ROBUST CANONICAL CORRELATION ANALYSIS: AUDIO-VISUAL FUSION FOR LEARNING CONTINUOUS INTEREST

Mihalis A. Nicolaou[1], *Yannis Panagakis*[1], *Stefanos Zafeiriou*[1] *and Maja Pantic*[1,2]

[1]Deparment of Computing, Imperial College London, UK
[2]EEMCS, University of Twente, NL

{mihalis, i.panagakis, s.zafeiriou, m.pantic}@imperial.ac.uk

ABSTRACT

The problem of automatically estimating the *interest* level of a subject has been gaining attention by researchers, mostly due to the vast applicability of interest detection. In this work, we obtain a set of continuous interest annotations for the SEMAINE database, which we analyse also in terms of emotion dimensions such as valence and arousal. Most importantly, we propose a robust variant of Canonical Correlation Analysis (RCCA) for performing audio-visual fusion, which we apply to the prediction of interest. RCCA recovers a low-rank subspace which captures the correlations of fused modalities, while isolating gross errors in the data without making any assumptions regarding Gaussianity. We experimentally show that RCCA is more appropriate than other standard fusion techniques (such as $l2$-CCA and feature-level fusion), since it both captures interactions between modalities while also decontaminating the obtained subspace from errors which are dominant in real-world problems.

Index Terms— Emotion Recognition, Interest Detection, Audio-visual Fusion, Multi-modal Fusion

1. INTRODUCTION

The automatic detection of interest in audiovisual sequences has been gaining rising attention amongst researchers, in both the fields of affective computing and pattern recognition and machine learning [1, 2, 3]. From a psychology perspective, interest has been extensively studied since 1910 [4], and has since then been considered as an *emotion* by various experts [5, 6]. Interest is commonly defined as *an emotion that causes the subject to focus his or hers attention to the event taking place* [6]. As can be understood, the detection of interest is crucial for a vast number of applications, ranging from virtual guides to interactive learning systems as well as enhancing the experience of human-computer interaction.

Although there has been previous work on the automatic detection of interest [3, 7, 2], most of previous work treats interest as a discrete emotion, focusing on classification in terms of discrimination between interest/non-interest, as well as discriminating amongst classes e.g., disinterest, indifference and interest. This is in line with traditional research in affective computing and emotion theory, which focuses only on a set of discrete emotions, such as anger and joy. In contrast, our paper follows the recent research path of employing a set of latent dimensions in order to describe the affective state of an individual [8, 9, 10, 11, 12, 13]. Based on Russell's seminal work [14], the dimensional, continuous representation of the emotional state of the subject is deemed much more expressive than confining to basic emotions and is well suited to emotional states that are commonly observed in routine, daily interactions of humans, with such emotional states falling well outside the spectrum of basic emotions [15, 16, 12]. In this paper, we attempt to treat *interest* similarly to an affective dimension, that is, to attain continuous (in both time and space) measurements of interest which describe the emotional state of the subject on a continuous scale. We firstly analyse the interest annotations obtained and attempt to evaluate the agreement between the interest annotations at hand and annotations already available in the SEMAINE database (namely, valence, arousal, power, intensity and expectation). Subsequently, we propose a novel, robust variant of Canonical Correlation Analysis (CCA), which is highly suitable for the fusion of multiple modalities under real-world scenarios, where gross noise can have a prominent presence. The contributions of our paper are summarised in what follows.

Continuous Interest & Emotion Dimensions. Evidence from the field of psychology points to various correlations between emotion dimensions and interest [17]. Nevertheless, this has remained unexplored in the field of affective computing and machine learning. In this paper (Sec. 4.1), we provide, to the best of our knowledge, the first empirical experimental evidence on continuous annotations which show that interest is highly correlated with specific emotion dimensions such as arousal, valence and intensity. Furthermore, our analysis reveals that although we use a disjoint set of annotators for interest, correlations between interest and other emotion dimensions are still high, thus motivating the utilisation of models exploiting output-correlations for detecting interest (c.f.,[18, 19, 9])

