Dimensional Affect Recognition using Continuous Conditional Random Fields

Tadas Baltrušaitis Ntombikayise Banda Peter Robinson
Computer Laboratory, University of Cambridge
tb346@cl.cam.ac.uk nb395@cl.cam.ac.uk pr10@cl.cam.ac.uk

***Abstract*— During everyday interaction people display various non-verbal signals that convey emotions. These signals are multi-modal and range from facial expressions, shifts in posture, head pose, and non-verbal speech. They are subtle, continuous and complex. Our work concentrates on the problem of automatic recognition of emotions from such multi-modal signals. Most of the previous work has concentrated on classifying emotions as belonging to a set of categories, or by discretising the continuous dimensional space. We propose the use of Continuous Conditional Random Fields (CCRF) in combination with Support Vector Machines for Regression (SVR) for modeling continuous emotion in dimensional space. Our Correlation Aware Continuous Conditional Random Field (CA-CCRF) exploits the non-orthogonality of emotion dimensions. By using visual features based on geometric shape and appearance, and a carefully selected subset of audio features we show that our CCRF and CA-CCRF approaches outperform previously published baselines for all four affective dimensions of valence, arousal, power and expectancy.**

I. Introduction

Reliable automated recognition of human emotions is crucial before the development of affect sensitive systems is possible [17]. Humans display affective behavior that is multi-modal, subtle and complex. People are adept at expressing themselves and interpreting others through the use of non-verbal cues such as vocal prosody, facial expressions, eye gaze, various hand gestures, head motion and posture. All of these modalities contain important affective information that can be used to infer the emotional state of a person automatically [8], [21], [32].

Most work in automated emotion recognition so far [32] has focused on analysis of the six discrete basic emotions [4] (happiness, sadness, surprise, fear, anger and disgust). However, a single label (or multiple discrete labels from a small set) might not describe the complexity of an affective state very well. There has been a move to analyse emotional signals along a set of small number of latent dimensions, providing a continuous rather than a categorical view of emotions. Examples of such affective dimensions are power (sense of control); valence (pleasant vs. unpleasant); activation (relaxed vs. aroused); and expectancy (anticipation). Fontaine *et al.* [6] argue that these four dimensions account for most of the distinctions between everyday emotion categories, and hence form a good set for automatic analysis.

Affective computing researchers have started exploring the dimensional representation of emotion [8]. The problem of dimensional affect recognition is often posed as a binary classification problem [8], [27] (active vs. passive etc.) or even as a four-class one (classification into quadrants of a 2D space). In our work, however, we represent the problem of dimensional affect recognition as a regression one.

In addition, most of the work so far has concentrated on analysing different modalities in isolation rather than looking for ways to fuse them [8], [32]. This is partly due to the limited availability of suitably labeled multi-modal datasets and the difficulty of fusion itself, as the optimal level at which the features should be fused is still an open research question [8], [32]. Our approach can fuse multiple modalities effectively, outperforming early SVR fusion.

Conditional Random Fields [10] (CRF) and various extensions have proven very useful for emotion recognition tasks [20], [30]. However, conventional CRF cannot be directly applied to continuous emotion prediction, as they model the output as being discrete rather than continuous. In our work, we propose the use of Continuous Conditional Random Fields [18] (CCRF) in combination with SVRs for the task of continuous emotion recognition.

We apply our CCRF model for the task of continuous dimensional emotion prediction on the AVEC 2012 subset of SEMAINE dataset [26]. We show the benefits of using this approach for emotion recognition by outperforming the SVR baseline. Furthermore, we present our Correlation Aware Continuous Conditional Random Field (CA-CCRF) model which exploits the correlations between the emotion dimensions, further improving the emotion prediction accuracy for some of the dimensions.

In our work we also demonstrate the benefit of using facial geometry/shape deformations of face for spontaneous affect recognition from video sequences. Such features are often ignored in favour of appearance based features, thus losing useful emotional information [9]. This is due to the difficulty of acquiring a neutral expression from which facial shape deformation can be measured. Our work shows how to extract a neutral expression and demonstrates the utility of geometry alongside appearance for emotion prediction.

The main contributions of our research are as follows:

- A fully continuous CCRF emotion prediction model that exploits temporal properties of the emotion signal
- Exploiting the correlations between the emotional dimensions using our CA-CCRF model
- A novel way to fuse multi-modal emotional data
- A demonstration of the utility of facial geometry for

continuous affect recognition
- Freely available implementation of CCRF[1]

II. Background

As this paper concentrates on the recognition of emotion in a dimensional space we present the previous work on this specific task. For recent surveys of dimensional and categorical affect recognition see Zeng *et al.* [32], Gunes and Pantic [8], and Gunes *et al.* [7].

Nicolaou *et al.* [11] present experiments for classification of spontaneous affect based on Audio-Visual features using coupled Hidden Markov Models which allow them to model temporal correlations between different cues and modalities. They also show the benefits of using the likelihoods produced from separate (C)HMMs as input to another classifier as a fusion approach, rather than picking the label with a maximum likelihood. In contrast with our work, they perform classification rather than regression.

Nicolaou *et al.* [12] propose the use of Output-Associative Relevance Vector Machine (OA-RVM) for dimensional and continuous prediction of emotions based on automatically tracked facial feature points. Their proposed regression framework exploits the inter-correlation between the valence and arousal dimensions by including in their mode the initial output estimation together with their input features. In addition, OA-RVM regression attempts to capture the temporal dynamics of output by employing a window that covers a set of past and future outputs. Our approach also captures temporal dynamics, is a regression one, and exploits correlations between dimensions.

