Weakly-Supervised Learning for Automatic Facial Behavior Analysis

Adrià Ruiz Ovejero

TESI DOCTORAL UPF / ANY 2017

DIRECTOR DE LA TESI
Xavier Binefa Valls. Departament de Tecnologies de la Informació i les Comunicacions.
Joost van de Weijer. Centre de Visió per Computador.

La infinita recursivitat de la cadena motiu-conseqüència ens empeny sempre a preguntar-nos sobre la causa primera de tot.
Per la religió: Déu.
Per la física: el "Big Bang".
Per mi: els meus avis i àvies.

A Paco, Joan, Pilar i Tinuca. Gràcies de tot cor.

Agraïments-Acknowledgments

Possiblement aquesta sigui la part més complicada d'escriure d'aquesta tesi. No és fácil resumir i expresar en paraules l'agraïment que sento cap a tantes persones. Sense el suport de tothom que mencionaré i molts altres, que malauradament m'hauré de deixar, tot hagués sigut molt més dur.

First of all, I thank my supervisor, Xavier Binefa, for the opportunity he gave me years ago and his support during all this time. He has taught me that a good supervisor needs to provide many other things than just technical knowledge or funding. Also, I would like to acknowledge Joost van de Weijer for his co-supervision and help in countless aspects of my PhD. I also want to mention all my colleagues in the CMTech: Marc, Ciro, Xavi, Pol, Lluis, Fede, Dima, Adriana, Decky and specially Oriol, who introduced me in the field when I was even more ignorant than now. Thanks to all of you for allowing me to learn new things every day. Finally, I would like to thank all the administrative staff in the DTIC: Lídia, Vanesa, Jana, Joana, Magda among others. Your work makes the daily life of a PhD student much easier.

En el terreny més personal, vull donar les gràcies a tots als meus amics i família. En especial a la meva mare, que sempre ha sabut "baixar-me a la terra" quan feia falta, i al meu pare, que desde petit ha intentat transmetre'm el seu amor pel coneixement. A vosaltres us correspon la major part de responsabilitat de totes les fites que he aconseguit i espero aconseguir. No vull tampoc deixar de mencionar a la meva germana Irene i a la Drina, que també han estat molt presents en els alts i baixos que tot doctorat comporta. Per últim i no per això menys important, vull agraïr a la meva parella, la Bet, la seva comprensió i el suport incondicional que m'ha donat durant tots aquests anys. Sé que estàs entre les dues persones que més s'han esforçat durant aquest doctorat. Moltes gràcies.

Abstract

In this Thesis we focus on Automatic Facial Behavior Analysis, which attempts to develop autonomous systems able to recognize and understand human facial expressions. Given the amount of information expressed by facial gestures, this type of systems has potential applications in multiple domains such as Human Computer Interaction, Marketing or Healthcare. For this reason, the topic has attracted a lot of attention in Computer Vision and Machine Learning communities during the past two decades. Despite the advances in the field, most of facial expression analysis problems can be considered far from being solved.

In this context, this dissertation is motivated by the observation that the vast majority of methods in the literature has followed the Supervised Learning paradigm, where models are trained by using data explicitly labelled according to the target problem. However, this approach presents some limitations given the difficult annotation process typically involved in facial expression analysis tasks. In order to address this challenge, we propose to pose Automatic Facial Behavior Analysis from a weakly-supervised perspective. Different from the fully-supervised strategy, weakly-supervised models are trained by using labels which are easy to collect but only provide partial information about the task that aims to be solved (i.e, weak-labels). Following this idea, we present different weakly-supervised methods to address standard problems in the field such as Action Unit Recognition, Expression Intensity Estimation or Affect Analysis. Our results obtained by evaluating the proposed approaches on these tasks, demonstrate that weakly-supervised learning may provide a potential solution to alleviate the need of annotated data in Automatic Facial Behavior Analysis. Moreover we also show how these approaches are able to facilitate the labelling process of databases designed for this purpose.

Resum

Aquesta tesi doctoral se centra en el problema de l'Anàlisi Automàtic del Comportament Facial, on l'objectiu és desenvolupar sistemes autònoms capaços de reconèixer i entendre les expressions facials humanes. Donada la quantitat d'informació que es pot extreure d'aquestes expressions, sistemes d'aquest tipus tenen multitud d'aplicacions en camps com la Interacció Home-Màquina, el Marketing o l'Assistència Clínica. Per aquesta raó, investigadors en Visió per Computador i Aprenentatge Automàtic han destinat molts esforços en les últimes dècades per tal d'aconseguir avenços en aquest sentit. Malgrat això, la majoria de problemes relacionats amb l'anàlisi automàtic d'expressions facials encara estan lluny de ser conisderats com a resolts.

En aquest context, aquesta tesi està motivada pel fet que la majoria de mètodes proposats fins ara han seguit el paradigma d'aprenentatge supervisat, on els models són entrenats mitjançant dades anotades explícitament en funció del problema a resoldre. Desafortunadament, aquesta estratègia té grans limitacions donat que l'anotació d'expressions en bases de dades és una tasca molt costosa i lenta. Per tal d'afrontar aquest repte, aquesta tesi proposa encarar l'Anàlisi Automàtic del Comportament Facial mitjançant el paradigma d'aprenentatge dèbilment supervisat. A diferència del cas anterior, aquests models poden ser entrenats utilitzant etiquetes que són fàcils d'anotar però que només donen informació parcial sobre la tasca que es vol aprendre. Seguint aquesta idea, desenvolupem un conjunt de mètodes per tal de resoldre problemes típics en el camp com el reconeixement d' "Action Units", l'Estimació d'Intensitat d'Expressions Facials o l'Anàlisi Emocional. Els resultats obtinguts avaluant els mètodes presentats en aquestes tasques, demostren que l'aprenentatge dèbilment supervisat pot ser una solució per tal de reduir l'esforç d'anotació en l'Anàlisi Automàtic del Comportament Facial. De la mateixa manera, aquests mètodes es mostren útils a l'hora de facilitar el procés d'etiquetatge de bases de dades creades per aquest propòsit.

Summary

List of Figures

Chapter 1

INTRODUCTION

Facial expressions are considered one of the most important channel of non-verbal communication. By observing people's facial behavior, we are able to infer their emotions, intentions [Ekman and Rosenberg, 1997] or other relevant treats such as psychiatric status [Cohn et al., 2009] and personality [Ponce-López et al., 2016]. The study and analysis of facial expressions has been addressed from different fields such as anthropology, psychology or biology. For example, in the seminal work of Charles Darwin [Darwin, 1998], he focused on the study of face and body gestures in mammals. His main goal was to find similarities in humans' and animals' facial displays, showing that they are genetically determined and providing an additional evidence to support his evolution theory.

More recently, the psychologist Paul Ekman developed a set of influential works [Ekman and Rosenberg, 1997, Ekman, 1993, Ekman and Friesen, 1971] setting the basis of modern research in the field. Among other findings, Ekman suggested that there exist six basic human emotions (anger, fear, disgust, sadness, happiness and surprise) which are universal and common across cultures. More importantly, he found that the facial gestures associated with the expression of these emotions were also universal (Fig. 1.1). These results were coherent with Darwin's findings, suggesting that non-verbal commu-

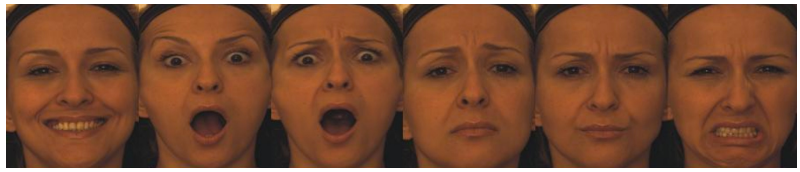

Figure 1.1: The six universal facial expressions. From left to right: happiness, surprise, fear, sadness, angry and disgust

nication has an important genetic factor independent from cultural issues.

Another relevant contribution of Ekman was the development of the Facial Action Coding System (FACS) [Ekman et al., 1978]. FACS defines a taxonomy for facial expressions describing 45 Action Units (AUs). AUs are atomic facial movements in the face caused by the activation of one or more muscles (see Fig. 1.2). For example, AU12 (Lip Corner Puller) is associated with the activation of the *Zygomaticus major* muscle. Given that any expression, including the six universal ones, can be defined by a concrete combination of Action Units, FACS provides an objective measure to describe human facial behavior.

Different from the aforementioned works, in this thesis we address the study of facial expressions from an Artificial Intelligence perspective. Concretely, we focus on *Automatic Facial Behavior Analysis*, which aims to develop autonomous systems able to recognize and understand human facial expressions. The remainder of this introductory Chapter is structured as follows. Firstly, in Sec. 1.1 we present an overview of Automatic Facial Behavior Analysis. Secondly, in Sec. 1.2 we describe current challenges in the field motivating the main research line developed in this thesis. Thirdly, we summarize our main contributions in 1.3. Finally, we conclude the Chapter with the list of publications resulting from the presented research.

