trained independent HMMs for each emotion category, and then performed emotion categorization by comparing the likelihoods of the HMMs. In (Otsuka & Ohya 1997), the input features are based on velocity vectors computed using the optical flow algorithm, while the observation probability, corresponding to the hidden states in the HMMs, is modeled using mixtures of Gaussians in order to account better for variation in facial expressions of different subjects. Likewise, (Shang & Chan 2009) used geometric features (i.e. locations of facial points) and a non-parametric estimate of the observation probability in the HMM model. While these methods perform the expression classification of the pre-segmented image sequences, corresponding to the target emotion category, (Cohen et al. 2003) presented a two-level HMM classifier that performs

expression classification by segmenting sequences of arbitrary length into the segments, corresponding to different emotion categories. This is accomplished by learning first the expression-specific HMMs, and then the transitions between the expression categories using another HMM, taking as an input the predictions of the expression-specific HMMs. Simultaneous classification of different AUs using HMMs was addressed in (Khademi et al. 2010) using a Hybrid HMM-ANN model. In this model, the temporal development of each AU is first modeled using AU-specific HMMs. Subsequently, the outputs of different HMMs are combined in the ANN to account for the AU dependencies.

Discriminative models based on CRFs have also been proposed (der Maaten & Hendriks 2012, Jain et al. 2011, Chang et al. 2009). In (der Maaten & Hendriks 2012), the authors trained one linear-chain CRF per AU. The model's states are binary variables indicating the AU activations. (Jain et al. 2011) proposed a generalization of the linear-chain CRF model, a Hidden Conditional Random Field (HCRF) (Wang et al. 2006), where additional layer of hidden variables is used to model temporal dynamics of facial expressions. The training of the model was performed using image sequences, but classification of the expressions was done by selecting the most likely class (i.e. emotion category) at each time instance. The authors showed that: (i) having the additional layer of hidden variables results in the model being more discriminative than the standard linear-chain CRF, and (ii) that modeling of the temporal unfolding of the facial shapes is more important for discrimination between different facial expressions than their spatial variation (based on comparisons with SVMs). Another modification of HCRF, named partially-observed HCRF, was proposed in (Chang et al. 2009). In this method, the appearance features based on the Gabor wavelets were extracted from image sequences, and linked to the facial expressions of the target emotion category via hidden variables in the model. The hidden variables represent subsets of AU combinations, encoded using the binary information about the AU activations in each image frame. In this way, classification of the emotion categories (sequence-based), and the AU combinations (frame-based), was accomplished simultaneously. This method outperformed the standard HCRF, which does not use a prior information about the AU combinations. Temporal consistency of AUs was also modeled in (Simon et al. 2010) using the structured-output SVM framework for detecting the starting and ending frames of each AU.

More complex graph structures within the DBN framework have been proposed in (Zhang & Ji 2005, Tong et al. 2007) for dynamic classification of facial expressions. In (Zhang & Ji 2005), the DBN was constructed from interconnected time slices of static Bayesian networks, where each static network was used to link the geometric features (i.e. locations of characteristic facial points) to the target emotion categories via a set of related AUs. Specifically, the relationships between the neighboring time slices in the DBN were modeled using the first-order HMMs. (Tong et al. 2007) modeled relationships between different AUs using another variant of a DBN. In this model, the Adaboost classifiers were first used for independent classification of AUs to select the AU-specific features.

These features were then passed as inputs to the DBN, used to model temporal unfolding of the AUs as well as their co-occurrences.

Finally, some authors attempted modeling of the facial expression dynamics on the expression-specific manifold (Hu et al. 2004, Shan et al. 2006, Lee & Elgammal 2005). For instance, (Hu et al. 2004) used a low dimensional Isomap embedding to build a manifold of shape variation across different subjects, and then used the I-condensation algorithm to simultaneously track and recognize target emotion categories within a common probabilistic framework. (Shan et al. 2006) used a Bayesian temporal model (with Markov property) for the expression classification on the manifold derived using a supervised version of the Locality Preserving Projections (LPP) method (He & Niyogi 2004). As with the models mentioned above, these models account for the context questions *how*, and implicitly for the context question *when*, due to their modeling of the temporal dynamics. Static modeling using the expression manifold can also be attained using multi-linear decomposable generative models, as done in (Lee & Elgammal 2005). The authors used these models to separate the subject identity from the facial expressions on a manifold, followed by the expression classification. In contrast to the dynamic manifold-based models mentioned above, this approach accounts only for the context question *how*. While it has potential for accounting for the context question *who*, as well as the other context questions due to its decomposable nature, this has not been explored so far.

1.2.2 Temporal Segmentation of Facial Expressions

Most of the works on facial expression analysis from image sequences implicitly answer the context question *when* as they focus only on classification of target expressions and/or AUs. For instance, in the HMM-based models for facial expression classification (Shang & Chan 2009, Cohen et al. 2003), the number of hidden states is set so that they correspond to the temporal segments (neutral/onset/apex/offset) of facial expressions. They do not, however, explicitly encode these dynamics (i.e. they do not perform classification of the temporal segments). Yet, both the configuration, in terms of AUs constituting the observed expressions, and their dynamics, in terms of timing and duration of the temporal segments of facial expressions, are important for categorization of, e.g., complex psychological states, such as various types of pain and mood (Pantic & Bartlett 2007). They also represent a critical factor in interpretation of social behaviors like social inhibition, embarrassment, amusement, and shame, and are a key parameter in differentiation between posed and spontaneous facial displays (Ekman et al. 2002).

The class of models that performs segmentation of the expression sequences into different temporal segments try to answer the context questions *how* (e.g. the information is passed on by the apex of a facial expression of emotion or AU) and *when* (i.e. when did it occur in the expression sequence?), thus accounting explicitly for this context question. For instance, in (Pantic & Patras 2005) and

(Pantic & Patras 2006), a static rule-based classifier and the geometric features (i.e. facial points) were used to encode temporal segments of AUs in near-frontal and profile view faces, respectively. The works in (Koelstra et al. 2010, Valstar & Pantic 2012) proposed modifications of standard HMMs to encode temporal evolution of the AU segments. Specifically, (Koelstra et al. 2010) proposed a combination of discriminative, frame-based GentleBoost ensemble learners, and HMMs for classification and temporal segmentation of AUs. Similarly, (Valstar & Pantic 2012) combined SVMs and HMMs in a Hybrid SVM-HMM model based on the geometric features for the same task. Classification and temporal segmentation of the emotion categories was also attempted in (Gunes & Piccardi 2009) using HMMs and SVMs.

