HAL
open science

Real-Time Facial Action Unit Intensity Prediction with Regularized Metric Learning

Jérémie Nicolle, Kevin Bailly, Mohamed Chetouani

To cite this version:

Jérémie Nicolle, Kevin Bailly, Mohamed Chetouani. Real-Time Facial Action Unit Intensity Prediction with Regularized Metric Learning. Image and Vision Computing, 2016, 52, pp.1-14. 10.1016/j.imavis.2016.03.004 . hal-01318177

HAL Id: hal-01318177

https://hal.sorbonne-universite.fr/hal-01318177v1

Submitted on 19 May 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ACCEPTED MANUSCRIPT

Title of the paper : Real-Time Facial Action Unit Intensity Prediction with Regularized Metric Learning

Authors names : Jérémie Nicolle, Kévin Bailly, Mohamed Chétouani.

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

Corresponding author's full contact details :
Mr. Jeremie Nicolle
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
Phone: +330144276375
E-mail: jeremie.nicolle@isir.upmc.fr

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.

Keywords : Facial Expression, Action Units, FACS, Metric Learning for Kernel Regression

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Real-Time Facial Action Unit Intensity Prediction with Regularized Metric Learning

Jérémie Nicolle, Kévin Bailly, Mohamed Chetouani

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.

Keywords: Facial Expression, Action Units, FACS, Metric Learning for Kernel Regression

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

ACCEPTED MANUSCRIPT

method lets us project data points into sparse and 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-occurences 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 [14] [11] [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 [16] [17] [15].

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 characteriza-

ACCEPTED MANUSCRIPT

tions. Rudovic et al. [23] used Local Binary Patterns (LBPs) that are invariant to illumination changes. Gabor wavelets are commonly used [10] [22] [13] 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 (HOG) 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 one of classification. Thus, Support Vector Machines (SVMs) have been widely used in the facial expression domain [22] [28] [6]. 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.

Recent works on intensity level estimation are mainly based on the newly released datasets with intensity labels to obtain a more accurate estimation (Bosphorus [33], CK+ [34], UNBC-McMaster [35], DISFA [11]). In the case of AUs, intensities are discrete values ranging from 0 to 5. Thus, intensity estimation can be viewed either as a 6-class recognition problem [36, 11, 37] or as a regression task [16, 38, 39, 40, 41]. In [36], binary SVM classifiers are used in a one-versus-one strategy to obtain a multi-class decision. In the same line, Ming et al. [37] extends the framework introduced in [42], based on LGBP features and multi-kernel SVM to tackle multi-class classification. The main drawback of classification approaches is that the training does not consider relative distances between labels.

In contrast, regression-based methods intrinsically consider the ordinal relation between labels. Large differences between prediction and ground truth will be more penalized than small errors. Given the good performance reached by SVM for AU classification, SVR [16] and RVM [38] are among the most widely used predictors for intensity estimations, but other machine learning algorithms have been investigated such as generative latent trees [39] or deep convolutional neural networks [40]. In that type of study, predictions in each frame are made independently based on observations in the current frame [16, 43, 13]. Temporal [44] and/or other contextual information [41] can be used to improve the prediction. On the one hand, the use of a sequence of frames can improve the prediction by removing some ambiguities. On the other hand, these methods can be applied only in video-based applications and required pre-segmented sequences during the training stage. Moreover, some of those graphical probabilistic models such as HMM can be challenging to learn given the number of features usually employed. The dimension of the input space can also be an issue for static methods. Savran et al. [13] investigated two regression-oriented versions of Adaboost to select features and noted the strong sensitivity of this approach to some hyper-parameters such as the threshold that convert the regression problem to a classification one or the power coefficient of the weighting function in AdaBoost.RT.

Our method combines a filter-based feature selection approach with a metric learning algorithm to best adapt the high-dimensional input feature space to the task to

ACCEPTED MANUSCRIPT

perform. Our choices regarding the issues highlighted by this state of the art are presented in the next section in an overview of our regression framework.

3. Overview

In figure 1, we present the architecture of our system. We use grayscale images as the raw data type to ensure a wide range of applications. We chose both geometric and appearance features. Our geometric features characterize relationships among triplets of fiducial points to avoid sensitivity to rotations and scaling. For appearance features, we use Histograms of Oriented Gradients (HOGs) on local patches because of their relevance for describing emotion-related wrinkles and their low computation time. Some of our patches are centered using the fiducial points. Other patches are located using only the Viola-Jones face detection area to ensure robustness in case of a landmark tracking failure. More details about our features can be found in Section 6.1. Those features and the associated labels are then used for learning our prediction system.

Labeling AUs is complex and time-consuming for several reasons. Only experts, with specific training, can precisely identify the activated muscles and their corresponding intensities in an image [45]. Thus, frame-by-frame annotation of an important number of AUs is difficult (there are more than 45 muscles in the human face). Moreover, in natural behavior, many AUs are very rarely activated. This explains why, even with several hours of video data, the number of positive activations can be small (*e.g.*, there are only 4 activations of maximal intensity for AU2 in the DISFA database). The data are said to be imbalanced (the number of unactivated samples is considerably higher than the number of activated ones). Thus, we decided to focus on overfitting when designing our method.

For each AU, we learn a low-dimensional space suited for a non-parametric Gaussian kernel regressor by using a Lasso-regularized version of MLKR within an iterative nonlinear feature selection process. The small number of dimensions in our representation spaces and the regularization aim at reducing a potential overfitting on the training data. Moreover, the imbalanced data distribution induces some issues for regression evaluation when using commonly used metrics such as Root Mean Square Error (RMSE) or Correlation Coefficient (CC). We discuss this and introduce a new evaluation metric, r-AUC, in Section 6.3.

More details about our regression framework can be found in Section 5. The main contributions of this paper are the following:

- A complete framework for real-time AU intensity prediction improving state-of-the-art results in prototypical and natural databases.
- A Lasso-regularized version of Metric Learning for Kernel Regression (Lasso-MLKR).
- A new evaluation metric (r-AUC), suited for regression tasks on imbalanced data, extending Area Under ROC Curve for regression, that we present in Section 6.3.

Our method is built upon Metric Learning for Kernel Regression (MLKR), which we introduce in the next section.

4. Metric Learning for Kernel Regression

Kernel regression has proved to be efficient in a wide range of applications (from image deblurring [46] or segmentation [47] to automatic human emotion prediction [48]). However, the performance of the regressors highly depends on the relevance of the space in which the samples lie, making appropriate dimensionality reduction a necessary initial step. Weinberger and Tesauro [49] proposed MLKR (Metric Learning for Kernel Regression), which aims at finding the optimal linear projection to minimize the kernel regression squared error on the training set.

In kernel regression, an instance label is predicted using the Nadaraya–Watson estimator [50], as an average of the training instance labels weighted using some similarity measure. If we consider n_s training samples $\{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_{n_s}\}$ associated with corresponding labels $\{y_1, y_2, .., y_{n_s}\}$, the label corresponding to a feature vector $\mathbf{x}_t$ will be approximated by

$$\hat{y}_t = \frac{\sum_{i=1}^{n_s} y_i k_{i,t}}{\sum_{i=1}^{n_s} k_{i,t}} \tag{1}$$

using a kernel $k_{i,t} = k(\mathbf{x}_i, \mathbf{x}_t)$ as a similarity metric between samples i and t.