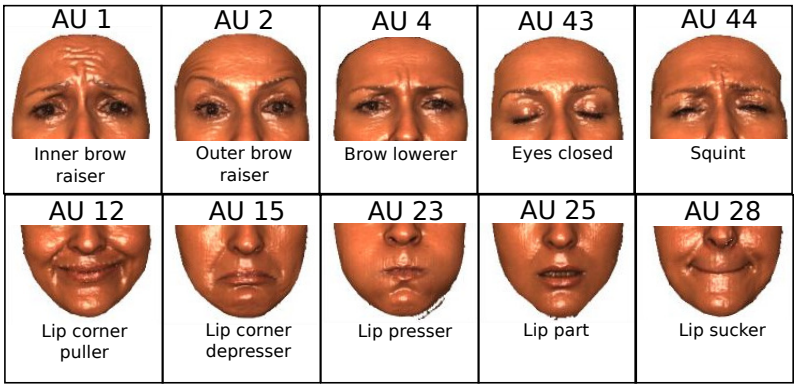

Figure 1.2: Examples of Action Units described in FACS. The samples are extracted from the Bosphorus 3D facial expression database [Savran et al., 2008]

1.1 Automatic Facial Behavior Analysis

In recent years, advances in artificial intelligence have allowed automatic systems to perform tasks which were easy for humans but very complex for machines. Nowadays, computers are able to detect objects in images [Krizhevsky et al., 2012], understand natural language [Bahdanau et al., 2016] or take decisions in very complex environments such as in autonomous driving [Geiger et al., 2012]. However, the automatic understanding of human facial behavior is still an open problem far from being solved [B. Martinez, 2016]. Given the amount of information carried by facial expressions, developing automatic systems to understand them could open a wide range of possible applications. In Human Computer Interaction, it could allow to create more naturalistic and rich interactions between humans and machines [Lisetti and Schiano, 2000]. In marketing, computers would be able to analyze consumer reactions [McDuff et al., 2013a]. In a clinical context, these systems could be used to monitor patients [Lucey et al., 2011] or diagnose mental illness such as depression [Cohn et al., 2009].

Automatic Facial Behavior Analysis (AFBA) uses Computer Vision and Machine Learning techniques in order to automatically interpret facial gestures from visual information (i.e, images or videos). The remainder of this section aims to give a brief overview of the AFBA field by describing the typical problems addressed and the standard pipeline followed by methods designed for this type of tasks. Specific works related with our main contributions will be reviewed in each particular chapter.

1.1.1 AFBA: Problems

In the following, we describe the standard problems addressed in Automatic Facial Behavior Analysis. Specifically, we differentiate between Discrete Expression Recognition, Expression Intensity Estimation and Affect Analysis. Despite the fact that this categorization is not the most standard in the literature, it is intended to clarify the relation between these problems and our particular contributions (see Section 1.3).

Discrete Expression Recognition: In this task, the goal is to detect a discrete set of facial gestures categories. Motivated by Ekman's studies, most research efforts have focused on the automatic recognition of the six universal expressions or the Action Units. These two problems are the most popular in the field and have attracted a lot of attention during the last two decades [Fasel and Luettin, 2003, De la Torre and Cohn, 2011]. For the validation of the proposed approaches, many databases have been collected containing images or videos of posed expressions [Lucey et al., 2010, Lyons et al., 1998b] or Action Units [Pantic et al., 2005, Valstar et al., 2012]. However, it is known that spontaneous facial behaviour differs from posed [Valstar et al., 2007, Littlewort et al., 2007]. This causes methods developed with these databases to not perform well in naturalistic conditions. For this reason, spontaneous facial behavior datasets have been collected more recently [McDuff et al., 2013b, Mollahosseini et al., 2016]. Apart from the six basic expressions, recent works have also addressed

the recognition of a much larger number of facial gesture categories [Du et al., 2014].

Facial Expression Intensity Estimation: Different from discrete expression recognition, several AFBA researchers have attempted to estimate expression intensity. The main motivation is that facial behavior is not a discrete phenomena but, in contrast, facial motion is smooth and is usually difficult to define the boundaries between discrete expression categories. For example, the Facial Action Coding System defines different Action Unit intensities depending on the activation level of each facial muscle. Specifically, these intensities are represented in an 6-point ordinal scale composed by a discrete set of levels. Intensity estimation has been addressed in the context of Universal Facial Expressions [Rudovic et al., 2012], [Zhao et al., 2016b], Action Units [Kaltwang et al., 2015],[Rudovic et al., 2015], or Pain expressions [Kaltwang et al., 2016],[Rudovic et al., 2013]. Despite the fact that the amount of available datasets is lower than for the discrete case, some have been collected for these particular tasks [Lucey et al., 2011, Mavadati et al., 2013]. These datasets are usually composed by videos where the expression intensity is annotated at frame-level either in a continuous domain (e.g in the range between 0 and 1) or in an ordinal scale (see Fig. 1.3).

Affect Analysis: One of the most interesting applications of AFBA is affect analysis. In this case, the problem is to infer people emotions by means of analyzing their facial expressions. Given the universality of the six basic emotions, their recognition can be considered a form of affect analysis. However, it has been shown that people may experience a larger variety of affect states [Du et al., 2014]. Therefore, the few discrete emotions defined by the universal facial expression may not reflect the real complexity of human affect. For this reason, AFBA researchers have addressed the estimation of Arousal and Valence levels. The concept of Arousal and Valence was defined by the psychologist James Russell in his Circumplex Model of affect. [Russell, 1980]. In this model, emotions are represented in a 2 dimensional space with two axes: Arousal, which refers to the level

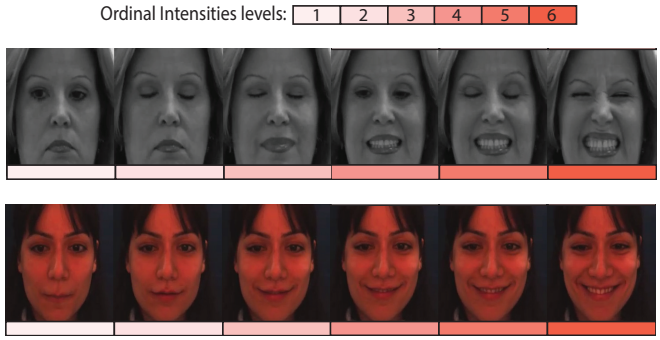

Figure 1.3: Illustration of the Pain and Action Unit Intensity Estimation problems. Top: Sequence showing different pain levels coded in an ordinal scale from 1 to 6. (Example extracted from the PAIN-UNBC Database [Lucey et al., 2011]). Bottom: Example of different intensities for Action Unit 12 (Lip-Corner Puller) also represented in an ordinal scale. (Example extracted from the DISFA Dataset [Mavadati et al., 2013])

of excitement and Valence, which is related with how unpleasant or pleasant is the emotion (see Fig. 1.4). Russel's studies suggested that the range of all possible human emotions can be represented in this space. Using similar techniques than the ones employed for intensity estimation, different works have attempted to estimate Arousal and Valence levels analyzing facial expressions [Zeng et al., 2009]. Moreover, several datasets have been collected for this particular problem [McKeown et al., 2012, Ringeval et al., 2013].

1.1.2 AFBA: Standard Pipeline

In order to solve the introduced problems, Automatic Facial Behaviour Analysis systems typically follow a pipeline composed by three main steps: Face preprocessing, Facial Feature Extraction and Machine Learning Analysis. They are described as follows:

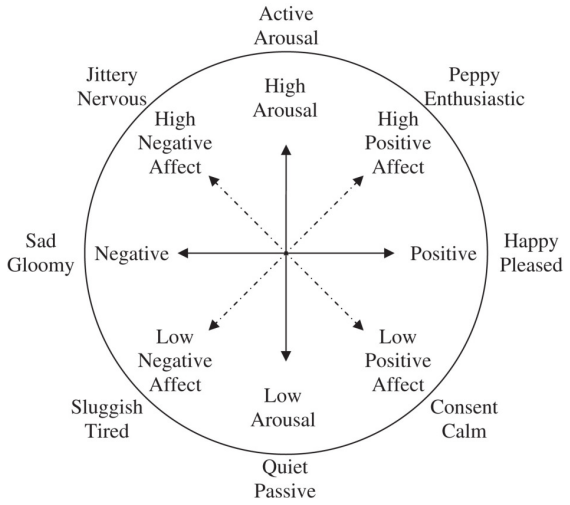

Figure 1.4: Circumplex model of affect. The x and y axis correspond to Valence and Arousal dimensions respectively

Face preprocessing

Given an input image, the goal of this first step is to obtain the region of interest where the face is located and represent it into a reference coordinate system. This prepossessing is of particular importance since it allows to remove non-relevant variations in face images such as rotation or scaling. This step is usually divided into three sub-tasks namely, face detection, landmark localization and face alignment:

Face Detection: For this task, the Viola & Jones algorithm [Viola and Jones, 2004] is usually employed in order to estimate a bounding box representing the face location in the image. This method has shown to provide reliable performance on near-to-frontal views and it is currently implemented in many commercial digital cameras. Even though some works have employed more sophisticated multi-view methods [Zhu and Ramanan, 2012] to deal with non-frontal faces, the Viola & Jones method remains as the standard approach for this step.

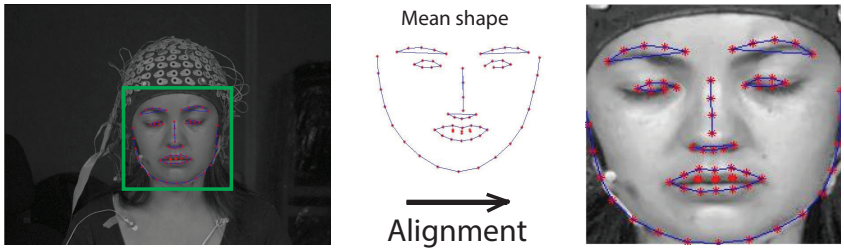

Figure 1.5: Face alignment process. (i) Face is detected in the image and facial landmark points are automatically extracted. (ii) Procrustes analysis is used to obtain an affine transformation which aligns the obtained points with a reference shape. (iii) The estimated transformation is applied to the original image

Landmark Localization: After the region of interest is obtained, the next step is to automatically find a set of landmark points in different regions (eyebrows, eyes, nose, mouth, etc..) which combined together define the face shape. This task has been traditionally performed by using different variants of the Active Appearance Models [Cootes et al., 2001, Matthews and Baker, 2004] where a statistical model of shape and texture variations is fitted ontp the face image. More recently, the Supervised Descent Method [Xuehan-Xiong and De la Torre, 2013] has been shown to outperform AAMs by posing the problem as a regression task and avoiding to explicitly compute a statistical model for the face shape and texture.