A variant of the linear-chain CRF, named the Conditional Ordinal Random Field (CORF), was proposed in (Kim & Pavlovic 2010) for temporal segmentation of the emotion categories. In this model, the node features of the linear-chain CRF model are set using the modeling strategy of the standard ordinal regression models, e.g. (Chu & Ghahramani 2005), in order to enforce the ordering of the temporal segments (neutral<onset<apex). The authors emphasize the importance of modeling the ordinal constraints, as well as the temporal constraints imposed by a transition model defined on the segments. On the target task, the proposed CORF model outperforms the static classifiers for nominal data such as SVMs, and ordinal data such as Support Vector Ordinal Regression (SVOR) (Chu & Keerthi 2005), as well as traditional dynamic models for nominal data such as HMMs and CRFs. An extension of this model was proposed in (Rudovic et al. 2012*b*), where the authors combined different emotion-specific CORF models in the HCRF framework. In contrast to the CORF model, this model performs simultaneous classification and temporal segmentation of the emotion categories. More recently, (Rudovic et al. 2012*a*) introduced a kernel extension of the CORF model and applied it to the AU temporal segmentation. Compared to the nominal temporal models such as Hybrid SVM-HMM (Valstar & Pantic 2012) and the linear CORF/CRF models, this model showed improved performance in the target task on most the AUs tested, which is mainly attributed to its non-linear feature functions.

1.2.3 Intensity Estimation of Facial Expressions

Facial expression dynamics can also be described in terms of their intensity. Explicit analysis of the expression intensity is important for accurate interpretation of facial expressions, and is also essential for distinguishing between spontaneous and posed facial expressions (Pantic & Bartlett 2007). For example, a full-blown smile and a smirk, both coded as AU12 but with different intensities, have very different meanings (e.g., enjoyment vs. sarcasm). However, discerning different intensities of facial expressions is a far more challenging task than the expression classification. This is mainly because the facial muscle contractions are combined with the individual's physical characteristics, producing changes in appearance

that can vary significantly between subjects (Ekman et al. 2002). As a consequence, the methods that work for intense expressions may generalize poorly to subtle expressions with low intensity.

While FACS (Ekman et al. 2002) provides a 5-point ordinal scale for coding the intensity AUs, there is no established standard for how to code the intensity of holistic facial expressions (e.g., those of the six basic emotions). Primarily for this reason and the observation in (Hess et al. 1997) that the expression decoding accuracy and the perceived intensity of the underlying affective state vary linearly with the physical intensity of a facial display, the existing works on intensity estimation of facial expressions of the basic emotions resort to an unsupervised approach to modeling of the expression intensity (e.g., (Amin et al. 2005, Shan 2007, Kimura & Yachida 1997, Lee & Xu 2003, Yang et al. 2009*b*)). The main idea in these works is that the variation in facial images due to the facial expressions can be represented on a manifold, where the image sequences are embedded as continuous curves. The distances from the origin of the manifold (corresponding to the embedding of the neutral faces) are then related to the intensity of the facial expressions. For instance, (Amin et al. 2005) used an unsupervised Fuzzy-K-Means algorithm to perform clustering of the Gabor wavelet features, extracted from expressive images, in a 2D eigenspace defined by the pairs of the features' principal components chosen so that the centroids of the clusters lie on a straight line. The cluster memberships are then mapped to three levels of intensity of a facial expression (e.g. less happy, moderately happy, and very happy). Similarly, (Shan 2007) first applied a supervised LPP technique (Shan et al. 2005) to learn a manifold of six basic expression categories. Subsequently, Fuzzy K-Means was used to cluster the embeddings of each expression category into three fuzzy clusters corresponding to a low, moderate and high intensity of target expressions. (Kimura & Yachida 1997) used a Potential Net model to extract the motion-flow-based features from images of facial expressions, which were used to estimate a 2D eigenspace of the expression intensity. (Lee & Xu 2003) and (Yang et al. 2009*b*) also performed the intensity estimation on a manifold of facial expressions. Specifically, (Lee & Xu 2003) used isometric feature mapping (Isomap) to learn a 1D expression-specific-manifold, and the distances on the manifold were then mapped into the expression intensity. The mapping of the input features to the expression intensity of three emotion categories (happiness, anger and sadness) was then modeled using either Cascade NNs or Support Vector Regression (SVR). On the other hand, (Yang et al. 2009*b*) treated the intensity estimation as a ranking problem. The authors proposed the RankBoost alghorithm for learning the expression-category-specific ranking functions that assign different scores to each image frame, assumed to correspond to the expression intensity. These scores are based on the pair-wise comparisons of the changes in the Haar-like features, extracted over time from facial images. The main criticism of these works is that the expression intensity is obtained as a byproduct of the learning method (and the features) used, which makes the comparison of the different methods difficult.

Recent release of the pain-intensity coded data (Lucey et al. 2011) has motivated research into automated estimation of the pain intensity levels (Hammal & Cohn 2012, Kaltwang et al. 2012, Rudovic et al. 2013*a*). For example, (Hammal & Cohn 2012) performed estimation of 4 pain intensity levels, with the levels greater than 3 on the 16-level scale being grouped together. The authors applied Log-Normal filters to the normalized facial appearance to extract the image features, which were then used to train binary SVM classifiers for each pain intensity level, on a frame-by-frame basis. Instead of quantizing the intensity levels for the classification, (Kaltwang et al. 2012) treated the pain intensity estimation as a regression problem. To this end, the authors proposed a feature-fusion approach based on the Relevance Vector Regression (RVR) model. While these works focus on static modeling of the pain intensity, (Rudovic et al. 2013*a*) proposed the Heteroscedastic CORF model for dynamic intensity estimation of six intensity levels of pain. In this CRF-like model, the authors model the temporal unfolding of the pain intensity levels in an image sequence, where the ordering of the image frames with different intensity levels is enforced. The heteroscedastic variance in the model also allows it to more easily adapt to different subjects.