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

RCCA for Audio-Visual Fusion. Although Canonical Correlation Analysis (CCA) has been often used for the fusion of multiple modalities in affective computing and pattern recognition in general [20, 21], the application of CCA is limited in real-world conditions where gross errors are observed in the measurements. We propose the *Robust Canonical Correlation Analysis* (RCCA, Sec. 3) for audio-visual fusion, which is able to isolate sparse errors in each modality, and learn an error-free low-rank subspace. With this robust variant of CCA, we can isolate non-Gaussian noise, thus obtaining a clean subspace, which as we experimentally show, can provide better results compared to standard fusion approaches such as $l2$ CCA and feature level (Sec. 4).

2. ANNOTATIONS, DATA & SETTING

SEMAINE Database. For this work, we employ the SEMAINE database [22], which contains a set of audio-visual recordings focusing on dyadic interaction scenarios. In more detail, each subject is conversing with an operator, who assumes the role of an avatar. Each operator assumes a specific personality, which is defined by the avatar he undertakes: happy, gloomy, angry or pragmatic. This is in order to elicit spontaneous emotional reactions by the subject that is conversing with the operator. SEMAINE has been annotated in terms of emotion dimensions, particularly in terms of valence, arousal, power, expectation and intensity. The interaction scenario employed in SEMAINE is though highly appropriate for analysing interest: since the behaviour of operators elicits naturalistic conversation, the subject can be interested in the conversation regarding some personal issue that the subject might be facing, or can become either annoyed or bored (i.e. disinterested) and e.g., request the conversation to finish or switch to another operator with different behaviour. We use a portion of the database running approximately 85 minutes, which has been annotated for emotion dimensions. We utilise 5 annotators, from which we use the averaged annotation[1]. Furthermore, following the procedure in the next section, we obtained interest annotations from 8 annotators.

Obtaining Interest Annotations. In this section, we detail the process which we followed in order to obtain continuous interest annotations. Firstly, the instructions given to the annotators were based on earlier work [2], and have been readjusted in order to fit to a continuous scale and enriched in order to correspond to the conversational setting of the SEMAINE database. They are as follows:

- *Interest Rating in* $[-1, -0.5)$: the subject is *disinterested* in the conversation, can be mostly passive or appear bored, does not follow the conversation and possibly wants to stop the session.
- *Interest Rating in* $[-0.5, 0)$: the subject appears passive, replies to the interaction partner, possibly with hesitation, just because he/she has to reply (unmotivated). The subject appears *indifferent*.
- *Interest Rating approx.* 0: the subject seems to follow the conversation with the interaction partner, but it can not be recognized if he/she is interested. The subject is *neutral*.
- *Interest Rating in* $(0, 0.5]$: The subject seems eager to discuss with the interaction partner, and interested in getting involved in the conversation. The subject is *interested*.
- *Interest Rating in* $(0.5, 1]$: The subject seems pleased to participate in the conversation, can show some signs of *enthusiasm*, is expressive in terms of (positive) emotions (e.g., laughing at a joke, curious to discuss a topic).

Feature Extraction & Experimental Setting. For extracting facial expression features, we employ an Active Appearance Model (AAM) based tracker [25], designed for simultaneous tracking of 3D head pose, lips, eyebrows, eyelids and irises in videos. For each 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 [26], obtaining a feature vector of dimensionality $d = 29$. Cross-validation is performed given the features and annotations. Regression was performed via a Relevance Vector Machine (RVM) [27]. Given the input-output pair $(\mathbf{x}_i, \mathbf{y}_i)$, RVM models the function $\mathbf{y}_i = \mathbf{w}^T \phi(\mathbf{x_i}) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. For the design matrix, we use an RBF Kernel, $\phi(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{l}\right\}$. Results are evaluated based on the Mean Squared Error (MSE) and the Correlation Coefficient (COR).