Face Alignment: Once the facial landmarks are located, the final task is to remove non-relevant transformations from the face such as scaling or rotation. This step is typically carried out by using Procrustes Analysis [Gower, 1975] in order to compute an affine transformation aligning the facial landmarks with a reference shape. Finally, this transformation is applied to the original image. Figure 1.5 shows a face alignment process following the described steps.

Facial Feature Extraction

Once the face is located and aligned, the next step is to obtain an abstract representation (*i.e* a numerical vector), encoding the information regarding the expression. This representation is known in the literature as the *facial features*. We can find two types of approaches employed to extract them: geometry-based and texture-based.

Geometry-based: In this case, facial features are obtained by analyzing the information regarding the face shape. For instance, the 2D coordinates of the aligned landmark points can be concatenated in order to obtain a numerical vector representation [Kotsia and Pitas, 2007]. Despite its simplicity, features constructed following this approach have shown reasonable performance on different facial expression analysis tasks. More sophisticated approaches make use of different statistics extracted from the face shape. For example, the angles and distances between landmark pairs can be computed in order to obtain more informative features [Valstar and Pantic, 2012]. We find another popular approach when target data is provided in the form of image sequences. In this case, geometric facial features can be constructed by computing the displacement of the landmark points along time [Pantic and Patras, 2006].

Texture-based: Different from geometry-based, texture-based features use the pixel intensity information of face images. Methods following this approach, typically extract a set of texture descriptors from local parts of the image (e.g, small pathes centered in facial landmarks) and concatenate them to create the final representation. For this purpose, standard gradient-based descriptors such as the Scale-Invariant Feature Transformation (SIFT) [Chu et al., 2013] or the Histogram of Oriented Gradients (HOG) [Jampour et al., 2015] are widely used. This type of descriptors are known to be robust to scale changes and are able to capture texture variations caused by subtle facial deformations. Apart from gradient-based, other popular texture-descriptors explored in this context include the Local Binary Patterns (LBPs) [Zhao and Pietikainen, 2007], the Gabor

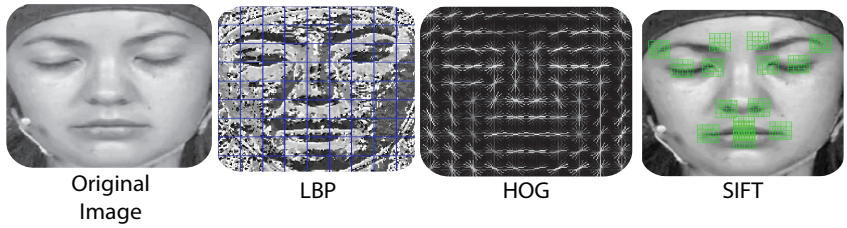

Figure 1.6: Examples of facial-descriptors used in Facial Behavior Analysis

Filters [Lyons et al., 1998a] or Haar-like features [Whitehill and Omlin, 2006]. Some examples of texture-based features are depicted in Figure 1.6.

In summary, geometry-based features are usually appealing for their computational simplicity and reasonable performance. However, they are unable to capture subtle changes caused by wrinkles, bulges and furrows [Shan, 2008]. On the other hand, texture-based features are able to encode such variations but are less robust to different factors including extreme head-pose or illumination changes. Moreover, the feature vectors resulting from texture-based approaches usually have a higher dimensionality which increases model complexity. Even tough the combination of both types of features has been also explored [Youssif and Asker, 2011], any of the proposed approaches has been shown to consistently perform well in a variety of applications. For this reason, the employed features are usually chosen taking into account their performance and robustness in the target task as well as other computational requirements.

Machine Learning Analysis

Once the facial features are computed, the last step is to design and train Machine Learning models in order to solve the target problems described in Sec. 1.1.1. Depending on whether the specific model takes into account the temporal information present in facial expressions, we can differentiate between static or dynamic approaches.

In static approaches, the model is aimed to predict expression labels on frame-by-frame basis. Specifically, for Discrete Expression Recognition, different binary or multi-class classification methods such as Support Vector Machines [Kotsia and Pitas, 2007], Boosting [Zhao and Pietikainen, 2007], Artificial Neural Networks [Tian, 2004] or Random Forests [El Meguid and Levine, 2014] have been explored. In the case of Expression Intensity Estimation or Affect Analysis, regression frameworks such as Relevance Vector Machines [Kaltwang et al., 2016], Ordinal Regression [Rudovic et al., 2012] or Gaussian Processes [Eleftheriadis et al., 2016] have been also used.

In contrast to the static case, dynamic approaches take into account the temporal information of gestures. These type of methods are more appealing in this context given the importance of dynamics in the interpretation of facial behavior [Ambadar et al., 2005]. To model temporal information, the vast majority of proposed solutions in the literature are based on Dynamic Bayesian Networks (DBNs) [Murphy and Russell, 2002]. DBNs are probabilistic graphical models modelling temporal dependencies between random variables (i.e, expression labels). Typical variants of DBNs used in facial expression analysis include Hidden Markov Models [Valstar and Pantic, 2007], Conditional Random Fields [Baltrušaitis et al., 2013] or Hidden Conditional Ordinal Random Fields [Kim and Pavlovic, 2010a].

It is worth to mention that recent works on Automatic Facial Behavior Analysis have attempted to combine the three described steps (i.e. Face prepossessing, Facial Feature Extraction and Machine Learning Analysis) employing end-to-end systems based on Deep Learning [Tősér et al., 2016, Zhao et al., 2016a]. Such an approach is out of the scope of this Thesis and we will follow the standard pipepline previously described. However, the contributions of the presented research are complementary to this new trend and, in Chapter 6, we provide a discussion about this issue.

1.2 Motivation: Weakly-Supervised Facial Behavior Analysis

During the last decade, research in Automatic Facial Behavior Analysis has mainly focused on proposing novel facial-features or Machine Learning methods. Typically, proposed approaches have been designed to answer questions such as: How to provide robustness to large head-pose variations? [Eleftheriadis et al., 2015]. What is the best methodology to model temporal dynamics? [Ding et al., 2016]. How to deal with individual differences among subjects in facial expression displays? [Chu et al., 2017]. Despite the advances in the field achieved by addressing these particular questions, most of these problems can be considered far from being solved. As a consequence, facial expression analysis methods have still not been extensively deployed in real-life applications as has been the case for other Computer Vision based systems.

Other than the previous questions, the presented thesis is motivated by the observation that most of the proposed Machine Learning methods in the field have followed the supervised-learning paradigm. Under this setting, models require to be trained using datasets explicitly labelled according to the target problem. For example, in Action Unit recognition, image sequences need to be annotated frame-by-frame according to binary labels indicating the presence of each Action Unit. In this scenario, it is reasonable to ask the following question: *What are the drawbacks of the supervised-learning strategy in the context of automatic Facial Behavior Analysis?*

It is well known that data annotation in AFBA is usually an expensive and time-consuming task. For example, labelling AU activations in one minute of video can require one hour for a specially trained coder [De la Torre et al., 2011]. As a consequence, standard databases are usually sub-optimal in terms of data variability and sample size. Therefore, is possible that limited training data may be decreasing the performance and generalization ability of learned models. In the literature, we can find multiple works providing evi-

Figure 1.7: Illustration of two classic Computer Vision problems which have been previously addressed using weakly-supervised approaches. Left: Image Semantic Segmentation (images extracted from the MRSC v2 dataset [Shotton et al., 2006]). Right: Object Dectection (examples obtained from the PASCAL VOC2011 Database [Everingham et al., 2015])

dences of this hypothesis. For example, in [Whitehill et al., 2009], it was shown that the performance of smile (Action Unit 12) detectors can be significantly increased by using larger datasets collected in naturalistic conditions. We find another example in [Girard et al., 2015], where exhaustive experiments revealed that subject variability in the training data plays an important role determining the quality of the learned models.

Apart from the cost of the labelling process, facial behavior annotations also suffers from low reliability. In Affect Analysis, for instance, annotations are inherently subjective even if they are performed by trained coders [Yannakakis and Martínez, 2015a]. Thus, a high inter-observer agreement is difficult to achieve while labelling datasets designed for this task. This is also common for other problems such as Action Unit recognition. In order to address this challenge, the standard solution consists in collecting annotations from multiple expert coders. However, this solution introduces an additional problem regarding how to obtain a more objective ground-truth from a pool of annotations in order to train supervised-models.

Given the described drawbacks of the supervised-learning strategy, the research presented in this thesis is motivated by the following question: *Can we address Facial Behavior Analysis problems by changing the fully-supervised paradigm by a weakly-supervised one?* Weakly-supervised learning has been explored in many different Computer Vision problems to alleviate the need of labelled data. In general terms, these type of models are trained by using labels which only provide partial information about the task that aims to be solved (i.e, weak-labels). In Object Detection, where the goal is to estimate the position of a given object in an image (Fig. 1.7), weakly-supervised approaches have achieved impressive performance when they are trained using only labels at image-level [Pandey and Lazebnik, 2011]. This supposes a huge advantage with respect to fully-supervised methods which require explicit annotations of the object bounding-boxes during learning [Azizpour and Laptev, 2012]. We find another example in Semantic Segmentation [Vezhnevets et al., 2011], where these type of methods are able to predict pixel-level labels (Fig. 1.7) by training them using only weak-annotations indicating the presence of the different semantic concepts in the image. Given the advantages of weakly-supervised learning in these scenarios, we aim to show that this paradigm is a potential solution to overcome the previously described limitations of fully-supervised approaches in the context of AFBA.