AU intensity estimation is a relatively recent problem within the field, and only a few works have addressed it so far. Based on the modeling approach, these can be divided into static methods (Mahoor et al. 2009, Mavadati et al. 2013, Savrana et al. 2012, Kaltwang et al. 2012, Jeni et al. 2013) and dynamic methods (Rudovic et al. 2013*b*). The static methods can further be divided into classification-based methods (e.g., (Mahoor et al. 2009, Mavadati et al. 2013)) and regression-based (e.g, (Savrana et al. 2012, Kaltwang et al. 2012, Jeni et al. 2013)). The static classification-based methods (Mahoor et al. 2009, Mavadati et al. 2013) perform multi-class classification of the intensity of AUs using the SVM classifier. For example, (Mahoor et al. 2009) performed the intensity estimation of AU6 (cheek raiser) and AU12 (lip corner puller) from facial images of infants. The input features were obtained by concatenation of the geometric and appearance features. Due to the excessive number of the features, the Spectral Regression (SR) (Cai et al. 2007) was applied to select the most relevant features for the intensity estimation of each AU. The intensity classification was performed using AU-specific SVMs. On the other hand, the static regression-based methods model the intensity of AUs on a continuous scale, using either logistic regression (Savrana et al. 2012), RVM regression (Kaltwang et al. 2012), or Support Vector Regression (SVR) (Jeni et al. 2013). For instance, (Savrana et al. 2012) used Logistic Regression for AU intensity estimation, where the input features were selected by applying an AdaBoost-based method to the Gabor wavelet magnitudes of 2D luminance and 3D geometry extracted from the target images. (Kaltwang et al. 2012) used the RVM model for intensity estimation of 11 AUs using image features such as Local Binary Patterns (LBPs), Discrete Cosine Transform (DCT) and the geometric features (i.e. facial points), as well as their fusion. (Jeni et al. 2013) proposed a sparse representation of the facial appearance obtained by applying Non-negative Matrix Factorization (NMF) filters to gray-scale im-

age patches extracted around facial points from the AU coded facial images, thus answering the context question *who* indirectly, in addition to the context question *how*, which is also answered in the other models mentioned above. The image patches were then processed by applying the personal mean texture normalization, and used as input to the SVR model for the intensity estimation. SVMs were also used to analyze the AU intensities in (Bartlett et al. 2006, Reilly et al. 2006, Delannoy & McDonald 2008), however, these works did not report any quantitative results.

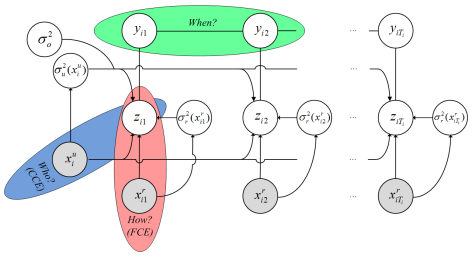

Figure 1.1 The cs-CORF model (Rudovic et al. 2013*b*) simultaneously accounts for the context questions *who*, *how* and *when*. x are the feature measurements, and the latent variable z is non-linearly related to the ordinal labels y via the ordinal probit function, used to define the node features in the cs-CORF model. For more details, see (Rudovic et al. 2013*b*).

So far, all the methods for intensity estimation of AUs, except that in (Jeni et al. 2013), account only for the context question *how*. Recently, (Rudovic et al. 2013*b*) proposed the Context-sensitive Conditional Ordinal Random Filed (cs-CORF) model for dynamic estimation of intensity of AUs, and facial expressions of pain. This model is a generalization of the CORF models (Kim & Pavlovic 2010, Rudovic et al. 2012*b*), proposed for expression classification and temporal segmentation. The cs-CORF provides means of accounting for all six context questions from the W5+ context model. In (Rudovic et al. 2013*b*), the authors demonstrate the influence of context on intensity estimation of facial expressions by modeling the context questions *who* (the observed person), *how* (the AU intensity-related changes in facial expressions), and *when* (the timing of the AU intensities). The context questions *who* and *how* are modeled by means of the newly introduced context and context-free covariate effects, while the context question *when* is modeled in terms of temporal correlation between the ordinal outputs, i.e., the AU intensity levels. To deal with skewed distributions of the AU intensity levels, the model parameters are adapted using a weighted softmax-margin learning approach. All these effects are summarized in the graphical representation of the cs-CORF model shown in Fig.1.1. In their experiments on spontaneously displayed facial expressions, the authors show that modeling of the ordinal relationships between the intensity levels, and their temporal unfolding, improves the estimation compared to that attained by static classification/regression models as well as the traditional nominal models for se-

quence classification (i.e. CRFs). More importantly, they show that the modeling of the context question *who* improves significantly the ability of the model to discriminate between the expression intensity levels of different subjects.

1.3 Annotations in Social Signal Processing

The urgency for obtaining meaningful annotations is crucial for any field which intersects with machine learning. Usually, the labelling task is performed manually, involving the cost of manual labour, where a set of experts or simple annotators is employed. This cost has though increased heavily during the past years, since the vast explosion of information in the so-called "Big Data" era led to the gathering of massive amounts of data to be annotated.

As an descriptive example, one can simply juxtapose Paul Ekman's seminal work on the six universal emotions (Pictures of Facial Affect) (Ekman et al. 1975), to one of the modern databases on affect, the SEMAINE database (McKeown et al. 2012). Ekman's work contained 110 black and white images, while approximately 2 seconds from one of the 959 sessions in SEMAINE contain approximately 100 color frames, accompanied with audio. It is no less than a fact that the task of annotating hours of audio-visual data is much more demanding than merely annotating 100 images.