3. METHODOLOGY: ROBUST CCA

Canonical Correlation Analysis (CCA) is typically used for fusing multiple modalities and views [20, 21]. The classical formulation of CCA, based on $l2$ regularisation, carries the assumption that the errors follow a Gaussian distribution with a small variance. Nevertheless, in problems dealing with real-world conditions where gross errors can be observed, the application of CCA is limited. In this paper, we propose using a robust (to gross errors) variant of CCA for audio-visual fusion. In more detail, let us say we have two modalities, with high-dimensional feature spaces $\mathbf{Z} \in \mathbb{R}^{dz \times T}$

[1] We note that more sophisticated methods for fusing annotations wrt. behaviour have been recently proposed, such as [23, 24].

and $\mathbf{A} \in \mathbb{R}^{da \times T}$[2], which can represent e.g., facial trackings and audio cues, corrupted by noise as is often the case in real-world scenarios. RCCA can be formulated as

$$\begin{aligned} &\underset{\mathbf{P}_z,\mathbf{P}_a,\mathbf{E}_z,\mathbf{E}_a}{\operatorname{argmin}} \quad \operatorname{rank}(\mathbf{P}_z) + \operatorname{rank}(\mathbf{P}_a) \\ &+\lambda_1\|\mathbf{E}_z\|_0 + \lambda_2\|\mathbf{E}_a\|_0 + \frac{\mu}{2}\|\mathbf{P}_z\mathbf{Z} - \mathbf{P}_a\mathbf{A}\|_F^2 \\ &\text{s.t. } \mathbf{Z} = \mathbf{P}_z\mathbf{Z} + \mathbf{E}_z, \mathbf{A} = \mathbf{P}_a\mathbf{A} + \mathbf{E}_a. \end{aligned} \tag{1}$$

where as can be seen, RCCA uncovers a low-rank subspace $\mathbf{P}_z$, $\mathbf{P}_a$, by estimating the gross errors for each modality, $\mathbf{E}_z$ and $\mathbf{E}_a$. λ_1, λ_2(which can be found via cross-validation.) and μ are non-negative parameters. Problem (1) is deemed difficult to solve due to the discrete nature of the rank function [28] and the ℓ_0 norm [29]. Nevertheless, it has been proved that the convex envelope of the ℓ_0 norm is the ℓ_1 norm [30], while the convex envelope of the rank function is the nuclear norm [31]. Therefore, convex relaxations of (1) can be obtained by replacing the ℓ_0 norm and the rank function with their convex envelopes. The resulting problem

$$\begin{aligned} &\underset{\mathbf{P}_z,\mathbf{P}_a,\mathbf{E}_z,\mathbf{E}_a}{\operatorname{argmin}} \quad \|\mathbf{P}_z\|_* + \|\mathbf{P}_a\|_* \\ &+\lambda_1\|\mathbf{E}_z\|_1 + \lambda_2\|\mathbf{E}_a\|_1 + \frac{\mu}{2}\|\mathbf{P}_z\mathbf{Z} - \mathbf{P}_a\mathbf{A}\|_F^2 \\ &\text{s.t. } \mathbf{Z} = \mathbf{P}_z\mathbf{Z} + \mathbf{E}_z, \mathbf{A} = \mathbf{P}_a\mathbf{A} + \mathbf{E}_a. \end{aligned} \tag{2}$$

can be solved by employing the Linearized Alternating Directions Method (LADM) [32], a variant of the *alternating direction augmented lagrange multiplier method* [33]. The algorithm is detailed in Alg. 1. We note that the singular value thresholding operator can be defined for any matrix $\mathbf{M}$ [34], as: $\mathcal{D}_\tau[\mathbf{M}] = \mathbf{U}\mathcal{S}_\tau\mathbf{V}^T$ where $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition (SVD) and $\mathcal{S}_\tau[q] = \operatorname{sign}(q)\max(|q| - \tau, 0)$ the shrinkage operator [35] (extended to matrices via element-wise application).

4. EXPERIMENTAL VALIDATION

4.1. Interest and Emotion Dimensions

In this section, we attempt to empirically evaluate the correlation of interest with other emotion dimensions. The question is of high interest for many algorithms which aim to model output-structure[18, 19]. Although this has been partly demonstrated for various emotion dimensions [19], in this case we examine the problem from a different perspective. The interest annotations differ from the annotations provided with SEMAINE by (i) the set of annotators are *disjoint* from the annotators for SEMAINE, and (ii) the annotation tool employed for interest is joystick-based, (with a neutral position of 0, i.e. when no force is applied on the joystick), while for SEMAINE, a mouse-based tool was used (FeelTrace [22]).