MLKR proposes a direct optimization of the kernel regression error for the commonly used Gaussian kernel, which can be defined as follows:

$$k_{i,j} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d_{i,j}^2}{\sigma^2}} \tag{2}$$

where σ is the Gaussian spread and $d_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$ is the euclidean distance between samples i and j. Let us consider an original space of dimension n_d and an output

ACCEPTED MANUSCRIPT

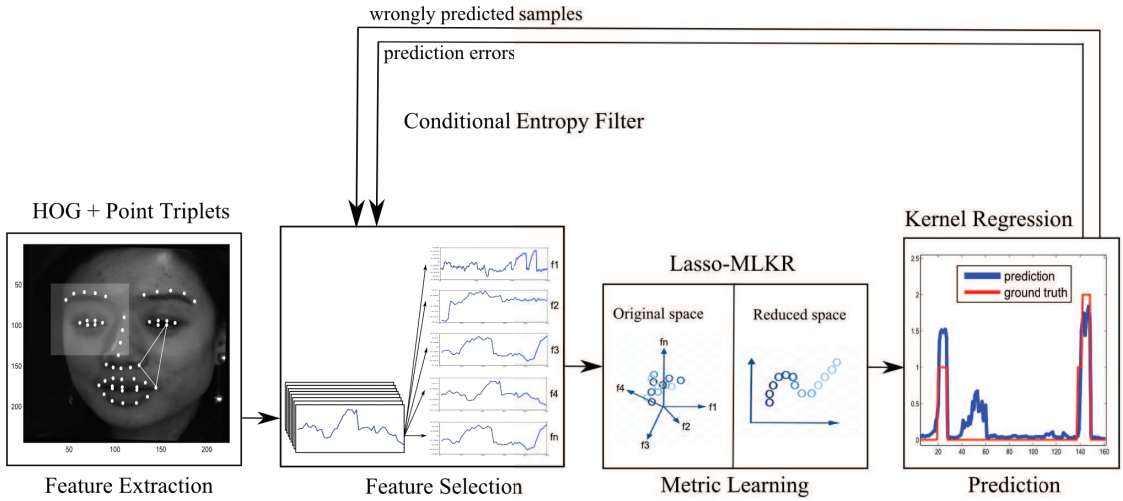

Figure 1: Architecture of the proposed framework : the most relevant geometric and appearance features are selected by evaluating the conditional entropy between each feature and the label to predict. For each Action Unit, a low-dimensional space suited for a non-parametric Gaussian kernel regressor is learned by using a Lasso-regularized version of Metric Learning for Kernel Regression. Complementary features are iteratively selected according to the prediction errors.

space of dimension n_r. MLKR aims at finding a projection matrix $\mathbf{A} \in \mathcal{M}_{n_d,n_r}(\mathbb{R})$ that minimizes the squared error $\mathcal{L}$ on the training samples:

$$\mathcal{L}(\mathbf{A}) = \sum_{i=1}^{n_s} (\hat{y}_i - y_i)^2 \tag{3}$$

where

$$\hat{y}_i = \frac{\sum_{j \neq i} y_j k_{j,i}(\mathbf{A})}{\sum_{j \neq i} k_{j,i}(\mathbf{A})}$$

where

$$k_{i,j}(\mathbf{A}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d_{i,j}(\mathbf{A})^2}{\sigma^2}}$$

$$d_{i,j}(\mathbf{A})^2 = \|\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\|^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{A}^\top \mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)$$

is the squared distance in the reduced subspace of dimension n_r. The optimization process of the squared error is performed with a gradient descent. We obtain by an analytical calculation

$$\frac{\partial \mathcal{L}(\mathbf{A})}{\partial \mathbf{A}} = 4\mathbf{A} \sum_i \frac{(\hat{y}_i - y_i)}{\sum_{j \neq i} k_{ij}} \sum_j (\hat{y}_i - y_j) k_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{4}$$

Metric Learning for Kernel Regression (MLKR) lets us project data points in a low-dimensional space suited for nonlinear prediction via Gaussian kernel regression.

In this paragraph, we explain our choice to use MLKR by discussing some advantages and limitations of the method.
First, MLKR does not directly learn a prediction function but learns a space in which a Nadaraya–Watson estimator is performed using a set of data and labels. We can then project new data points in the learned space for predicting without relearning the system (for instance, to easily adapt to a new database or to a specific subject). Second, the Nadaraya–Watson estimator is able to adapt easily to heterogeneous point distribution because of the normalization by $\sum_{i=1}^{n_s} k_{i,t}$, which helps AU intensity prediction when trying to predict an unknown subject lying in a sparsely populated part of the space. However, MLKR has several drawbacks. First, it is non-convex. However, experiments have shown that local minima lead to accurate predictions on standard regression datasets [49]. We observed a similar behavior with our experiments on AU databases. Second, it has a quadratic complexity relatively to both the number of features and the number of samples, which makes it difficult to use on high-dimensional and large datasets.
In the next section, we introduce our regression method, which is based on an adaptation of MLKR for high-dimensional spaces.

ACCEPTED MANUSCRIPT

5. Iterative Regularized Kernel Regression (IRKR)

In the MLKR algorithm, the number of estimated parameters when reducing a space of dimension n_d into a subspace of dimension n_r is $n_{par} = n_d.n_r$. If the number of training samples is too small compared with the number of model parameters, the risk of overfitting increases. We propose in Section 5.1 to modify the original formulation by regularizing it using a Lasso-penalty for the reduction of overfitting risk.

Moreover, the gradient computation for a projection of n_s samples into a space of dimension n_d has a complexity of $O(n_s^2.n_d^2)$, making it difficult to use in high-dimensional spaces. We propose a complete framework improving the original MLKR formulation to make it efficient in high-dimensional datasets.

A widely used step for supervised dimensionality reduction is filter feature selection [51], which aims at characterizing the relevance of the features independently of the predictor's choice, often one by one, for predicting the label. In other words, it computes a similarity (or dissimilarity) measure between each feature and the label and selects the highest ones (or the smallest ones, respectively). We propose the use of a conditional entropy measure that is able to find nonlinear relationships between features and labels. Details are given in Section 5.2.

Furthermore, we propose that it be included within an iterative framework because filter-based methods have a high risk of selecting redundant information (see paragraph 5.3).

5.1. Lasso-MLKR

Original MLKR minimizes the training reconstruction error with respect to the coefficients of the projection matrix $\mathbf{A}$. We propose to regularize this original MLKR formulation using a Lasso-penalty, meaning that we add a weight to the cost function corresponding to the L^1-norm of the matrix $\mathbf{A}$ (which is the sum of the absolute values of its coefficients). This penalty has been proved to induce sparsity in the estimated parameters, reducing the risk of overfitting [52]. Some of the coefficients are shrunk all the way to zero. Corresponding solutions, with multiple values that are identically zero are said to be sparse. The new energy formulation becomes

$$\mathcal{L}(\mathbf{A}) = \sum_{i=1}^{n_s} (\hat{y}_i - y_i)^2 + \lambda.L_1(\mathbf{A}) \tag{5}$$

where λ controls the regularization rate. The associate gradient becomes

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 4\mathbf{A} \sum_i \frac{(\hat{y}_i - y_i)}{\sum_{j \neq i} k_{ij}} \sum_j (\hat{y}_i - y_j) k_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top + \lambda.s(\mathbf{A}) \tag{6}$$

where s is the sign function.

The optimization of the regularization rate λ can be performed in cross-validation. In Iterative Regularized Kernel Regression (IRKR), this Lasso-MLKR method is used after a filter-based selection method and within an iterative process. This corresponds to the Metric Learning step in our system's schema (see figure 1).

It is common in face-related machine learning problems to extract tens of thousands of features for characterizing face appearance. However, the complexity of each step of the MLKR algorithm, quadratic with respect to the number of features, makes it complicated to use with such a high number of features. This motivates the feature selection we perform using a nonlinear dissimilarity metric described in the next section.

5.2. Conditional Entropy Feature Selection

The purpose of supervised filter-based feature selection is to identity features that contain relevant information for predicting a label. Different prior assumptions can be made on the functional relationship between features and labels. The simplest prior assumption that can be made between features and labels is linear dependency. The similarity measure associated with that dependency is the Correlation Coefficient. Because our regression method is nonlinear, we chose to use conditional entropy, which is able to discover nonlinear relationships between features and labels. The conditional entropy of a label l given a feature f is defined as follows:

$$H(l|f) = -\sum_{x \in \mathcal{F}} p(x) \sum_{y \in \mathcal{L}} p(y|x) \log(p(y|x))$$

where $\mathcal{F}$ and $\mathcal{L}$ are the sample spaces in which the feature and label are defined, respectively. Because fine estimations of conditional probabilities can be time-consuming with a high number of samples, we decided to compute the probabilities using six-level quantization of the features.