The exponential increase of data availability functioned as a catalyst for the adoption of a new direction in *Social Signal Processing* (SSP). Since a large amount of audiovisual material was now available, instead of assigning one class label to a set of pre-defined episodes, researchers started to adopt continuous annotations in terms of the temporal dimension, i.e. instead of labelling a set of frames as "happy", now we can have one label per frame. Furthermore, if the label is a real number indicating the "magnitude" of happiness, the labels are continuous in both *space and time*. Most related research is based on the seminal work of Russel (Posner et al. 2005), where affect is described via a set of latent dimensions, which capture the emotional state of the subject *beyond* the basic, discrete classes of emotion introduced by Ekman (anger, disgust, fear, happiness, sadness and surprise). The most commonly used dimensions are valence, indicating the emotional state as positive or negative) and arousal, indicating the emotion intensity, while continuous annotations have been employed for other social signals such as pain and conflict. The shift from discrete classes of emotion to continuous annotations is part of an ongoing change in the field of affective computing and SSP, where the locus of attention was changing to more real-world problems, outside heavily controlled laboratory conditions, focusing on spontaneous emotion expressions instead of posed. By adopting a dimensional description of emotions, we are now able to represent emotional states that are commonly found in everyday life, e.g., being bored or interested (Gunes et al. 2008).

1.3.1 Challenges

The challenges arising from the recent focus of SSP on spontaneous, naturalistic data, along with the adoption of continuous annotations and the exponential increase in to-be-annotated data are many. Firstly, an issue inherent to annotation tasks related to SSP is *label subjectivity*. When measuring quantities such as subject *interest* or emotion dimensions such as *valence*, it is natural for some ambiguity to arise, especially when utilising spontaneous data in naturalistic, interactive scenarios (as in most state-of-the-art databases such as SEMAINE). While this issue manifests regardless of the label type, be it continuous, discrete or ordinal, the most tricky scenario is when dealing with continuous *in space* annotations. This is mostly due to the fact that instead of pre-defined classes (e.g., happy, neutral, sad), the annotation is in terms of the *magnitude* of e.g., happiness, leading to essentially infinite (upto machine/input device accuracy) classes. Essentially, this is a trade-off situation, since capturing a larger spectrum of expressions leads to increased label ambiguity.

As aforementioned, many modern databases such as SEMAINE[1] adopt continuous annotations *in time*. This entails that the annotation task is performed on-line, i.e. while each annotator is watching/listening to the audio/visual data, he or she is also moving the input device, usually a mouse (Cowie et al. 2000) or a joystick, according to his or her understanding of the emotional state of the subject. A prominent implication of the latter is that each annotator will demonstrate a time-varying, *person-specific lag*. Although one can claim that, due to the efficacy of the human brain, the realisation of the emotional state of the subject can be near-instant, the lag can be due to the time it takes for the annotator to actually perform the annotation (e.g., move the mouse), or can even depend on the input device itself or on how alert the annotator is at the time (e.g., the annotator can become tired and less responsive when annotating large amounts of data). Furthermore, the annotator is called to make an on-the-spot decision regarding the annotation, i.e. the annotation is no longer per-frame/per-image, making the processes more prone to errors.

In an effort to minimize person-specific bias, databases such as SEMAINE are annotated by multiple expert psychologists, who were trained in annotating such behaviour. Still, as one can easily verify by examining the provided annotations (Fig. 1.2), the subjectivity bias, annotator lag and other issues are still prominent. Other issues, which we do not comment on extensively here, can arise from weaknesses of physical input device which affect the accuracy of the annotation (e.g., moving the mouse can be highly inaccurate and can cause the appearance of spikes and other artefacts in the annotation). Some of the issues mentioned in this section are illustrated in Fig. 1.2.

[1] Besides SEMAINE, other examples of databases which incorporate continuous annotations include the Belfast Naturalistic Database, the Sensitive Artificial Listener (Douglas-Cowie et al. 2003), (Cowie et al. 2005) as well as the CreativeIT database (Metallinou et al. 2010).

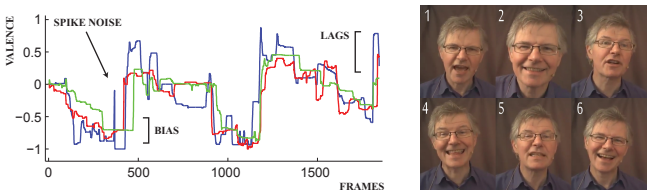

Figure 1.2 Example valence annotations from multiple annotators.

1.3.2 The sub-optimality of Majority Voting and Averaging

Due to the challenges discussed (Sec. 1.3.1), it is clear that the task of obtaining a "gold standard" (i.e. the "true" annotation, given a set of possibly noisy annotations), is a quite tedious task, and researchers in the field have not been agnostic regarding this in previous work (Metallinou et al. 2011, Nicolaou et al. 2012). In the majority of past research related to SSP though, the average annotation is usually used as an estimation of the underlying true annotation, either in the form of a weighted average by e.g., the correlations of each annotator to the rest (Nicolaou et al. 2011) or a simple, unweighted average (Wöllmer et al. 2008).

Majority voting (for discrete labels) or averaging (for continuous in space annotations) makes a set of explicit assumptions, namely that all annotators are *equally* good, and that the majority of the annotators *will* identify the correct label eliminating any ambiguity/subjectivity. Nevertheless, in most in real-world problems these assumptions typically do not hold. So far in our discussion we have assumed that all annotators are considered experts[2], a common case for labels related to SSP. In many cases though, annotators can be inexperienced, naive or even uninterested in the annotation task. This phenomenon has been amplified by the recent trend of *crowdsourcing* annotations (via services such as Mechanical Turk), which allows gathering labels from large groups of people, who usually have no formal training in the task-at-hand, shifting the annotation processes from a small group of experts to a massive but weak-annotator scale. In general, besides experts, we can consider that annotators can be assigned to classes such as *naive* which commonly make mistakes, *adversarial* or *malicious* annotators, that provide erroneous annotations on purpose, or *spammers* that do not even pay attention at the sequence they are annotating. It should be clear that if e.g., the majority of annotators are adversarial then majority voting will always obtain the wrong label. This is also the case if the majority of annotators are *naive*, and on a difficult/subjective data all make the same mistake. This phenomenon led to particular interest manifesting in modelling annotator performance, c.f.(Dai et al. 2010, Dai et al. 2011, Raykar et al. 2010, Yan et al. 2012).

