Accepted Manuscript

Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals

Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, Björn Schuller

PII: S0262-8856(16)30217-7
DOI: doi:10.1016/j.imavis.2016.11.020
Reference: IMAVIS 3582

To appear in: *Image and Vision Computing*

Received date: 24 June 2016
Revised date: 22 October 2016
Accepted date: 28 November 2016

Please cite this article as: Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, Björn Schuller, Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals, *Image and Vision Computing* (2016), doi:10.1016/j.imavis.2016.11.020

Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals

Jing Han[a], Zixing Zhang[a,*], Nicholas Cummins[a], Fabien Ringeval[a,b], Björn Schuller[a,c]

[a]*Chair of Complex and Intelligent Systems, University of Passau, Innstr. 41, 94032 Passau, Germany*
[b]*Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, 700 Avenue Centrale, 38058 Grenoble, France*
[c]*Department of Computing, Imperial College London, 180 Queens' Gate, London SW7 2AZ, UK*

Abstract

Automatic continuous affect recognition from audiovisual cues is arguably one of the most active research areas in machine learning. In addressing this regression problem, the advantages of the models, such as the global-optimisation capability of Support Vector Machine for Regression and the context-sensitive capability of memory-enhanced neural networks, have been frequently explored, but in an isolated way. Motivated to leverage the individual advantages of these techniques, this paper proposes and explores a novel framework, Strength Modelling, where two models are concatenated in a hierarchical framework. In doing this, the strength information of the first model, as represented by its predictions, is joined with the original features, and this expanded feature space is then utilised as the input by the successive model. A major advantage of Strength Modelling, besides its ability to hierarchically explore the strength of different machine learning algorithms, is that it can work together with the conventional feature- and decision-level fusion strategies for multimodal affect recognition. To highlight the effectiveness and robustness of the proposed approach, extensive experiments have been carried out on two time- and value-continuous spontaneous emotion databases (RECOLA and SEMAINE) using audio and video signals. The experimental results indicate that employing Strength Modelling can deliver a significant performance improvement for both arousal and valence in the unimodal and bimodal settings. The results further show that the proposed systems is competitive or outperform the other state-of-the-art approaches, but being with a simple implementation.

Keywords: Strength Modelling, support vector regression, memory-enhanced recurrent neural networks, audiovisual affective computing

1. Introduction

Automatic affect recognition plays an essential role in smart conversational agent systems that aim to enable natural, intuitive, and friendly human–machine interaction. Early works in this field have focused on the recognition of prototypic expressions in terms of basic emotional states, and on the data collected in laboratory settings, where speakers either act or are induced with predefined emotional categories and content [9, 29, 30, 47]. Recently, an increasing amount of research efforts have converged into dimensional approaches for rating naturalistic affective behaviours by continuous dimensions (e. g., arousal and valence) along the time continuum from audio, video, and music signals [8, 10, 24, 39, 46, 16, 32, 33]. This trend is partially due to the benefits of being able to encode small difference in affect over time and distinguish the subtle and complex spontaneous affective states. Furthermore, the affective computing community is moving toward combining multiple modalities (e. g., audio and video) for the analysis and recognition of human emotion [19, 23, 34, 43, 49], owing to (i) the easy access to various sensors like camera and microphone, and (ii) the complementary information that can be given from different modalities.

In this regard, this paper focuses on the realistic time- and value-continuous affect (emotion) recognition from audiovisual signals in the arousal and valence dimensional space. To handle this regression task, a variety of models have been investigated. For instance, *Support Vector Machine for Regression* (SVR) is arguably the most frequently employed approach owing to its mature theoretical foundation. Further, SVR is regarded as a baseline regression approach for many continuous affective computing tasks [27, 31, 36]. More recently, memory-enhanced *Recurrent Neural Networks* (RNNs), namely *Long Short-Term Memory RNNs* (LSTM-RNNs) [14], have started to receive greater attention in the sequential pattern recognition community [7, 26, 48, 50]. A particular advantage offered by LSTM-RNNs is a powerful capability to learn longer-term contextual information through the implementation of three memory gates in the hidden neurons. Wöllmer et al. [41] was amongst the first to apply LSTM-RNN on acoustic features for continuous affect recognition. This technique has also been successfully employed for other modalities (e. g., video, and physiological signals) [2, 21, 26].

Numerous studies have been performed to compare the advantages offered by a wide range of modelling techniques,

[*]corresponding author: zixing.zhang@uni-passau.de, Tel.: +49 851 509-3359, Fax.: +49 851 509-3352