This metric allows relevant feature selection for predicting labels by assuming nonlinear functional relationships between features and labels. It is used in the Feature Selection step of IRKR (see figure 1).

ACCEPTED MANUSCRIPT

5.3. Iterative Feature Selection

Filter-based methods have been commonly included in iterative frameworks to select feature sets containing uncorrelated information [53]. In our framework, we first select a set of features and apply our regression to the learning database. We then select features correlated (in terms of conditional entropy) to the prediction error (for selecting uncorrelated information) and features correlated to the samples with the highest errors (to rapidly reduce the prediction error). The final framework, Iterative Regularized Kernel Regression (IRKR), which combines our three proposed contributions (Lasso-MLKR, conditional entropy and iterative feature selection), is presented in algorithm 1.

Algorithm 1 Iterative Regularized Kernel Regression

select a subset of n_s features $\{F\}$ calculating $H(l|f_j)$ for all j
$v_{sel} = \{F\}$
compute $\mathbf{A}$ using Lasso-MLKR on the feature set v_{sel}
calculate the prediction $\hat{l}^1$ on the training set
calculate the prediction squared error $e^1 = (\hat{l}^1 - l)^2$ on the training set
calculate the sum of errors of training samples s^1
identify the subset S^1 of samples whose errors are superior to the mean of e^1
$u = 1$
repeat
    $u = u + 1$
    select a subset of n_s features $\{F_e\}$ calculating $H(e^{u-1}|f_j)$ for all j
    select a subset of n_s features $\{F_m\}$ calculating $H(l(S^{u-1})|f_j(S^{u-1}))$ for all j
    $v_{sel} = v_{sel} \cup \{F_e\} \cup \{F_m\}$
    compute $\mathbf{A}$ using Lasso-MLKR on the feature set v_{sel}
    calculate the prediction $\hat{l}^u$ on the training set
    calculate the prediction error $e^u = (\hat{l}^u - l)^2$ on the training set
    calculate the sum of errors of training samples s^u
    identify the subset S^u of samples whose errors are superior to the mean of e^u
until $\frac{s^u}{s^{u-1}} < 0.99$
perform Kernel Regression on the test samples in the projected space defined by the matrix $\mathbf{A}$ learned last

In the next section, we present the application of IRKR to the task of AU intensity prediction.

6. Application to AU prediction

The feature extraction process is described in Section 6.1, followed by a presentation of the databases we used in Section 6.2. Different metrics commonly used for measuring AU system performance are discussed in Section 6.3. Finally, we detail our evaluation protocol in Section 6.4 and present the evaluations of the different parts of IRKR in Section 6.5 followed by our results in Section 6.6.

6.1. Feature Extraction

Most of the methods in facial-related information prediction combine two types of features: shape-based features and appearance-based features. Shape-based features are information relative to the positions of key landmarks in faces (eyes, nose, eyebrows and mouth contours), and appearance-based ones aim at describing image texture (globally or locally). For our task, landmark positions contain particularly interesting information because some AU activations directly induce important key point movements (such as raising the eyebrows or smiling). However, it is important to combine shape-related features with appearance-related ones for at least two reasons. First, shape-based features cannot encode some crucial information for AU prediction such as expression-relative wrinkle characterization. Second, current trackers may suffer from a lack of precision or robustness in challenging conditions, and appearance information may compensate for those errors in landmark prediction.

Shape-based features

We use Intraface tracker [54], which localizes 49 facial landmarks in real-time. To be insensitive to scaling and rotation in the image plane, we extract features relative to point triplets (as in [55]). Some works on facial expression analysis have proposed using features obtained after projection onto a manifold learned by PCA [56] [48]. However, those features encode global information. AU prediction is a local task because each AU corresponds to one facial muscle. Thus, we chose to extract information relative to point triplets. For each triplet of points $\mathbf{t}_{k_1k_2k_3} = (\mathbf{p}_{k_1}, \mathbf{p}_{k_2}, \mathbf{p}_{k_3})$, we calculate the ratio of both vectors

$$\mathbf{v}_{k_2k_3} = \mathbf{p}_{k_3} - \mathbf{p}_{k_2} = (\mathbf{p}^x_{k_3} - \mathbf{p}^x_{k_2}) + i.(\mathbf{p}^y_{k_3} - \mathbf{p}^y_{k_2})$$

and

$$\mathbf{v}_{k_2k_1} = \mathbf{p}_{k_1} - \mathbf{p}_{k_2} = (\mathbf{p}^x_{k_1} - \mathbf{p}^x_{k_2}) + i.(\mathbf{p}^y_{k_1} - \mathbf{p}^y_{k_2})$$

ACCEPTED MANUSCRIPT

to form

$$f(\mathbf{t}_{k_1k_2k_3}) = \frac{\mathbf{v}_{k_2k_1}}{\mathbf{v}_{k_2k_3}} = \frac{\|\mathbf{v}_{k_2k_1}\|}{\|\mathbf{v}_{k_2k_3}\|}.e^{i(\widehat{\mathbf{v}_{k_2k_3},\mathbf{v}_{k_2k_1}})}$$

that indicates the location of $\mathbf{p}_{k_1}$ relatively to $\mathbf{p}_{k_2}$ and $\mathbf{p}_{k_3}$. In this work, we take the real part and the imaginary part of $f(\mathbf{t}_{k_1k_2k_3})$ as features.

Appearance-based features

Before extracting appearance features, we cancel the rotation in the image plane and normalize the image using the estimation of the centers of the eyes. We then extract HOG descriptors (Histograms of Oriented Gradients) on different patches in the image. Some of them are centered on the landmarks to describe local texture and be able to capture expression-related wrinkles, and others are obtained by a 4x4 division of the image (see figure 2), giving us the possibility of catching up for potential point tracking errors. The patches centered using the landmarks we chose are presented in figure 3.

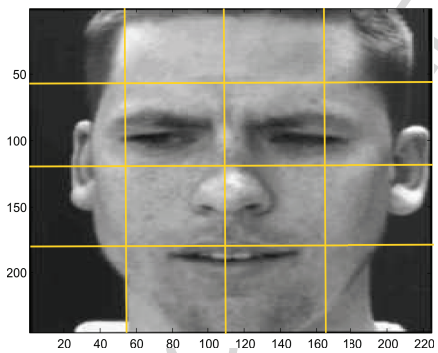

Figure 2: Patches located without the landmarks

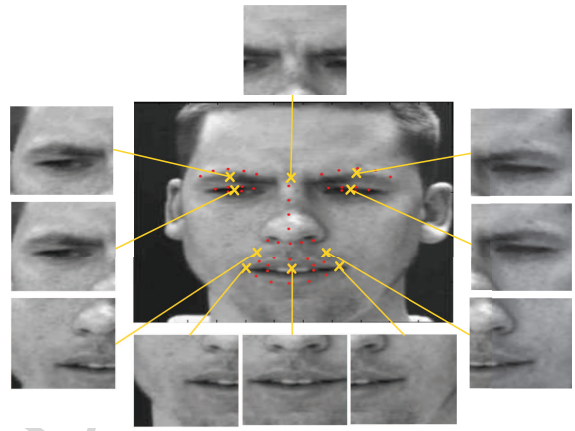

Figure 3: Patches centered using the landmarks

6.2. Databases

We used the Enhanced Cohn–Kanade Dataset for evaluating the different key points of our framework. This database contains prototypical behavior recorded in controlled conditions. We compared our algorithm results with state-of-the-art methods using the more natural DISFA Dataset.

Enhanced Cohn-Kanade Dataset

The CK dataset [25] consists of small video sequences in which subjects change their facial expressions from neutral to expressive. Each sequence is labeled in discrete emotions and FACS. A second version with more sequences (CK+) has been released [34], increasing the number of different subjects to 123. However, the labels are available only for the last frames of the sequences. The Intelligent Systems Lab of Rensselear Polytechnic Institute added manual relabeling of the dataset, frame by frame, with three different intensity levels for each AU (Enhanced Cohn–Kanade dataset). The different intensity levels are 0 if the AU is not activated, 1 if it is activated with small intensity, and 2 if it is completely activated. Image samples of the database are presented in figure 4.

