Facial Action Unit Intensity Prediction via Hard Multi-Task Metric Learning for Kernel Regression

Jeremie Nicolle, Kevin Bailly and Mohamed Chetouani
Univ. Pierre & Marie Curie
ISIR - CNRS UMR 7222
F-75005, Paris - France

***Abstract*— The problem of learning several related tasks has recently been addressed with success by the so-called multi-task formulation, that discovers underlying common structure between tasks. Metric Learning for Kernel Regression (MLKR) aims at finding the optimal linear subspace for reducing the squared error of a Nadaraya-Watson estimator. In this paper, we propose two Multi-Task extensions of MLKR. The first one is a direct application of multi-task formulation to MLKR algorithm and the second one, the so-called Hard-MT-MLKR, lets us learn same-complexity predictors with fewer parameters, reducing overfitting issues. We apply the proposed method to Action Unit (AU) intensity prediction as a response to the Facial Expression Recognition and Analysis challenge (FERA'15). Our system improves the baseline results on the test set by 24% in terms of Intraclass Correlation Coefficient (ICC).**

I. INTRODUCTION

Automatic understanding of human behavior is a key element in many domains such as social robotics. Analyzing facial expressions has been proven useful for inferring human mental states [1] and has recently been very rapidly evolving. In 1976, Ekman proposed the Facial Action Coding System (FACS [2]), characterizing activations of face muscles (AUs) for describing facial expressions. In order to be able to compare methods for AUs detection, the first Facial Expression Recognition and Analysis Challenge has been organized in 2011 [3]. However, being able to predict AUs more precisely by inferring their intensity levels can improve facial expression description, and as a consequence more high-level tasks relative to human behavior understanding. In this context, the second Facial Expression Recognition and Analysis Challenge (FERA'15 [4]) proposed a sub-challenge for AUs intensity prediction. It is based on BP4D database [5] and concerns five AUs (AU6, AU10, AU12, AU14 and AU17). This paper presents our response to this sub-challenge.

The AUs to be predicted in this challenge are all linked to the lower part of the face, all related to mouth movements. Thus, some features are potentially relevant for all five tasks corresponding to separate predictions of each AU. In order to respond to such learning problematic, Evgeniou and Pontil [6] proposed an extension of Support Vector Machines for multi-task learning, which aims at learning common representation for several related classification tasks. In [7], Zhang et al. show that multi-task extension of Multi-Kernel Support Vector Machines improves the detection of AUs.

In order to predict the intensities of AUs, we chose to use the Nadaraya-Watson estimator, which is a non-parametric regressor able to adapt efficiently to heterogeneous point distribution. In [8], Weinberger et al. proposed MLKR, a metric learning algorithm that directly optimizes the error of a Gaussian kernel Nadaraya-Watson regressor on the training data.

In this context, we propose two multi-task extensions of MLKR for AUs intensity prediction. The first one is a direct application of multi-task regularization for MLKR, letting us discover an underlying common representation between tasks and the second one, the so-called Hard-MT-MLKR, lets us force learned subspaces to share common axis, resulting in less parameters to learn and a reduction of overfitting.

In section II, we present related works (MLKR, multi-task SVM). In section III, we introduce our two multi-task extensions of MLKR. In section IV, we present the application of Hard-MT-MLKR to AUs intensity prediction, before concluding in section V.

II. PREVIOUS WORKS

In this section, we present MLKR and multi-task SVM extension, on which our work is built upon.

A. MLKR

In kernel regression, a test label is predicted using a Nadaraya-Watson estimator [9], as an average of the training labels weighted by a similarity measure between test sample and training samples. Considering n_s training samples $\{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_{n_s}\}$ and their corresponding labels $\{y_1, y_2, .., y_{n_s}\}$, the label associated to a feature vector $\mathbf{x}_t$ is approximated using:

$$\hat{y_t} = \frac{\sum_{i=1}^{n_s} y_i k_{i,t}}{\sum_{i=1}^{n_s} k_{i,t}} \quad (1)$$