It is important to note that the case of fusing continuous *in time* annotations comes with particular difficulties, since as discussed in Sec. 1.3.1, there is increased ambiguity and, most importantly, an annotator-specific lag, which in

[2] but not infallible when it comes to a *subjective*, online annotation process (Sec. 1.3.1).

turn leads to the misalignment of samples, as can be seen in Fig. 1.2. By simply averaging, we are essentially integrating these temporal discrepancies into the estimated ground truth, possibly giving rise to both *phase* and *magnitude* errors (e.g., false peaks). The idea of shifting the annotations in time in order to attain maximal agreement has been touched upon in (Nicolaou et al. 2010) and (Mariooryad & Busso 2013). Nevertheless, these works refer to a constant time-shift, which assumes that the annotator-lag is constant. This does not appear to be the case, as the annotator-lag depends on time-varying conditions (Sec. 1.3.1). The work of Nicolaou et al. (2012) is the first approach in the field which formally introduces a time alignment component into the ground truth estimation in order to tackle this issue. We will discuss the work of Nicolaou et al. (2012) along with other works on fusing multiple annotations in what follows.

1.3.3 Beyond Majority Voting & Averaging: Fusing Multiple Annotations

As mentioned in the previous section, the sub-optimality of majority voting given the challenges mentioned led to much interest in designing models to better fuse labels. In (Raykar et al. 2009), an attempt is made to model the performance of annotators, who assign a possibly noisy label. The latent "true" (binary) annotation is not known, and should be discovered in the estimation process. By assuming independence of all annotators and futhermore, assuming that annotator performance does not intrinsically depend on the annotated sample, each annotator can be characterised by his/her sensitivity and specificity. In this naive Bayes scenario, the annotator scores are essentially used as weights for a weighted majority rule, where if all annotators have the same annotator characteristics it collapses to the majority rule[3]. Note that the more general approach of (Raykar et al. 2009) indicates that, in the presence of data that is being labeled, neither simple nor weighted majority voting is optimal. In fact majority voting can be seen only as a first guess aimed at assigning an uncertain consensus "true" label, which is then further refined using an iterative EM process, where both the "true" label and the annotator performance are recursively estimated.

Spatio-temporal Fusion of Continuous Annotations

In general, Canonical Correlation Analysis (CCA) is a fitting paradigm for fusing annotations. CCA can find maximally correlating projections for the set of variables involved, and in a way, this can translate to the goal of fusing multiple annotations: find maximally correlating projections for the fused annotations, in order to minimise subject-dependent bias. CCA has been extended to a probabilistic formulation in (Bach & Jordan 2005), while Klami & Kaski (2008)[4] have extended Probabilistic CCA (PCCA) to a *private-shared* space model. In

[3] Detailed analysis of majority voting, including its weighted version, can be found in (Lam & Suen 1997, Ruta & Gabrys 2005).

[4] This formulation is closely related to (Tucker 1958), while the model of (Raykar et al. 2010) for fusing continuous annotations can considered a special case of (Bach & Jordan 2005).

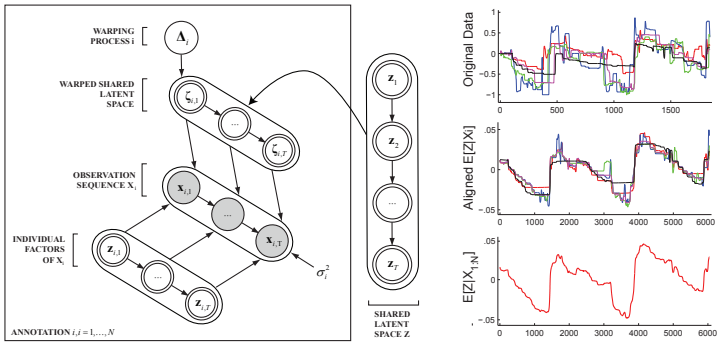

Figure 1.3 (a) Graphical model of (Nicolaou et al. 2012). The shared space $\mathbf{Z}$ generates all annotations $\mathbf{X}_i$, while also modelling the individual factors $\mathbf{Z}_i$, specific only to annotation i. The time-warping process $\mathbf{\Delta}_i$ temporally aligns the shared space given each annotation in time. (b) Applying the model of (Nicolaou et al. 2012) on a set of annotations. From top to bottom: original annotations, aligned shared space, derived annotation.

effect, by applying the model of Klami & Kaski (2008) on a set of signals, we obtain an estimation of the common characteristics of the signal (projected onto a maximally correlated space), while also isolating uninteresting factors which are signal-specific. Practically, this model is computationally efficient as it can lead to a closed-form SVD-based solution for a simple Gaussian noise model. Nevertheless, in order to apply this model on annotations, it is highly desirable that (i) the model takes dynamics into account, since temporally continuous annotations are rich in dynamics, and (ii) somehow alleviate temporal discrepancies, which appear due to e.g., annotator-specific lags. These extensions are proposed and implemented in (Nicolaou et al. 2012), where Markovian dependencies are imposed on both the shared and private latent spaces, while annotations are temporally aligned in order to alleviate for lags by introducing a time-warping process based on Dynamic Time Warping (DTW) on the sampled shared space of each annotation. Thus, the model is able to isolate uninteresting parts of the annotation (which are defined, in this context, as factors specific to an annotation and not shared) and learn a latent representation of the common, underlying signal which should express the "true annotation", ideally being free of all nuisances such as annotator bias and spike noise. The graphical model of (Nicolaou et al. 2012) is illustrated in Fig. 1.3, along with an example application. We note that both the model of (Nicolaou et al. 2012) and (Raykar et al. 2010) are able to incorporate data points (to which the annotations correspond) in the learning process. Furthermore, the application of CCA-related models to handle discrete/categorical annotations is still an open issue. This would require using similar methodologies such as (De Leeuw 2006, Niitsuma & Okada 2005), the

CCA model described in (Hamid et al. 2011) or by modifying the generative model used in (Klami & Kaski 2008, Nicolaou et al. 2012).

1.4 Future Directions

In this chapter we identified two key challenges in data-driven SSP, the joint signal-context and the annotation-annotator modeling. While modeling of the signal context and W5+ is crucial, few approaches to date have focused on this task and none have solved it in a satisfactory manner. The key difficulty is the lack of models for W5+ and the corresponding learning algorithms that are robust and scalable enough to produce models that generalize from posed or even real-world training datasets to arbitrary real-world, spontaneous query instances. Models that explicitly encode W5+ factors, such as the cs-CORF (Rudovic et al. 2013*b*) have the potential to generalize beyond training sets, but face difficulty in estimation. Related approaches based on tensor/multilinear decomposition (Lu et al. 2011) provide one avenue but face similar algorithmic and modeling (in particular, out-of-sample prediction) challenges. One practical direction to address the generalization problem has been to use the so-called domain-adaptation or transfer learning techniques (Pan & Yang 2010). These methods work well on simpler models but may face difficulty on full blown W5+. How to effectively integrate multifactor W5+ modeling, temporal information, and generalization ability remains a significant challenge.