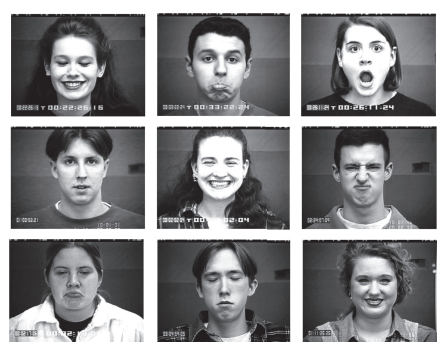

Figure 4: Examples of images extracted from the CK+ dataset

DISFA Dataset

The Denver Intensity of Spontaneous Facial Actions (DISFA) dataset [11] contains videos of 27 subjects (12 females and 15 males) with different ethnicities recorded watching a 4-minute emotive video stimulus. Data have been manually labeled frame by frame for 12 AUs on a six-level scale by a human FACS expert and verified by a second FACS coder. Image samples are presented in figure 5.

ACCEPTED MANUSCRIPT

Figure 5: Examples of images extracted from the DISFA dataset

6.3. Metrics

Different metrics exist for evaluating the performance of regression systems. In this paragraph, we define three commonly used metrics–namely, Root Mean Squared Error (RMSE), Pearson's Correlation Coefficient (CC) and Intraclass Correlation Coefficient (ICC)–and introduce a new metric called r-AUC and empirically show its advantages over other metrics on two examples.

The most commonly used metric for regression evaluation is RMSE, defined as follows for a label l and an estimated label $\hat{l}$:

$$RMSE(\hat{l}, l) = \sqrt{\frac{\sum_{i=1}^{n}(\hat{l}(i) - l(i))^2}{n}} \tag{7}$$

RMSE is often combined with Pearson's CC for evaluating the performance of a regression system. CC is defined as follows:

$$CC(\hat{l}, l) = \frac{\sum_{i=1}^{n}(l(i) - \bar{l})(\hat{l}(i) - \bar{\hat{l}})}{\sqrt{\sum_{i=1}^{n}(l(i) - \bar{l})^2}\sqrt{\sum_{i=1}^{n}(\hat{l}(i) - \bar{\hat{l}})^2}} \tag{8}$$

where $\bar{l}$ denotes the mean of the label.
Another commonly used metric is ICC, which, for k judges (i.e. a label l and an estimated label $\hat{l}$ in our case), is defined as

$$ICC = \frac{W - S}{W + (k-1)S} \tag{9}$$

where W is the Within-target Mean Squares, and S is the Residual Sum of Squares. Details of the computation can be found in [57].

Note that for our computations of the previously introduced evaluation metrics, we compare the ground truth to the regression system output with no quantization step.

Data used for learning AU prediction systems have a particular characteristic: they are highly imbalanced, meaning that there are in many cases very few positive samples compared with zero-valued samples (AU unactivated). In this context, and considering AU prediction as a classification task, Jeni et al. [58] investigated different performance measures (accuracy, F1 measure, AUC score, etc.) and concluded that Area Under ROC Curve (AUC) was the most robust and reliable metric for this task. The Receiver Operating Characteristic (ROC) curve represents the rate of true positives (positive samples that are correctly detected) as a function of the rate of false positives (negative samples that are incorrectly detected).

To take advantage of the robustness of this AUC metric for imbalanced data in the context of regression, we propose a new metric, called regression Area Under ROC Curve (r-AUC), defined as a mean of AUC scores for different binary quantizations of the label. Let us consider a label l varying from 0 to 1. We define a set $\{l_j, j \in [\![1; n_s]\!]\}$ of n_s binary quantizations of the label. $l_j(i)$ is 0 if $l(i) < \frac{j}{n_s+1}$ and 1 otherwise. r-AUC corresponds to the mean of the n_s AUCs calculated using the prediction and the different binary quantizations of the label. For a label l and a prediction p, we can obtain an explicit r-AUC score in a continuous manner as

$$r\text{-}AUC(l, p) = \frac{1}{max(l) - min(l)} \int_{min(l)}^{max(l)} AUC(p, l_s) ds$$

where l_s is the binary quantization of label l using the threshold s.

Let us consider two examples to illustrate the interest of r-AUC. If the system predicts a linear transformation of the label $\hat{l} = \alpha.l + \beta$, RMSE can be high, even for α near one and β near zero. We illustrate this issue in figure 6, where we can see that the noisy prediction on the lower part leads to a smaller RMSE than the prediction on the upper part. For many applications, this latter prediction would nevertheless be of great value because it contains all the dynamic information. We can notice that by using r-AUC metric, as well as CC or ICC, the first prediction is evaluated as the most relevant one. Note that a random prediction leads to an r-AUC of 0.5, and random predictions in binary classification lead to an AUC of 0.5.

Pearson's Correlation Coefficient (CC) lets us consider

ACCEPTED MANUSCRIPT

that linear transformations of the label are accurate predictions. However, in some cases, CC can be misleading. On the upper part of figure 7, the prediction is successful for the four main activations of the AU, but with wrong intensities, and on the lower part, the system succeeds only in predicting the most important activation. We can notice that all three metrics (RMSE, CC and ICC) indicate in this example that the second prediction is the best. The proposed r-AUC metric in this case would evaluate the first prediction more favorably because more activations are detected.

We believe that this metric lets us overcome important limitations of other standard metrics in the context of imbalanced data. We decided to use it along with CC for evaluating the different parts of our method. We used RMSE and CC for comparing our system's performance with recent state-of-the-art methods because those metrics were reported in the corresponding papers.

6.4. Experimental setup

All presented results for both datasets correspond to a subject-independent 4-fold cross-validation. All evaluations are performed on a global prediction signal corresponding to the concatenation of the 4 predictions. We extract 22,960 features from each frame (19,632 geometric features extracted from triplets of points and 3328 appearance features). Our HOG features are extracted with 8 directions on a 4x4 grid for each of the 26 patches. The λ regularization rate optimal value we found is 0.06. We add 10 features at each step of our iterative feature selection strategy, obtaining 70 final selected features for each AU. Our Lasso-MLKR algorithm performs projections on 4-dimensional spaces.

For the Enhanced Cohn–Kanade dataset, we used 2600 images for each of the four training folds and for each AU. We selected them to have 1300 unactivated samples (corresponding AU of value 0) and 1300 activated samples (randomly selected). The total training for the 14 AUs takes approximately 8 h on an Intel Core i7-3770 at 3.4 GHz.

For DISFA database, we used 6000 images for each of the four training folds and each AU. We selected them to have 3000 unactivated samples (corresponding AU of value 0) and 3000 activated samples (randomly selected). The total training for the 12 AUs takes approximately 14 h on an Intel Core i7-3770 at 3.4 GHz.

6.5. Evaluations of the Enhanced CK

In this section, we evaluate the contribution of the different key points of our framework on the Enhanced Cohn–Kanade Dataset (which is annotated using three different intensity levels).

Conditional entropy

In this paragraph, we compare feature selection with the conditional entropy similarity measure and Pearson's Correlation Coefficient on the Enhanced CK dataset. We consider the simplest configuration, without iterative feature selection or regularization of the MLKR formulation. We present the results obtained in terms of CC and r-AUC scores in table 1. We can observe a global improvement of 1.7 % when using the conditional entropy metric for feature selection in terms of CC, which is consistent for an important number of AUs (12 of 14 AUs are better predicted). This improvement is significant for several AUs. The main improvements correspond to AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU23 (Lip Tightener), AU24 (Lip Pressor) and AU25 (Lips Part). Most of these AUs have the common characteristic of provoking small landmark displacements, making appearance-based information of primary interest. We can explain those improvements by the important amount of nonlinearities between appearance-based features and labels, making conditional entropy particularly relevant for those AUs.