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

Kernel regression has proven its efficiency in a large spectrum of applications (image deblurring [10], segmentation [11], automatic human emotion prediction [12]). However, the space in which the samples lie has an important impact on the prediction performance, making dimensionality reduction a relevant initial step. Weinberger and Tesauro [8] proposed MLKR, whose goal is to find the optimal linear subspace for

minimizing Nadaraya-Watson squared error on the training set for the commonly used Gaussian kernel, defined as follows:

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

with σ the Gaussian spread and $d_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$ the euclidean distance between samples i and j. Considering an original space of dimension n_d and an output space of dimension n_r, MLKR estimates the projection matrix $\mathbf{A} \in \mathcal{M}_{n_r,n_d}(\mathbb{R})$ that minimizes the squared error of the training samples, defined as:

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

where

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

with

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

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

being the squared distance in the reduced subspace of dimension n_r. The optimization of the squared error is done with Rasmussen's implementation of Polak-Ribiere Conjugate Gradients method [1], with:

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

In this paragraph, we discuss a few positive and negative aspects of MLKR.

First, MLKR does not directly learn a prediction function but learns a space in which a Nadaraya-Watson estimator gives successful prediction for a set of data and labels. As a consequence, one can easily identify most important features by studying space axis. Moreover, new data points can be projected in the learned space and the prediction can be performed without re-training the system (*e.g.* in order to adapt to a new database, to a specific subject, or for online updates).

Second, on the contrary to Gaussian kernel SVR, it is unnecessary to optimize the spread of the Gaussian kernel in cross-validation during learning. In MLKR, the Gaussian spread is fixed *a priori* and the learned space scales according to it.

Finally, the Nadaraya-Watson estimator used as the prediction function in MLKR is able to adapt efficiently to heterogeneous data point distribution because of the normalization by $\sum\limits_{i=1}^{n_s} k_{i,t}$ which can be very important for AUs intensity prediction when trying to analyze an unknown subject lying in a sparsely populated part of the space. When using a Gaussian kernel SVR, a data point localized far away from support vectors will tend to be predicted as zero. When using the Nadaraya-Watson estimator, it will be predicted approximately as a mean of closest training sample labels, which can be considered as a more relevant extrapolation.

However, MLKR has several drawbacks. First, it is non convex but has been shown to converge towards interesting local minima experimentally. Second, it has an important complexity with respect to the original space dimension, which makes it difficult to apply on data with numerous features. In our application to AUs intensity prediction, we preselect features before learning (details can be found in Section IV-B).

[1] http://learning.eng.cam.ac.uk/carl/code/minimize/

B. MT-SVM

Learning a task with an insufficient amount of labeled data can result in overfitting issues. Reducing the number of learned parameters or including *a priori* knowledge can reduce this effect. Thus, including the *a priori* knowledge that different tasks have a great probability of sharing a common representation because they are related together can result in a more efficient learning.

In [6], Evgeniou and Pontil proposed an extension of Support Vector Machines for multi-task learning. The goal is to discover common representation between different related tasks. Considering T tasks, the algorithm aims at learning T classifiers $\{\mathbf{w}_t, t \in [\![1;T]\!]\}$ one for each task. A common representation is introduced between tasks by decomposing each hyperplane into:

$$\mathbf{w}_t = \mathbf{v}_0 + \mathbf{v}_t$$

A test label y_i for task t corresponding to a feature vector $\mathbf{x}_i$ is then classified using:

$$\hat{y}_i = sign(\mathbf{x}_i^\top (\mathbf{v}_0 + \mathbf{v}_t))$$

A regularization term is then added to the standard SVM cost function in order to encourage $\mathbf{v}_0$ to contain a representation that is shared by the different tasks and each vector $\mathbf{v}_t$ to contain a particular representation for task t. The minimization problem becomes:

$$\min_{\mathbf{v}_0,\ldots\mathbf{v}_T} \sum_{t=0}^{T} \gamma_t \|\mathbf{v}_t\|^2 + \sum_{t=1}^{T} \sum_i [1 - y_i(\mathbf{v}_0 + \mathbf{v}_t)^\top \mathbf{x}_i]_+$$

with $[a]_+ = max(a, 0)$ and γ_t controlling multi-task penalties.

III. MULTI-TASK EXTENSIONS OF MLKR

In this section, we propose two extensions of MLKR for multi-task learning. The first one is a direct application of multi-task regularization for MLKR and the second one is a modification we propose that enforces common features between tasks in a hard manner, letting us reduce overfitting by learning fewer parameters for same dimensionality projection spaces.