[2]in case of $dz \neq da$, one can reduce the signals with maximum dimensionality to $min(dz, da)$ by applying e.g., PCA or k-SVD

Algorithm 1 Solving (2) via LADM.

Input: Modality Features: $\mathbf{Z} \in \mathbb{R}^{d \times T}$ and $\mathbf{A} =\in \mathbb{R}^{d \times T}$, parameters: λ_1, λ_2.
Output: Projection/error matrices: $\mathbf{P}_z, \mathbf{P}_a, \mathbf{E}_z, \mathbf{E}_a$.

Initialize: $\mathbf{P}_{z[0]}, \mathbf{P}_{a[0]}, \mathbf{E}_{z[0]}, \mathbf{E}_{a[0]}$ are set to zero matrices of compatible dimensions, $\mu_{[0]} = \mu_{z[0]} = \mu_{a[0]} = 10^{-6}$, $t = 0$, $\rho = 1.9$, $\eta_z = 1.02\sigma_z^2$, $\eta_a = 1.02\sigma_a^2$.
while not converged **do**
  Fix other variables, update $\mathbf{P}_{z[t+1]}$ by:
  $\nabla_{\mathbf{P}_z}\mathcal{L} = \mu_z(\mathbf{P}_{z[t]}\mathbf{Z}\mathbf{Z}^T + \mathbf{E}_{z[t]}\mathbf{Z}^T - \mathbf{Z}\mathbf{Z}^T) + \mu(\mathbf{P}_{z[t]}\mathbf{Z}\mathbf{Z}^T - \mathbf{P}_{a[t]}\mathbf{A}\mathbf{Z}^T) - \mathbf{\Lambda}_{1[t]}\mathbf{Z}^T$.
  $\mathbf{P}_{z[t+1]} \leftarrow \mathcal{D}_{\frac{1}{\mu_{z[t]}}}[\mathbf{P}_{z[t]} - 1/(\mu_{z[t]} \cdot \eta_z)\nabla_{\mathbf{P}_z}\mathcal{L}]$.
  Fix other variables, update $\mathbf{E}_{z[t+1]}$ by:
  $\mathbf{E}_{z[t+1]} = \mathcal{S}_{\frac{\lambda_1}{\mu_{z[t]}}}[\mathbf{Z} - \mathbf{P}_{z[t+1]}\mathbf{Z} + \frac{1}{\mu_{z[t]}}\mathbf{\Lambda}_{1[t]}]$.
  Fix other variables, update $\mathbf{P}_{a[t+1]}$ by:
  $\nabla_{\mathbf{P}_a}\mathcal{L} = \mu_z(\mathbf{P}_{a[t]}\mathbf{A}\mathbf{A}^T + \mathbf{E}_{a[t]}\mathbf{A}^T - \mathbf{A}\mathbf{A}^T) + \mu(\mathbf{P}_{a[t]}\mathbf{A}\mathbf{A}^T - \mathbf{P}_{z[t]}\mathbf{Z}\mathbf{A}^T) - \mathbf{\Lambda}_{2[t]}\mathbf{A}^T$.
  $\mathbf{P}_{a[t+1]} \leftarrow \mathcal{D}_{\frac{1}{\mu_{a[t]}}}[\mathbf{P}_{a[t]} - 1/(\mu_{a[t]} \cdot \eta_a)\nabla_{\mathbf{P}_a}\mathcal{L}]$.
  Fix other variables, update $\mathbf{E}_{a[t+1]}$ by:
  $\mathbf{E}_{a[t+1]} = \mathcal{S}_{\frac{\lambda_2}{\mu_{a[t]}}}[\mathbf{A} - \mathbf{P}_{a[t+1]}\mathbf{A} + \frac{1}{\mu_{a[t]}}\mathbf{\Lambda}_{2[t]}]$.
  Update the Lagrange multipliers by:
  $\mathbf{\Lambda}_{1[t+1]} \leftarrow \mathbf{\Lambda}_{1[t]} + \mu_{z[t]}(\mathbf{Z} - \mathbf{P}_{z[t+1]}\mathbf{Z} - \mathbf{E}_{z[t+1]})$.
  $\mathbf{\Lambda}_{2[t+1]} \leftarrow \mathbf{\Lambda}_{2[t]} + \mu_{a[t]}(\mathbf{A} - \mathbf{P}_{a[t+1]}\mathbf{A} - \mathbf{E}_{a[t+1]})$.
  Update $\mu_{z[t+1]}$ by:
  if $\mu_{z[t]}\|\mathbf{P}_{z[t+1]} - \mathbf{P}_{z[t]}\|_F \leq \epsilon_2$ **then**
    $\mu_{z[t+1]} \leftarrow \min(\rho \cdot \mu_{z[t]}, 10^6)$.
  end if
  if $\mu_{a[t]}\|\mathbf{P}_{a[t+1]} - \mathbf{P}_{a[t]}\|_F \leq \epsilon_2$ **then**
    $\mu_{a[t+1]} \leftarrow \min(\rho \cdot \mu_{a[t]}, 10^6)$.
  end if
  Update $\mu_{[t+1]}$ by: $\mu_{[t+1]} \leftarrow \min(\mu_{z[t+1]}, \mu_{a[t+1]})$
  Check convergence conditions.
  $t \leftarrow t + 1$.
end while

