Regularized Multi-Concept MIL for weakly-supervised facial behavior categorization

Adria Ruiz[1]
adria.ruiz@upf.edu
Joost Van de Weijer[2]
joost@cvc.uab.es
Xavier Binefa[1]
xavier.binefa@upf.edu

[1] Universitat Pompeu Fabra (DTIC)
Barcelona, Spain
[2] Centre de Visió per Computador
Barcelona, Spain

Abstract

In this work, we address the problem of estimating high-level semantic labels for videos of recorded people by means of analysing their facial expressions. This problem, to which we refer as facial behavior categorization, is a weakly-supervised learning problem where we do not have access to frame-by-frame facial gesture annotations but only weak-labels at the video level are available. Therefore, the goal is to learn a set of discriminative expressions appearing during the training videos and how they determine these labels. Facial behavior categorization can be posed as a Multi-Instance-Learning (MIL) problem and we propose a novel MIL method called Regularized Multi-Concept MIL to solve it. In contrast to previous approaches applied in facial behavior analysis, RMC-MIL follows a Multi-Concept assumption which allows different facial expressions (concepts) to contribute differently to the video-label. Moreover, to handle with the high-dimensional nature of facial-descriptors, RMC-MIL uses a discriminative approach to model the concepts and structured sparsity regularization to discard non-informative features. RMC-MIL is posed as a convex-constrained optimization problem where all the parameters are jointly learned using the Projected-Quasi-Newton method. In our experiments, we use two public data-sets to show the advantages of the Regularized Multi-Concept approach and its improvement compared to existing MIL methods. RMC-MIL outperforms state-of-the-art results in the UNBC data-set for pain detection.

1 Introduction

The face is one of the main human non-verbal communication channels. By analysing facial gestures performed by people in a given situation, we can infer their mood, feelings and intentions. Most effort in facial expression analysis has focused on proposing supervised methods to detect a set of predefined gestures such as Action Units [9]. However, supervised AU detection is far from being solved [27] and requires a huge labelling effort to annotate spontaneous behavior databases. In contrast, we focus on a weakly-supervised problem which we call facial behavior categorization. As an example, consider a set of videos of

people recorded while watching an advertisement. The videos are labelled with the subject's appreciation of the advertisement, revealing whether or not he liked it. The task of facial behavior categorization is to analyse the set of subject facial expressions during the whole recording and estimate the "Like/Not Like" label. We consider it as a weakly-supervised learning problem because we do not have access to frame-level annotations of gestures such as AUs, but we only know a high-level label at the video-level. Thus, the goal is to learn a set of discriminative expressions and how they determine these video weak-labels.

Facial behavior categorization can be naturally posed as a Multi-Instance Learning (MIL) problem. In MIL, the training set $\mathcal{T} = \{(X_1, y_1), (X_i, y_i), \ldots, (X_N, y_N)\}$ is formed by N pairs of bags $X_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$. Every $X_i = \{\mathbf{x}_{i1}, \mathbf{x}_{ij}, ..., \mathbf{x}_{iM}\}$ is a set of M instances $\mathbf{x}_{ij} \in R^D$. The labels $y_i \in \{0,1\}$ are typically binary variables indicating whether the class of the bag is positive or negative. In facial behavior categorization, we consider a video as a bag X_i, its instances x_{ij} correspond to facial-descriptors extracted at each video-frame and y_i refers to the video weak-label. Using the training set $\mathcal{T}$, the goal is to obtain a classifier $F(X_*) = y_*$ able to predict a label y_* from a new test bag X_*. In order to learn the bag-classifier, MIL methods assume that there exist an underlying relation between the bag label and its instances distribution [10]. In this work, we differentiate between Single-Concept and Multi-Concept MIL methods.

Single-Concept MIL methods assume that there exist a single target-concept in the instance space. The probability of a bag to be positive is defined as the maximum probability among its instances given this concept. In facial behavior analysis, Single-Concept approaches have been recently applied to supervised AU localization [26] and weakly-supervised pain detection [25]. However, for general facial behavior categorization problems, the Single-Concept assumption does not take into account that different types of expressions can appear during the video which contribute differently to its label. For example, in the case of subjects watching an advertisement, the subject can express different combinations of smiles or neutral faces which will determine the "Like/Not Like" label.

Multi-Concept MIL methods can be considered a generalization of Single-Concept approaches. They assume that there exist a set of concepts in the instance space whose combined presence in the bag determine its label. However, existing Multi-Concept methods are limited in facial behavior categorization because they assume that the concepts can be modelled by isotropic Gaussians where all the features have the same importance. In constrast, facial-descriptors (instances) are typically highly dimensional and contain a low number of informative features related with facial expression changes [31].