1.3 Contributions and Thesis Outline

Given the previously explained motivation, in this thesis we develop a set of weakly-supervised learning methods in order to address some of the Facial Behavior Analysis problems described in Sec. 1.1.1. We build the proposed approaches upon different technical frameworks such as Multiple Instance Learning [Amores, 2013] or Probabilistic Graphical Models [Barber, 2012]. These frameworks account for the particular idiosyncrasies of each problem and will be briefly

introduced in the corresponding chapter. Our main contributions, together with the outline of this manuscript, are detailed as follows:

- In **Chapter 2**, we explore weakly-supervised learning in the context of Discrete Expression Recognition. Specifically, we focus on a novel task which we refer as Facial Behavior Categorization. In this problem, the goal is to estimate high-level semantic labels for videos of recorded people by means of analyzing their facial expressions. Different from the standard supervised scenario, we do not have access to frame-by-frame annotations of discrete expression categories, but only weak-labels at the video level are available. Therefore, the goal is to automatically discover a set of discriminative gestures appearing in the sequences and how they determine the high-level labels. We show how Facial Behavior Categorization can be posed as a Multi-Instance-Learning (MIL) problem and we propose a novel method called Regularized Multi-Concept MIL to solve it. In contrast to previous approaches, RMC-MIL follows a Multi-Concept assumption which allows to discover multiple facial expressions (concepts) and how they determine the video weak-label. Moreover, to handle with the high-dimensional nature of facial-features, RMC-MIL uses a discriminative approach to model the concepts and structured sparsity regularization to discard non-informative features. In our experiments, we use two public data-sets to show the advantages of RMC-MIL in different Facial Behavior Categorization problems and to compare it with standard MIL methods previously applied in other domains.

- Also related with weakly-supervised Discrete Expression Recognition, in **Chapter 3**, we investigate how the use of large databases labelled only according to the six universal facial expressions can increase the quality of learned Action Unit classifiers. Our motivation is that most AU datasets are typically obtained in controlled laboratory conditions and have limita-

tions in terms of variability and positive samples. This is due to the tedious and expensive task involved in their annotation. In contrast, labelling large prototypical facial expression databases is much easier. In this context, we propose a novel weakly-supervised learning framework: Hidden-Task Learning. HTL aims to learn a set of Hidden-Tasks (Action Units) for which samples are not available but, in contrast, training data is easier to obtain from a set of related Visible-Tasks (Facial Expressions). To that end, HTL exploits prior knowledge about the relation between Hidden and Visible-Tasks. In our case, we base this prior knowledge on empirical psychological studies providing statistical correlations between Action Units and universal facial expressions. Additionally, we extend HTL to Semi-Hidden Task Learning (SHTL) assuming that Action Unit training samples are also provided. Performing exhaustive experiments over four different datasets, we show that HTL and SHTL improve the generalization ability of AU classifiers by training them with additional facial expression data. Additionally, we show that SHTL achieves competitive performance compared with previous Transductive Learning approaches which face the problem of limited training data by using unlabelled test samples during training.

- In **Chapter 4**, we address Facial Expression Intensity estimation from a weakly-supervised perspective. Specifically, we focus on a novel problem which we refer as Multi-Instance Dynamic Ordinal Regression. In this task, the goal is to predict an ordinal label (expression intensity) for each instant of a sequence (image frame). The weakly-supervised setting is given because no frame-by-frame annotations are available during training. In contrast, a single label provides weak-information about the set of intensity levels within the sequence (e.g the maximum expression intensity within the video frames). To address this problem, we propose Multi-Instance

Dynamic Ordinal Random Fields (MI-DORF). In this framework, frame-labels are treated as temporally-dependent latent variables in a graphical model. The weak-information provided by sequence-labels is modelled by incorporating a high-order potential into the model energy function. Moreover, we extend the proposed framework for Partially-Observed MI-DOR problems, where a subset of frame intensity labels can be also available during training. We show that the proposed framework significantly outperforms alternative approaches in the tasks of weakly-supervised Action Unit and Pain Intensity estimation.

- In **Chapter 5**, we address the problem of fusing manual annotations from multiple observers when labels are given in an ordinal scale and annotated items are structured as temporal sequences. This problem is of special interest in Affect Analysis, where collected data is typically formed by videos of human interactions where frames are annotated according to the Valence and Arousal (V-A) dimensions. Moreover, different works have shown that inter-observer agreement of V-A annotations can be considerably improved if these are given in a discrete ordinal scale. Note that annotation fusion can be considered a weakly-supervised learning problem given that we aim to estimate a common ground-truth (main task) from a set of subjective labels from multiple observers (weak-labels). In this context, we propose a novel probabilistic framework which explicitly introduces ordinal constraints to model the subjective perception of annotators. We also incorporate dynamic information to take into account temporal correlations between ground-truth labels. In our experiments on synthetic and real data with V-A annotations, we show that the proposed method outperforms alternative approaches which do not take into account either the ordinal structure of labels or their temporal correlation.

  Finally, in **Chapter 6** we conclude this thesis by giving some final remarks and pointing out potential future research lines.

1.4 Publications

The research developed during this thesis has resulted in the following list of publications:

Journals

1. **A. Ruiz**, , O. Rudovic, X. Binefa, M. Pantic, "Multi-Instance Dynamic Ordinal Random Fields for Weakly-supervised Facial Behavior Analysis", *IEEE Transactions on Image Processing*, *(Under Review)*.

International Conferences

1. **A. Ruiz**, O. Martinez, X. Binefa, F. Sukno, "Fusion of Valence and Arousal Annotations through Dynamic Subjective Ordinal Modelling", *International Conference on Automatic Face and Gesture Recognition*, 2017.

2. **A. Ruiz**, O. Rudovic, X. Binefa, M. Pantic, "Multi instance Dynamic Ordinal Random Fields for weakly-supervised pain Intensity Estimation", *Asian Conference on Computer Vision*, 2016, *(Oral Presentation)*.

3. **A. Ruiz**, J. van de Weijer, X. Binefa, "From Emotions to Action Units with Hidden and Semi-Hidden-Task Learning", *International Conference on Computer Vision*, 2015.

4. **A. Ruiz**, J. van de Weijer, X. Binefa, "Regularized Multi-Concept MIL for weakly supervised facial behavior categorization", *Brittish Machine Vision Conference*, 2014, *(Oral presentation)*.

International Workshops

1. D. Derkach, **A. Ruiz**, F. Sukno, "Head Pose Estimation Based on 3-D Facial Landmarks Localization and Regression", *In FG 2017 Workshop on Dominant and Complementary Emotion Recognition Using Micro Emotion Features and Head-Pose Estimation*, 2017,*(Winner of the Head Pose Estimation Challenge)*

2. F. Sukno, M. Dominguez, **A. Ruiz**, D. Schiller, F. Lingenfelser, L. Pragst, E. Kamateri, S. Vrochidis, "A multimodal annotation schema for non-verbal affective analysis in the health-care domain", *International Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction*, 2016.

Chapter 2

REGULARIZED MULTI-CONCEPT MIL FOR WEAKLY-SUPERVISED FACIAL BEHAVIOR CATEGORIZATION

2.1 Introduction and motivation

As introduced in Chapter 1, the vast majority of research addressing Discrete Expression Recognition has focused on designing supervised-learning models for this task. Following this approach, models are trained using datasets labelled according to a set of predefined gesture categories (e.g. Action Units), whose annotation is typically a laborious and expensive task. Opposite to the supervised-learning strategy, in this Chapter we focus on a related weakly-supervised problem which we call Facial Behavior Categorization. To illustrate it, consider a set of videos of people recorded in a given context, e.g.

watching an advertisement. For each of these videos we know a high-level semantic label related with this context: Did he/she like the advertisement? The task in Facial Behavior Categorization is to analyze the subject facial behavior during the whole recording and estimate the "Like/Not Like" label. This problem can be considered a weakly-supervised learning task because frame-level annotations of gestures are not available during training. In contrast, only a high-level label at the video-level is provided and the goal is to automatically learn and recognize the set of expression categories determining it. For instance, in the previously described scenario, the model should learn to recognize smiles as an expression revealing whether the subject liked or not the advertisement. Learning to recognize facial expressions categories through Facial Behavior Categorization has a relevant advantage with respect to the traditional fully-supervised approach. In this case, the model automatically learns to interpret facial behavior by using only the context weak-labels which, for many applications, are much easier to obtain than frame-by-frame expression annotations.

2.2 Facial Behavior Categorization as Multiple Instance Learning

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 $\mathbf{x}_{ij} \in \mathbb{R}^D$. 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 $\mathbf{x}_{ij}$ correspond to facial-features 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 meth-

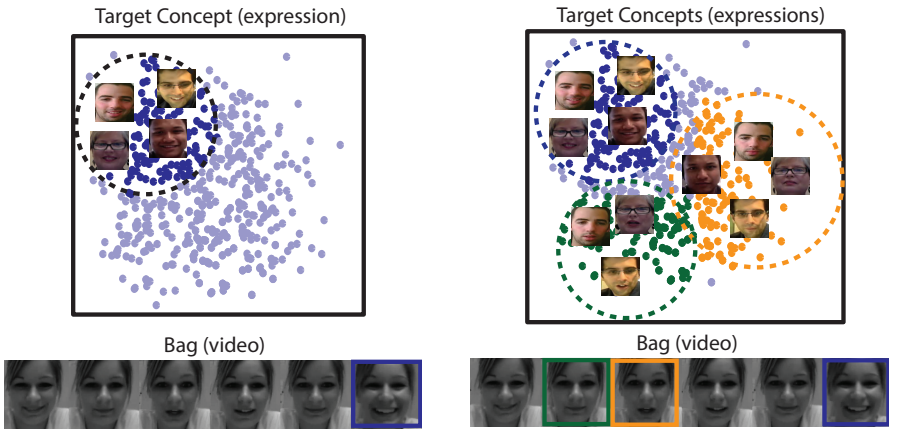

Figure 2.1: Illustration of the MIL Single-Concept (left) and Multi-Concept (right) assumptions in the context of Facial Behavior Categorization. The Single-Concept approach defines an unique expression whose presence in a video determines the video weak-label. On the other hand, the Multi-Concept assumption is able to take into account different expressions which can contribute differently to the estimation of the video label.

ods assume that there exist an underlying relation between the bag label and its instances distribution [Foulds and Frank, 2010]. In this work, we differentiate between Single-Concept and Multi-Concept MIL methods (see Fig. 2.1).