Of special relevance to our work is the work done by Wöllmer *et al.* [30] which uses Conditional Random Fields (CRF) for discrete emotion recognition by quantising the continuous labels for valence and arousal based on a selection of acoustic features. In addition, they use Long Short-Term Memory Recurrent Neural Networks to perform regression analysis on these two dimensions. Both of the approaches demonstrate the benefits of including temporal information when predicting emotion.

More recently Ramirez *et al.* [20] proposed the use of Latent Dynamic Conditional Random Fields (LDCRF). Their approach attempts to learn the hidden dynamics between input features by incorporating hidden state variables that can model the sub-structure of gesture sequences. Their approach was particularly successful in predicting dimensional emotions from the visual signal. However, the LDCRF model can model only discrete output variables, hence the problem was posed as a classification one.

III. Continuous CRF

We want to model the affect continuously rather than turning this problem into a classification one by discretising the signal as done by many previous approaches [8]. Furthermore, we want to model the temporal relationships between each time step, since emotion has temporal properties and is not instantaneous. A recent and promising approach that would allow us to model such temporal relationships is the Continuous Conditional Random Fields [18] (CCRF). It is an extension of the classic Conditional Random Fields [10] (CRF) to the continuous case. We extend the original CCRF model so it can be used for continuous emotion prediction.

A. Model definition

CCRF is an undirected graphical model where conditional probability $P(y|x)$ is modeled explicitly. It is a discriminative approach, which has shown promising results for sequence labeling and segmentation [29]. This is in contrast to generative models where a joint distribution $P(y, x)$ is modeled instead. The graphical model that represents our CCRF for emotion prediction is shown in Figure 1.

In our discussion we will use the following notation: $\{\mathbf{x}_1^{(q)}, \mathbf{x}_2^{(q)}, \ldots, \mathbf{x}_n^{(q)}\}$ is a set of observed input variables $\{y_1^{(q)}, y_2^{(q)}, \ldots, y_n^{(q)}\}$ is a set of output variables that we wish to predict, $\mathbf{x}_i^{(q)} \in \mathcal{R}^m$ and $y_i^{(q)} \in \mathcal{R}$, n is the number of frames/time-steps in a sequence, m is the number of predictors used, q indicates the q^{th} sequence of interest. When there is no ambiguity, q is omitted for clarity.

Our CCRF model for a particular sequence is a conditional probability distribution with the probability density function:

$$P(\mathbf{y}|\mathbf{X}) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi) d\mathbf{y}} \tag{1}$$

$$\Psi = \sum_i \sum_{k=1}^{K1} \alpha_k f_k(y_i, \mathbf{X}) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j, \mathbf{X}) \tag{2}$$

Above $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is the set of input feature vectors (can be represented as a matrix with per frame observations as rows), $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ is the unobserved variable. $\int_{-\infty}^{\infty} \exp(\Psi) d\mathbf{y}$ is the normalisation (partition) function which makes the probability distribution a valid one (by making it sum to 1). Following the convention of Qin *et al.* [18] we call f_k vertex features, and g_k edge features. The model parameters $\alpha = \{\alpha_1, \alpha_2, \ldots \alpha_{K1}\}$, and $\beta = \{\beta_1, \beta_2, \ldots \beta_{K2}\}$ would be provided for inference and need to be estimated during learning.

B. Feature functions

We define two types of features for our CCRF model, vertex features f_k and edge features g_k.

$$f_k(y_i, \mathbf{X}) = -(y_i - \mathbf{X}_{i,k})^2, \tag{3}$$

$$g_k(y_i, y_j, \mathbf{X}) = -\frac{1}{2} S_{i,j}^{(k)} (y_i - y_j)^2. \tag{4}$$

Vertex features f_k represent the dependency between the $\mathbf{X}_{i,k}$ and y_k, for example dependency between a static emotion prediction from a regressor and the actual emotion label. Intuitively, the corresponding α_k for vertex feature f_k represents the reliability of the k^{th} predictor. This is particularly useful for multimodal fusion, as it models the reliability of a particular signal for a particular emotion,

[1] http://www.cl.cam.ac.uk/research/rainbow/projects/ccrf/

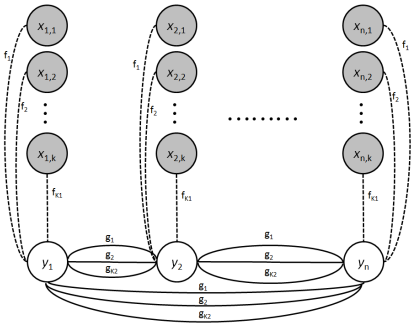

Fig. 1. Graphical representation of the CCRF model. $x_{i,k}$ represents the k^{th} feature of the j^{th} observation (corresponding to the i^{th} observation of the sequence), and y_i is the unobserved variable we want to predict. Dashed lines represent the connection of observed to unobserved variables (f_k vertex features), so the first predictor is connected using f_1, whilst the k^{th} predictor is connected using f_k. The solid lines show connections between the unobserved variables (edge features), the first connection is controlled by g_1, the k^{th} connection is controlled by g_k. In our model all the output variables y_i are connected to each other (edge functions can break the connections by setting the appropriate $S_{i,j}$ to 0)

for example the CCRF model could learn that the facial appearance might be more important in predicting valence than the audio signal.