A. MT-MLKR

We recall that $\mathbf{A}$ in eq. 3 is the projection matrix defining the linear transformation used for passing from the original space to the reduced space. If $\mathbf{X}$ is a matrix containing data points in the original space, data points in the reduced space become:

$$\mathbf{X}_r = \mathbf{A}\mathbf{X}$$

In the multi-task extension, we aim at learning T projection spaces $\mathbf{A}_t$, one for each task. We introduce common representation between tasks by decomposing spaces into:

$$\mathbf{A}_t = \mathbf{B}_0 + \mathbf{B}_t$$

Let $\mathcal{L}_t$ be the error associated to task t. The cost function of our multi-task extension of MLKR is defined as follows:

$$\mathcal{L}_{MT} = \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{B}_0 + \mathbf{B}_t) + \gamma \sum_{t=0}^{T} \|\mathbf{B}_t\|^2$$

with $\|.\|$ the Frobenius norm. To simplify derivative expressions, let

$$\mathbf{D}_t = \sum_i \frac{(\hat{y_i} - y_i)}{\sum\limits_{j \neq i} k_{ij}} \sum_j (\hat{y_i} - y_j) k_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$$

associated with task t labels and estimators. We obtain:

$$\frac{\partial \mathcal{L}_{MT}}{\partial \mathbf{B}_0} = 4 \sum_{t=1}^{T} (\mathbf{B}_0 + \mathbf{B}_t)\mathbf{D}_t + 2\gamma \mathbf{B}_0$$

and

$$\frac{\partial \mathcal{L}_{MT}}{\partial \mathbf{B}_t} = 4(\mathbf{B}_0 + \mathbf{B}_t)\mathbf{D}_t + 2\gamma \mathbf{B}_t$$

B. Hard-MT-MLKR

By decomposing each task prediction parameters as a sum of common and specific ones and by adding an adequate regularization, the multi-task formulation presented in the previous paragraph lets us find an underlying common space. We propose in this paragraph a more constrained multi-task formulation of MLKR replacing the sum by a concatenation in the subspaces decomposition. Thus, we force a certain number of axis to be shared by the different spaces. Let n_c be the number of common axis and n_r be the projection space dimension. The matrices $\mathbf{A}_t$ can be defined as follows:

$$\mathbf{A}_t = \left[\frac{\mathbf{B}_0}{\mathbf{B}_t} \right]$$

with $\mathbf{B}_0 \in \mathcal{M}_{n_c, n_d}(\mathbb{R})$ and $\mathbf{B}_t \in \mathcal{M}_{n_r - n_c, n_d}(\mathbb{R})$

The derivatives become:

$$\frac{\partial \mathcal{L}_{MT}}{\partial \mathbf{B}_0} = 4\mathbf{B}_0 \sum_{t=1}^{T} \mathbf{D}_t + 2\gamma \mathbf{B}_0$$

and

$$\frac{\partial \mathcal{L}_{MT}}{\partial \mathbf{B}_t} = 4\mathbf{B}_t \mathbf{D}_t + 2\gamma \mathbf{B}_t$$

The proposed multi-task regularization lets us reduce overfitting by learning fewer parameters for same dimensionality projection spaces. In MT-MLKR, the number of learned parameters is

$$n_{par}^{MT} = n_d.n_r.(T+1)$$

while in Hard-MT-MLKR, it is

$$n_{par}^{HMT} = n_d.(n_c + T.(n_r - n_c))$$

Example: for a number of common axis $n_c = 3$ and subspaces dimension $n_r = 5$ with $n_d = 80$ features in the original space, we obtain, for 5 tasks, $n_{par}^{MT} = 2400$ and $n_{par}^{HMT} = 1040$ parameters to optimize. We are able to learn same complexity predictors with less than half parameters, resulting in a reduction of overfitting.

IV. APPLICATION TO ACTION UNIT INTENSITY PREDICTION

In this section, we present the application of Hard-MT-MLKR to AUs intensity prediction. First, we present the features we extracted, followed by the selection method we used. Then, we introduced BP4D, the database used in the FERA'15 contest, before presenting an analysis of the main features used for predicting each task and obtained results.