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 determined by the maximum probability of this concept given its constituent instances. In general Facial Behavior Analysis problems, Single-Concept approaches have been applied to supervised AU localization [Tax et al., 2010] and weakly-supervised pain detection [Sikka et al., 2013]. However, for more general Facial Be-

havior Categorization problems, the Single-Concept assumption does not take into account that different expressions categories can appear during the video and 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, standard 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 contrast, facial-features (instances) are typically highly dimensional and contain a low number of informative features related with facial expression changes [Zhong et al., 2012].

2.3 Contributions

In this Chapter, we address the introduced Facial Behavior Categorization problem by proposing a novel MIL method called Regularized Multi-Concept MIL (RMC-MIL). The novelties of the presented model are summarized as follows:

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

- **Structured Sparsity Regularization:** RMC-MIL applies $L_{2,1}$-norm regularization over the concept-classifiers. This reg-

ularization forces common sparsity across them and, as a consequence, they only use a common subset of dimensions belonging to the high-dimensional facial-features. These dimensions are expected to be related with the information regarding facial expressions. 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 constrained optimization problem where all the parameters are jointly learned and efficiently solved using the Projected-Quasi-Newton method [Schmidt et al., 2009]. In our experiments, we test the proposed method in two different Facial Behavior Categorization problems to show the advantages of our Regularized Multi-Concept approach. RMC-MIL achieves better performance than previously proposed Single-Concept and Multi-Concept MIL methods.

2.4 Related work on Multiple Instance Learning

Most of MIL research approaches follow the Single-Concept assumption. The various methods differ on how the target-concept is obtained from the training set. Diverse-Density [Maron and Lozano-Pérez, 1998] model it as a Gaussian and learn its mean and diagonal covariance using gradient-descent optimization. Bayesian-MIL [Raykar et al., 2008] adapts Logistic Regression to MIL and incorporates a prior over the parameters in order to perform feature selection. MM-MIL [Wang et al., 2012] uses a mixture of linear classifiers to represent a muti-modal target-concept. Other approaches reformulate standard supervised methods such as AnyBoost [Zhang et al., 2005a], SVM [Andrews et al., 2002, Bunescu and Mooney, 2007], Gaussian Processes [Kim and Torre, 2010] or Random Forests [Leistner et al., 2010] and adapt them to the MIL assumption. As discussed above, Single-Concept approaches can not model that different con-

cepts (expressions) can appear inside a bag (video) and that they can contribute differently 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 probability that the k-th concept appear in the bag by following the Single-Concept assumption. In a seminal work, DD-SVM [Chen and Wang, 2004] 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 [Chen et al., 2006] considers all the training instances in the data-set as potential prototypes and selects the most relevant with $l1$-norm SVM. MILIS [Fu et al., 2011] uses a coordinate descent procedure to iteratively learn the most relevant prototypes and the bag-classifier. More recently, [Hong et al., 2014] proposed an algorithm based on AdaBoost to select a set of prototypes from different information sources.

2.5 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 2.5.1 and then we extend it to the regularized case in 2.5.2.

2.5.1 MC-MIL

Let us denote $\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \dots \ \mathbf{z}_K]$ as a $\mathbb{R}^{D \times K}$ 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 $p(\mathbf{z}_k|\mathbf{x}_{ij}) = \sigma({\mathbf{z}_k}^T \mathbf{x}_{ij})$ where $\sigma(s)$ corresponds to the sigmoid function.

Following the standard MIL assumption, the probability of a concept k given a bag X_i is defined as $p(\mathbf{z_k}|X_i) = \max_j p(\mathbf{z}_k|\mathbf{x}_{ij})$. Since max(.) is not differentiable, we approximate it using the Generalized Mean (GM) function defined as:

$$p(\mathbf{z_k}|X_i) = \left(\sum_{j=1}^{M} p(\mathbf{z}_k|\mathbf{x}_{ij})^r \right)^{\frac{1}{r}} \tag{2.1}$$

GM have been previously used in MIL methods [Sikka et al., 2013] 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:

$$g(X_i, \mathbf{Z}) =< p(\mathbf{z}_1|X_i), p(\mathbf{z}_2|X_i), \dots, p(\mathbf{z}_K|X_i) >^\intercal . \tag{2.2}$$

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 concept k given the bag i. 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 $\mathbf{w} = [w_1, w_2, \dots, w_K]$ are the parameters of a linear classifier separating positive and negative bags embedded in the K dimensional space. Figure 2.2 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 classifiers $\mathbf{Z}$ 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:

$$\min_{\mathbf{w},\mathbf{Z}} \quad \sum_{i=1}^{N} \ell(\mathbf{w}^T g(X_i, \mathbf{Z}), y_i) = -\sum_{i=1}^{N} y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \quad (2.3)$$

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.

2.5.2 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 [Ng, 2004]. For this purpose, we introduce in Eq. 2.3 a regularizer over $\mathbf{Z}$:

$$\min_{\mathbf{w},\mathbf{Z}} \quad \mathcal{L}(\mathcal{T}, \mathbf{Z}, \mathbf{w}) = \sum_{i=1}^{N} \ell(\mathbf{w}^T g(X_i, \mathbf{Z}), y_i) + \lambda \Omega_{\mathbf{Z}}(\mathbf{Z}) \quad (2.4)$$

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 $\Omega_{\mathbf{Z}}(\mathbf{Z}) = ||\mathbf{Z}||_{2,1}$

It is known that $L_{2,1}$-norm encourages sparsity across the rows of $\mathbf{Z}$ [Argyriou et al., 2007]. The use of structured sparsity regularization is motivated by a previous work [Zhong et al., 2012] 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 classifiers to use a common subset of features expected to be related with facial expression changes.

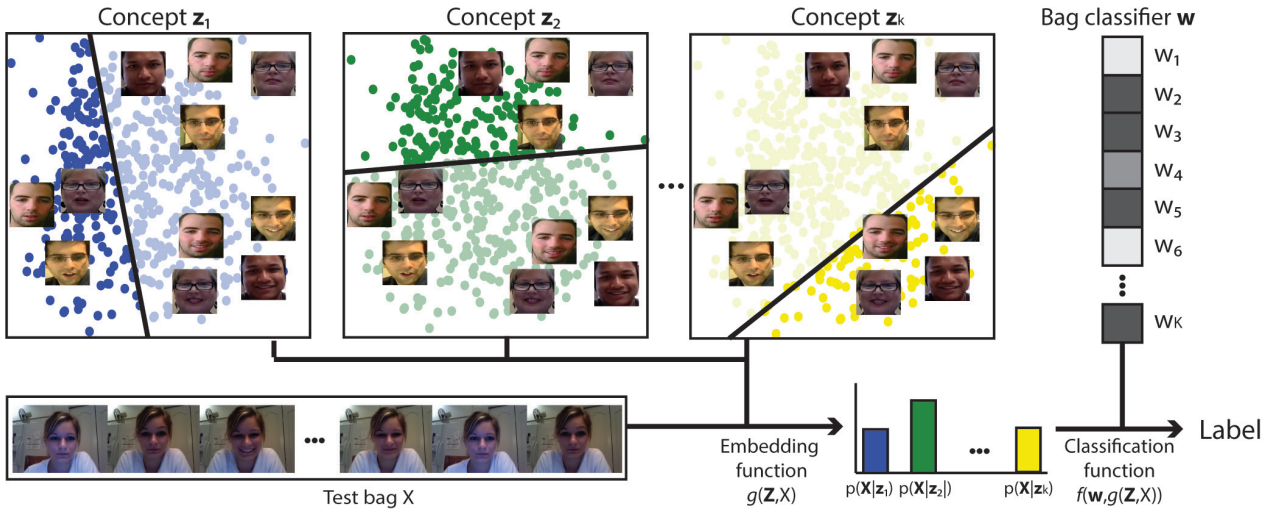

Figure 2.2: Overview of the proposed Multi-Concept MIL method. Concepts are modelled as a set of K linear classifiers $\mathbf{z_k}$ in instance space. Given a bag, it is represented using the probability of each concept given its instances. The bag-classifier $\mathbf{w}$ maps this bag-representation into high-level labels. Both $\mathbf{Z}$ and $\mathbf{w}$ parameters are jointly optimized during training.

2.5.3 RMC-MIL optimization

In order to efficiently minimize Eq. 2.4 including the loss and the non-smooth $L_{2,1}$ regularization terms, we propose to use the Projected Quasi-Newton method (PQN) [1] presented in [Schmidt et al., 2009]. In 2.5.3 we briefly describe PQN and in 2.5.3 we explain how we apply it to RMC-MIL. It is worth mentioning that Eq. 2.4 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 [Li and Sminchisescu, 2010], and local-optimal solutions are shown to achieve good results.