Edge features g_k represent the dependencies between observations y_i and y_j, for example how related is the emotion prediction at time step j to the one at time step i. This is also affected by the similarity measure $S^{(k)}$. Because we are using a fully connected model, the similarities $S^{(k)}$ allow us to control the strength or existence of such connections. We define two types of similarities in our work:

$$S_{i,j}^{(\text{neighbor})} = \begin{cases} 1, & |i-j| = n \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$S_{i,j}^{(\text{distance})} = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{\sigma}) \tag{6}$$

By varying n for neighbor similarity we can construct a family of similarities, this allows us to connect the observation y_i not only to y_{i-1}, but also to y_{i-2} and so on. By varying σ for the distance similarity we create another set of similarities that control how related are the y terms based on how similar the x terms are. Our framework allows for easy creation of different similarity measures which could be appropriate for other applications.

The learning phase of CCRF will determine which of the similarities is important for the dataset of interest. For example, it can learn that for one emotion neighbor similarities are more important than for others.

Similarly to Radosavljevic *et al.* [19] and Qin *et al.* [18], our feature functions model the square error between prediction and a feature. Therefore, each element of the feature vector $\mathbf{x}_i$ should be already predicting the unobserved variable y_i. This can be achieved using Support Vector Regression (used in our work), linear regression, neural networks etc.

C. Learning

In this section we describe how to estimate the parameters $\{\alpha, \beta\}$ of a CCRF with quadratic vertex and edge functions. We are given training data $\{\mathbf{x}^{(q)}, \mathbf{y}^{(q)}\}_{q=1}^{M}$ of M sequences, where each $\mathbf{x}^{(q)} = \{\mathbf{x}_1^{(q)}, \mathbf{x}_2^{(q)}, \ldots, \mathbf{x}_n^{(q)}\}$ is a sequence of inputs and each $\mathbf{y}^{(q)} = \{y_1^{(q)}, y_2^{(q)}, \ldots, y_n^{(q)}\}$ is a sequence of real valued outputs. We also use the matrix $\mathbf{X}$ to denote the concatenated sequence of inputs.

In learning we want to pick the α and β values that optimise the conditional log-likelihood of the CCRF on the training sequences:

$$L(\alpha, \beta) = \sum_{q=1}^{M} \log P(\mathbf{y}^{(q)}|\mathbf{x}^{(q)}) \tag{7}$$

$$(\bar{\alpha}, \bar{\beta}) = \arg\max_{\alpha,\beta}(L(\alpha, \beta)) \tag{8}$$

As the problem is convex [18], the optimal parameter values can be determined using standard techniques such as stochastic gradient ascent.

It helps with the derivation of the partial derivatives of Eq.(7) and with explanation of inference to convert the Eq.(1) into multivariate Gaussian form (see Appendix A in supplementary material for a detailed derivation).

$$P(\mathbf{y}|\mathbf{X}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)), \tag{9}$$

$$\Sigma^{-1} = 2(A + B) \tag{10}$$

The diagonal matrix A represents the contribution of α terms (vertex features) to the covariance matrix, and the symmetric B represents the contribution of the β terms (edge features).

$$A_{i,j} = \begin{cases} \sum_{k=1}^{K1} \alpha_k, & i = j \\ 0, & i \neq j \end{cases} \tag{11}$$

$$B_{i,j} = \begin{cases} (\sum_{k=1}^{K2} \beta_k \sum_{r=1}^{n} S_{i,r}^k) - (\sum_{k=1}^{K2} \beta_k S_{i,j}^k), & i = j \\ -\sum_{k=1}^{K2} \beta_k S_{i,j}^k, & i \neq j \end{cases} \tag{12}$$

We also define a further vector $\mathbf{b}$, that describes the linear terms in the distribution, and a helpful variable μ which is the mean value of the Gaussian CCRF distribution.

$$\mathbf{b}_i = 2\sum_{k=1}^{K1} \alpha_k \mathbf{X}_{i,k} \tag{13}$$

$$\mu = \Sigma\mathbf{b} \tag{14}$$

We can now write down the partial derivatives of the $\log P(\mathbf{y}|\mathbf{X})$. Please refer to Appendix B in supplementary material for details.

$$\frac{\partial \log(P(\mathbf{y}|\mathbf{X}))}{\partial \alpha_k} = -\mathbf{y}^T\mathbf{y} + 2\mathbf{y}^T\mathbf{X}_{*,k}^T - 2\mathbf{X}_{*,k}\mu + \mu^T\mu + \text{tr}(\Sigma) \tag{15}$$

$$\frac{\partial \log(P(\mathbf{y}|\mathbf{X}))}{\partial \beta_k} = -\mathbf{y}^T B^{(k)}\mathbf{y} + \mu^T B^{(k)}\mu + \text{Vec}(\Sigma)^T\text{Vec}(B^{(k)}) \tag{16}$$

$$B^{(k)} = \begin{cases} (\sum_{r=1}^{n} S_{i,r}^{(k)}) - S_{i,j}^{(k)}, & i = j \\ -S_{i,j}^{(k)}, & i \neq j \end{cases} \tag{17}$$

In order to guarantee that our partition function is integrable we constrain $\alpha_k > 0$ and $\beta_k > 0$ [18], [19]. Such

constrained optimisation can be achieved by using partial derivatives with respect to $\log \alpha_k$ and $\log \beta_k$ instead of just α_k and β_k. We also add a regularisation term in order to avoid overfitting. The regularisation is controlled by λ_α and λ_β hyper-parameters (determined during cross-validation).

$$\frac{\partial \log(P(\mathbf{y}|\mathbf{X}))}{\partial \log \alpha_k} = \alpha_k\left(\frac{\partial \log(P(\mathbf{y}|\mathbf{X}))}{\partial \alpha_k} - \lambda_\alpha \alpha_k\right) \quad (18)$$

$$\frac{\partial \log(P(\mathbf{y}|\mathbf{X}))}{\partial \log \beta_k} = \beta_k\left(\frac{\partial \log(P(\mathbf{y}|\mathbf{X}))}{\partial \beta_k} - \lambda_\beta \beta_k\right) \quad (19)$$

Using these partial derivatives we can write down the CCRF learning algorithm that uses stochastic gradient ascent.

