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

Introduction: Most efforts in facial behavior analysis have focused on proposing supervised methods to detect a set of predefined gestures such as the Action Units. However, supervised AU detection is a difficult task which requires a huge labelling effort to annotate spontaneous behavior databases. In contrast, we focus on a different problem which we call facial behavior categorization. The goal is to estimate high-level semantic labels for videos of recorded people by means of analysing their facial expressions. 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. This problem can be considered a weakly-supervised learning problem because we do not have access to frame-by-frame facial gesture annotations but only weak-labels at the video level are available. From this weak-annotations, we aim to learn a set of discriminative expressions and how they determine the high-level labels. Similar to [5], we pose facial behavior categorization as a Multiple Instance Learning 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 binary variables indicating whether the class of the bag is positive or negative. The goal is to lean a classifier $F(X_*) = y_*$ able to predict a label y_* from a new test bag X_*. 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.

Contributions: We propose a novel MIL method called Regularized Multi-Concept MIL for facial behavior categorization. In contrast to previous MIL methods applied to facial behavior analysis which use a Single-Concept approach, 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 potential large number of non-informative features present in the high-dimensional facial-descriptors, RMC-MIL uses a discriminative approach to model the concepts and structured sparsity regularization. As a consequence, the concepts use only a common subset of features expected to be related with facial expression changes.

Rregularized Multi-Concept MIL: An overview of RMC-MIL is illustrated in Fig. 1. Our model learns a set of K hyperplanes $\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \ldots \ \mathbf{z}_K]$ in the instance space which classify instances depending when they belong or not to the k-th concept. This concepts are expected to represent different types of discriminative facial expressions. A bag (video) is represented as a K dimensional vector:

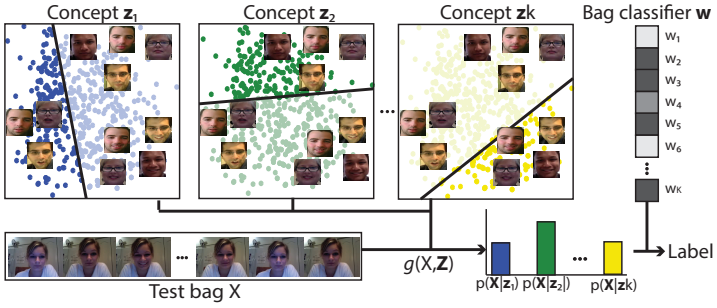

Figure 1: Overview of RMC-MIL. Concepts are modelled as a set of K classifiers $\mathbf{z_k}$ in instance space. A bag is represented using the probability of its instances given each concept. The bag-classifier $\mathbf{w}$ maps this bag-representation into high-level labels.

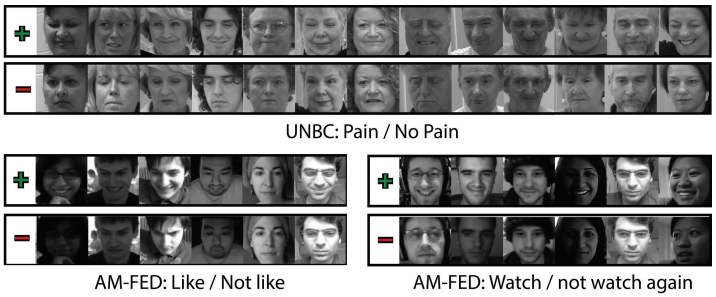

Figure 2: Most positive and negative instances estimated by RMC-MIL in a set of randomly selected videos for different facial behavior categorization problems

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

where the value in the k-th dimension is the probability of that a concept k appears in the bag X_i. This probability is defined as the maximum probability $p(X_i|\mathbf{z_k}) = \max_j p(\mathbf{x}_{ij}|\mathbf{z}_k)$ among all the bag-instances. Finally, the bag-classifier is defined as $F(X) = \text{sgn}(\mathbf{w}^T g(X, \mathbf{Z}))$, where $\mathbf{w} = [w_1, w_2, \ldots, w_K]$ are the parameters of a linear classifier separating positive and negative bags embedded in the K dimensional space. In the training stage, RMC-MIL jointly optimize the bag-classifier $\mathbf{w}$ and concept-classifiers $\mathbf{Z}$ by using a logistic-loss function ℓ and solving:

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

The use of $L_{2,1}$ regularization is motivated by previous work [1] 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.

Eq. 2 is a convex-constrained optimization problem and we use the Projected-Quasi-Newton [3] method to efficiently solve it.

Experiments: In our experiments, we evaluate the proposed approach in two different facial behavior categorization problems. Using the AMFED [4] and UBNC [2] public datasets, we attempt to categorize viewer's responses to advertisements and detect pain from patients from weakly-labelled videos. We demonstrate the advantages of using multiple concepts in facial behavior categorization and the effectiveness of structured sparsity regularization in this context. Moreover, the results show 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 (Fig. 2).

[1] Lin Zhong et al. Learning active facial patches for expression analysis. In *CVPR*, June 2012.

[2] Lucey et al. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG*, 2011.

[3] M. Schmidt et al. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *AISTATS*, 2009.

[4] McDuff et. al. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected in-the-wild. In *CVPR Workshops*, 2013.

[5] Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. In *FG*, 2013.