The main contributions of the paper are the following. Other than previous facial behavior analysis work which uses Single-Concept methods, we use a Multi-Concept approach to solve the facial behavior categorization problem. Moreover, we address the limitations of current Multi-Concept approaches proposing Regularized Multi-Concept MIL (RMC-MIL), a novel MIL method adapted to the highly dimensional nature of facial-descriptors:

- **Discriminative concepts:** RMC-MIL follows the Multi-Concept MIL assumption and jointly learns a set of concepts (facial gestures) and a higher-level classifier defining their contribution to the bag (video) label. In contrast to current Multi-Concept approaches, the concepts are not assumed to follow any fixed distribution and we model them as discriminative hyperplanes in instance space.

- **Structured Sparsity Regularization:** RMC-MIL applies matrix $L_{2,1}$-norm regularization over the concept-hyperplanes. This regularization forces common sparsity

across them and, as a consequence, they only use a common subset of features which are expected to be related with facial gestures. To the best of our knowledge, this is the first work to introduce this type of regularization in the context of MIL.

RMC-MIL learning process is posed as a convex-constrained optimization problem where all the parameters are jointly learned and efficiently solved using the Projected-Quasi-Newton method [23]. In our experiments, we test the proposed method in different public data-sets and facial behavior categorization problems to show the advantages of our Regularized Multi-Concept approach. RMC-MIL achieves better performance than existing Single-Concept and Multi-Concept MIL methods and outperforms state-of-the-art results in the UNBC dataset for pain detection.

2 Related work on Multiple Instance Learning

Most of MIL research follows the Single-Concept assumption. The various methods differ on how the target-concept is obtained from the training set. Diverse-Density [18] model it as a Gaussian and learn its mean and diagonal covariance using gradient-descent optimization. Bayesian-MIL [22] adapts Logistic Regression to MIL and incorporates a prior over the parameters in order to perform feature selection. MM-MIL [28] uses a mixture of linear classifiers to represent a muti-modal target-concept. Other approaches reformulate standard supervised methods such as AnyBoost [30], SVM [1, 4], Gaussian Processes [14] or Random Forests [15] and adapt them to the MIL assumption. As discussed above, Single-Concept approaches can not model that different concepts (expressions) can appear inside a bag (video) and that they can have different contribution to its label.

Multi-Concept MIL methods learn a set of concepts in the instance space and a bag-classifier defining how their presence define the label. For this purpose, the bags are embedded into a K dimensional space where standard classifiers can be used. Each dimension in this space contains the bag probability given the k-th concept following the Single-Concept assumption. In a seminal work, DD-SVM [6] proposed to learn the set of concepts by using multiple runs of Diverse Density initialized from all the instances in the training set. However, its high computational cost makes it impractical for large data-sets as in the case of facial behavior categorization. Posterior works have considered to model the concepts as hyper-spheres (isotropic Gaussians) centered in a set of training instances (prototypes). MILES [7] consider all the training instances in the data-set as potential prototypes and selects the most relevant with $l1$-norm SVM. MILIS [11] uses a coordinate descent procedure to iteratively learn the most relevant prototypes and the bag-classifier. Recently, [12] proposed an algorithm based on AdaBoost to select a set of prototypes from different information sources.

3 Regularized Multi-Concept Multi-Instance Learning

In this section we describe the proposed Regularized Multi-Concept Multi-Instance Learning approach (RMC-MIL). We firstly explain the non-regularized version of the method in 3.1 and then we extend it to the regularized case in 3.2.