Algorithm 1 Our CCRF learning algorithm

Require: $\{\mathbf{X}^{(q)}, \mathbf{y}^{(q)}, S_q^{(1)}, S_q^{(2)}, \ldots, S_q^{(K)}\}_{q=1}^{M}$
Params: number of iterations T, learning rate ν, $\lambda_\alpha, \lambda_\beta$
Initialise parameters $\{\alpha, \beta\}$
for r = 1 **to** T **do**
  for i = 1 **to** N **do**
    Compute gradients of current query (Eqs.(18),(19))
    $\log \alpha_k = \log \alpha_k + \nu \frac{\partial \log(P(\mathbf{y}|\mathbf{X}))}{\partial \log \alpha_k}$
    $\log \beta_k = \log \beta_k + \nu \frac{\partial \log(P(\mathbf{y}|\mathbf{x}))}{\partial \log \beta_k}$
    Update $\{\alpha, \beta\}$
  end for
end for
return $\{\bar{\alpha}, \bar{\beta}\} = \{\alpha, \beta\}$

D. Inference

Because our CCRF model can be viewed as a multivariate Gaussian, inferring $\mathbf{y}$ values that maximise $P(\mathbf{y}|\mathbf{x})$ is straightforward. The prediction is the mean value of the distribution.

$$\mathbf{y}' = \arg\max_{\mathbf{y}}(P(\mathbf{y}|\mathbf{X})) = \mu = \Sigma \mathbf{b} \quad (20)$$

As a practical note, depending on the regularisation terms when training, and the types of similarity functions chosen, and especially if the input variables $\mathbf{x}$ are very noisy, the learning algorithm can sometimes oversmooth the data, as CCRF learns to trust temporal consistency much more than the $\mathbf{x}$ observations. An effect of this is inference producing a dampened signal. To combat this, one could use very high λ_β values to force the training to rely on the α predictions more than on temporal elements controlled by β, this would, however, be at a cost of retaining a noisy signal. Alternatively, we recommend learning a scaling term s from the same training data (after the CCRF training is finished). Then the inference becomes $\mathbf{y} = s \cdot \mu$, leading to a correctly scaled signal.

Furthermore, if multiple CCRF models are to be trained (as is the case for dimensional emotions), we recommend using the Z-score of both input $\mathbf{x}$ and output $\mathbf{y}$ variables. Then the same learning rate can be used on all of them. This also helps if we want to use predictions from other dimensions in a single CCRF, as is done in our CA-CCRF.

IV. Video Features

In order to analyse the geometry of the face, in addition to knowing where to analyse the appearance features, we need to track landmark points on the face together with the head pose. For tracking faces we use a modified version of the CLM-GAVAM tracker [2] for facial expression tracking.

A. CLM Tracker

The CLM tracker is combined with a Generalised Adaptive View Based Appearance Model to help with pose estimation [2]. In addition, we extend the CLM tracker to handle identity/morphology and expressions separately. This is needed in order to decouple shape deformations due to identity from deformations due to expression.

The CLM model we use can be described by parameters $\mathbf{p} = [s, \mathbf{R}, \mathbf{q_m}, \mathbf{q_e}, \mathbf{t}]$ that can be varied to acquire various instances of the model: the scale factor s; object rotation $\mathbf{R}$ (first two rows of a 3D rotation matrix); 2D translation $\mathbf{t}$; a vector describing non-rigid variation of the identity shape $\mathbf{q_m}$; and expression shape $\mathbf{q_e}$ (similar to a model used by Amberg *et al.* [1]). Our point distribution model (PDM) is:

$$\mathbf{x}_i = s \cdot \mathbf{R}(\bar{\mathbf{x}}_\mathbf{i} + \mathbf{\Phi}_i \mathbf{q_m} + \mathbf{\Psi}_i \mathbf{q_e}) + \mathbf{t}. \quad (21)$$

Here $\mathbf{x}_i = (x, y)$ denotes the 2D location of the i^{th} feature point in an image, $\bar{\mathbf{x}}_i = (X, Y, Z)$ is the mean value of the i^{th} element of the PDM in the 3D reference frame, and the vector $\mathbf{\Phi}_i$ is the i^{th} eigenvector obtained from the training set that describes the linear variations of non-rigid shape of this feature point in morphology space (constructed from the Basel 3DMM dataset [15]), and the vector $\mathbf{\Psi}_i$ is the i^{th} eigenvector obtained from the training set that describes the linear variations of non-rigid shape in expression space (constructed from BU-4DFE [31]).

In CLM we estimate the maximum *a posteriori* probability (MAP) of the face model parameters $\mathbf{p}$:

$$p(\mathbf{p}|\{l_i{=}1\}_{i=1}^n, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=1}^{n} p(l_i{=}1|\mathbf{x}_i, \mathcal{I}), \quad (22)$$

where $l_i \in \{1, -1\}$ is a discrete random variable indicating if the i^{th} feature point is aligned or misaligned, $p(\mathbf{p})$ is the prior probability of the model parameters $\mathbf{p}$, and $\prod_{i=1}^{n} p(l_i = 1|\mathbf{x}_i, \mathcal{I})$ is the joint probability of the feature points $\mathbf{x}$ being aligned at a particular point $\mathbf{x}_i$, given an intensity image $\mathcal{I}$.