Lasso-MLKR

In this paragraph, we evaluate the contribution of MLKR Lasso-regularization. We consider a configuration with conditional entropy-based feature selection without iterative feature selection. We present the results obtained in table 2. We can observe a global improvement of 2.1% in terms of CC when adding regularization that is consistent with an important number of AUs (11 of 14 AUs are better predicted). The regularization, which lets us reduce the overfitting and increase the generalization power of our models, has a significant impact on some AUs, such as AU4 (Brow Lowerer), AU15 (Lip Corner Depressor), AU17 (Chin Raiser), AU23 (Lip Tightener) and AU24 (Lip Pressor). We can observe an important negative correlation between scores without regularization and the gain provided by the regularization, meaning a greater improvement for AUs that are the most difficult to predict. This can be explained by the regularization, which is useful when the training samples are insufficient to learn models without overfitting. For AUs with high scores without regularization, the training samples were sufficient

ACCEPTED MANUSCRIPT

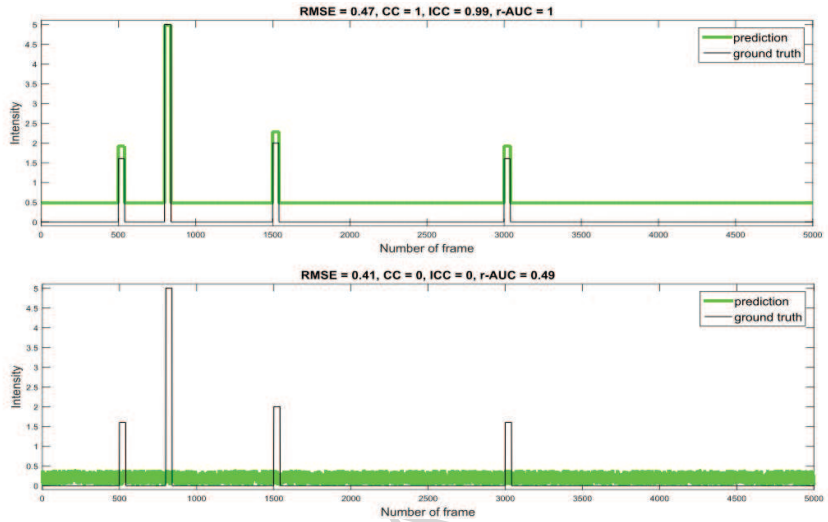

Figure 6: Comparison of different evaluation metrics on synthetic data, first example : r-AUC, ICC and CC metrics did succeed to discriminate a good AU intensity predictor (the upper figure) from a poor one (the lower figure) contrary to RMSE that gives approximately the same value in both cases

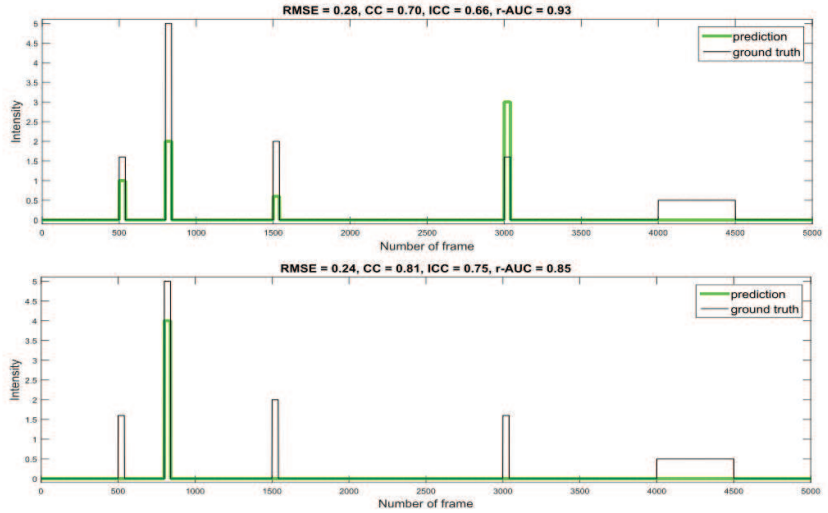

Figure 7: Comparison of different evaluation metrics on synthetic data, second example : the proposed r-AUC metric would evaluate the first prediction more favorably because more activations are detected. On the contrary, RMSE, CC and ICC favor the system that only predicts the strongest peak of intensity

in number and contained enough variability for learning models. In those cases, the gain provided by the Lasso-regularization is less important.

Iterative Feature Selection

In this paragraph, we evaluate the contribution of our iterative feature selection framework. We consider a configuration with conditional entropy-based feature selection and Lasso-MLKR. We present in figure 8 the re-

ACCEPTED MANUSCRIPT

Table 1: Comparison of Pearson's Correlation Coefficient (CC) and Conditional Entropy (C-Ent) for feature selection on Enhanced CK dataset

Evaluation measure	CC (%)		r-AUC (%)	
Feature selection	CC	C-Ent	CC	C-Ent
AU1	82.4	**83.1**	**94.4**	94
AU2	**89.4**	87.7	**96.4**	96.2
AU4	79.4	**80.8**	90.7	**92**
AU5	68.5	**73.1**	91.7	**92.9**
AU6	73.9	**75.2**	93	**94.2**
AU7	66.7	**67.9**	89.6	**89.9**
AU9	**79.7**	74.6	**95.3**	93.8
AU12	88.9	**90**	96.5	**97.2**
AU15	68.7	**69.7**	91	**91.4**
AU17	73.8	**74.7**	**92**	**92**
AU23	50.7	**53.8**	88.8	**90.7**
AU24	50.4	**53.2**	80.2	**82.9**
AU25	79.5	**85.8**	95.5	**95.6**
AU27	92.8	**92.9**	99.2	**99.3**
Mean	74.6	**75.9**	92.4	**93**

Table 2: Comparison of MLKR (M) and Lasso-MLKR (L-M) on Enhanced CK dataset

Evaluation measure	CC (%)		r-AUC (%)	
Algorithm	M	L-M	M	L-M
AU1	83.1	**83.4**	94	**94.1**
AU2	87.7	**88.5**	96.2	**96.5**
AU4	80.8	**84.2**	92	**93.6**
AU5	73.1	**73.8**	92.9	**93.5**
AU6	**75.2**	75	**94.2**	93.9
AU7	67.9	**68.8**	89.9	**90.7**
AU9	74.6	**74.9**	**93.8**	93.5
AU12	90	**91.2**	97.2	**98**
AU15	69.7	**71.3**	91.4	**92.3**
AU17	74.7	**79.7**	92	**94.8**
AU23	53.8	**57**	90.7	**92.3**
AU24	53.2	**60.2**	82.9	**87.9**
AU25	**85.8**	84.5	**95.6**	95.1
AU27	**92.9**	92.6	99.3	**99.4**
Mean	75.9	**77.5**	93	**94**

sults obtained for CC scores averaged over all 14 AUs. For learning the first model, we selected 10 features, and then we added 10 features at each iteration. We can observe that the models learned using our iterative framework lead to a greater CC score at every iteration. In applications where very fast predictions are needed, the number of features must be restricted to save time during kernel computations. The iterative process we propose lets us perform better with the same number of features (for instance, using only 30 features selected in 2 iterations, we see an improvement of 2 % compared with a direct selection of 30 features). This iterative feature selection process leads to a more efficient and compact representation by avoiding the selection of redundant information.

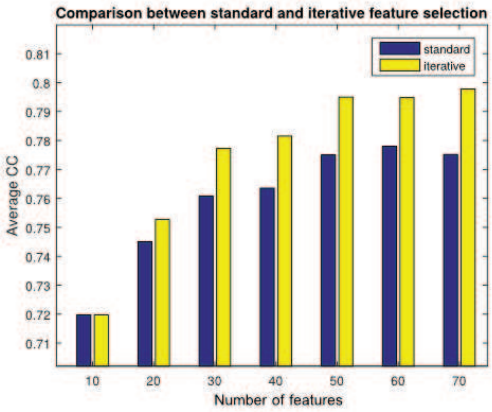

Figure 8: Comparison of standard conditional entropy feature selection and iterative conditional entropy feature selection on Enhanced CK dataset with different number of selected features

The evaluations using the prototypical Enhanced Cohn–Kanade Dataset prove the relevance of the different key points of IRKR: conditional entropy similarity metric, our L^1-regularization of MLKR original formulation and our iterative framework for feature selection.