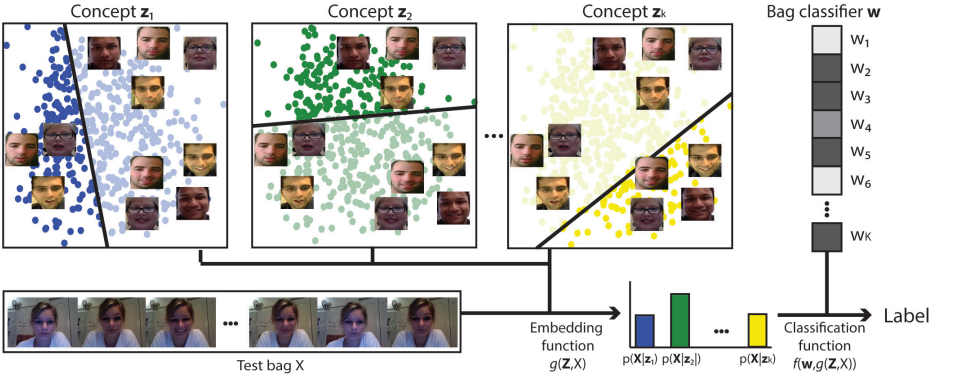

Figure 1: Overview of the proposed Multi-Concept MIL method. Concepts are modelled as a set of K hyperplanes $\mathbf{z_k}$ in instance space. Given a bag, it is represented using the probability of its instances given each concept. The bag-classifier $\mathbf{w}$ maps this bag-representation into high-level labels. Both $\mathbf{Z}$ and $\mathbf{w}$ parameters are jointly optimized during training.

3.1 MC-MIL

Let us denote $\mathbf{Z} = [\mathbf{z}_1 \ \ \mathbf{z}_2 \ \ \ldots \ \ \mathbf{z}_K]$ as a $R^{D\times K}$ matrix where each column $\mathbf{z}_k$ is a D-dimensional hyperplane classifying instances depending whether they belong or not to the k-th concept. Now we define the probability of an instance $\mathbf{x}_{ij}$ given a concept k as $p(\mathbf{x}_{ij}|\mathbf{z}_k) = \sigma(\mathbf{z}_k{}^T\mathbf{x}_{ij})$ where $\sigma(s)$ corresponds to the sigmoid function.

Following the standard MIL assumption, the probability of a bag X_i given a concept k is defined as $p(X_i|\mathbf{z_k}) = \max_j p(\mathbf{x}_{ij}|\mathbf{z}_k)$. Since $\max(.)$ is not differentiable, we approximate it using the Generalized Mean (GM) function defined as: $p(\mathbf{z_k}|X_i) = (\sum_{j=1}^{M} p(\mathbf{z}_k|\mathbf{x}_{ij})^r)^{\frac{1}{r}}$. GM have been previously used in MIL methods [25] and is equivalent to the arithmetic mean when $r = 1$ and to *max* function when r tends to ∞.

Following the main idea of current Multi-Concept approaches, we define:

$$g(X_i,\mathbf{Z}) =< p(X_i|\mathbf{z}_1), p(X_i|\mathbf{z}_2), \ldots, p(X_i|\mathbf{z}_K) >. \tag{1}$$

Intuitively, $g(X_i,\mathbf{Z})$ embeds the bag X_i into a K-dimensional space, where the value in the k-th dimension is the probability of the bag i given the concept k. Given $g(X_i,\mathbf{Z})$, the bag-classifier is defined as $F(X_i) = sign(\mathbf{w}^T g(X_i,\mathbf{Z}))$, where $\mathbf{w} = [w_1, w_2, \ldots, w_K]$ are the parameters of an hyperplane which separates positive and negative bags embedded in the K dimensional space. Figure 1 shows an overview of the proposed MC-MIL method.

The goal of MC-MIL is to learn the classifier $F(X)$ estimating the optimal concept hyperplanes $\mathbf{Z}$ and the bag-classifier $\mathbf{w}$ given the training set $\mathcal{T}$. For this purpose, we use a classification loss function ℓ and solve the following optimization problem:

$$\min_{\mathbf{w},\mathbf{Z}} \quad \mathcal{L}(\mathcal{T},\mathbf{Z},\mathbf{w}) = \sum_{i=1}^{N} \ell(\mathbf{w}^T g(X_i,\mathbf{Z}), y_i) = -\sum_{i=1}^{N} y_i \ln(p_i) + (1-y_i)\ln(1-p_i) \tag{2}$$

where p_i is defined as $\sigma(\mathbf{w}^T g(X_i,\mathbf{Z}))$ and can be understood as the probability of the bag X_i to be positive. Note that we used logistic loss similar to other existing MIL methods. However, any differentiable classification loss can be used instead.

3.2 Regularized MC-MIL

In facial behavior categorization, the instances $\mathbf{x}_{ij}$ lie in a high dimensional space and there is a high number of potential non-informative features. In this scenario, it is required to incorporate regularization mechanisms in order to find the discriminative features and reduce the risk of overfitting [21]. For this purpose, we introduce in Eq. 2 a regularizer over $\mathbf{Z}$:

$$\min_{\mathbf{w},\mathbf{Z}} \quad \mathcal{L}(\mathcal{T},\mathbf{Z},\mathbf{w}) = \sum_{i=1}^{N} \ell(\mathbf{w}^T g(X_i,\mathbf{Z}),y_i) + \lambda \Omega_{\mathbf{Z}}(\mathbf{Z}) \tag{3}$$

where λ is a positive scalar controlling the importance of the regularization term. In this work, we explore the use of the matrix $L_{2,1}$-norm regularization $\Omega_{\mathbf{Z}}(\mathbf{Z}) = ||\mathbf{Z}||_{2,1}$ defined as $||\mathbf{Z}||_{2,1} = \sum_{d=1}^{D} ||\mathbf{z}^d||_2$, where $\mathbf{z}^d$ denotes the d-th row of matrix $\mathbf{Z}$.