Projected Quasi-Newton Method

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

$$\min_x f(x) \quad s.t \;\; x \in \mathcal{C} \tag{2.5}$$

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

$$q_k(x) \triangleq f(x_k) + (x - x_k)^T \nabla f(x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k) \tag{2.6}$$

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 [Byrd et al., 1994] (see also Appendix A.1) 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 (2.6), PQN finds a better x_{k+1} by solving:

$$x_{k+1} = \min_x q_k(x) \;\; s.t \;\; x_k \in C \tag{2.7}$$

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

This sub-problem is solved by using Spectral Projected Gradient (SPG) [Birgin et al., 2000] which computes the solution to Eq. 2.7 with a gradient descent approach but, at each iteration, the solution x is projected into the convex set $\mathcal{C}$ using a projection function $\mathcal{P}_{\mathcal{C}}(x)$:

$$\mathcal{P}_{\mathcal{C}}(x) = \min_{c} ||c - x||_2 \;\; s.t \;\; c \in \mathcal{C} \tag{2.8}$$

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 [Schmidt et al., 2009], the PQN method is particularly interesting when we are minimizing a function such as Eq. (2.4) 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.

RMC-MIL optimization via PQN method

In order to solve RMC-MIL optimization, we firstly reformulate Eq. 2.4 as the following equivalent constrained-convex optimization problem:

$$\min_{\mathbf{w},\mathbf{Z}} \quad \mathcal{L}(\mathbf{w}, \mathbf{Z}) = \sum_{i=1}^{N} \ell(\mathbf{w}^T g(X_i, \mathbf{Z}), y_i) \;\; s.t. \; ||\mathbf{Z}||_{2,1} \leq \tau_Z \tag{2.9}$$

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:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = -\sum_{i=1}^{N} (y_i - p_i) g(X_i, \mathbf{Z}) \tag{2.10}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}_k} = -\sum_{i=1}^{N} \frac{w_k}{M_i}(y_i - p_i)\left(\frac{1}{M_i}\sum_{j=1}^{M} p_{ijk}\right)^{\frac{1}{r}-1} \sum_{j=1}^{M_i} p_{ijk}^{r}(1-p_{ijk})\mathbf{x_{ij}} \quad (2.11)$$

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. 2.5.3. 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 [Schmidt et al., 2009].

2.6 Experiments

Other than existing work on Facial Behavior Analysis [Tax et al., 2010, Sikka et al., 2013] 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-features. In this section, we firstly describe the facial-features and the data-sets used in our experiments. In 2.6.2 we analyze the impact of the number of concepts and the regularization term on RMC-MIL. In 2.6.3, we compare our approach with previously proposed Single-Concept and Multi-Concept MIL methods. Finally, we illustrate the ability of RMC-MIL to discover discriminative facial expressions from weakly-labelled videos.

2.6.1 Datasets and experimental setup

Facial-features: Given a video (bag), we extract a facial-descriptor (instance) for each frame. The whole process is illustrated in Figure 2.3. Firstly, we obtain a set of 49 landmark facial-points with the Supervised Descent method described in [Xuehan-Xiong and De la Torre, 2013]. 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 [Scovanner et al., 2007] [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 [Chu et al., 2013]. 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.

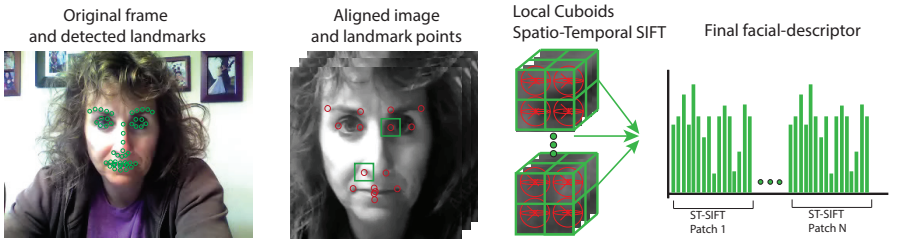

Figure 2.3: (i) 49 extracted landmark points. (ii) image aligned with the obtained affine transformation (ii) Spatial-Temporal SIFT descriptors extracted from each local cuboid. Red points corresponds to the subset of landmarks used.

AM-FED: The AM-FED dataset [McDuff et al., 2013b] 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 [McDuff et al., 2013a], 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 cor-

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

responding 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 [Lucey et al., 2011] 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 pain). The work in [Sikka et al., 2013] 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.

2.6.2 Multiple Concepts and Structural Sparsity Regularization for Facial Behavior Categorization

In this experiment, we investigate 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-classifiers computed as the Gini Coefficient [Hurley and Rickard, 2009] over the $L2$-norms of Z rows. This measure indicates how many features have a very low contribution defining the concepts.

Figure 2.4 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.

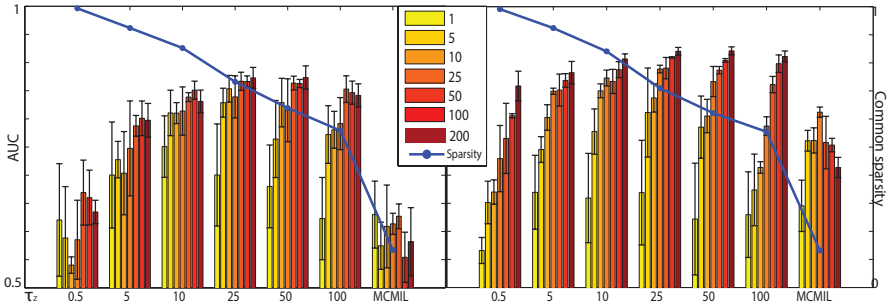

Figure 2.4: AUC obtained by RMC-MIL and MC-MIL in "UNBC-Pain/No pain" (left) and "AM-FED-Watch/Not watch again" (right) problems. Bar colors indicates the number of concepts used and X axis refers to different values for τ_Z . Blue line corresponds to the mean common sparsity coefficient for all K given 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-classifiers. 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.

2.6.3 Comparison with other MIL methods

In this experiment, we compare the performance of RMC-MIL with four popular MIL methods: MilBoosting [Zhang et al., 2005a], MI-Forest [Leistner et al., 2010], MILES [Chen et al., 2006] and MILIS [Fu et al., 2011]. 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 applied to weakly-supervised pain detetion [Sikka et al., 2013] 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 [Sikka et al., 2013], 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 [Sikka et al., 2013]. For MI-Forest and RMC-MIL the results are computed as the mean obtained over five different runs.

As the reader can observe, RMC-MIL achieves better performance in all the problems. Given these results and the reported in Sec. 2.6.2, our hypothesis is that RMC-MIL outperforms Single-Concept

	MILES	MILIS	MilBoosting	MI-Forest	MS-MIL	RMC-MIL
AM-FED: Like	0.62	0.63	0.61	0.68	-	**0.72**
AM-FED: Watch again	0.76	0.73	0.83	0.78	-	**0.87**
UNBC: Pain	0.85 / 78.2	0.82 / 76.9	0.78 / 76.9	0.81 / 75.8	- / 83.7	**0.92 / 85.7**

Table 2.1: Results obtained by Multi-Concept, Single-Concept MIL methods and RMC-MIL in the AM-FED and UNBC data-sets. See text for details

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-features. 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 [Sikka et al., 2013] 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-features compare favorably to their approach.

2.6.4 Applying RMC-MIL to discover discriminative facial expressions

To provide more insights 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 2.5 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.

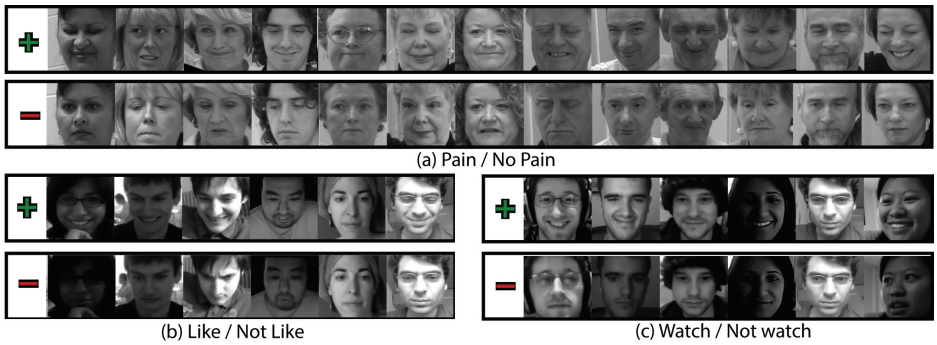

Figure 2.5: Most positive and negative instances estimated by RMC-MIL in a set of randomly selected videos for the different facial behavior categorization problems

2.7 Summary

In this Chapter, we have addressed Facial Behavior Categorization. Different from the fully-supervised approach for Discrete Expression Recognition, in this problem the model needs to learn and recognize different facial expression categories by using only the information

provided by high-level labels at video-level. We have shown that Facial Behavior Categorization can be naturally posed as a Multiple Instance Learning problem and we have presented a novel method to address it: Regularized-Multi-Concept MIL. Other than previous MIL methods used in the context of Facial Behavior Analysis, RMC-MIL does not follow a Single-Concept assumption. This allows to learn multiple discriminative facial expressions and how they determine the video label. Moreover, in contrast to existing Multi-Concept MIL methods, RMC-MIL can learn more optimal concepts from high-dimensional facial-features by using structured sparsity regularization. We have evaluated the proposed method in two different Facial Behavior Categorization problems. Specifically, we have considered the scenario where video sequences were labelled according to the reaction of people while watching an advertisement. On the other hand, we have also evaluated RMC-MIL when video-labels are related with the absence or presence of pain during the recording. In our experiments, we have shown the improvement of RMC-MIL over existing Single-Concept and Multi-Concept MIL methods in these problems, and its ability to learn discriminant facial expressions from weakly-labeled videos. Future work may be focused on exploring other applications which could be naturally posed as Facial Behavior Categorization problems. For example, it has been shown that micro-expressions [Pfister et al., 2011] provide a reliable cue to detect deception [Warren et al., 2009]. These micro-expressions are very subtle, sparse and typically occur in fractions of a second. However, posing deception detection as a Facial Behavior Categorization problem, would not require to collect databases with annotated micro-expressions. In contrast, the proposed model would be able to automatically discover these subtle gestures only from videos labelled as "deceptive" or "non deceptive".