Patch experts are used to calculate $p(l_i = 1|\mathbf{x}_i, \mathcal{I})$, which is the probability of a feature being aligned at point $\mathbf{x}_i$ (Eq. 21). As a probabilistic patch expert we use an SVM classifier combined with a logistic regressor.

$$p(l_i|\mathbf{x}_i, \mathcal{I}) = \frac{1}{1 + e^{d\mathcal{C}_i(\mathbf{x}_i;\mathcal{I}) + c}} \quad (23)$$

Here $\mathcal{C}_i$ is the output of the SVM classifier, for the i^{th} feature, c is the logistic regressor intercept, and d the regression coefficient. The classifier is thus:

$$\mathcal{C}_i(\mathbf{x}_i; \mathcal{I}) = \mathbf{w}_i^T \mathcal{P}(\mathcal{W}(\mathbf{x}_i; \mathcal{I})) + b_i, \quad (24)$$

where $\{\mathbf{w}_i, b_i\}$ are the weights and biases associated with a particular SVM. Here $\mathcal{W}(\mathbf{x}_i; \mathcal{I})$ is a vectorised version of a

local $n \times n$ image patch centered around $\mathbf{x}_i$. $\mathcal{P}$ normalises the vectorised patch to zero mean and unit variance.

We employ a common two step CLM fitting strategy [3], [22]; performing an exhaustive local search around the current estimate of feature points leading to a response map around every feature point, and then iteratively updating the model parameters to maximise Eq.(22) until a convergence metric is reached. For fitting we use Regularised Landmark Mean-Shift (RLMS) [22].

In order to fit CLM using our split PDM we first optimise with respect to the morphology parameters q_m, followed by expression parameters q_e. After a frame is successfully tracked in a video sequence the morphology parameters are fixed, and only expression parameters are optimised.

As a prior $p(\mathbf{p})$ for parameters $\mathbf{p}$, we assume that the non-rigid shape parameters $\mathbf{q_m}$, $\mathbf{q_e}$ and vary according to a Gaussian distribution with the variance of the i^{th} parameter corresponding to the eigenvalue of the i^{th} mode of non-rigid deformation; the rigid parameters $s, \mathbf{R}$, and $\mathbf{t}$ follow a non-informative uniform distribution.

We use Baltrušaitis *et al.*'s freely available CLM implementation [2] that is based on Saragih *et al.*'s RLMS algorithm [22]. There are several differences between the available implementation and the algorithm described in Saragih *et al.* [22]: the model is trained using face images at various orientations in addition to frontal ones, secondly, CLM is coupled with a GAVAM head-pose tracker for more accurate estimates of head-pose. For more details on fitting see Baltrušaitis *et al.* [2] and Saragih *et al.* [22].

B. Geometric features

In order to extract the geometry features of facial expressions one needs to establish the neutral facial expression from which the expression is measured. We cannot rely on the geometric configuration of the initial frame, as not all of them start with neutral expressions. In order to extract a neutral expression we use our PDM from Eq.(21) which separates the expression and morphology subspaces.

After the fitting has been performed we can use the expression parameters $\mathbf{q}_e$ for describing the deformations due to expression. In our PDM $\mathbf{q}_e$ has 27 dimensions, this feature vector can be used with SVR to predict emotion.

C. Appearance-based features

We augment the geometric features with appearance-based ones, specifically local binary patterns (LBPs) which have been widely used in facial analysis tasks due to their tolerance against illumination variations, and their computational simplicity [28]. The original LBP operator introduced by Ojala *et al.* [13] is formulated as follows:

The local binary code for each pixel, which assumes centre position (x_c, y_c) with respect to its neighbours, is

$$\begin{aligned} LBP_P(x_c, y_c) &= \sum_{n=0}^{P-1} s(i_n - i_c)2^n, \\ s(x) &= \begin{cases} 1 & x >= 0 \\ 0 & x < 0 \end{cases} \end{aligned} \tag{25}$$

where P represents the number of neighbouring pixels, i_n the intensity value of a neighbour pixel and i_c the intensity value of the centre pixel.

We employ an extension of the LBP operator which seeks to combine motion features with appearance features thus incorporating the temporal dynamics of an image sequence [33]. This is achieved by concatenating local binary pattens on three orthogonal planes (LBP-TOP): XY, XT and YT. The operator is expressed as:

$$LBP - TOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$$

where the notation $(P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T)$ denotes a neighbourhood of P points equally sampled on a circle of radius R on XY, XT and YT planes respectively. An LBP code is extracted from the XY, XT and YT planes for all pixels, and statistics of the three different planes are obtained and then concatenated into a single histogram. This is demonstrated in Figure 2. This technique incorporates spatial domain information through the XY plane, and spatio-temporal co-occurrence statistics through the XT and YT planes. We refer the reader to Zhao *et al.*'s [33] article for derivation of the LBP-TOP descriptor.