It is known that $L_{2,1}$-norm encourages sparsity across the rows of $\mathbf{Z}$ [2]. The use of structured sparsity regularization is motivated by a previous work [31] in Multi-Task Learning for supervised facial expression recognition. That work uses $L_{2,1}$ regularization to force joint sparsity between independent facial expressions classifiers. Similarly, in the case of RMC-MIL, this regularization encourages the concept hyperplanes to use a common subset of features expected to be related with facial expression changes.

4 RMC-MIL optimization

In order to efficiently minimize Eq. 3 including loss and the non-smooth $L_{2,1}$ regularization terms, we propose to use the Projected Quasi-Newton method (PQN) [1] presented in [23]. In 4.1 we briefly describe PQN and in 4.2 we explain how we apply it to RMC-MIL. It is worth mentioning that Eq. 3 is not convex and is not guaranteed to converge into a global minimum using PQN. However, most of state-of-the-art MIL methods are non-convex [16], and local-optimal solutions are shown to achieve good results.

4.1 Projected Quasi-Newton Method

Projected-Quasi-Newton is a generalization of standard Quasi-Newton method which minimize convex-constrained problems of the form:

$$\min_x f(x) \quad s.t \quad x \in \mathcal{C} \tag{4}$$

where $f(x)$ is any continuous differentiable function and C is a convex set.

Similarly as Quasi-Newton method, PQN minimize $f(x)$ using iterative 2nd-order gradient descent. At the k-th iteration, a second-order approximation of $f(x)$ is computed as:

$$q_k(x) f(x_k) + (x - x_k)^T \nabla f(x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k) \tag{5}$$

where x_k is the solution at iteration k and B_k is a positive definite approximation of the Hessian matrix $\nabla^2 f(x_k)$. PQN uses the Limited-memory-BFGS strategy [5] to approximate B_k using a diagonal plus low-rank compact form. This approach is convenient when the number of variables in x is large. Using (5), PQN finds a more optimal x_{k+1} by solving:

$$x_{k+1} = \min_x q_k(x) \quad s.t \quad x_k \in C \tag{6}$$

[1] Code available at: http://www.cs.ubc.ca/ schmidtm/Software/PQN.html

This sub-problem is solved by using Spectral Projected Gradient (SPG) [3] which computes the solution to Eq. 6 with a gradient descent approach but, at each iteration, the solution x is projected into the convex set $\mathcal{C}$ using a projection function $\mathcal{P}_{\mathcal{C}}(x)$:

$$\mathcal{P}_{\mathcal{C}}(x) = \min_{c} ||c-x||_2 \quad s.t \quad c \in \mathcal{C} \tag{7}$$

Intuitively, c is the nearest point to x in terms of the euclidean distance which belongs to the set $\mathcal{C}$ representing feasible solutions. As explained in [23], the PQN method is particularly interesting when we are minimizing a function such as Eq. (3) with matrix $L_{2,1}$-norm regularization. In this cases, the soft-regularizer $\lambda\Omega(x)$ is non-smooth but induces the solution to be in the convex norm-ball: $\mathcal{C} = \{x \mid ||x||_{2,1} \leq \tau\}$. Note that τ is the ball radius and it is directly related with the original parameter λ. In this case, the projection $\mathcal{P}_{\mathcal{C}}(x)$ for a given x can be efficiently computed. For more details about SPG and the Projected-Quasi-Newton algorithms, the reader is referred to the original papers.

4.2 RMC-MIL optimization via PQN method

In order to solve RMC-MIL optimization, we firstly reformulate Eq. 3 as the following equivalent constrained-convex optimization problem:

$$\min_{\mathbf{w},\mathbf{Z}} \quad \mathcal{L}(\mathbf{w},\mathbf{Z}) = \sum_{i=1}^{N} \ell(\mathbf{w}^T g(X_i,\mathbf{Z}), y_i) \quad s.t. \; ||\mathbf{Z}||_{2,1} \leq \tau_Z \tag{8}$$

where the constraint forces $\mathbf{Z}$ to lie in the convex $L_{2,1}$-norm ball with radius τ_Z.