Chapter 3

FROM EMOTIONS TO ACTION UNITS WITH HIDDEN AND SEMI-HIDDEN TASK LEARNING

3.1 Introduction and Motivation

As discussed in Chapter 1, many works on Facial Behavior Analysis have focused on the recognition of the 6 Universal Facial Expressions. This problem is motivated by the studies of the psychologist Paul Ekman who showed that there exist 6 universal emotions (anger, happiness, fear, surprise, sadness, and disgust) and that each of them has a corresponding prototypical facial expression [Ekman and Friesen, 1971]. Despite their cross-cultural universality, it has been demonstrated that people can perform many other non-basic expressions and that their combination are usual in our every-day life [Du et al., 2014]. For these reasons, a more objective method to categorize expressions is the Facial Action Coding System (FACS) [Ekman and

Friesen, 1978]. In FACS, Ekman defined a set of 45 Action Units which are atomic gestures caused by the activation of one or more facial muscles. Since any expression that humans can do can be characterized by a concrete combination of Action Units, its automatic recognition is one of the most interesting problems in Facial Behavior Analysis.

Action Unit recognition is a challenging problem due to different factors such as illumination changes, pose variations or individual subject differences. All these factors cause large variations in the appearance of the same Action Unit across different face images. However, AU annotation is an expensive and laborious task even for expert coders. As a consequence, collected Action Unit datasets are typically obtained in controlled laboratory conditions and have limitations in terms of variability and positive samples. In this Chapter, we aim to address Action Unit recognition from a weakly-supervised perspective by asking the following question: *Can we use additional samples labelled with the six prototypical facial expressions in order to learn better Action Unit classifiers?*. The motivation behind this question is twofold:

- Firstly, the recognition of universal expressions and Action Units can be considered closely related problems. Many psychological studies have empirically shown their strong relation [Lewis et al., 2010]. For instance, Ekman developed the EMFACS dictionary [Friesen and Ekman, 1983], a set of rules mapping Action Unit activation patterns to emotions. Other studies have shown that the expression of a given emotion does not always follow a fixed pattern but that there exist a statistical correlation with concrete Action Unit activations [Gosselin et al., 1995],[Scherer and Ellgring, 2007].

- Secondly, most works addressing AU recognition have followed the supervised- learning paradigm. However, we hypothesize that the limited data in current AU datasets can decrease the performance and generalization ability of the learned models.

Even though one solution would be to add larger and more varied data sets, this approach is not practical given the expense of the annotation process. In contrast, collecting universal facial expression databases is much easier. For instance, the FER2013 Challenge Dataset [Goodfellow et al., 2013] provides thousands of facial expression samples semi-automatically collected from the Google Images search engine. Moreover, facial expression annotations does not require expert coders as in the case of Action Units. Therefore, ground-truth labels for larger and more varied facial expression datasets are much more easy to obtain compared to Action Units annotations.

3.2 Contributions

Given the previous described motivation, the contributions of the presented Chapter are summarized as follows:

- We propose a novel weakly-supervised learning framework called Hidden-Task Learning (HTL), that allows to learn a set of Hidden-Tasks when no annotated data is available. For this purpose, HTL exploits prior knowledge about the relation between these Hidden-Tasks and a set of Visible-Tasks for which annotations are provided. Additionally, we extend HTL to Semi-Hidden-Task Learning (SHTL) which is able to use additional training samples belonging to the Hidden-Tasks.

- We show how HTL and SHTL can be used to improve the generalization ability of Action Unit classifiers (Hidden-Tasks) by using additional training data labelled according to prototypical facial expressions (Visible-Tasks). The prior knowledge defining the relation between the AU and Facial Expression recognition tasks is based on empirical results of psychological studies [Gosselin et al., 1995]. Even though previous work has used this knowledge for facial expression analysis [Valstar and

Pantic, 2006], to the best of our knowledge, this is the first work which exploits it in order to investigate how additional training data of facial expressions can be used to learn better AU classifiers. An overview of our method is provided in Fig.3.1.

- Performing exhaustive experiments over four different Action Unit databases, our results demonstrate that using SHTL, we can improve AU recognition performance by using additional data from Facial Expression Datasets. In cross-database experiments, HTL generally achieves better performance than standard Single-Task-Learning even when no Action Unit annotations are used. Moreover, SHTL achieves competitive results compared with Transductive Learning approaches which use test data during training in order to learn personalized models for each subject. Our results suggest that the limitation of training data in AU recognition is an important factor which can be effectively addressed with the proposed HTL and SHTL frameworks.

3.3 Related Work

Action Unit recognition: As discussed in Chapter 1, most works on AU recognition have focused on proposing different types of facial-descriptors and classification models. Popular descriptors are based on LBP [Jiang et al., 2011], SIFT [Chu et al., 2013], Active Appearance Models [Lucey et al., 2009] or face-geometry [Pantic and Patras, 2006] features. On the other hand, different classifiers based on SVM [Mahoor et al., 2009], AdaBoost [Yang et al., 2007] or HMM [Valstar and Pantic, 2012] have been used to recognize Action Units in images or sequences. However, these approaches do not explicitly face the problem of limited training data in Action Unit recognition. In this Chapter, we show that using simple linear classifiers and standard facial-features, the proposed HTL and SHTL frameworks can increase the generalization ability of AU classifiers (Hidden-Tasks)

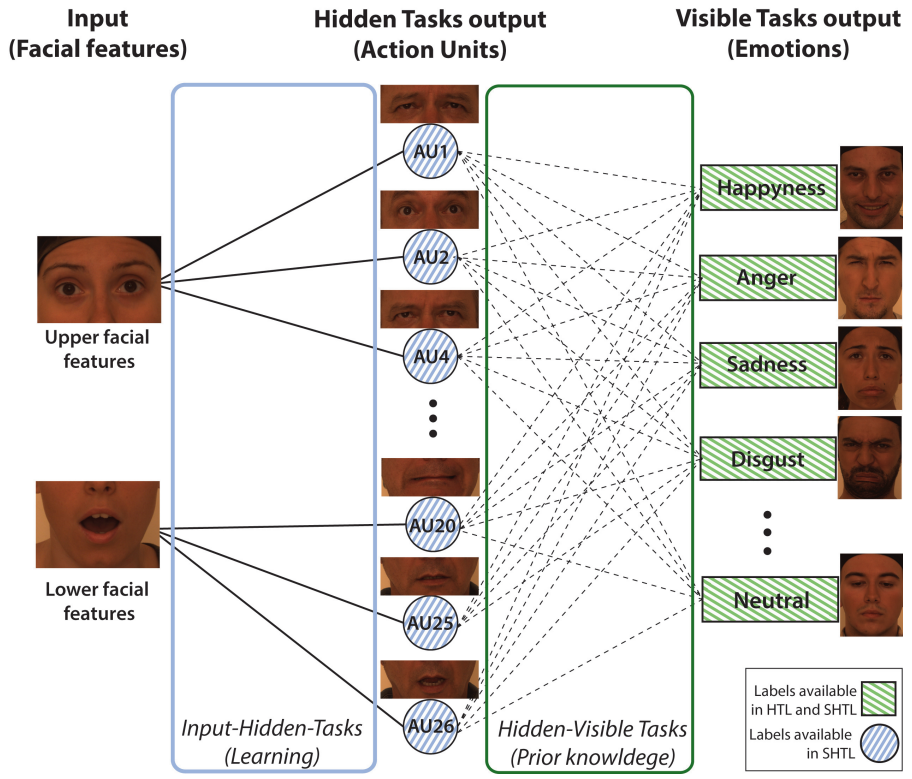

Figure 3.1: Hidden-Task Learning and Semi-Hidden-Task Learning frameworks applied to Action Unit recognition. HTL aims to learn AU classifiers (Hidden-Tasks) by using only training samples labelled with universal facial expressions (Visible-Tasks). For this purpose, HTL exploits prior knowledge about the relation between Hidden and Visible-Task outputs. In this Chapter, the relation between Action Unit and facial expressions is modelled based on empirical results obtained in psychological studies. SHTL is an extension of HTL assuming that samples from the Hidden-Tasks (Action Units) can also be provided. We show that the use of additional facial expression training samples increases the generalization ability of the learned AU classifiers.

by just providing additional training samples labelled with facial expressions (Visible-Tasks).

Transductive learning for AU recognition: Individual subject differences suppose one of the main challenges in Action Unit recognition . For instance, [Girard et al., 2015] showed that the variability of subjects in the training set plays an important role determining the generalization ability of learned models. Therefore, the limited number of subjects in current databases complicates the learning process. In order to address this problem, some works have used Transductive Learning to train personalized AU classifiers by using unlabelled data from the test subject. [Chu et al., 2013] proposed a method called Selective Transfer Machine. STM learns a penalized SVM by weighting training samples according to their similarity to unlabelled test data. Similarly, Transductive Parameter Transfer [Sangineto et al., 2014, Zen et al., 2014] learns a mapping from the sample distribution of the test subject to the parameters of a personalized AU classifier. Note that Transductive Learning can be considered an opposite solution to ours. Instead of training specific models for each subject, our approach can use samples from additional subjects present in the facial expressions data in order to learn more generic AU classifiers. Although Transductive Learning approaches have achieved promising results, they are limited in real applications where training classifiers for each subject in testing time is not practical.

