Correlated-Spaces Regression for Learning Continuous Emotion Dimensions

Mihalis A. Nicolaou
Department of Computing
Imperial College London, U.K.
mihalis@imperial.ac.uk

Stefanos Zafeiriou
Department of Computing
Imperial College London, U.K.
s.zafeiriou@imperial.ac.uk

Maja Pantic
Imperial College London, U.K.
U. of Twente, The Netherlands
m.pantic@imperial.ac.uk

ABSTRACT

Adopting continuous dimensional annotations for affective analysis has been gaining rising attention by researchers over the past years. Due to the idiosyncratic nature of this problem, many subproblems have been identified, spanning from the fusion of multiple continuous annotations to exploiting output-correlations amongst emotion dimensions. In this paper, we firstly empirically answer several important questions which have found partial or no answer at all so far in related literature. In more detail, we study the correlation of each emotion dimension (i) with respect to other emotion dimensions, (ii) to basic emotions (e.g., happiness, anger). As a measure for comparison, we use video and audio features. Interestingly enough, we find that (i) each emotion dimension is more correlated with other emotion dimensions rather than with face and audio features, and similarly (ii) that each basic emotion is more correlated with emotion dimensions than with audio and video features. Motivated by these findings, we present a novel regression algorithm (Correlated-Spaces Regression, CSR), inspired by Canonical Correlation Analysis (CCA) which learns output-correlations and performs supervised dimensionality reduction and multimodal fusion by (i) projecting features extracted from all modalities and labels onto a common space where their inter-correlation is maximised and (ii) learning mappings from the projected feature space onto the projected, uncorrelated label space.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition Applications

Keywords

Continuous and dimensional emotion descriptions, valence, arousal, output-correlations, multi-modal fusion, component analysis, feature selection

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502201.

1. INTRODUCTION

In recent years, the field of dimensional continuous emotion analysis has gained rising attention, and a significant number of works has been published on this topic [2, 3, 12, 10]. Introduced by Russel [11], this emotion description originated a radically different approach on describing emotional states. Instead of the traditional approach of discrete emotions (e.g., anger, joy), the emotional state of an individual is described by measurements on a set of latent dimensions. Most of the past-research has focused on the first two dimensions, valence and arousal, signifying respectively how negative/positive and active/inactive the subject's emotional state is. Claims from the field of psychology show that the dimensional descriptions of emotions are much more expressive than basic emotions, and better describe emotions expressed during our everyday lives, e.g., interest, boredom [2].

The contribution of our paper is twofold. Firstly we provide empirical answers to several important questions related to the correlations of emotion dimensions which so-far have found partial or no answer at all. Secondly, we present a regression algorithm which correlates both labels and multi-modal features by projecting them on a common space, eliciting an elegant framework for multi-modal fusion, dimensionality reduction and output-correlations learning. These contributions are discussed in detail in what follows.

Analysing emotion dimension correlations. The occurrence of inter-correlations amongst emotion dimensions such as valence and arousal has been well-supported by various research in psychology [5], and has recently been explored in affective computing in terms of valence and arousal [8]. Nevertheless, to the best of our knowledge, none of the previous work studies (i) correlation between emotion dimensions in isolation, (i.e. without including features), and (ii) the correlations of emotion dimensions to *basic emotions* such as joy and sadness. Furthermore, most works only employ valence and arousal without addressing dimensions such as power and expectation. We address all of these points in our paper. Firstly, by using a set $\mathbf{R}_s$ of 5 dimensions (Valence, Arousal, Power, Expectation and Intensity) [6], in our first experiment (Sec. 3.1), we essentially pose the problem of predicting dimension k given the rest. We also perform experiments using face/audio features for comparison. Interestingly enough, we show that the correlation of the $k-1$ other dimensions to dimension k is *higher* than the correlation of audio/face features to dimension k.