Another related difficulty is the lack of sufficiently comprehensive spontaneous affect labeled datasets that can be used to estimate W5+ models. Databases such as MAHNOB `http://mahnob-db.eu` or SEMAINE are initial efforts in this direction. Nevertheless, providing comprehensive labeled data is challenging. Most current SSP models take into account neither the stimulus itself (a part of W5+) nor the annotators, including the errors and bias they may be imposing in the annotation process. We have described some initial approaches in the SSP domain that attempt to model the annotation process, annotator performance, bias, and temporal lag. However, many challenges continue to exist, including how to couple the predictive model estimation with the annotator modeling, how to track changes in annotator performance over time, how to select new or avoid underperforming experts, etc. Some of these and related problems are already being addressed in the domain of crowdsourcing (Quinn & Bederson 2011) and data-driven SSP can leverage those efforts. Related efforts have ensued in the context of multi-label learning(Tsoumakas et al. 2010), that focuses on learning a model that partitions the set of labels into relevant and irrelevant with respect to a query instance or orders the class labels according to their relevance to a query. Multi-label learning approaches have not yet been directly applied to problems in SSP, although they carry great potential.

References

Amin, A. M., Afzulpurkar, N. V., Dailey, M. N., Esichaikul, V. & Batanov, D. N. (2005), Fuzzy-c-mean determines the principle component pairs to estimate the degree of emotion from facial expressions, *in* 'Fuzzy Systems and Knowledge Discovery', Vol. 3613, pp. 484–493.

Bach, F. R. & Jordan, M. I. (2005), 'A probabilistic interpretation of canonical correlation analysis'.

Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. & Movellan, J. (2005), Recognizing facial expression: machine learning and application to spontaneous behavior, *in* 'CVPR', Vol. 2, pp. 568–573 vol. 2.

Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. & Movellan, J. (2006), 'Fully automatic facial action recognition in spontaneous behavior', *IEEE FG* pp. 223 –230.

Bazzo, J. & Lamar, M. (2004), Recognizing facial actions using gabor wavelets with neutral face average difference, *in* 'IEEE FG', pp. 505–510.

Black, M. J. & Yacoob, Y. (1997), 'Recognizing facial expressions in image sequences using local parameterized models of image motion', *Int'l J. of Computer Vision* **25**, 23–48.

Cai, D., He, X. & Han, J. (2007), 'Spectral regression for efficient regularized subspace learning', *IEEE ICCV* pp. 1–8.

Chang, K.-Y., Liu, T.-L. & Lai, S.-H. (2009), Learning partially-observed hidden conditional random fields for facial expression recognition, *in* 'IEEE CVPR', pp. 533–540.

Chew, S., Lucey, P., Lucey, S., Saragih, J., Cohn, J., Matthews, I. & Sridharan, S. (2012), 'In the pursuit of effective affective computing: The relationship between features and registration', *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **42**(4), 1006–1016.

Chu, W. & Ghahramani, Z. (2005), 'Gaussian processes for ordinal regression', *JMLR* **6**, 1019–1041.

Chu, W. & Keerthi, S. S. (2005), 'New approaches to support vector ordinal regression', *ICML* pp. 145–152.

Chu, W.-S., De la Torre, F. & Cohn, J. (2013), Selective transfer machine for personalized facial action unit detection, *in* 'IEEE CVPR', pp. 3515–3522.

Cohen, I., Sebe, N., Chen, L., Garg, A. & Huang, T. S. (2003), Facial expression recognition from video sequences: Temporal and static modelling, *in* 'Computer Vision and Image Understanding', pp. 160–187.

Cowie, R., Douglas-Cowie, E. & Cox, C. (2005), 'Beyond emotion archetypes: Databases for emotion modelling using neural networks', *Neural networks* **18**(4), 371–388.

Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M. & Schröder, M. (2000), 'feeltrace': An instrument for recording perceived emotion in real time, *in* 'ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion'.

Dai, P., Weld, D. S. et al. (2010), Decision-theoretic control of crowd-sourced workflows, *in* 'Twenty-Fourth AAAI Conference on Artificial Intelligence'.

Dai, P., Weld, D. S. et al. (2011), Artificial intelligence for artificial artificial intelligence, *in* 'Twenty-Fifth AAAI Conference on Artificial Intelligence'.

De Leeuw, J. (2006), 'Principal component analysis of binary data by iterated singular value decomposition', *Computational statistics & data analysis* **50**(1), 21–39.

Delannoy, J. & McDonald, J. (2008), Automatic estimation of the dynamics of facial expression using a three-level model of intensity, *in* 'IEEE FG', pp. 1–6.

Deng, L. & Li, X. (2013), 'Machine learning paradigms for speech recognition: An overview', *Audio, Speech, and Language Processing, IEEE Transactions on* **21**(5), 1060–1089.

der Maaten, L. V. & Hendriks, E. (2012), 'Action unit classification using active appearance models and conditional random fields', *Cognitive Processing* **13**(2), 507–518.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. (2003), 'Emotional speech: Towards a new generation of databases', *Speech communication* **40**(1), 33–60.

Ekman, P., Friesen, W. & Hager, J. (2002), *Facial Action Coding System (FACS): Manual*, A Human Face.

Ekman, P., Friesen, W. V. & Press, C. P. (1975), *Pictures of facial affect*, Consulting Psychologists Press.

Fasel, B. & Luettin, J. (2000), Recognition of asymmetric facial action unit activities and intensities, *in* 'ICPR', Vol. 1, pp. 1100–1103 vol.1.

Gholami, B., Haddad, W. M. & Tannenbaum, A. R. (2009), Agitation and pain assessment using digital imaging., *in* 'Int'l Conf. of the Engineering in Medicine and Biology Society', pp. 2176–2179.