A. Feature extraction

Most methods in facial analysis combine both shape-based features and appearance-based ones. Shape-based features contain information relative to facial landmark locations (centers and contours of the eyes, the nose, the eyebrows and the mouth), and appearance-based ones aim at describing texture. For AUs intensity prediction, landmark locations seem particularly relevant because some AUs activations directly induce key point movements, for instance when rising the eyebrows or smiling. However, it is important to combine shape-related features with appearance-related ones for several reasons. First, shape-based features cannot characterize expression-relative wrinkles. Second, appearance can make up for potential errors of current landmark trackers in challenging conditions.

Shape-based features

We first localize 49 facial landmarks using Supervised Descent Method (SDM [13]). For being insensitive to scaling and rotation in the image plane, the features we extract are relative to point triplets (as in [14]). For each triplet of points $\mathbf{t}_{k_1 k_2 k_3} = (\mathbf{p}_{k_1}, \mathbf{p}_{k_2}, \mathbf{p}_{k_3})$ we calculate the ratio of both vectors

$$\mathbf{v}_{k_2 k_3} = \mathbf{p}_{k_3} - \mathbf{p}_{k_2} = (\mathbf{p}_{k_3}^x - \mathbf{p}_{k_2}^x) + i.(\mathbf{p}_{k_3}^y - \mathbf{p}_{k_2}^y)$$

and

$$\mathbf{v}_{k_2 k_1} = \mathbf{p}_{k_1} - \mathbf{p}_{k_2} = (\mathbf{p}_{k_1}^x - \mathbf{p}_{k_2}^x) + i.(\mathbf{p}_{k_1}^y - \mathbf{p}_{k_2}^y)$$

to form

$$f(\mathbf{t}_{k_1 k_2 k_3}) = \frac{\mathbf{v}_{k_2 k_1}}{\mathbf{v}_{k_2 k_3}} = \frac{\|\mathbf{v}_{k_2 k_1}\|}{\|\mathbf{v}_{k_2 k_3}\|} . e^{i(\widehat{\mathbf{v}_{k_2 k_3}, \mathbf{v}_{k_2 k_1}})}$$

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 norm and angle of $f(\mathbf{t}_{k_1 k_2 k_3})$ as features.

Appearance-based features

The first step we perform is to cancel the rotation in the image plane and normalize the image to a 128x128 pixels image using eye locations (in most landmark tracking systems, those points appear to be the most reliable ones). Then, we extract Histograms of Oriented Gradients descriptors (HOG [15]) on different image patches. Some of them are centered on and around the landmarks (for describing local texture and be able to capture expression-related wrinkles), and others are obtained by a 8x8 division of the image (see figure 1), letting us the possibility to catch up for potential point tracking errors. The centers of the patches defined using the landmarks we chose are presented figure 2. The size of those patches is the same as those obtained by the 8x8 regular division of the image.