In our second experiment (Sec. 3.2), we attempt to answer an interesting question which has not been explored so far: *how correlated are emotion dimensions to basic emo-*

tions? Intuitively, the correlation should be high, since in theory there is a (rather abstract and relatively ambiguous) mapping from these dimensions to basic emotions (e.g., high valence, positive arousal can point to joy, excitement etc.). To verify this intuition empirically, we use a set of basic emotions $\mathbf{L}_s$ (e.g., anger, happiness). Using the set of dimensions $\mathbf{R}_s$, we evaluate how correlated the emotion dimensions are to basic emotions, in comparison to facial points and audio cues. Our findings are in line with the previous experiment: Emotion dimensions are positively correlated with the intensity of basic emotions, exhibiting higher correlations than face/audio features.

Exploiting emotion dimension correlations. An important contribution of our paper lies in the introduction of the Correlated-Spaces Regression (CSR), a principled, novel framework based on canonical correlation analysis, which elegantly combines multi-modal fusion, the learning of output-correlations and supervised dimensionality reduction. Our algorithm, heavily motivated by conclusions drawn from our empirical study, is shown to increase the accuracy of both single-cue and fused experiments and up to a point, "heal" the relatively weak correlation of face/audio features to the emotion dimensions[1].

2. DATA & FEATURE EXTRACTION

For evaluation, we employ the SEMAINE database [6], which contains a set of audio-visual recordings of subjects interacting with operators. Each operator assumes a certain personality, i.e. happy, gloomy, angry and pragmatic, with a goal of inducing spontaneous emotions during a naturalistic conversation. We use a portion of the database running approximately 85 minutes, which has been annotated for the emotion dimensions at hand by 5 raters, from which we use the averaged annotation[2]. For extracting facial expression features, we employ an Active Appearance Model (AAM) based tracker [9], designed for simultaneous tracking of 3D head pose, lips, eyebrows, eyelids and irises in video sequences. For each video frame, we obtain 113 2D-points, resulting in an 226 dimensional feature vector. To compensate for translation variations, we center the coordinate system to the fixed point of the face (average of inner eyes and nose), while for scaling we normalise by dividing with the inter-ocular distance. Regarding audio features, we utilise MFCC and MFCC-Delta coefficients along with prosody features (energy, RMS Energy and pitch). We used 13 cepstrum coefficients for each audio frame, essentially employing the typical set of features used for automatic affect recognition [14]. We obtain a feature vector with dimensionality $d = 29$, obtaining a frame-rate equivalent to 100-fps. To match the video fps, the audio features used are vertically concatenated for each pair of consecutive frames, thus obtaining 58 dimensional feature vectors. For feature-level fusion, the vectors are concatenated, resulting to 284 dimensions.

3. ANALYSIS OF EMOTION DIMENSIONS

In this section we present several experiments evaluating the correlations of emotion dimensions. For regression, we employ the Relevance Vector Machine (RVM [13]), which given the input-output pair $(\mathbf{x}_i, \mathbf{y}_i)$ models the function $\mathbf{y}_i = \mathbf{w}^T\phi(\mathbf{x_i}) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\phi(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{l}\right\}$ being the RBF kernel. Using the extracted features and annotations (Sec. 2) we perform cross-validation. For evaluation, we use the mean-squared error (MSE) to measure bias error and the correlation coefficient (COR) to measure the correlation deviation. We mostly refer to COR, since (i) it is most commonly used in related work [12], and (ii) the MSE bias errors are relatively very small.

[1]Regarding dimensionality reduction for regression, c.f. [4].
[2]For the basic emotion experiments, we use only the subset of this data which was annotated in terms of basic emotions.