Secondly, we define the gradient $\nabla\mathcal{L}(\mathbf{w},\mathbf{Z})$ using the first order derivatives of $\mathcal{L}$ w.r.t $\mathbf{w}$ and $\mathbf{z}_k$. Being ℓ defined as the logistic-loss function, they can be expressed as:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = -\sum_{i=1}^{N} (y_i - p_i) g(X_i,\mathbf{Z}) \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}_k} = -\sum_{i=1}^{N} \frac{w_k}{M_i} (y_i - p_i) \left(\frac{1}{M_i} \sum_{j=1}^{M} p_{ijk} \right)^{\frac{1}{r}-1} \sum_{j=1}^{M_i} p_{ijk}^r (1 - p_{ijk}) \mathbf{x_{ij}} \tag{10}$$

where $p_i = \sigma(\mathbf{w}^T g(X_i,\mathbf{Z}))$, $p_{ijk} = \sigma(\mathbf{z}_k^T \mathbf{x}_{ij})$, M_i is the total number of instances in X_i and r is the parameter used in the Generalized-Mean function.

With the above definitions, we apply the Projected-Quasi-Newton explained in Sec. 4.1. During the Spectral Projection Gradient steps, the projection of $\mathbf{Z}$ into the $L_{2,1}$ norm-ball with radius τ_Z can be computed in linear time [23]. RMC-MIL source code will be available at *http://cmtech.upf.edu/research/projects/rmc-mil*.

5 Experiments

Other than existing work on facial behavior analysis [25, 26] we propose a a Multi-Concept MIL approach. In addition, our method proposes the usage of discriminative concepts and structured sparsity regularization to handle the highly dimensional nature of facial-descriptors. In this section, we first describe the facial-descriptors and the data-sets used in the experiments. In 5.2 we analyze the impact of the number of concepts and the impact

of the regularization on the final results. In 5.3 , we compare RMC-MIL to standard Single-Concept and Multi-Concept MIL methods. Finally, we illustrate the ability of RMC-MIL to discover discriminative facial expressions from weak-labelled videos.

5.1 Datasets and experimental setup

Facial-descriptors: Given a video (bag), we extract a facial-descriptor (instance) for each frame. The whole process is illustrated in Figure 2. Firstly, we obtain a set of 49 landmark facial-points with the method described in [29]. Then, the face is aligned and re-sized (250x250) by estimating an affine transformation from the obtained landmark points and a mean-shape computed from all video-frames. Finally, a set of 3D-Temporal-SIFT descriptors [24] [2] are extracted from local patches placed in 16 landmark points (8 for eyes and eyebrows, 2 for nose wings and 6 for mouth). The final facial-descriptor is obtained by concatenating the 3D-SIFT features extracted from each patch resulting in a total of 2560 dimensions. This patch-based facial-descriptor is similar to the used in other works such as [8]. However, we use 3D-Temporal-SIFT instead of SIFT in order to encode the temporal information present in facial expressions. The size of the local patches has been set to 30 by 30 pixels and a temporal window of 0.5 seconds has been used.

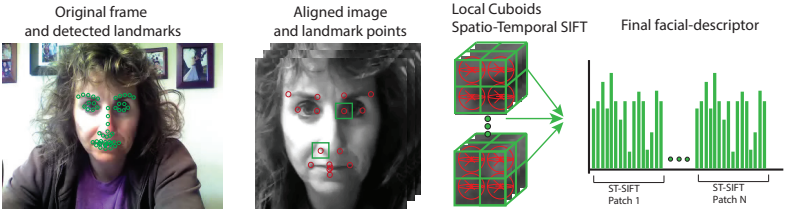

Figure 2: (i) 49 extracted landmark points. (ii) image aligned with the obtained affine transformation (ii) Spatial-Temporal SIFT descriptors extracted from each local cuboid. Red points corresponds to the subset of landmarks used.

AM-FED: The AM-FED dataset [19] contains 242 on-line web-cam recordings from different subjects watching three TV advertisements. After watching a video, subjects were asked two questions: "Did you like the video?" and "Do you want to view this video again?". The subjects chose between positive (1), neutral (0) or negative responses (-1). The goal is to analyse the subject's facial behavior during the advertisement and predict these labels. Similar as [20], only videos where the subjects reported positive (1) and negative (-1) answers to the questions are considered. A 3-fold cross validation is used for evaluation where the videos corresponding to one advertisement are used for testing. 26 videos were discarded for the experiments since the detection of landmark points failed. A total of 158 and 94 videos for the "Watch/Not Watch again" and "Like/Does not like" problems respectively are used. To the best of our knowledge, we are the first work in applying weakly-supervised learning to this data-set without previous supervised detection of AUs.

