Image and Vision Computing 52 (2016) 1–14

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Real-time facial action unit intensity prediction with regularized metric learning[☆]

Jérémie Nicolle*, Kévin Bailly, Mohamed Chetouani

Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7222, Institut des Systèmes Intelligents et de Robotique (ISIR), 4 place Jussieu, 75005 Paris, France

ARTICLE INFO

Article history:
Received 10 December 2014
Received in revised form 19 February 2016
Accepted 4 March 2016
Available online 3 May 2016

Keywords:
Facial expression
Action Units
FACS
Metric Learning for Kernel Regression

ABSTRACT

The ability to automatically infer emotional states, engagement, depression or pain from nonverbal behavior has recently become of great interest in many research and industrial works. This will result in the emergence of a wide range of applications in robotics, biometrics, marketing and medicine. The Facial Action Coding System (FACS) proposed by Ekman features objective descriptions of facial movements, characterizing activations of facial muscles. Achieving an accurate intensity prediction of Action Units (AUs) has a significant impact on the prediction quality of more high-level information regarding human behavior (e.g. emotional states). Real-time AU intensity prediction, in many image-related machine learning tasks, is a high-dimensional problem. For solving this task, we propose adapting the Metric Learning for Kernel Regression (MLKR) framework focusing on overfitting issues induced in high-dimensional spaces. MLKR aims at estimating the optimal linear subspace for reducing the squared error of a Gaussian kernel regressor. We introduce Iterative Regularized Kernel Regression (IRKR), an iterative nonlinear feature selection method combined with a Lasso-regularized version of the original MLKR formulation that improves on the state-of-the-art results on several AU databases, ranging from prototypical to natural and wild data.

1. Introduction

Automatic facial expression recognition has recently become a very active and rapidly evolving research domain. To precisely describe facial expressions, the Facial Action Coding System (FACS [1]) encodes Action Units (AUs), which correspond to the activation of facial muscles.

The ability to accurately predict AU intensity has a significant impact on human behavior assessment. During a video, the ability to describe in each frame what and to what extent facial muscles are activated gives us a complete description of a subject's facial movements. This would contain precious information regarding mental states [2], depression [3] and pain [4,5] prediction, for instance. Industrial applications that take advantage of AU predictions are numerous as well. Applications in marketing [6] or Human–Computer Interaction [7] have recently emerged.

In this paper, we address three main issues: First, AU automatic prediction has mainly been seen as a classification problem. However, the ability to predict muscle activation more precisely is essential. Very small and short activations of AUs (called micro-expressions) can be of great value for emotion assessment [8]. Moreover, the dynamics of AUs have an important impact on the meaning of facial expressions. In [9], the authors worked on classifying two different types of smiles (frustrated and delighted) showing the relevance of temporal pattern analysis for this task. For those reasons, multilevel annotated databases have recently been released (enhanced CK+ [10], DISFA dataset [11], AM-FED dataset [6]), thus making it possible to build and evaluate new methods suited for regression tasks. The second issue is that the algorithms should be run in real time, which is an important constraint for many domains such as personal robotics and car passenger security. This constraint encourages fast-to-compute features and fast regression methods. Finally, some AUs are very rarely activated in natural behavior such as the Nose Wrinkler (AU9) or Lip Stretcher (AU20). This makes the number of positive examples small, even when the amount of acquired video data is important. Thus, a particular focus on the risk of overfitting on the training data must be made.

We propose a regression method based on a Lasso-regularization of MLKR included within an iterative nonlinear feature selection framework. This method lets us project data points into sparse and

[☆] This paper has been recommended for acceptance by Georgios Tzimiropoulos, PhD.

[*] Corresponding author at: UPMC (Université Pierre et Marie Curie) ISIR (Institut des Systèmes Intelligents et de Robotique) Pyramide, T55/65 CC 173, 4 Place Jussieu 75005 Paris, France. Tel.: +33144276375.
E-mail address: jeremie.nicolle@isir.upmc.fr (J. Nicolle).

http://dx.doi.org/10.1016/j.imavis.2016.03.004
0262-8856/© 2016 Elsevier B.V. All rights reserved.

low-dimensional spaces, allowing us to reduce overfitting issues. In Section 2, we present a brief state of the art of AU prediction methods. Section 3 contains an outline of our framework and the paper contributions. In Section 4, we present MLKR, on which our regression method is built, and discuss some of its advantages. Section 5 describes our proposed regression method. Its application to AU intensity prediction and the associated results are presented in Section 6. Finally, we conclude and discuss a few issues and perspectives in Section 7.

2. Related works

Numerous AU prediction methods have been proposed during the past decade along with the growing interest in this domain. Detecting AUs is a supervised machine learning problem. Face-centered data are acquired (gray-level, RGB and/or depth-map) and labeled manually. The labels indicate the different muscles activated by the subject. We then must extract features describing data before learning a prediction model. Because AUs are related to local changes in facial expression, it is common to use a facial landmark detector to localize the different parts of the face (mouth, eyes, nose, eyebrows). The features can subsequently be extracted on different facial areas. Those features characterizing data samples are then used for predicting labels with a supervised machine learning algorithm. Along the entire data processing chain, from the acquisition sensors to the prediction method, many questions have been highlighted by past works. First, the availability of affordable 3D sensors has attracted many researchers to focus on the utility and contribution of depth-related data for facial muscle activation predictions and has made the data type a relevant question. Second, the choice of the areas used for feature extraction has an important impact. Third, the inclusion of prior human knowledge when designing high-level features relevant to the task can increase performance but leads to less generic methods. Similarly, including prior knowledge within the models (e.g. regarding AU co-occurrences in natural facial expressions) has also raised questions. Finally, the choice of the learning machines used to model the data has also been an active topic in past works. In this section, we will briefly review and discuss some of the main AU prediction methods recently proposed.