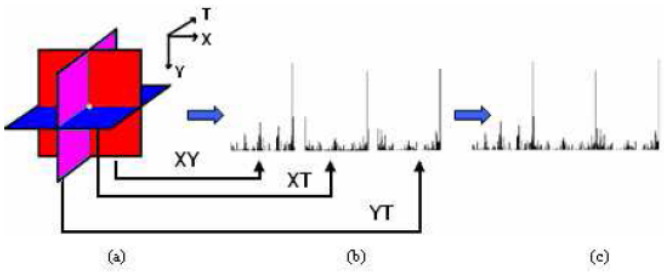

Fig. 2. (a) Three planes from which spatio-temporal local features are extracted (b) LBP histogram from each plane (c) Concatenated feature histogram. [33]

In our approach, we use the facial feature points from the CLM-GAVAM tracker to extract frontal faces from an image sequence. In order to extract a frontal face we use perspective warping from the current tracked points to the neutral reference frame, this also ensures size uniformity. The extracted faces are divided into a 3x3 non-overlapping grid, and LBP-TOP features are extracted for each block in the grid. We apply uniform patterns which produce $P(P-1)+3$ output labels (instead of 2^P) resulting in a significant dimension reduction to a 59-dimensional histogram per image block (for $P = 8, R = 3$). A complete feature vector is obtained by concatenating the block histograms for each plane resulting in a 1593-dimensional vector.

D. Motion features

Head gestures are an integral part of human communication as they convey a range of meanings and emotion. They involve a range of dynamics such as head orientation, rhythmic patterns, amplitude and speed of movement which act as indicators of affective states.

TABLE I

DESCRIPTION OF THE AUDIO FEATURES USED IN THIS WORK.

Feature	Description	Motivation
Energy (in dB)	reflects the perceived loudness of the speech signal	has been found to have a high, positive correlation with arousal [16], with increased intensity correlating well with valence [25]
Articulation rate	is calculated by identifying the number of syllables per second	has been found to be positively correlated with arousal [25]
Fundamental frequency (f0)	is the base frequency of the speech signal (that is, the frequency the vocal folds are vibrating at during voiced speech segments)	has been found to have a high, positive correlation with arousal [16]; and a positive correlation between lower f0 and power [25]
Peak slope	is a measure suitable for the identification of breathy to tense voice qualities	there is evidence of a positive correlation between warm voice quality and valence [25]
Spectral stationarity	captures the fluctuations and changes in the voice signal; a measure of the speech monotonicity	monotonicity in speech is associated with low activity and negative valence [24]

The CLM-GAVAM tracker estimates 6 degrees-of-freedom of head pose corresponding to head rotation and translation. We track the intensity variation of rigid head motion by calculating the standard deviation of the rotational and translational parameters, a measure which takes into account the amplitude range and speed of change in head motion. In addition to these statistics, the Euclidean norm of all rotational parameters and that of translational parameters are added to describe the overall head movement. This results in the following 8-dimensional feature vector:

$$\left[\sigma_{r_x}, \sigma_{r_y}, \sigma_{r_z}, \sigma_{t_x}, \sigma_{t_y}, \sigma_{t_z}, \sigma_{r_{xyz}}, \sigma_{t_{xyz}}\right]$$

where r corresponds to rotation parameters and t to translation parameters.

V. AUDIO FEATURES

Vocal affect recognition analyses how things are said by extracting non-verbal information from speech. Scherer [23] states that emotion may produce changes in respiration, phonation and articulation, which in turn affect the acoustic features of the signal. It is therefore the variations in the acoustic measures that makes it possible to discriminate between different emotional states. We adopt prosodic features used in [14]. Table I lists the adopted features and provides motivations for their choice. Details of their extraction algorithms can be found in [14].

VI. FINAL SYSTEM

The final emotion prediction system proposed in our work can be seen in Figure 3. Our model depends on the per time step predictions from the previous layer. We use SVR, but this could be replaced by any other continuous predictor, such as linear regression, artificial neural networks or others. The features that are used with each SVR are explained in more detail in the Sections IV and V. The CCRF model used is explained in Section III.

CCRF can use any number of SVR predictors, and we explore various combinations of them in our evaluation section. First, we have a system that just uses a prediction from an audio-visual SVR as its input ($K = 1$). Secondly, we use four SVR predictors (audio, shape, appearance, pose) of the same dimension ($K = 4$). Finally, as the emotional dimensions do not form an orthogonal set, we exploit the correlations between them using our Correlation Aware CCRF (CA-CCRF). We do this by including SVR predictions from other dimensions alongside the corresponding SVRs. We include both the original and negated SVR predictions from valence, arousal, expectancy and power dimensions when training the four CA-CCRFs ($K = 32$ for each). This allows us to exploit both positive and negative correlations. In order to account for the fact that the dimensions have different scalings and different offsets, we used the Z-scores of the $\mathbf{X}^{(\mathbf{q})}_{*,\mathbf{k}}$ and $\mathbf{y}^{(\mathbf{q})}$ instead of raw values for training and inference.

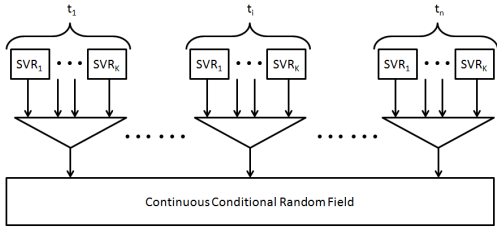

Fig. 3. Final continuous emotion recognition system, that combines support vector regressors with continuous conditional random fields. The number of SVRs used can be varied, and depends on the experiment.

VII. EVALUATION

A. Database

The proposed CCRF framework was evaluated using the dataset distributed through the AVEC 2012 Emotion Challenge [26]. This dataset forms part of the Solid SAL section of SEMAINE database, which contains naturalistic dialogues between two human participants, with one of the participants simulating an artificial listener agent. The dataset was, however, partitioned differently from the challenge. The recordings were split into three partitions: training set I (for SVR training), training set II (for CCRF training) and a test set (for evaluation) with 21, 20 and 18 video sessions in each partition, respectively. The interactions were annotated by at least two raters along the dimensions *arousal, valence, power* and *expectancy*.