6.6. Evaluations of DISFA dataset

In this section, we present and compare three versions of IRKR learned with three different sets of features and compare IRKR with two recent state-of-the-art methods on the natural DISFA dataset.

Comparison between different sets of features
In this paragraph, we present the results obtained by learning IRKR with only the geometric features (I1), only the appearance features (I2) and the complete set of geometric and appearance features (I3). We present in table 3 the results in terms of CC and r-AUC. In the DISFA dataset, AUs are labeled on a six-level scale

ACCEPTED MANUSCRIPT

from 0 (no activation of AU) to 5 (activation with maximal intensity). For continuous signals, r-AUC can be calculated as follows:

$$r\text{-}AUC(l,p) = \frac{1}{max(l) - min(l)} \int_{min(l)}^{max(l)} AUC(p, l_s) ds$$

where l_s is the binary quantization of label l using the threshold s. For the six-level labels of DISFA, by considering the following vector of thresholds

$$\mathbf{v} = \{0.5, 1.5, 2.5, 3.5, 4.5\}$$

r-AUC can be simplified as

$$r\text{-}AUC(l,p) = \frac{1}{5} \sum_{i=1}^{5} AUC(p, l_{v_i})$$

which corresponds to an average of 5 AUC scores for the different thresholds (the first one separating the samples of values 0 from the others, the second one separating the samples of values 0 and 1 from the others, and so on).

Table 3: Comparison among three versions of IRKR on the DISFA dataset. I1 corresponds to a model learned using only shape features, I2 to a model learned only with appearance features and I3 to a model learned with both shape and appearance features

Evaluation measure	CC (%)			r-AUC (%)		
Feature set	I1	I2	I3	I1	I2	I3
AU1	57	60	**70**	91	91	**94**
AU2	61	60	**68**	**94**	93	**94**
AU4	54	61	**68**	87	89	**91**
AU5	36	39	**49**	77	96	**98**
AU6	63	58	**65**	93	91	**94**
AU9	40	34	**43**	90	88	**92**
AU12	**83**	75	**83**	**96**	94	**96**
AU15	19	33	**34**	74	77	**83**
AU17	29	26	**35**	79	82	**86**
AU20	10	**25**	21	65	68	**74**
AU25	**87**	78	86	**94**	90	**94**
AU26	55	37	**62**	90	88	**92**
Mean	50	49	**57**	86	87	**91**

We observe a gain of 14% of the average CC score when adding appearance features to geometric ones. We can see that for AU12 (Lip Corner Puller) and AU25 (Lips Part), appearance features did not improve the prediction. Geometric features were sufficient to obtain relatively precise predictions for those AUs. However, for more subtle AUs inducing smaller facial movements, appearance features have considerably improved the predictions for AU5 (Upper Lid Raiser), AU15 (Lip Corner Depressor), AU17 (Chin Raiser) and AU20 (Lip Stretcher).

In the next paragraph, to compare our method, we use the complete version of IRKR (I3), which includes both geometric and appearance features.

Comparison with state-of-the-art methods

IRKR is compared to the method proposed by Sandbach et al. [59] and that proposed by Baltrušaitis et al. [44] in terms of Root Mean Square Error (RMSE) and Correlation Coefficient (CC) in table 4. In [59], the authors used Support Vector Regression on Local Binary Pattern (LBP) features and included priors via Markov Random Fields (MRF). In [44], Continuous Conditional Neural Fields (CCNF) are used after modeling the appearance with Nonnegative Matrix Factorization (NMF) on local patches.

In [59], only the AUs corresponding to the upper face have been predicted. Considering this subset of AUs, the average RMSE of IRKR is 0.58 compared to 0.66 for [59] and 0.71 for [44]. The average CC of IRKR is 60.3 compared to 34.2 for [59] and 46.5 for [44].The statistical significance of these results has also been evaluated using the Friedman test [60]. With the 3 methods, and 6 upper face AUs the p-values are 0.015 and 0.009 for RMSE and CC respectively which highlight the strong significance of the results. If we consider the whole set of 12 AUs and we compare our results with [44], the test is less significant (the p-value is equal to 0.08 for both CC and RMSE). However, if we compare the two means 0.59 vs 0.66 and 56.9 vs 49 for RMSE et CC respectively, this improvement of our proposed method is highly significant (the p-values of the t-tests are 0.036 and 0.015 for RMSE and CC respectively).

We can notice, for AU9, that the RMSE error of [59] is lower than that of IRKR but that the CC score of IRKR is higher. This contradiction illustrates the metric problem we discussed in Section 6.3. We proposed an r-AUC score to overcome this issue. In figure 9, we present the 5 AUC scores we obtain for the different thresholds and the average r-AUC for each AU. For most AUs, we can see that the AUC scores corresponding to the low thresholds are lower than the AUC scores for higher thresholds. This means that the algorithm succeeds more easily at separating high-intensity activations from the others and has more difficulties in separating the non-activated ones from the rest. If we

ACCEPTED MANUSCRIPT

consider that an AU is activated when its intensity is equal to or higher than level 3 (corresponding to the third threshold), 8 of 12 AUs are predicted with an AUC score higher than 0.9 (except AU9, AU15, AU17 and AU20).

In figure 10, we show the prediction of IRKR algorithm on a part of the third sequence of the DISFA dataset for AU4 (Brow Lowerer). We can observe that the algorithm succeeds well at predicting two brow lowering actions in the middle and at the end of the sequence. In those actions, the intensity reaches level 4. For the beginning of the sequence, our algorithm succeeds at differentiating level 3 from non-activation but with a certain amount of noise. The small activation reaching level 2 around frame 1450 is not predicted by our system. This example illustrates the AUC scores of AU4 on figure 9. It is more difficult to differentiate activations of levels 0, 1 and 2, but recognition is relevant from level 3 (AUC score is higher than 0.9 for those thresholds).

The computational complexity is $O(n_s.n_r)$, where n_s is the number of samples, and n_r is the number of selected features. Using 6000 samples of dimension $n_r = 70$, our Matlab implementation of the IRKR algorithm is able to simultaneously predict 14 AUs at a frequency of 16 frames per second on an Intel Core i7-3770 at 3.4 GHz, making it usable in real-time applications.

7. Discussion and conclusion

In this paper, we presented the Iterative Regularized Kernel Regression (IRKR) framework, a generic regression method built upon Metric Learning for Kernel Regression (MLKR) [49]. We applied it to real-time prediction of AU intensity, improving state-of-the-art results with several databases. In this work, we propose an L^1-regularization of the original MLKR formulation to reduce overfitting. We use conditional entropy for selecting features with nonlinear functional relationships with labels. We then perform an iterative framework to avoid selecting redundant information. Finally, we introduce r-AUC, a new evaluation metric for regression in the context of imbalanced data.

We evaluated and compared our method using two AU databases containing multilevel annotations. The first one, the enhanced Cohn–Kanade dataset, is a widely used prototypical database upon which we evaluated the different key points of our method. We compared IRKR with state-of-art methods on the natural DISFA dataset on 12 AUs, leading to mean improvements of 10.3% and 11.6% for Root Mean Squared Error (RMSE) and Correlation Coefficient (CC), respectively.

The most accurate predictions were obtained for AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU12 (Lip Corner Puller) and AU25 (Lips Part), which are frequently activated in natural behavior. Other AUs appear to be more complex to model and predict. This can be explained by the small number of positive samples for some AUs in natural databases. Indeed, some AUs are activated only in particular and rare emotional states, which can be difficult to induce in natural setups when acquiring data. Considering this imbalanced data distribution, it is particularly important to focus on overfitting reduction, which is the purpose of the Lasso-penalty we added to the original cost function of MLKR, which leads to important gains, especially for the complex AUs.