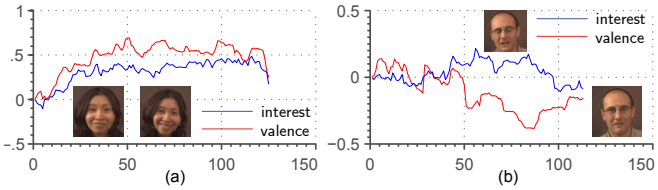

Fig. 1. Examples from SEMAINE where (a) interest is positively correlated with valence, since the subject is in a joyful mood, (b) interest is negatively correlated with valence since the subject is angry/sad but interested in the conversation.

Firstly, we study the correlations of other emotion dimensions included in SEMAINE to the obtained interest annotations. By analysing the entire annotation set based on the correlation coefficient, we find that interest seems to be highly correlated firstly with arousal (.74), and secondly with valence (.49) and intensity (.48). We note that these findings are in accordance to previous work on evaluating the dependen-

cies between interest, valence and arousal [17]. Plots comparing valence and interest annotations can be seen in Fig. 1.

Secondly, we perform experiments to evaluate the correlations between emotion dimensions and interest based on prediction accuracy. In what follows, we denote $\boldsymbol{\mathcal{S}}$ as the set of emotion dimensions (valence, arousal, power, intensity and expectation), and $\boldsymbol{\mathcal{I}}$ as the interest annotation. For each emotion dimension k in $\boldsymbol{\mathcal{S}}$, we learn the mapping $f : \boldsymbol{\mathcal{S}}_{\setminus k} \rightarrow k$, where $\boldsymbol{\mathcal{S}}_{\setminus k}$ is the set of all emotion dimensions in $\boldsymbol{\mathcal{S}}$ except k. We repeat the experiment with $\boldsymbol{\mathcal{SI}} = \boldsymbol{\mathcal{S}} \cup \boldsymbol{\mathcal{I}}$ in place of $\boldsymbol{\mathcal{S}}$, i.e. we also use interest along with emotion dimensions. Results are presented in Tab. 1. As can be seen, the correlation (COR) for most emotion dimensions increases when also using interest as a feature. As expected, the most significant increase occurs for arousal. Interestingly, this experimentally validates that although the annotations have been obtained via different tools and a disjoint set of annotators, still the obtained signals exhibit linear and non-linear correlations. In Sec. 4.2, we also examine the prediction of interest and evaluate how well *interest* is predicted by using emotion dimensions as features, as compared to face/audio features.