Grauman, K. & Leibe, B. (2011), 'Visual Object Recognition', *Synthesis Lectures on Artificial Intelligence and Machine Learning* **5**(2), 1–181.

Gunes, H. & Piccardi, M. (2009), 'Automatic temporal segment detection and affect recognition from face and body display', *IEEE Trans. on Systems, Man, and Cybernetics* **39**(1), 64–84.

Gunes, H., Piccardi, M. & Pantic, M. (2008), 'From the lab to the real world: Affect recognition using multiple cues and modalities'.

Hamid, J., Meaney, C., Crowcroft, N., Granerod, J., Beyene, J. et al. (2011), 'Potential risk factors associated with human encephalitis: application of canonical correlation analysis', *BMC medical research methodology* **11**(1), 120.

Hamm, J., Kohler, C. G., Gur, R. C. & Verma, R. (2011), 'Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders', *Journal of Neuroscience Methods* **200**(2), 237 – 256.

Hammal, Z. & Cohn, J. F. (2012), 'Automatic detection of pain intensity', *ICMI* pp. 47–52.

He, X. & Niyogi, P. (2004), 'Locality preserving projections', *NIPS* .

Hess, U., Blairy, S. & Kleck, R. (1997), *Journal of Nonverbal Behavior* **21**(4), 241–257.

Hu, C., Chang, Y., Feris, R. & Turk, M. (2004), Manifold based analysis of facial expression, *in* 'IEEE CVPR'W', pp. 81–81.

Jain, S., Hu, C. & Aggarwal, J. (2011), Facial expression recognition with temporal modeling of shapes, *in* 'ICCV'W', pp. 1642–1649.

Jeni, L. A., Girard, J. M., Cohn, J. F. & Torre, F. D. L. (2013), 'Continuous au intensity estimation using localized, sparse facial feature space', *IEEE FG* pp. 1–7.

Kaltwang, S., Rudovic, O. & Pantic, M. (2012), 'Continuous pain intensity estimation from facial expressions', *ISVC* **7432**, 368–377.

Kapoor, A., Qi, Y. A. & Picard, R. W. (n.d.), *in* 'AMFG', IEEE Computer Society, pp. 195–202.

Khademi, M., Manzuri-Shalmani, M. T., Kiapour, M. H. & Kiaei, A. A. (2010), Recognizing combinations of facial action units with different intensity using a mixture of hidden markov models and neural network, *in* 'Proc. of the 9th Int'l Conf. on Multiple Classifier Systems', pp. 304–313.

Kim, M. & Pavlovic, V. (2010), 'Structured output ordinal regression for dynamic facial emotion intensity prediction', *ECCV* pp. 649–662.

Kimura, S. & Yachida, M. (1997), Facial expression recognition and its degree estimation, *in* 'IEEE ICPR', pp. 295–300.

Klami, A. & Kaski, S. (2008), 'Probabilistic approach to detecting dependencies between data sets', *Neurocomputing* **72**(1), 39–46.

Koelstra, S., Pantic, M. & Patras, I. (2010), 'A dynamic texture based approach to recognition of facial actions and their temporal models', *IEEE TPAMI* **32**, 1940–1954.

Lam, L. & Suen, S. (1997), 'Application of majority voting to pattern recognition: an analysis of its behavior and performance', *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* **27**(5), 553–568.

Lee, C. S. & Elgammal, A. (2005), Facial expression analysis using nonlinear decomposable generative models, *in* 'FG', pp. 17–31.

Lee, K. K. & Xu, Y. (2003), Real-time estimation of facial expression intensity, *in* 'IEE ICRA', Vol. 2, pp. 2567–2572 vol.2.

Lu, H., Plataniotis, K. N. & Venetsanopoulos, A. N. (2011), 'A survey of multilinear subspace learning for tensor data', *Pattern Recognition* .

Lucey, P., Cohn, J., Prkachin, K., Solomon, P. & Matthews, I. (2011), 'Painful data: The unbc-mcmaster shoulder pain expression archive database', *IEEE FG* pp. 57–64.

Mahoor, M., Cadavid, S., Messinger, D. & Cohn, J. (2009), 'A framework for automated measurement of the intensity of non-posed facial action units', *IEEE CVPR'W* pp. 74–80.

Mariooryad, S. & Busso, C. (2013), Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations, *in* 'in Affective Computing and Intelligent Interaction (ACII 2013)'.

Mavadati, S., Mahoor, M., Bartlett, K., Trinh, P. & Cohn, J. (2013), 'Disfa: A spontaneous facial action intensity database', *IEEE Trans. on Affective Comp.* **4**(2), 151–160.

McKeown, G., Valstar, M., Cowie, R., Pantic, M. & Schroder, M. (2012), 'The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent', *Affective Computing, IEEE Transactions on* **3**(1), 5–17.

Metallinou, A., Katsamanis, A., Wang, Y. & Narayanan, S. (2011), Tracking changes in continuous emotion states using body language and prosodic cues, *in* 'Acoustics,

Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on', IEEE, pp. 2288–2291.

Metallinou, A., Lee, C.-C., Busso, C., Carnicke, S., Narayanan, S. & Tx, D. (2010), The usc creativeit database: a multimodal database of theatrical improvisation, *in* 'Proceedings of the multimodal corpora workshop: advances in capturing, coding and analyzing, multimodality (MMC 2010)', pp. 64–68.

Nicolaou, M. A., Gunes, H. & Pantic, M. (2010), 'Automatic segmentation of spontaneous data using dimensional labels from multiple coders'.

Nicolaou, M. A., Gunes, H. & Pantic, M. (2011), 'Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space', *Affective Computing, IEEE Transactions on* **2**(2), 92–105.

Nicolaou, M. A., Pavlovic, V. & Pantic, M. (2012), Dynamic probabilistic CCA for analysis of affective behaviour, *in* 'Computer Vision–ECCV 2012', Springer, pp. 98–111.

Niitsuma, H. & Okada, T. (2005), Covariance and pca for categorical variables, *in* 'Advances in Knowledge Discovery and Data Mining', Springer, pp. 523–528.

Otsuka, T. & Ohya, J. (1997), Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences, *in* 'ICIP', Vol. 2, pp. 546–549 vol.2.

Padgett, C. & Cottrell, G. W. (1996), Representing face images for emotion classification, *in* 'NIPS', MIT Press, pp. 894–900.