B. Methodology

The video features used in the experiments were extracted at a frame rate of 50 frames per second and downsampled by employing the block averaging technique with a block size of

25 frames. The audio features were computed at 100Hz and downsampled for alignment purposes. We used linear kernel L2 loss ϵ-SVRs with L2 regularisation. The training was performed using the Liblinear package [5]. The SVR hyper-parameters were optimized using five-fold cross validation on training set I. Prediction labels were generated from each feature-type SVR model for the remaining two partitions for further CCRF training and inference. The training set II was used to determine the CCRF and CA-CCRF parameters ($\bar{\alpha}$, $\bar{\beta}$) and to cross-validate the regularization hyper-parameters λ_α and λ_β (ranging in $[10^{-2}, 10^0, 10^2, 10^4, 10^6]$). Ten edge features ($g_k$) were used for all experiments; 5 neighbour $n = \{1, 2, \cdots, 5\}$ and 5 distance $\sigma = \{2^{-6}, 2^{-7}, \cdots, 2^{-11}\}$ similarities. The learned $\bar{\beta}$ weights that model the temporal and spatial similarities of the signals, together with the channel reliability measures $\bar{\alpha}$ and SVR predictions were used to predict unseen data (test set). The continuous emotion label predictions were then up-sampled to the original video frame rate through linear interpolation. An example of a CCRF prediction is shown in Figure 4.

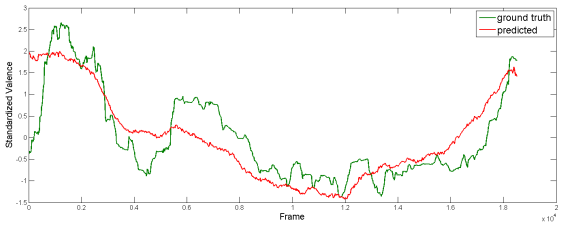

Fig. 4. A plot of a standardized CCRF valence prediction against the ground truth from the test partition

Baseline SVR models were trained using both training sets I and II to ensure that the baseline and the CCRF models have been exposed to the same training data. We trained unimodal SVR models and a multimodal SVR model (through early fusion) for comparison with the CCRF framework.

C. Results

We measure performance using Pearson's correlation coefficient (r) following the AVEC2012 emotion challenge evaluation strategy. The results are obtained by computing the correlation coefficient between the predicted labels and ground truth labels per character interaction and per dimension, and calculating the average over all sessions. In the following sections we present the results of experiments conducted in this work.

1) Feature-type analysis: Figure 5 illustrates the performance of the feature SVR models for each dimensional emotion. The plots suggest that the appearance based features (temporal LBPs) are a better estimator of valence and arousal, and that audio features provide better predictions for power and expectancy. Apart from the arousal dimension, the shape (geometry) features do not perform much worse than the appearance ones, highlighting their potential as comparable estimators of emotion. The generally low correlation results are indicative of the challenging task of working with naturalistic data due to the variety of expressions that can be associated with an affective state.

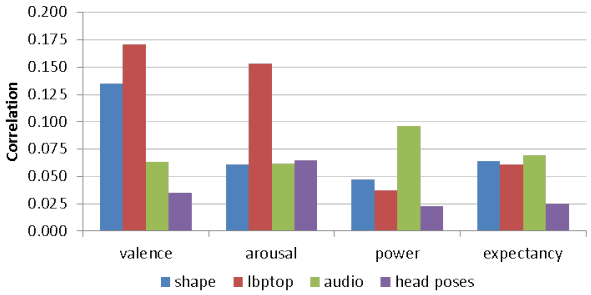

Fig. 5. Comparison of the correlation results for each feature SVR model per dimensional affect

TABLE II
CORRELATION RESULTS OF BASELINE SVR AND CCRF MODELS EVALUATED ON THE TEST PARTITION

	VAL.	AROUS.	POW.	EXPECT.	MEAN
video features					
SVR	0.176	0.234	0.100	0.120	0.158
CCRF	**0.311**	**0.294**	**0.171**	**0.214**	**0.248**
audio features					
SVR	0.062	0.053	0.103	0.104	0.081
CCRF	**0.064**	**0.166**	**0.297**	**0.277**	**0.201**
audio-visual features					
SVR	0.170	0.241	0.132	0.127	0.168
CCRF	**0.326**	**0.341**	**0.273**	**0.248**	**0.297**

2) Model and modality comparisons: Three types of modalities were investigated for both the baseline SVR and CCRF models: audio, video and audio-visual. Table II presents a comparative view of the three modalities for each model type. The results show that the CCRF model significantly outperforms the baseline SVR in all modalities and dimensions. This attests to the importance of temporal data in the analysis and recognition of emotion, and the success of the CCRF model in capturing these dynamics.

Consistent with other studies ([12], [14]), we found that visual features are better predictors of valence and that audio features perform better for the power dimension. However, in contrast to previous findings the arousal state was better predicted by visual rather than audio features.

Furthermore, our CCRF model succeeded in fusing valence and arousal dimension; with overall results of the audio-visual CCRF outperforming the individual CCRFs.

3) Fusion strength of CCRF: Table III contrasts the use of a fused audio-visual SVR ($K = 1$) over the use of several SVR predictors ($K = 4$) as input to the CCRF model. The results show that fusing within the CCRF framework is better than providing fused predictors, therefore highlighting one of the strengths of the CCRF model: the information gain from using signal dynamics for fusion.