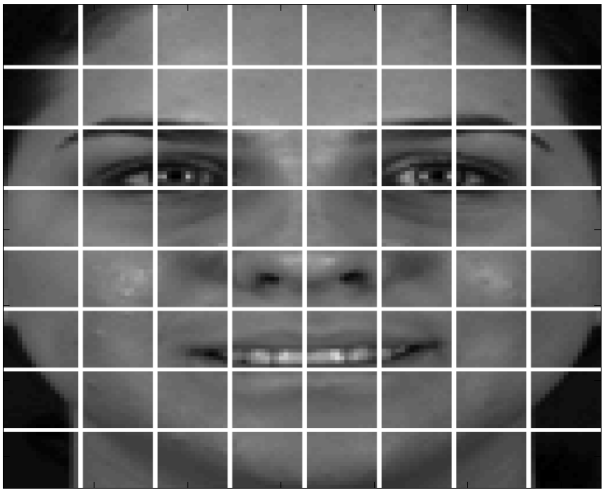

Fig. 1. Patches located without the landmarks

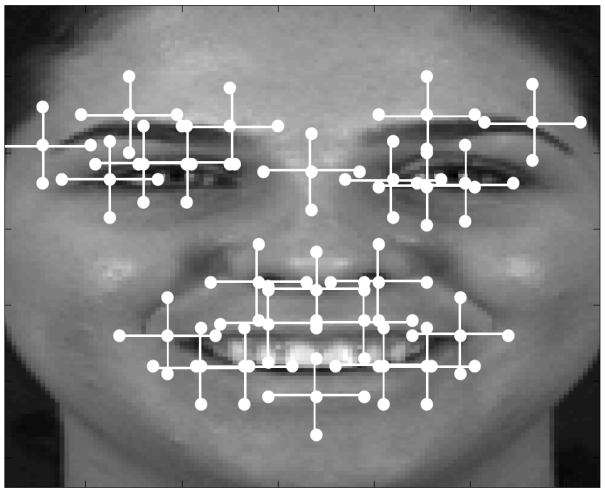

Fig. 2. Centers of the patches defined using the landmarks

B. Feature selection

In MLKR, the gradient computation for a projection of n_s samples lying in a space of dimension n_d has a complexity in $O(n_s^2.n_d^2)$ making it difficult to use in high dimension spaces. As an initial step, we perform supervised dimensionality reduction by filter feature selection [16], which aims at characterizing the relevance of features independently of the predictor's choice, often one by one, for predicting the label. In other words, it means to compute a similarity (or dissimilarity) measure between each feature and the label and select the highest (or respectively the smallest) ones. We chose to use a nonlinear metric for feature selection because of the nonlinearity of the Nadaraya-Watson estimator. We use conditional entropy for selecting relevant features, which quantifies the prediction performance of label y considering that a feature f_i is known.

C. Database

The BP4D-Original dataset includes digital video of 41 participants (56.1% female, 49.1% white, ages 18-29). To elicit target emotional expressions and conversational behavior, different tasks were administered by an experimenter who was a professional actor/director of performing arts. The procedures were designed to elicit a range of emotions and facial expressions that include happiness/amusement, sadness, surprise/startle, embarrassment, fear/nervous, physical pain, anger/upset and disgust.

AUs were annotated by a team of experts, frame-by-frame for the intensity of a subset of the AUs. Five AUs were annotated: AU6 (Cheek Raiser), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller), AU14 (Dimpler) and AU17 (Chin Raiser).

The data has been divided for the challenge into training and development sets, each containing approximately half of the participants, male and female mixed up.

D. Experimental setup

In this paragraph, we present the experimental setup and the hyper-parameters used for learning the different systems whose results are presented in Section IV-F. The optimization of those parameters has been performed on a subject independent four-folds cross-validation on the concatenation of training and development datasets. Our HOG calculations are performed in 8 directions. We obtain a total of 2768 features. We found that selecting $n_d = 80$ features for each experiment using conditional entropy was an adequate choice. For multi-task formulations, the selection is done using the sum of conditional entropies over the different tasks. We standardized each feature before learning the optimal linear subspace. We randomly select 10.000 images because of the memory cost of the kernel calculations needed in MLKR. The dimension of the projection spaces we learn is $n_r = 5$. For Hard-MT-MLKR, the number of common axis is $n_c = 3$. The optimal hyper-parameter we found for regularization is $\gamma = 0.9$.

E. Analysis of feature impact

In this paragraph, we discuss the impact of features on the different tasks by analyzing learned subspaces. To do this analyze, we learned a system on the concatenation of training

and development sets. In figure 3, we represented the main features for the common subspace $\mathbf{B}_0$, as well as for the 5 particular spaces $\mathbf{B}_t$. White lines between fiducial points indicate that the angle between corresponding vectors had important impact. Black arrows indicate that HOG extracted in the area along the indicated direction had important impact. We can observe that the angle between the extremity of right eye, the center of inferior lip and the right mouth corner is an important feature for all tasks, which can be explained by the fact that this angle varies a lot when AU12, AU10 or AU17 are activated. Appearance around nasolabial folds appears to be important for all tasks too. As for specific spaces, we can notice appearance between right ear and right cheek for AU6 (cheek raiser), or external to the right mouth corner for AU14 (Dimpler) that both seem relevant. We can also observe that the eyebrows seem to be useful for predicting AU06, AU10 and AU14. This can be explained by correlations between AUs in natural facial expressions (for instance, it seem rare to frown eyebrows when raising cheeks).