3.1 Inter-Correlations and Multimedia

In this section we pose the problem of predicting an emotion dimension given a set of annotated dimensions. Let us assume we have a set of ρ annotations $\mathbf{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_\rho\}$ with $\mathbf{r}_i \in \mathbb{R}^{1\times T}$. In this experiment, we assume that $\mathbf{R}$ consists of dimensions valence, arousal, power, expectation and intensity, i.e. $\rho = 5$. Our problem can then be defined as

$$f : \mathbf{R}_{\setminus k} \rightarrow \hat{\mathbf{r}}_k, \forall k \in \{1, \ldots, \rho\} \quad (1)$$

where $\mathbf{R}_{\setminus k}$ denotes the entire set of annotations excluding dimension k and $\hat{\mathbf{r}}_k$ the estimated values of dimension k. The performance of the learnt functions is then compared against the performance obtained when using facial expressions and audio cues as features, in order to obtain a comparative measure of performance. By this experiment, we essentially ask the following question: *Which signal is most correlated with a specific emotion dimension k, the features extracted from audio/video cues or the annotations for the rest of the dimensions,* $\mathbf{R}_{\setminus k}$*?* Results are presented in Tab. 1 and Fig. 1. It is very interesting to observe that by using all the emotion dimensions except the one being tested provides better results for *all* dimensions at hand. This important observation empirically confirms that each and every emotion dimension has higher correlation with the rest of the dimensions than with the audio/face features. It is also interesting to observe that for the arousal and the intensity dimensions, the audio cues appear to perform better than the facial features in terms of correlation coefficient, a conclusion that confirms previous findings [7].

3.2 Correlations to Basic Emotions

Another question we address in this work refers to the correlations amongst the dimensional emotion descriptions, as perceived by Russel [11] and a set of emotions which are of discrete nature (e.g., basic emotions). Although emotion dimensions can be inherently more expressive in comparison to discrete emotions such as joy and sadness, no explicit mapping between the two descriptions has been established. One would of course assume that e.g., negative valence with negative arousal maps to sadness or boredom, nevertheless this is more of an abstract and relatively ambiguous correspondence. In this section we evaluate the correlations of emotion dimensions when learning to predict emotions such as anger, happiness, sadness, surprise etc. In more detail, given the set $\mathbf{R}$, as defined in Section 3.1 (consisting of dimensions valence, arousal, power, expectation and intensity) we aim to predict a specific emotion belonging in the set $\mathbf{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_\nu\}$, i.e.

$$f : \mathbf{R} \rightarrow \hat{\mathbf{L}}_k, \forall k \in \{1, \ldots, \nu\} \quad (2)$$

Results are presented in Tab, 2 and Fig. 1, where we also use face/audio features for comparison. The first conclusion is that the emotion dimensions (namely valence, arousal, power, expectation and intensity) are highly correlated with

Table 1: Results for predicting each emotion dimension, using the other four dimensions as features ($\mathbf{R}_s$), compared to using facial features (F), audio features (A) and the feature-level fusion of face and audio (F+A).

	Valence		Arousal		Power		Expectation		Intensity	
	MSE	COR	MSE	COR	MSE	COR	MSE	COR	MSE	COR
$\mathbf{R}_{s\setminus k}$	0.074	**0.28**	0.051	**0.47**	0.088	**0.28**	0.037	**0.15**	0.067	**0.30**
Face	0.088	0.14	0.061	0.41	0.131	0.06	0.024	0.02	0.066	0.17
Audio	0.072	0.14	0.050	0.44	0.082	0.05	0.018	0.01	0.042	0.26
F+A	0.880	0.16	0.055	0.44	0.080	0.06	0.020	0.02	0.058	0.20

Table 2: Predicting each basic emotion using the five emotion dimensions as features ($\mathbf{R}_{s\setminus k}$), compared to using facial features (F) and audio features (A).