However, the results show that for the Lip Corner Depressor (AU15) and Lip Stretcher (AU20), the number of positive samples may not be sufficient for modeling the activations accurately. The amount of available labeled data still remains as a hindrance to AU intensity prediction, and natural protocols inducing rare facial expressions may be very important for continued increasing accuracy in automatic face-centered human behavior analysis.

The comparison between IRKR learned only with geometric features and IRKR learned with geometric and appearance features stresses the importance of appearance characterization for AU intensity prediction. Recent advances in facial landmark tracking considerably improves AU prediction scores, but research on appearance features remains of great interest in this domain, as underscored by these results. Although the relationship between facial landmarks and AU activations have an important chance of being near linear, the same is not the case for appearance features. Thus, it is important to model those relations nonlinearly. This is why we decided to use the nonlinear conditional entropy metric for selecting features. The obtained results show that using this metric, compared with the Correlation Coefficient, improves predictions of AUs that are linked to appearance characterizations, such as AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU23 (Lip Tightener) and AU24 (Lip Pressor).

We used this metric within an iterative framework for feature selection to avoid selecting redundant information. This leads to a more compact representation, obtaining higher scores with a reduced set of features. This compact representation can be interesting for several reasons. First, reducing the number of parameters

ACCEPTED MANUSCRIPT

Table 4: Comparison between our algorithm and those proposed by Sandbach et al. [59] and Baltrušaitis [44] on the DISFA dataset in terms of Root Mean Square Error (R) and Correlation Coefficient (CC)

AU	R, IRKR	R, [59]	R, [44]	CC, IRKR	CC, [59]	CC, [44]
1	**0.57**	0.63	0.74	**69.7**	56.3	48
2	**0.49**	0.58	0.63	**68.2**	54.1	50
4	**0.85**	1.10	1.13	**67.7**	43.8	52
5	**0.29**	0.30	0.33	**49.2**	22.6	48
6	**0.63**	0.77	0.75	**64.7**	11.9	45
9	0.62	**0.58**	0.67	**42.6**	16.8	36
12	**0.58**	-	0.71	**83.2**	-	70
15	0.47	-	**0.46**	34.2	-	**41**
17	**0.62**	-	0.67	35	-	**39**
20	0.59	-	**0.58**	**21.0**	-	11
25	0.69	-	**0.63**	86.2	-	**89**
26	0.69	-	**0.63**	**61.5**	-	57
Mean	**0.59**	-	0.66	**56.9**	-	49

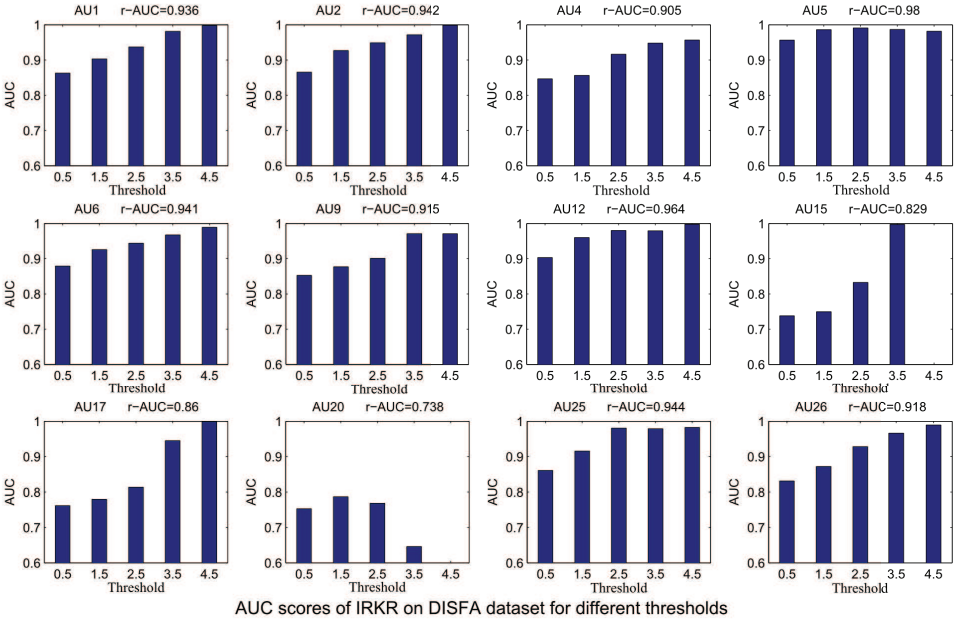

Figure 9: AUC scores of IRKR on the DISFA dataset for different thresholds. There are only 4 thresholds for AU15 and AU20 because the database does not include samples of intensity 5 for those AUs.

in the model can reduce overfitting, and second, compact representations lead to faster predictions, which is useful considering the real-time constraints of many applications related to automatic AU prediction.

Several metrics exist for evaluating a regression method. The most commonly used in AU intensity prediction are Root Mean Square Error (RMSE) and Correlation Coefficient (CC). However, some problems are encountered when using those metrics on imbalanced data. To solve those problems, we propose r-AUC, an

ACCEPTED MANUSCRIPT

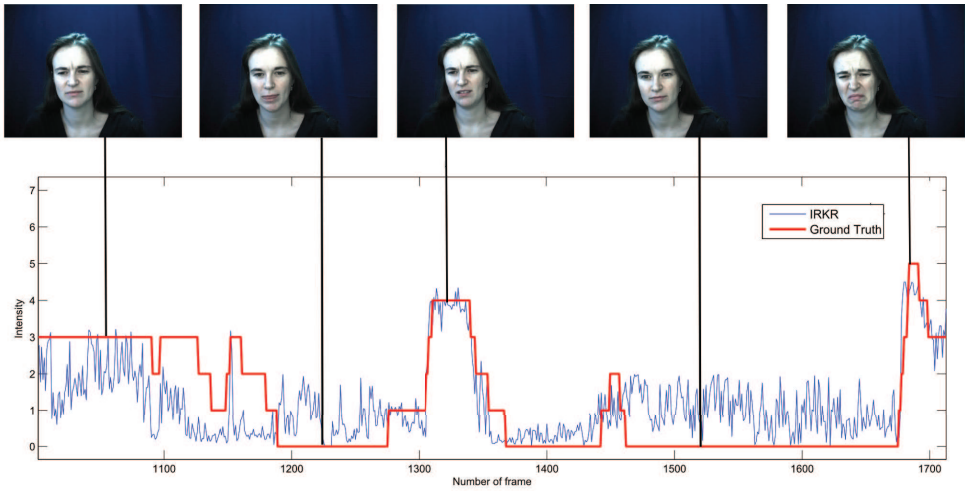

Figure 10: Prediction of AU4 on a part of sequence 3 of the DISFA dataset

adaptation of Area Under ROC Curve (AUC) that is suited for regression problems in an imbalanced context.

The results obtained on the natural DISFA dataset are very promising, especially for the most frequently activated facial muscles. However, AU intensity prediction is a particularly difficult task, and many improvements could still be made–for instance, by proposing multi-task methods including other information such as age or head pose, which play a crucial role in facial appearance deformations, or by thinking about new database acquisition protocols to naturally induce rare facial expressions.

8. Acknowledgements

This work was partially supported by the French National Agency (ANR) in the frame of its Technological Research CONTINT program (JEMImE, project number ANR-13-CORD-0004)

[1] P. Ekman, W. V. Friesen, Measuring facial movement, Environmental Psychology and Nonverbal Behavior 1 (1) (1976) 56–75.

[2] R. El Kaliouby, P. Robinson, Real-time inference of complex mental states from facial expressions and head gestures, in: Real-time vision for human-computer interaction, Springer, 2005, pp. 181–200.

[3] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, F. De la Torre, Detecting depression from facial actions and vocal prosody, in: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, IEEE, 2009, pp. 1–7.

[4] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, P. E. Solomon, The painful face–pain expression recognition using active appearance models, Image and Vision Computing 27 (12) (2009) 1788–1796.