In next Section, we present the gain brought by the different multi-task MLKR extensions we proposed and compare our results to the FERA'15 baseline system.

F. Results

In table I, we evaluate the performance of the different systems (MLKR, MT-MLKR and Hard-MT-MLKR) in a subject independent four-folds cross-validation on the concatenation of the training and development datasets of FERA'15 competition. We observe improvements when using multi-task formulation of MLKR for four of the five considered AUs. For AU12, which is the most accurately predicted, the results seem unchanged when using multi-task regularization. Hard MT-MLKR, the more constrained formulation we introduced, significantly improves results for four of the five considered AUs with a mean improvement of 5% over classical MLKR formulation. The highest improvement (more than 15%) is for AU14, that appears to be the hardest to predict for our system.

TABLE I

COMPARISON BETWEEN ORIGINAL MLKR AND THE TWO PROPOSED MULTI-TASK EXTENSIONS OF MLKR IN TERMS OF PEARSON'S CORRELATION COEFFICIENT

AU	MLKR	MT-MLKR	Hard-MT-MLKR
6	74.2	75.4	**76.3**
10	70.9	72.8	**75.2**
12	**86.6**	86.4	86.5
14	41.4	44.3	**47.7**
17	52.6	52.7	**54.8**
Mean	65.1	66.3	**68.1**

In table II, we compare the results we obtain to the baseline results of FERA'15 AUs intensity prediction subchallenge. The baseline geometric and appearance features are detailed in the baseline paper [4]. The machine learning system used for the baseline is a Support Vector Regressor (SVR). The evaluation is done by learning on the training dataset and testing on the development dataset. Results are given in Pearson's correlation coefficient. In the baseline paper, the results are presented separately for geometric and appearance features. For being able to compare our system to the baseline, we present results of three Hard-MT-MLKR systems learned with only geometric (G), only appearance (A), and both geometric and appearance (F) features. We can observe a mean improvement of 14% with our system.

TABLE II

COMPARISON BETWEEN BASELINE SYSTEM (B) AND PROPOSED HARD-MT-MLKR (H) IN TERMS OF PEARSON'S CORRELATION COEFFICIENT FOR DIFFERENT FEATURE SUBSET (ONLY GEOMETRIC (G) ONLY APPEARANCE (A) AND FUSION OF BOTH (F))

AU	B, G	H, G	B, A	H, A	H, F
6	69.9	**74**	72	**78.2**	**78.2**
10	71.5	**72.3**	68.3	**70.4**	**75.4**
12	70.6	**85.2**	69.5	**84.8**	**85.6**
14	47.2	**50**	39.6	**47.4**	**54.2**
17	36.5	**56.2**	30.3	**58**	**60.6**
Mean	59.2	**67.5**	55.9	**67.8**	**70.8**

In table III, we compare the results we obtain on the test partition to the baseline results in terms of Intraclass Correlation Coefficient (ICC) and Mean Squared Error (MSE).

TABLE III

COMPARISON BETWEEN GEOMETRIC AND APPEARANCE BASELINE SYSTEMS (G AND A) AND PROPOSED HARD-MT-MLKR (H) IN TERMS OF ICC (I) AND MSE (M)

AU	G, I	A, I	H, I	G, M	A, M	H, M
6	0.67	0.622	**0.787**	1.004	1.366	**0.829**
10	0.732	0.656	**0.801**	0.897	1.209	**0.801**
12	0.78	0.767	**0.86**	0.738	1.092	**0.622**
14	0.586	0.389	**0.71**	1.227	1.526	**1.14**
17	0.144	0.168	**0.443**	**0.806**	0.819	0.844
Mean	0.582	0.52	**0.72**	0.934	1.202	**0.847**

V. CONCLUSION AND FUTURE WORK

In this paper, we presented Hard Multi-Task regularization for MLKR, that lets us introduce a more constrained common representation between tasks compared to standard multi-task regularization. By replacing the sum by a concatenation in the subspaces decomposition, we are able to learn same dimensionality subspaces with less parameters. We applied the proposed method for AUs intensity prediction as a response to the FERA'15 challenge, evaluating the gain brought by the new regularization. Our system performance is largely higher than the baseline results obtained with SVR. For future work, we aim at extending and evaluating the relevance of the proposed method for transfer learning. Regarding AUs intensity prediction, we can observe that results for AU14 and AU17 are far from being accurate. Lots of work on database acquisition protocols and machine learning systems have to be done for being able to exploit precise facial expression automatic description in real-world applications (where various head poses occur and large morphological differences are present).