COR	Anger	Happiness	Sadness	Contempt	Amusement
$\mathbf{R}_s$	**0.74**	**0.48**	**0.67**	**0.33**	**0.49**
F	0.06	0.11	0.13	0.05	0.06
A	0.02	0.10	0.10	0.11	0.02
MSE	**Anger**	**Happiness**	**Sadness**	**Contempt**	**Amusement**
$\mathbf{R}_s$	**0.07**	**0.10**	**0.06**	**0.02**	**0.07**
F	0.21	0.21	0.26	0.34	0.15
A	0.17	0.17	0.10	0.21	0.09

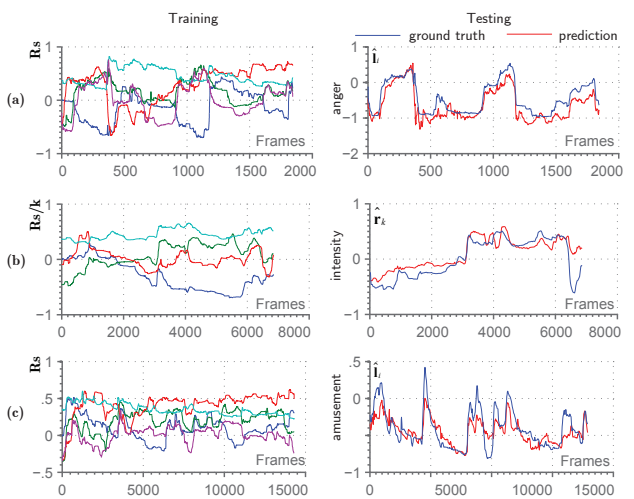

Figure 1: (a,c) Using emotion dimensions ($\mathbf{R}_s$) for predicting basic emotions, (b) using $k-1$ emotion dimensions ($\mathbf{R}_{s\setminus k}$) for predicting dimension k.

the discrete emotions we study. Similarly to the results regarding the previous experiment, the dimension to discrete-emotion correlation is quite higher compared to face or audio features. The most correlated discrete emotion to emotion dimensions appears to be anger.

4. CORRELATED-SPACES REGRESSION

Inspired by the results described in previous sections, we demonstrate a method which exploits output-correlations, while performing multi-modal fusion and dimensionality reduction. Note that the latter experiments also motivate the idea of dimensionality reduction on this problem: In the experiments in Sec. 3.1, $\mathbf{R}_{\setminus k}$ consists of 4-dimensional feature vectors and attains better performance than, i.e. the 226-dimensional facial expression vectors. We show how by exploiting feature-label, inter-feature and inter-label correlations we can significantly improve the results.

Let us assume that for a training sequence s, we have a set of annotations for emotion dimensions $\mathbf{R}_s$, containing the five dimensions used in Sec. 3.1, along with a given set of features, $\mathbf{F}_{j,s}, j=\{1,\dots,\mu\}$ which can contain e.g., video or/and audio cues. Canonical Correlation Analysis (CCA) enables the discovery of projections of the features onto a space where they are maximally correlated. We reformulate the problem to match our context as follows

$$\begin{aligned} &\underset{\mathbf{V}_{F_s},\mathbf{V}_R}{\arg\min} ||\mathbf{F}_s\mathbf{V}_{F_s} - \mathbf{R}_s\mathbf{V}_R||_2^F \\ &s.t.\ \mathbf{F}_s\mathbf{V}_{F_s}\mathbf{V}_{F_s}^T\mathbf{F}_s^T = \mathbf{R}_s\mathbf{V}_R\mathbf{V}_R^T\mathbf{R}_s^T = \mathbf{I} \\ &\mathbf{F}_s = [\mathbf{F}_{1,s},\dots,\mathbf{F}_{j,s}],\ \mathbf{V}_{F_s}^T = [\mathbf{V}_{F_{1,s}}^T,\dots,\mathbf{V}_{F_{\mu,s}}^T]^T, \end{aligned} \quad (3)$$