UNBC-McMaster: The UNBC-McMaster Shoulder Pain Expression Archive Database [17] contains 200 recordings of 25 different subjects undergoing some kind of shoulder pain. During the sessions, the subjects performed active and passive arm movements and expert coders annotate the different levels of pain felt. Levels are between 0 (no pain) to 5 (strong

[2] Code available at: http://crcv.ucf.edu/source/3D

pain). The work in [25] reported the state-of-the-art results in this data-set for weakly-supervised pain detection. In our experiments, we follow the same experimental setup: the sequence pain levels are converted into no pain (-1) and pain (1) binary labels and the task is to classify the sequences by analysing the subjects facial gestures during the session. A Leave-One-Subject-Out Cross-Validation is used for evaluation. Only subjects with more than one sequence are used resulting in a total of 147 videos and 23 subjects.

5.2 Multiple Concepts and Structural Sparsity Regularization for Facial Behavior Categorization

In this experiment, we investigates the dependence on the number of concepts (as determined by K) and the impact of the proposed regularization (controlled by τ_Z) on RMC-MIL performance. We also evaluate the results when no regularization is used (MC-MIL). In addition, we measure the common sparsity between the concept-hyperplanes computed as the Gini Coefficient [13] over the $L2$-norms of Z rows. This measure indicates how many features have a very low contribution defining the concepts. Figure 3 shows the Area Under the Curve obtained in "Watch/Not watch again" and "Pain/No pain" problems. Since RMC-MIL and MC-MIL parameters are randomly initialized, we report the mean and variance over five runs.

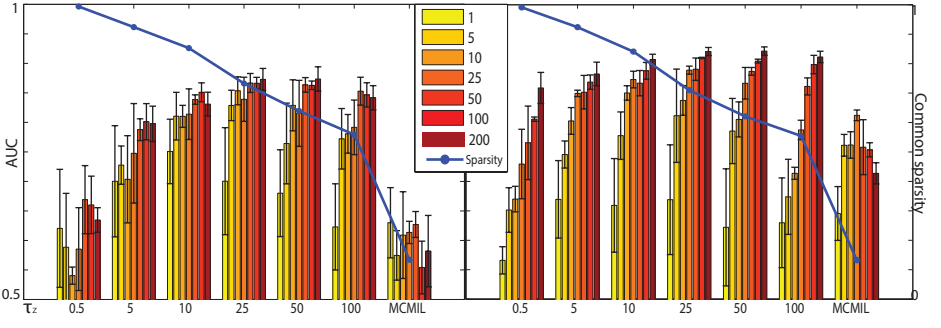

Figure 3: AUC obtained by RMC-MIL and MC-MIL in "UNBC-Pain/No pain" (left) and "AM-FED-Watch/Not watch again" (right) problems. Bar colors indicates the number of concepts used and X axis refers to different values for τ_Z . Blue line corresponds to the mean common sparsity coefficient for all K given a fixed τ_Z value.

As expected, the performance decreases for too small and too large τ_Z values (including for MC-MIL). In the first case, too much sparsity is imposed on Z whereas in the second one, large values cause the regularization to have no impact. Note that the best results are obtained with a high common sparsity between concept-hyperplanes. This indicates that only a small subset of facial-descriptor features is useful to discriminate discriminative facial expressions. The results also show that the use of more concepts consistently improves the performance except in the case of unregularized MC-MIL. This can be explained because the more concepts are used the more parameters need to be learned. Therefore, the regularization has a critical importance in order to reduce overfitting. The variance over different runs shows a stable behavior of RMC-MIL despite random initialization. In conclusion, the results demonstrate the advantages of using multiple concepts in facial behavior categorization and the effectiveness of structured sparsity regularization in this context.

5.3 Comparison with other MIL methods

In this experiment, we compare the performance of RMC-MIL with four popular MIL methods: MilBoosting [30], MI-Forest [15], MILES [7] and MILIS [11]. MilBoosting and MI-Forest follow the Single-Concept assumption and implicitly incorporate feature selection by using single-feature decision-stumps to model the target-concept. Note that MilBoosting has been recently applied to weakly-supervised pain detetion [25] and MI-Forest has achieved comparable or better performance than other MIL methods. On the other hand, MILES and MILIS are popular Multi-Concept approaches modelling the concepts as isotropic Gaussians in the instance space where all the features have the same importance.