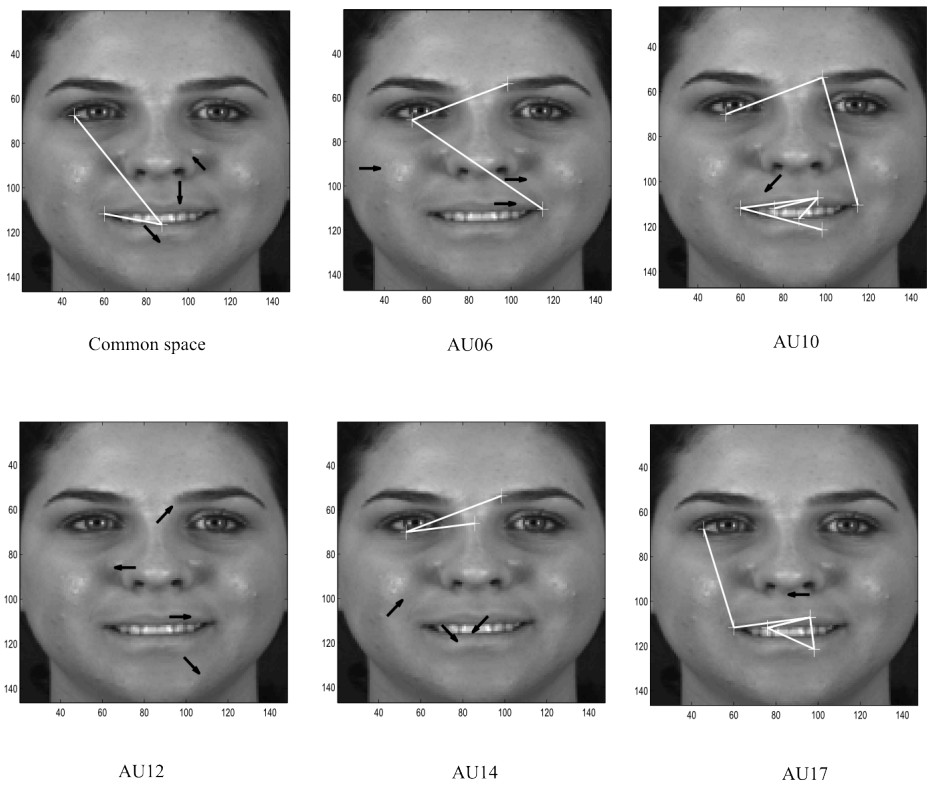

Fig. 3. Analyze of feature impact for AUs intensity prediction. The four principal features are represented for each learned subspace. White lines indicate point triplet angles and black arrows indicate HOG features.

VI. AKNOWLEDGMENTS

This work has been partially supported by the French National Agency (ANR) in the frame of its Technological Research CONTINT program (JEMImE, project number ANR-13-CORD-0004). This work was also performed within the Labex SMART supported by French state funds managed by the ANR within the Investissements dAvenir program under reference ANR-11-IDEX-0004-02.

REFERENCES

[1] R. El Kaliouby and 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.

[2] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, vol. 1, no. 1, pp. 56–75, 1976.

[3] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and 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.

[4] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn, "Fera 2015 - second facial expression recognition and analysis challenge," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, 2015.

[5] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[6] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 109–117.

[7] X. Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn, "A l p-norm mtmkl framework for simultaneous detection of multiple facial action units," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*. IEEE, 2014, pp. 1104–1111.

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

[9] E. A. Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

[10] H. Takeda, S. Farsiu, and P. Milanfar, "Deblurring using regularized locally adaptive kernel regression," *Image Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 550–563, 2008.

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

[12] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust 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.

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

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

[15] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *International Workshop on Automatic Face and Gesture Recognition*, vol. 12, 1995, pp. 296–301.

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