Table 1. Results for each emotion dimension, using (i) other emotion dimensions as features ($\boldsymbol{\mathcal{S}}_{\setminus k}$), and (ii) other emotion dimensions and interest dimension as features ($\boldsymbol{\mathcal{SI}}_{\setminus k}$).

	Valence		**Arousal**		**Power**		**Expectation**		**Intensity**	
	MSE	**COR**	**MSE**	**COR**	**MSE**	**COR**	**MSE**	**COR**	**MSE**	**COR**
$\boldsymbol{\mathcal{S}}_{\setminus k}$	0.074	0.28	0.051	0.47	**0.088**	**0.28**	**0.037**	0.15	0.067	0.30
$\boldsymbol{\mathcal{SI}}_{\setminus k}$	**0.063**	**0.30**	**0.052**	**0.56**	**0.088**	0.23	0.039	**0.16**	**0.052**	**0.330**

4.2. RCCA Fusion and Predicting Interest

In this section, we will focus on predicting interest, and in this way evaluate the performance of the proposed RCCA, as well as derive some more conclusions on the relationship between interest and emotion dimensions. Firstly, in order to evaluate the performance of RCCA, we learn the mapping

$$f : [\mathbf{P}_z\mathbf{Z}, \mathbf{P}_a\mathbf{A}] \rightarrow \boldsymbol{\mathcal{I}}, \tag{3}$$

where the matrices $\mathbf{Z}, \mathbf{A}$ represent the facial trackings/audio features, $\mathbf{P}_z$ and $\mathbf{P}_a$ the projections recovered by RCCA, and $\boldsymbol{\mathcal{I}}$ represents the interest annotation. For comparison, we evaluate using (i) the annotations for emotion dimensions as features ($\boldsymbol{\mathcal{S}}$=Valence, Arousal, Power, Expectation, Intensity), (ii) single modalities separately, i.e. facial tracings and audio features, (iii) feature-level fusion, where the features from different modalities are simply concatenated, (iv) classical CCA with $l2$ regularisation, and (v) RCCA. Results from this experiment can be found in Table 2. There are several interesting results we can observe. Firstly, audio cues appear better for predicting interest in contrast to facial features. This is expected, since according to theory [17] as well as the evaluation performed in this paper (Sec. 4.1), interest is more correlated with arousal, which is the primary dimension for which audio cues are known to perform better [15, 12], while this has also been confirmed by other works on interest recognition (c.f., [2]). Furthermore, it is clear that feature level fusion and classical CCA fusion are not able to outperform single-cue prediction. In fact, CCA fusion merely manages to achieve equal accuracy to using simply audio cues. It is clear that RCCA outperforms all compared techniques, by correctly estimating a low-rank subspace where the input modalities are maximally correlated, free of gross noise contaminations, capturing both intra and inter-cue correlations. Two final observations regard the interest annotations themselves. In previous work [19], it has been shown that by using other emotion dimensions as features, one could obtain better results than by just using facial trackings or audio cues as features. This conclusion does not hold for interest, as can be seen here. This could be an indication that joystick-based annotations can provide more accurate, better correlated results with respect to audio/visual features. Furthermore, from Fig. 1 and Tab 1 and 2, we can conclude that although interest appears to have an overlap with other emotion dimensions, the interest annotation seems to hold information which is not entirely captured in other dimensions.

Table 2. Results for predicting interest from emotion dimensions in the SEMAINE database ($\mathcal{S}$), facial trackings (Face), audio cues (Audio), feature-level fusion (F_l), CCA-based fusion (CCA$_f$) and Robust CCA fusion (RCCA$_f$).