Pan, S. J. & Yang, Q. (2010), 'A survey on transfer learning', *Knowledge and Data Engineering, IEEE Transactions on* **22**(10), 1345–1359.

Pantic, M. & Bartlett, M. (2007), *Machine Analysis of Facial Expressions*, I-Tech Education and Publishing, pp. 377–416.

Pantic, M. & Patras, I. (2005), 'Detecting facial actions and their temporal segments in nearly frontal-view face image sequences', *Proc. of IEEE Int'l Conf. Systems, Man and Cybernetics* pp. 3358–3363.

Pantic, M. & Patras, I. (2006), 'Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences', *IEEE Trans. on Systems, Man and Cybernetics - Part B* **36**(2), 433–449.

Pantic, M. & Rothkrantz, L. J. (2004), 'Facial action recognition for facial expression analysis from static face images', *Trans. Sys. Man Cyber. Part B* **34**(3), 1449–1461.

Pentland, A. (2007), 'Social Signal Processing', *Signal Processing Magazine, IEEE* **24**(4), 108–111.

Poppe, R. (2010), 'A survey on vision-based human action recognition', *Image and vision computing* **28**(6), 976–990.

Posner, J., Russell, J. A. & Peterson, B. S. (2005), 'The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology', *Development and psychopathology* **17**(03), 715–734.

Quinn, A. J. & Bederson, B. B. (2011), Human computation: a survey and taxonomy of a growing field, *in* 'CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM Request Permissions.

Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L. & Moy, L. (2009), Supervised learning from multiple experts: whom to trust when everyone lies a bit, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ACM, pp. 889–896.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L. & Moy, L. (2010), 'Learning from crowds', *The Journal of Machine Learning Research* **99**, 1297–1322.

Reilly, J., Ghent, J. & McDonald, J. (2006), 'Investigating the dynamics of facial expression', *Lecture Notes in Computer Science* **4292**, 334–343.

Rudovic, O., Pavlovic, V. & Pantic, M. (2012*a*), 'Kernel conditional ordinal random fields for temporal segmentation of facial action units', *IEEE ECCV'W* .

Rudovic, O., Pavlovic, V. & Pantic, M. (2012*b*), 'Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation', *IEEE CVPR* pp. 2634–2641.

Rudovic, O., Pavlovic, V. & Pantic, M. (2013*a*), 'Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields', *ISVC* .

Rudovic, O., Pavlovic, V. & Pantic, M. (2013*b*), Context-sensitive conditional ordinal random fields for facial action intensity estimation, *in* 'ICCV'W', pp. 492–499.

Ruta, D. & Gabrys, B. (2005), 'Classifier selection for majority voting', *Information fusion* **6**(1), 63–81.

Savrana, A., Sankur, B. & Bilgeb, M. (2012), 'Regression-based intensity estimation of facial action units', *Image and Vision Computing* .

Shan, C. (2007), Inferring facial and body language, PhD thesis, Queen Mary, University of London.

Shan, C., Gong, S. & Mcowan, P. W. (2005), 'Appearance manifold of facial expression', *Lecture Notes in Comp. Science* **3766**, 221–230.

Shan, C., Gong, S. & Mcowan, P. W. (2006), Dynamic facial expression recognition using a bayesian temporal manifold model, *in* 'BMVC', pp. 297–306.

Shan, C., Gong, S. & McOwan, P. W. (2009), 'Facial expression recognition based on local binary patterns: A comprehensive study', *Image and Vision Comp.* **27**(6), 803–816.

Shang, L. & Chan, K.-P. (2009), Nonparametric discriminant hmm and application to facial expression recognition, *in* 'IEEE CVPR', pp. 2090–2096.

Simon, T., Nguyen, M. H., De la Torre, F. & Cohn, J. F. (2010), Action unit detection with segment-based svms, *in* 'IEEE CVPR'.

Tian, Y.-L. (2004), Evaluation of face resolution for expression analysis, *in* 'IEEE CVPR'W', pp. 82–82.

Tong, Y., Liao, W. & Ji, Q. (2007), 'Facial action unit recognition by exploiting their dynamic and semantic relationships', *IEEE TPAMI* **29**(10), 1683–1699.

Tsoumakas, G., Katakis, I. & Vlahavas, I. (2010), Mining multi-label data, *in* 'Data mining and knowledge discovery handbook', Springer, pp. 667–685.

Tucker, L. R. (1958), 'An inter-battery method of factor analysis', *Psychometrika* **23**(2), 111–136.

Valstar, M. F. & Pantic, M. (2012), 'Fully automatic recognition of the temporal phases of facial actions', *IEEE Trans. on Syst., Man and Cyber.* **42**, 28–43.

Vinciarelli, A., Pantic, M. & Bourlard, H. (2009), 'Social signal processing: Survey of an emerging domain', *Image and Vision Computing* .

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F. & Schroeder, M. (2012), 'Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing', *Affective Computing, IEEE Transactions on* **3**(1), 69–87.

Wang, S., Quattoni, A., Morency, L.-P., Demirdjian, D. & Darrell, T. (2006), 'Hidden conditional random fields for gesture recognition', *IEEE CVPR* pp. 1097–1104.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E. & Cowie, R. (2008), Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies., *in* 'INTERSPEECH', pp. 597–600.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S. & Yan, S. (2010), 'Sparse representation for computer vision and pattern recognition', *Proceedings of the IEEE* **98**(6), 1031–1044.

Yan, Y., Rosales, R., Fung, G. & Dy, J. (2012), 'Modeling multiple annotator expertise in the semi-supervised learning scenario', *arXiv preprint arXiv:1203.3529* .

Yang, P., Liu, Q. & Metaxas, D. N. (2009*a*), 'Boosting encoded dynamic features for facial expression recognition', *Pattern Recognition Letters* (2), 132 – 139.

Yang, P., Liu, Q. & Metaxas, D. N. (2009*b*), 'Rankboost with l1 regularization for facial expression recognition and intensity estimation', *IEEE ICCV* pp. 1018–1025.

Zhang, Y. & Ji, Q. (2005), 'Active and dynamic information fusion for facial expression understanding from image sequences', *IEEE TPAMI* **27**(5), 699–714.