where $\mathbf{I}$ is the identity matrix. Therefore, by applying CCA on *both* the labels and the features, we are in a sense employing supervision on the feature projections, i.e. performing supervised component analysis. This is due to the fact that the labels and features are projected into a common space where they maximally correlate. In fact, for problems where labels are discrete classes, it has been shown that applying CCA on both features and binary labels collapses to applying Linear Discriminant Analysis [1], where $\mathbf{F}_s\mathbf{V}_F$ are the discriminant projections. Furthermore, as an implication of the orthogonality constraints of the problem statement in Eq. 3, the projected label space will be uncorrelated, thus enabling regessors to learn output-correlations which exist in the label space. Finally, due to the block-matrix formulation we learn correlated features from *all* feature sets, i.e. we perform multi-modal supervised fusion. Our model is described in Alg. 1, and visually depicted in Fig. 2. During training, the projection vectors for the continuous label space $\mathbf{V}_R$ and the feature sets employed $\mathbf{F}_{1:\mu}$ are obtained.

Algorithm 1: Correlated-Spaces Regression

Data: Train=$(\mathbf{R}_s, \mathbf{F}_{1,s},\dots,\mathbf{F}_{\mu,s})$
Test=$(\mathbf{F}_{1,t},\dots,\mathbf{F}_{\mu,t})$
Result: $\hat{\mathbf{R}}_t$
train
1 Set $[\mathbf{V}_R, \overbrace{\mathbf{V}_{F_1},\dots,\mathbf{V}_{F_\mu}}^{\mathbf{V}_{F_s}}]$ to the leading eigenvectors of
$$\begin{bmatrix} \mathbf{0} & \mathbf{F}_s\mathbf{R}_s^T \\ \mathbf{R}_s\mathbf{F}_s^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{F_s} \\ \mathbf{V}_R \end{bmatrix} = \begin{bmatrix} \mathbf{F}_s\mathbf{F}_s^T & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_s\mathbf{R}_s^T \end{bmatrix} \begin{bmatrix} \mathbf{V}_{F_s} \\ \mathbf{V}_R \end{bmatrix} \Lambda$$
(Problem defined in Eq. 3)
2 $\mathbf{F}_{i,s}^c = \mathbf{F}_{i,s}\mathbf{V}_{F_i}, \forall i \in \{1,\dots,\mu\}$
3 $f : \mathbf{F}_{1:\mu,s}^c \rightarrow \mathbf{R}_s\mathbf{V}_R$
test
4 $\mathbf{F}_{i,t}^c = \mathbf{F}_{i,t}\mathbf{V}_{F_i}, \forall i \in \{1,\dots,\mu\}$
5 $\hat{\mathbf{R}}_t^c \leftarrow f(\mathbf{F}_{1:\mu,t}^c)$
6 $\hat{\mathbf{R}}_t = \hat{\mathbf{R}}_t^c\mathbf{V}_R^{-1}$

Table 3: Results for predicting each emotion dimension using Correlated-Spaces Regression (CSR) utilising facial features ($\mathbf{F}^{CSR}$), audio features ($\mathbf{A}^{CSR}$) and the fusion of face and audio ($\{\mathbf{F}+\mathbf{A}\}^{CSR}$) using CSR.

	Valence		Arousal		Power		Expectation		Intensity	
	MSE	COR	MSE	COR	MSE	COR	MSE	COR	MSE	COR
$\mathbf{F}^{CSR}$	0.070	0.20	**0.046**	**0.46**	0.080	0.11	**0.020**	0.06	**0.044**	**0.29**
$\mathbf{A}^{CSR}$	0.070	0.15	0.510	0.45	0.075	**0.11**	0.022	0.02	0.040	**0.29**
$\{\mathbf{F},\mathbf{A}\}^{CSR}$	**0.056**	**0.21**	0.050	**0.46**	**0.063**	**0.12**	**0.020**	**0.07**	**0.044**	**0.29**