The relevance of using 3-dimensional data for facial expression recognition has been investigated by several researchers. Sun et al. [12] used 3D motion vectors and Hidden Markov Models (HMMs) for predicting AUs and discrete emotions in a Dynamic 3D Facial Expression Database. Savran et al. [13] extracted local 3D shape features (mean and Gaussian curvatures, shape index and curvedness among others) and use an SVM for predicting AUs in a Bosphorus database. However, 3D sensors are not yet widely democratized, and many applications have a need for 2D data solutions, which explains the numerous recent 2D approaches for AU prediction [11,14,15]. Most of those 2D approaches can be easily extended to 3D approaches by extracting complementary features using depth maps in the same way as grayscale or color images.

Before extracting features from images, a common first step in many face-centered machine learning systems is to detect fiducial points, which are some key points in faces (centers and corners of the eyes, contours of the nose, the mouth and the eyebrows). In Jeni et al. [16] and Chu et al. [17], those fiducial points are used to define local patches for feature extraction to predict AUs. However, a few methods [18,19] avoid this part of fiducial point localization, extracting features on somewhat global regions defined only using the area obtained with the face detector (commonly using the Viola and Jones algorithm [20]). Yang et al. [18] directly extracted dynamic Haar-like features after a rescaling the detected face image and then encoded it with binary patterns before classification using Adaboost [21]. Chuang and Shih [19] divided the face region in upper and lower parts before using the Support Vector Machine (SVM) on Independent Component Analysis (ICA) projections. Other methods use only eye localization for defining feature extraction areas [10,22]. By definition, AUs are characterized by local movements of face appearance. This is why the extraction of features in local areas defined from fiducial points lead to relevant information for our task. However, using more global areas defined using only the face region or the centers of the eyes (which are the most accurately located points in most landmark detection methods) can avoid the spread of possible errors in facial point tracking. The recent improvement of facial point localization systems can explain the fact that local areas are increasingly used in AU prediction systems [15,16,17].

AU prediction methods also differ regarding the amount of human knowledge included in the feature choice. Some methods use data-driven features, which often makes the framework more generic; for example, Chuang and Shih [19] used Independent Component Analysis (ICA), and Jeni et al. [16] used Non-negative Matrix Factorization (NMF). Even if it introduces a loss of genericity, other methods use handcrafted features, which may lead to relevant invariance and characterizations. Rudovic et al. [23] used Local Binary Patterns (LBPs) that are invariant to illumination changes. Gabor wavelets are commonly used [10,13,22] and have shown promising results for AU prediction as noted by Littlewort et al. [24]. However, dense computation of those features for different scales and orientations quickly becomes time-consuming and unsuited for real-time algorithms. This can explain the choice of Histograms of Oriented Gradients (HOGs) made by McDuff et al. [6], which encode relevant information for expression-relative wrinkle characterization while being less time-consuming to extract.

Prior knowledge can also be included in data modeling. Several researchers have focused on learning dynamic relationships and co-occurrences between AUs to increase algorithm performance, such as Tong et al. [10] and Li et al. [14], using Dynamic Bayesian Networks (DBNs). These approaches are able to consider correlations between AUs in natural facial expressions. For instance, eyebrow raising (AU1+AU2) and upper lid raising (AU5) are often activated simultaneously. However, AUs correspond to facial muscles and can be activated independently, making the prior knowledge about dynamic relations between AUs inadequate in some applications. For instance, in the context of facial reeducation for patients who had a cerebrovascular accident (CVA), different muscles may need to be separately activated by the patient and thus separately recognized. A prior knowledge inclusion in this case could bias the prediction system.

Finally, there is the question of the machine learning algorithms used for building prediction models. In many databases (Cohn–Kanade [25], Carnegie Mellon University PIE database [26], Fera-Gemep [27]) AUs are labeled as activated or not, stating the problem as a classification problem. Thus, Support Vector Machines (SVMs) have been widely used in the facial expression domain [6,22,28]. However, information given by AU detectors is limited, and many applications require more comprehensive information—i.e., the intensity of the AU. In the first few attempts to estimate intensities of facial expression [29,30,31,32], only binary labels were used to train classifiers such as SVM or AdaBoost. Intensities were thus inferred from the output of the classifier (e.g., the signed distance from the sample to the separating hyperplane of the SVM [29,31] or the confidence of the decision in the case of AdaBoost classifier [30,32]). These approaches assume that facial expression intensity is directly related to the distance from the decision boundary. The idea is that samples corresponding to low intensities are more difficult to classify and are thus more likely to be near the boundary. This point is questionable because the difficulty of classifying a sample can be due to other unrelated factors such as lighting conditions and morphological characteristics.