[5] S. Kaltwang, O. Rudovic, M. Pantic, Continuous pain intensity estimation from facial expressions, in: Advances in Visual Computing, Springer, 2012, pp. 368–377.

[6] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, R. Picard, Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected" in-the-wild", in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, IEEE, 2013, pp. 881–888.

[7] B. Zaman, T. Shrimpton-Smith, The facereader: Measuring instant fun of use, in: Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, ACM, 2006, pp. 457–460.

[8] T. Pfister, X. Li, G. Zhao, M. Pietikainen, Recognising spontaneous facial micro-expressions, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1449–1456.

[9] M. E. Hoque, D. J. McDuff, R. W. Picard, Exploring temporal patterns in classifying frustrated and delighted smiles, Affective Computing, IEEE Transactions on 3 (3) (2012) 323–334.

[10] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, Pattern Analysis and Machine Intelligence, IEEE Transactions on 29 (10) (2007) 1683–1699.

[11] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, J. F. Cohn, Disfa: A spontaneous facial action intensity database, IEEE Transaction on Affective Computing 4 (2) (2013) 151–160.

[12] Y. Sun, M. Reale, L. Yin, Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition, in: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, IEEE, 2008, pp. 1–8.

[13] A. Savran, B. Sankur, M. Taha Bilge, Regression-based intensity estimation of facial action units, Image and Vision Computing 30 (10) (2012) 774–784.

[14] Y. Li, J. Chen, Y. Zhao, Q. Ji, Data-free prior model for facial action unit recognition, Affective Computing, IEEE Transactions on 4 (2) (2013) 127–141.

ACCEPTED MANUSCRIPT

[15] S. Wan, J. Aggarwal, Spontaneous facial expression recognition: A robust metric learning approach, Pattern Recognition.

[16] L. A. Jeni, J. M. Girard, J. F. Cohn, F. De La Torre, Continuous au intensity estimation using localized, sparse facial feature space, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–7.

[17] W.-S. Chu, F. De la Torre, J. F. Cohn, Selective transfer machine for personalized facial action unit detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[18] P. Yang, Q. Liu, D. N. Metaxas, Boosting coded dynamic features for facial action units and facial expression recognition, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–6.

[19] C.-F. Chuang, F. Y. Shih, Recognizing facial action units using independent component analysis and support vector machine, Pattern recognition 39 (9) (2006) 1795–1798.

[20] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, Vol. 1, IEEE, 2001, pp. I–511.

[21] Y. Freund, R. E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: Computational learning theory, Springer, 1995, pp. 23–37.

[22] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, L. Prevost, Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 860–865.

[23] O. Rudovic, V. Pavlovic, M. Pantic, Kernel conditional ordinal random fields for temporal segmentation of facial action units, in: Computer Vision–ECCV 2012. Workshops and Demonstrations, Springer, 2012, pp. 260–269.

[24] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image and Vision Computing 24 (6) (2006) 615–625.

[25] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, 2000, pp. 46–53.

[26] T. Sim, S. Baker, M. Bsat, The cmu pose, illumination, and expression database, Pattern Analysis and Machine Intelligence, IEEE Transactions on 25 (12) (2003) 1615–1618.

[27] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, K. Scherer, The first facial expression recognition and analysis challenge, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 921–926.

[28] M. F. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42 (1) (2012) 28–43.

[29] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image and Vision Computing 24 (6) (2006) 615–625.

[30] S. Koelstra, M. Pantic, Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics, in: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, IEEE, 2008, pp. 1–8.

[31] W.-Y. Chang, C.-S. Chen, Y.-P. Hung, Analyzing facial expression by fusing manifolds, in: Computer Vision–ACCV 2007, Springer, 2007, pp. 621–630.

[32] J. Hamm, C. G. Kohler, R. C. Gur, R. Verma, Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders, Journal of neuroscience methods 200 (2) (2011) 237–256.

[33] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, L. Akarun, Bosphorus database for 3d face analysis, in: Biometrics and Identity Management, Springer, 2008, pp. 47–56.

[34] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010, pp. 94–101.

[35] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, I. Matthews, Painful data: The unbc-mcmaster shoulder pain expression archive database, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 57–64.

[36] M. H. Mahoor, S. Cadavid, D. S. Messinger, J. F. Cohn, A framework for automated measurement of the intensity of non-posed facial action units, in: Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, IEEE, 2009, pp. 74–80.

[37] Z. Ming, A. Bugeau, J.-L. Rouas, T. Shochi, Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, Vol. 6, IEEE, 2015, pp. 1–6.

[38] S. Kaltwang, O. Rudovic, M. Pantic, Continuous pain intensity estimation from facial expressions, in: Advances in Visual Computing, Springer, 2012, pp. 368–377.

[39] S. Kaltwang, S. Todorovic, M. Pantic, Latent trees for estimating intensity of facial action units, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 296–304.

[40] A. Gudi, H. E. Tasli, T. M. den Uyl, A. Maroulis, Deep learning based facs action unit occurrence and intensity estimation, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, Vol. 6, IEEE, 2015, pp. 1–5.

[41] O. Rudovic, V. Pavlovic, M. Pantic, Context-sensitive dynamic ordinal regression for intensity estimation of facial action units, Pattern Analysis and Machine Intelligence, IEEE Transactions on 37 (5) (2015) 944–958.

[42] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, L. Prevost, Facial action recognition combining heterogeneous features via multikernel learning, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42 (4) (2012) 993–1005.

[43] J. M. Girard, J. F. Cohn, F. De la Torre, Estimating smile intensity: A better way, Pattern Recognition Letters.

[44] T. Baltrušaitis, P. Robinson, L.-P. Morency, Continuous conditional neural fields for structured regression, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 593–608.

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

[46] H. Takeda, S. Farsiu, P. Milanfar, Deblurring using regularized locally adaptive kernel regression, Image Processing, IEEE Transactions on 17 (4) (2008) 550–563.

[47] M. Schaap, L. Neefjes, C. Metz, A. van der Giessen, A. Weustink, N. Mollet, J. Wentzel, T. van Walsum, W. Niessen, Coronary lumen segmentation using graph cuts and robust kernel regression, in: Information Processing in Medical Imaging, Springer, 2009, pp. 528–539.

[48] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, M. Chetouani, Robust

ACCEPTED MANUSCRIPT

continuous prediction of human emotions using multiscale dynamic cues, in: Proceedings of the 14th ACM international conference on Multimodal interaction, ACM, 2012, pp. 501–508.

[49] K. Q. Weinberger, G. Tesauro, Metric learning for kernel regression, in: International Conference on Artificial Intelligence and Statistics, 2007, pp. 612–619.

[50] E. A. Nadaraya, On estimating regression, Theory of Probability & Its Applications 9 (1) (1964) 141–142.

[51] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: ICML, Vol. 3, 2003, pp. 856–863.

[52] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[53] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research 3 (2003) 1157–1182.

[54] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 532–539.

[55] J. Nicolle, K. Bailly, V. Rapp, M. Chetouani, Locating facial landmarks with binary map cross-correlations., in: ICIP, 2013, pp. 2978–2982.

[56] G. Murthy, R. Jadon, Effectiveness of eigenspaces for facial expressions recognition, International Journal of Computer Theory and Engineering 1 (5) (2009) 1793–8201.

[57] P. E. Shrout, J. L. Fleiss, Intraclass correlations: uses in assessing rater reliability., Psychological bulletin 86 (2) (1979) 420.

[58] L. A. Jeni, J. F. Cohn, F. De La Torre, Facing imbalanced data–recommendations for the use of performance metrics, in: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE, 2013, pp. 245–251.

[59] G. Sandbach, S. Zafeiriou, M. Pantic, Markov random field structures for facial action unit intensity estimation, in: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, IEEE, 2013, pp. 738–745.

[60] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research 7 (2006) 1–30.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Highlights

1) We present a framework for real-time Action Unit intensity prediction.
2) We introduce a Lasso-regularized version of Metric Learning for Kernel Regression.
3) We propose a new evaluation metric (r-AUC) designed for regression tasks.

ACCEPTED MANUSCRIPT