	$\boldsymbol{\mathcal{S}}$	**Face**	**Audio**	$\mathbf{F}_l$	**CCA**$_f$	**RCCA**$_f$
MSE	0.032	0.033	0.031	0.031	0.031	0.029
COR	0.378	0.432	0.460	0.443	0.458	**0.490**

5. CONCLUSIONS

In this work, we analyse a set of continuous *interest* annotations corresponding to audio-visual data. Amongst other findings, we experimentally demonstrate that despite the fact that interest annotations were obtained utilising different tools and a disjoint set of annotators, there still exist strong correlations between interest and other emotion dimensions, thus motivating the utilisation of models which exploit output-correlations for detecting interest. Most significantly, we introduce a robust Canonical Correlation Analysis (RCCA) for audio-visual fusion, which is able to learn low-rank projections and isolate gross errors in the fused modalities. We experimentally show that RCCA provides features which outperform $l2$ CCA, feature-level fusion as well as single-cue features.

6. ACKNOWLEDGEMENTS

This work has been funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG). The work of Y. Panagakis is funded by the European Research Council under the FP7 Marie Curie Intra-European Fellowship. The work of S. Zafeiriou is funded by the EPSRC project EP/J017787/1 (4DFAB).

7. REFERENCES

[1] Alex Pentland and Anmol Madan, "Perception of social interest," in *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*, 2005.

[2] Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[3] Björn Schuller and Gerhard Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features.," in *INTERSPEECH*, 2009, pp. 1999–2002.

[4] Felix Arnold, *Attention and interest: A study in psychology and education*, Macmillan, 1910.

[5] Silvan S Tomkins, "Affect, imagery, consciousness: Vol. i. the positive affects.," 1962.

[6] Paul J Silvia, *Exploring the psychology of interest*, Oxford University Press, 2006.

[7] Björn Schuller, Niels Köhler, Ronald Müller, and Gerhard Rigoll, "Recognition of interest in human conversational speech.," in *INTERSPEECH*, 2006.

[8] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies.," in *INTERSPEECH*, 2008, pp. 597–600.

[9] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[10] Angeliki Metallinou and Shrikanth S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Apr. 2013.

[11] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Björn Schuller, and Shrikanth Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

[12] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 827–834.

[13] Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction*, pp. 396–406. Springer, 2011.

[14] James A Russell, Maria Lewicka, and Toomas Niit, "A cross-cultural study of a circumplex model of affect.," *Journal of personality and social psychology*, vol. 57, no. 5, pp. 848, 1989.

[15] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 92–105, 2011.

[16] Hatice Gunes and Björn Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, 2012.

[17] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[18] Tadas Baltrušaitis, Ntombikayise Banda, and Peter Robinson, "Dimensional affect recognition using continuous conditional random fields," in *IEEE FG*, 2013.

[19] Mihalis A. Nicolaou, Stefanos Zafeiriou, and Maja Pantic, "Correlated-spaces regression for learning continuous emotion dimensions," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 773–776.

[20] Caifeng Shan, Shaogang Gong, and Peter W McOwan, "Beyond facial expressions: Learning human emotion from body gestures.," in *BMVC*, 2007, pp. 1–10.

[21] Nicolle M Correa, Yi-Ou Li, Tülay Adali, and Vince D Calhoun, "Fusion of fmri, smri, and eeg data using canonical correlation analysis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 385–388.

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

[23] Mihalis A. Nicolaou, Vladimir Pavlovic, and Maja Pantic, "Dynamic Probabilistic CCA for Analysis of Affective Behaviour," in *Proceedings of the 12th European Conference on Computer Vision, ECCV 2012.*, Florence, Italy, October 2012, pp. 98–111.

[24] Soroosh Mariooryad and Carlos Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 85–90.

[25] 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.

[26] Zhihong Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE TPAMI*, 2009.

[27] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *JMLR*, vol. 1, pp. 211–244, 2001.

[28] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[29] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[30] D. Donoho, "For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[31] M. Fazel, *Matrix Rank Minimization with Applications*, Ph.D. thesis, Dept. Electrical Engineering, Stanford University, CA, USA, 2002.

[32] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. 2011 Neural Information Processing Systems Conf.*, Granada, Spain, 2011, pp. 612–620.

[33] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, Belmont, MA, 2nd edition, 1996.

[34] J. F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal Optimization*, vol. 2, no. 2, pp. 569–592, 2009.

[35] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of ACM*, vol. 58, no. 3, pp. 1–37, 2011.