For MI-Forest, we have used the code provided by the authors and the same parameters used in all the original paper experiments. For the other methods, we have developed our own implementation. Same as [25], our MilBoosting implementation use single-feature decision stumps as weak-classifiers and Generalized Mean to approximate the max function. In MILES and MILIS, the parameters σ and C (see original paper) have been optimized using 4-fold-cross-validation over the training set. For RMC-MIL, the parameter τ_Z and the number of concepts have been fixed to 50 and 200 respectively for all the experiments. Table 1 shows the AUC obtained in AM-FED and UNBC data-sets. In the case of UNBC, we also report the accuracy computed at the Equal Error Rate point of the Receiver Operating Characteristic curve in order to compare our results to [25]. For MI-Forest and RMC-MIL the results are computed as the mean obtained over five different runs.

	MILES[7]	MILIS[11]	MilBoosting[30]	MI-Forest[15]	MS-MIL [25]	RMC-MIL
AM-FED: Like/Do not Like	0.62	0.63	0.61	0.68	-	**0.72**
AM-FED: Watch / Not watch again	0.76	0.73	0.83	0.78	-	**0.87**
UNBC: Pain/No Pain	0.85 / 78.2	0.82 / 76.9	0.78 / 76.9	0.81 / 75.8	- / 83.7	**0.92 / 85.7**

Table 1: Results obtained by Multi-Concept, Single-Concept MIL methods and RMC-MIL in the AM-FED and UNBC data-sets. See text for details

As the reader can observes, RMC-MIL achieves better performance in all the problems. Given these results and the reported in Sec. 5.2, our hypothesis is that RMC-MIL outperforms Single-Concept approaches because it does not assume that the presence of a unique concept (facial expression) in a bag determine the video label. This allows RMC-MIL to learn different types of discriminative gestures which can appear during the video and contribute to the video-label. On the other hand, the better results of RMC-MIL compared to Multi-Concept approaches can be explained because RMC-MIL can better handle the highly-dimensional nature of facial-descriptors. The concepts are not assumed to follow a Gaussian distribution and the incorporation of matrix $L_{2,1}$ regularization is able to discard non-informative features. State-of-the-art results reported in [25] for the UNBC data-set, are not directly comparable since they use Bag-of-Words-based features and a ensemble of MilBoost classifiers trained with bootstrapped data. However, the results using RMC-MIL and 3D-SIFT-based facial-descriptors compare favorably to their approach.

5.4 Applying RMC-MIL to discover discriminative facial expressions

To provide insight into what RMC-MIL is actually learning, we visualize the expressions which determine the video labels. Using RMC-MIL, we can consider a video frame as a bag with only one instance and classify it. Note that instance classification can be understood as a weighted sum (determined by $\mathbf{w}$) of the instance probabilities for each concept $\mathbf{z}_k$. In this experiment, we have trained RMC-MIL for the different problems and applied the learned

model to all the frames in the data-set. Again, the parameter τ_Z and the number of concepts have been fixed to 50 and 200 respectively. Figure 4 shows the most positive and most negative frames in a set of random selected videos from both data sets. For the UNBC dataset, different kind of facial expressions representing pain are considered more positive whereas neutral faces obtain less probability. For the AM-FED problems, RMC-MIL learns that smiles contribute positively to the bag label whereas neutral faces are considered negative. Note that these discriminative facial expressions represent different appearances with varying intensity and which depend on the subject. RMC-MIL can effectively handle these facial expressions differences since it is able to model them by using a Multi-Concept approach.

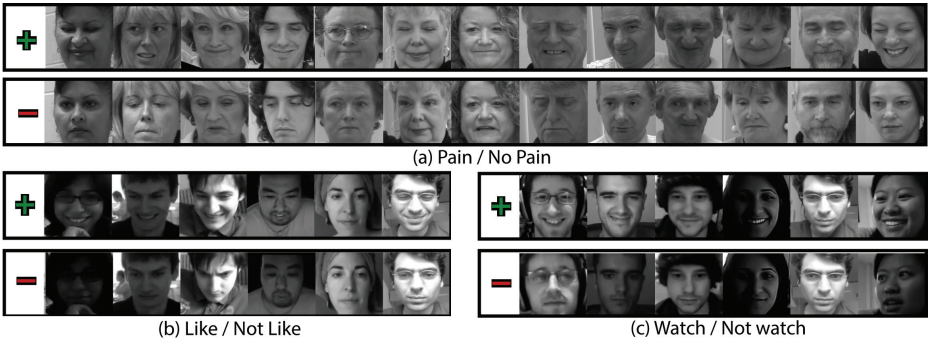

Figure 4: Most positive and negative instances estimated by RMC-MIL in a set of randomly selected videos for the different facial behavior categorization problems