4) Correlations between dimensions: With reference to Table IV, it can be seen that our CA-CCRF model outperforms the regular CCRF for some dimensions. The effect of using CA-CCRF is especially beneficial for power dimension. This is not suprising as in the dataset used, power

TABLE III

INVESTIGATING THE FUSION ABILITY OF CCRF WITH FUSED AND NON-FUSED PREDICTOR INPUTS

CCRF INPUTS	VAL.	AROUS.	POW.	EXPECT.	MEAN
1 fused SVR	0.305	0.239	0.110	**0.275**	0.232
4 feature SVRs	**0.326**	**0.341**	**0.273**	0.248	**0.297**

TABLE IV

COMPARISON OF CCRF AND CA-CCRF MODEL PERFORMANCES ON TEST PARTITION

	VAL.	AROUS.	POW.	EXPECT.	MEAN
CCRF	0.326	**0.341**	0.273	**0.248**	0.297
CA-CCRF	**0.343**	0.333	**0.309**	0.218	**0.301**

correlates with other dimensions ($r = 0.25$ with valence, $r = 0.43$ with arousal and $r = -0.46$ with expectancy).

VIII. CONCLUSION

We presented a CCRF model that can be used to model continuous dimensional emotion. The model can easily incorporate multiple simple predictors and exploits temporal correlations between time steps and different modalities. The model can be easily extended to include various other similarity functions that capture the dynamic nature of the signals. It also allows for high-order paths to be defined, exploiting long and short range dependencies of time series. During learning it determines the reliability of the different channels and reflects this knowledge through the learned weights allowing for an insight into what is happening in the system. The resulting model has shown significant improvement over the baseline SVR results. We also demonstrate how to useour model to exploit the correlations between emotional dimensions leading to better prediction for some dimensions. The compact and simple CCRF design allows for applications in other domains with dynamic properties.

IX. ACKNOWLEDGEMENTS

We would like to thank Louis-Philippe Morency and Lech Świrski for useful discussions and suggestions. We acknowledge funding support from Thales Research and Technology (UK), Qualcomm Innovation Fellowship, Bradlow Foundation and National Research Fund (SA).

REFERENCES

[1] B. Amberg, R. Knothe, and T. Vetter. *Expression invariant 3D face recognition with a Morphable Model*, pages 1–6. Ieee, Sep 2008.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking. In *CVPR*, 2012.

[3] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[4] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200, 1992.

[5] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth. The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

[7] H. Gunes, M. A. Nicolaou, and M. Pantic. *Continuous Analysis of Affect from Voice and Face*, pages 255 – 291. Springer London, 2011.

[8] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *Int'l Journal of Synthetic Emotion*, 1(1):68–99, 2010.

[9] L. Jeni, A. Lörincz, T. Nagy, Z. Palotai, J. Sebök, Z. Szabó, and D. Takács. 3D shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing*, 2012.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001.

[11] M. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *ICPR*, 2010.

[12] M. Nicolaou, H. Gunes, and M. Pantic. Output-Associative RVM Regression for Dimensional and Continuous Emotion Prediction. In *IEEE FG'11*, 2011.

[13] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[14] D. Ozkan, S. Scherer, and L.-P. Morency. Step-wise Emotion Recognition Using Concatenated-HMM. In *2nd International Audio/Visual Emotion Challenge and Workshop*, 2012.

[15] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *AVSS*, 2009.

[16] C. Pereira. Dimensions of emotional meaning in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[17] R. W. Picard and J. Klein. *Computers that recognise and respond to user emotion: Theoretical and practical implications*. 2001.

[18] T. Qin, T.-y. Liu, X.-d. Zhang, D.-s. Wang, and H. Li. Global ranking using continuous conditional random fields. In *NIPS*, 2008.

[19] V. Radosavljevic, S. Vucetic, and Z. Obradovic. Continuous conditional random fields for regression in remote sensing. In *European Conference on Artificial Intelligence*, pages 809–814, 2010.

[20] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *ACII*, Berlin, Heidelberg, 2011. Springer-Verlag.

[21] P. Robinson and R. el Kaliouby. Computation of emotions in man and machines. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3441–3447, 2009.

[22] J. Saragih, S. Lucey, and J. Cohn. Deformable Model Fitting by Regularized Landmark Mean-Shift. *Int. J. Comput. Vision*, 91(2):200–215, Jan. 2011.

[23] K. Scherer. Vocal correlates of emotional arousal and affective disturbance. *Handbook of social psychophysiology*, pages 165–197, 1989.

[24] K. Scherer, T. Johnstone, and G. Klasmeyer. Vocal expression of emotion. *Handbook of affective sciences*, pages 433–456, 2003.

[25] M. Schröder. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. *Affective dialogue systems*, pages 209–220, 2004.

[26] B. Schuller, M. Valstar, R. Cowie, and M. Pantic. Avec 2012–the continuous audio/visual emotion challenge. In *Proceedings 2nd International Audio/Visual Emotion Challenge and Workshop*, 2012.

[27] B. Schuller, M. F. Valstar, R. Cowie, and M. Pantic. The first audio/visual emotion challenge and workshop - an introduction. In *ACII*, page 322, 2011.

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

[29] C. Sutton and A. Mccallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006.

[30] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*. ISCA, 2008.

[31] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3D dynamic facial expression database. In *FG*, 2008.

[32] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *TPAMI*, 31(1), 2009.

[33] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE TPAMI*, 29(6):915–928, 2007.