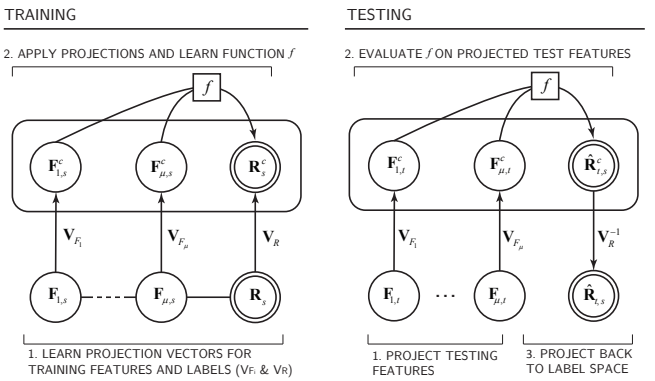

Figure 2: Correlated-Spaces Regression model, following Algorithm 1.

Using these projection matrices, the training features $\mathbf{F}_{1:\mu,s}$ and labels $\mathbf{R}_s$ are projected onto the space where they maximally correlate, obtaining the matrices $\mathbf{F}^c_{1:\mu,s}$ and $\mathbf{R}^c_s$. The regressor is subsequently optimised on this space

$$f : \mathbf{F}^c_{1:\mu,s} \rightarrow \mathbf{R}^c_s \tag{4}$$

For testing, we obtain a set of features $\mathbf{F}_{1:\mu,t}$, which we project as $\mathbf{F}^c_{i,t} = \mathbf{F}_{i,t}\mathbf{V}_{F_i}$. The learnt function f is evaluated on $\mathbf{F}^c_{i,t}$, obtaining the predictions $\hat{\mathbf{R}}^c_t$, which are then projected back to the annotation space. Results with our method are presented in Tab. 3. As can be clearly seen, our method performs much better than using simply the raw features or performing feature-level fusion, as seen in Tab. 1. In fact, it is interesting to observe that in some dimensions, our method achieves comparable correlation to using all the other annotations/labels as features ($\mathbf{R}_s$, Sec. 3.1). Essentially this means that the model manages to capture output-correlations and in addition propagate this information during dimensionality reduction onto the projected features.

5. CONCLUSIONS

In this work, we performed a thorough investigation on the inter-correlation of emotion dimensions and their correlation to basic emotions. We have shown that there are more dominant correlations within emotion dimensions rather than to face or audio features. Most importantly, we presented CSR, a CCA-based algorithm which learns output-correlations while performing multi-modal fusion and supervised dimensionality reduction. Our algorithm increases the accuracy both in terms of multi-modal fusion and single-cue regression, successfully learning output structure and maximising input-output correlations. Our algorithm can be straight-forwardly applied to any learning problem with a set of feature modalities and multi-dimensional output vectors.

6. ACKNOWLEDGEMENTS

This work has been funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG) and by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

7. REFERENCES

[1] F. R. Bach and M. I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis.

[2] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120 – 136, 2013.

[3] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. of IEEE FG'11-W*, Santa Barbara, CA, USA, March 2011.

[4] M. Kim and V. Pavlovic. Central subspace dimensionality reduction using covariance operators. *IEEE TPAMI*, 33(4):657–670, 2011.

[5] R. Lane et al. *Cognitive Neuroscience of Emotion*. Oxford University Press, 2000.

[6] G. McKeown et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE TAC*, 2012.

[7] M. A. Nicolaou et al. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE TAC*, 2011.

[8] M. A. Nicolaou et al. Output-associative rvm regression for dimensional and continuous emotion prediction. In *Proceedings of IEEE FG'11*, pages 16–23, Santa Barbara, CA, USA, March 2011.

[9] J. Orozco et al. Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, February 2013.

[10] G. A. Ramirez et al. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Proc. of ACII'11*, 2011.

[11] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1980.

[12] B. Schuller, M. Valstar, et al. Avec 2012: the continuous audio/visual emotion challenge - an introduction. In *ICMI*, pages 361–362, 2012.

[13] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *JMLR*, 1:211–244, 2001.

[14] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE TPAMI*, 2009.