6 Conclusions and future work

In this work, we have presented Regularized-Multi-Concept MIL and its application to facial behavior categorization. Other than previous MIL methods used in facial behavior analysis, RMC-MIL does not follow a Single-Concept assumption. Moreover, in contrast to existing Multi-Concept MIL methods, RMC-MIL can learn more optimal concepts from high-dimensional facial-descriptors by using a discriminative approach and structured sparsity regularization. In the experiments, we have shown the improvement of RMC-MIL over existing Single-Concept and Multi-Concept MIL methods and its ability to learn discriminative facial gestures from weakly-labeled data. Future work will focus on incorporating temporal information in the model in order to take into account how different concepts appear in time-domain. This information is expected to be important to solve complex facial behavior categorization problems.

Acknowledgements

This work has been partially funded by Spanish government under projects IPT-2012-0630-020000, IPT-2011-1015-430000 and CICYT grant TIN2012-39203.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing*

Systems, 2002.

[2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *Advances in Neural Information Processing Systems*, 2007.

[3] Ernesto G Birgin, José Mario Martínez, and Marcos Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 2000.

[4] Razvan C Bunescu and Raymond J Mooney. Multiple instance learning for sparse positive bags. In *Proc. International Conference on Machine Learning*. ACM, 2007.

[5] Richard H Byrd, Jorge Nocedal, and Robert B Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 1994.

[6] Yixin Chen and James Z Wang. Image categorization by learning and reasoning with regions. *J. of Machine Learning Research*, 5, 2004.

[7] Yixin Chen, Jinbo Bi, and James Ze Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[8] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F. Cohn. Selective transfer machine for personalized facial action unit detection. *CVPR*, 2013.

[9] Fernando De la Torre and Jeffrey F Cohn. Facial expression analysis. In *Visual Analysis of Humans*, pages 377–409. Springer, 2011.

[10] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(01):1–25, 2010.

[11] Zhouyu Fu, Antonio Robles-Kelly, and Jun Zhou. Milis: Multiple instance learning with instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[12] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, and X. Wu. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *Cybernetics, IEEE Transactions on*, 2014.

[13] Niall Hurley and Scott Rickard. Comparing measures of sparsity. *Information Theory, IEEE Transactions on*, 55(10):4723–4741, 2009.

[14] Minyoung Kim and Fernando Torre. Gaussian processes multiple instance learning. In *Proc. International Conference on Machine Learning*, 2010.

[15] Christian Leistner, Amir Saffari, and Horst Bischof. Miforests: Multiple-instance learning with randomized trees. In *Proc. European Conf. on Computer Vision*. 2010.

[16] Fuxin Li and Cristian Sminchisescu. Convex multiple-instance learning by estimating likelihood ratio. In *Advances in Neural Information Processing Systems*, 2010.

[17] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG*. IEEE, 2011.

[18] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, 1998.

[19] D. McDuff, R. Kaliouby, T. Senechalz, M. Amrz, J.F. Cohn, and 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*, 2013.

[20] Daniel McDuff, Rana el Kaliouby, David Demirdjian, and Rosalind Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *FG*, 2013.

[21] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proc. International Conference on Machine Learning*. ACM, 2004.

[22] Vikas C Raykar, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, and R Bharat Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proc. International Conference on Machine Learning*. ACM, 2008.

[23] Mark W Schmidt, Ewout Berg, Michael P Friedlander, and Kevin P Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *International Conference on Artificial Intelligence and Statistics*, page None, 2009.

[24] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *International conference on Multimedia*, 2007.

[25] Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. In *Automatic Face and Gesture Recognition (FG)*. IEEE, 2013.

[26] David MJ Tax, E Hendriks, Michel François Valstar, and Maja Pantic. The detection of concept frames using clustering multi-instance learning. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2917–2920. IEEE, 2010.

[27] Michel François Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 2012.

[28] Qifan Wang, Luo Si, and Dan Zhang. A discriminative data-dependent mixture-model approach for multiple instance learning in image classification. In *Proc. European Conf. on Computer Vision*. Springer, 2012.

[29] Xuehan-Xiong and Fernando De la Torre. Supervised descent method and its application to face alignment. In *Proc. Computer Vision and Pattern Recognition*, 2013.

[30] Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems*, 2005.

[31] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and D.N. Metaxas. Learning active facial patches for expression analysis. In *Proc. Computer Vision and Pattern Recognition*, June 2012.