Combining AU with Facial Expressions: Exploiting the relation between Action Units and Facial Expressions has been previously explored in the field. Some works have considered to classify expressions by using Action Unit information. For instance, [Valstar and Pantic, 2006] proposed to use a set of rules based on the EMFACS dictionary in order to recognize facial expressions from estimated AU outputs. Similarly, [Velusamy et al., 2011] used the Longest Common Subsequence algorithm in order to classify expressions by measuring the similarity between Action Unit patterns in testing and training images. Our work differs from these approaches because we do not

use this relation for facial expression recognition but we use it to learn better AU classifiers. Following this idea, some other works have used probabilistic graphical models such as Restricted Boltzmann Machines [Wang et al., 2013b] or Partially-Observed HCRF [Chang et al., 2009] in order to include facial expression annotations during AU classifiers learning. However, these approaches use samples labelled with both facial expressions and Action Units requiring even more annotation effort. Therefore, they can not be used in order to evaluate how additional training data from facial expression databases can improve Action Unit recognition.

3.4 Hidden Task Learning and Semi- Hidden Task Learning

Hidden-Task and Semi-Hidden-Task Learning are general purpose frameworks. They can be used in problems where we want to learn a set of Hidden-Tasks for which training data is limited but training samples are easier to obtain from a set of related Visible-Tasks. Note that we consider the set of Hidden and Visible-Tasks disjoint. The use of additional training data from the Visible-Tasks is expected to increase Hidden-Tasks performance. In this section, we formalize the proposed frameworks.

3.4.1 Hidden-Task Learning

In HTL, we are provided with a training set $\mathbf{X}^v = \{(\mathbf{x}_1^v, \mathbf{y}_1^v), (\mathbf{x}_n^v, \mathbf{y}_n^v), ..., (\mathbf{x}_N^v, \mathbf{y}_N^v)\}$. Each $\mathbf{x}_n \in \mathbb{R}^d$ represents the sample features and $\mathbf{y}_n^v = [y_{n1}^v, y_{nk}^v, ..., y_{nK}^v] \in \{0,1\}^K$ is a vector indicating its label for a set of K binary Visible-Tasks. Using $\mathbf{X}^v$, our goal is to learn a set of T Hidden-Tasks for which training data is not provided.

We denote a Hidden-Task t as a function $\mathbf{h}(\mathbf{x}, \theta_t)$ mapping a feature vector $\mathbf{x}$ to an output according to some parameters θ_t. Given the set of task parameters $\Theta = \{\theta_1, \theta_t, ..., \theta_T\}$, we define the Input-

Hidden-Task function:

$$\mathbf{H}(\mathbf{x}, \Theta) =< \mathbf{h}(\mathbf{x}, \theta_1), \mathbf{h}(\mathbf{x}, \theta_t), ..., \mathbf{h}(\mathbf{x}, \theta_T) >^T, \tag{3.1}$$

mapping $\mathbf{x}$ to a vector containing the outputs of all the T Hidden-Task.

Similarly to Θ, we denote $\Phi = \{\phi_1, \phi_k, ..., \phi_K\}$ as a set of parameters for the K Visible-Tasks. For a given ϕ_k, the Hidden-Visible-Task function $\mathbf{v}(\mathbf{H}(\mathbf{x}_n, \Theta), \phi_k)$ maps $\mathbf{H}(\mathbf{x}_n, \Theta)$ to the output for the Visible-Task k. We assume that Φ can be obtained before the training stage by exploiting prior knowledge about the relation between Hidden and Visible-Task outputs (see Sec. 3.5.2 for the case of Action Unit and Facial Expressions recognition tasks)

Given the previous definitions, HTL aims to learn the optimal Hidden-Task parameters Θ by minimizing:

$$\min_{\Theta, \mathbf{X}^v} \ \mathcal{L}^v(\Theta, \mathbf{X}^v) + \beta \mathcal{R}(\Theta). \tag{3.2}$$

Here, $\mathcal{R}(\Theta)$ refers to a regularizer over the parameters Θ preventing over-fitting, $\mathcal{L}^v$ is the empirical-risk over the Visible-Task training set $\mathbf{X}^v$ defined in Eq. 3.3 and ℓ can be defined as any classification loss function. The parameter β controls the impact of the regularization term.

$$\mathcal{L}^v(\Theta, \mathbf{X}^v) = \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \ell(\mathbf{v}(\mathbf{H}(\mathbf{x}_n^v, \Theta), \phi_k), y_{nk}^v), \tag{3.3}$$

Note that HTL shares some relation with weakly-supervised structured learning [Vezhnevets et al., 2012]. In our case, the goal is to learn a set of Hidden-Tasks predicting a latent structured output $\mathbf{H}(\mathbf{x}, \Theta)$ by using only the visible weak-labels y^v. As discussed, HTL is able to solve this problem by pre-defining the relation between Hidden and Visible-Tasks based on prior knowledge.

3.4.2 Semi-Hidden Task Learning

In SHTL, we assume that additional training data for the Hidden-Tasks is provided. Similarly to $\mathbf{X}^v$, we denote $\mathbf{X}^h = \{(\mathbf{x}_1^h, \mathbf{y}_1^h), (\mathbf{x}_m^h, \mathbf{y}_m^h), ..., (\mathbf{x}_M^h, \mathbf{y}_M^h)\}$ as a training set of M samples where $\mathbf{y}_m^h \in \{0,1\}^T$ indicates the sample class label for each Hidden-Task t. Following the definitions in the previous section, now we are interested in learning the optimal parameters Θ by minimizing:

$$\min_{\Theta} (1-\alpha)\mathcal{L}^h(\Theta, \mathbf{X}^h) + \alpha\mathcal{L}^v(\Theta, \mathbf{X}^v) + \beta\mathcal{R}(\Theta) \tag{3.4}$$

where $\mathcal{L}^h(\mathbf{\Theta}, \mathbf{X}^h)$ represents the empirical-risk function over the Hidden-Task training set $\mathbf{X}^h$:

$$\mathcal{L}^h(\mathbf{\Theta}, \mathbf{X}^h) = \frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}\ell(\mathbf{h}(\mathbf{x}_m^h, \theta_t), y_{mt}^h). \tag{3.5}$$

The parameter $\alpha \in [0,1]$ controls the trade-off between the minimization of the Hidden-Task and Visible-Task losses. Concretely, note that when $\alpha = 1$ the minimization is the same as HTL. In contrast, when $\alpha = 0$, SHTL is equivalent to learning the Hidden-Tasks without taking into account the Visible-Tasks training data, i.e., traditional Single-Task Learning (STL). Therefore, SHTL can be considered a generalization of both HTL and STL.

An interesting interpretation of SHTL is to understand the term $\alpha\mathcal{L}^v(\Theta, \mathbf{X}^v)$ in Eq. 3.4 as a regularization function. Concretely, it penalizes cases where the Hidden-Task-outputs in x^v are not coherent with its label y^v according to the known relation between Hidden and Visible tasks. To the best of our knowledge, this is a novel idea which can be useful in different problems than AU recognition where training data is limited but samples are easier to annotate for a set of related tasks.

3.5 From universal emotions to Action Units

The use of HTL and SHTL allow us to evaluate how larger training sets can improve Action Unit recognition. Using the relation between AUs and universal facial expressions, we can learn Action Unit classifiers (Hidden-Tasks) by training them using additional samples labelled with prototypical facial expressions (Visible-Tasks). As previously discussed, the use of additional training data is expected to improve classifier performance by increasing their generalization ability. Following, we describe how we apply both HTL and SHTL frameworks to this particular problem.

3.5.1 Defining HTL and SHTL for AU recognition

For HTL, we assume that we are only provided with a facial expressions training set $\mathbf{X}^v$ composed by N samples. Each $\mathbf{x}_n^v \in \mathbb{R}^D$ is a facial-descriptor extracted from a face image and $\mathbf{y}_n^v \in \{0,1\}^K$ indicates its expression label. In this case, K=7 because we consider the 6 universal facial expressions plus the neutral face. In SHTL, we are also provided with an Action Unit training set $\mathbf{X}^h$ of M samples. The label vector $\mathbf{y}_m^h \in \{0,1\}^T$ indicates what Action Units are present in $\mathbf{x}_m^h$. Note that T refers to the number of Action Units considered.

The Hidden-Task parameters Θ are defined as $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_t, ..., \mathbf{a}_T]$. Each $\mathbf{a}_t \in \mathbb{R}^D$ is a linear classifier and the Hidden-Task function $\mathbf{h}(\mathbf{x}, \mathbf{a_t})$:

$$p_t(\mathbf{x}) = \mathbf{h}(\mathbf{x}, \theta_t) = (1 + \exp(-\theta_t^T \mathbf{x}))^{-1}, \tag{3.6}$$

represents the probability of the Action Unit t given an input feature $\mathbf{x}$ modelled with a sigmoid function.

Now we define $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_k, ..., \mathbf{e}_K]$ as the set of Visible-Task parameters Φ. Each $\mathbf{e}_k \in R^T$ is also a linear classifier mapping the set of T Action Unit probabilities to an output for the facial expression

k. Concretely, the Hidden-Visible-Task function $\mathbf{v}(\mathbf{H}(\mathbf{x}, \mathbf{A}), \mathbf{e}_k)$ is defined as:

$$p_k(\mathbf{x}) = \mathbf{v}(\mathbf{H}(\mathbf{x}, \mathbf{\Theta}), \phi_k) = \frac{\exp(\phi_k^T \mathbf{H}(\mathbf{x}, \Phi))}{\sum_{r=1}^{K} \exp(\phi_r^T \mathbf{H}(\mathbf{x}, \Theta))} \tag{3.7}$$

and denotes the probability of the facial expression k given the set of Action Unit outputs $\mathbf{H}(\mathbf{x}, \mathbf{A})$.

Given the previous definitions, the Visible-Task Loss is defined as the cross-entropy error function over the Facial-Expression-Recognition tasks as:

$$\mathcal{L}^v(\mathbf{A}, \mathbf{X}^v) = \frac{-1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk}^v \ln(p_k(\mathbf{x}_n^v)). \tag{3.8}$$

Similarly, the Hidden-Task Loss is defined as the log-loss function over the set of Action Unit classification tasks:

$$\mathcal{L}^h(\mathbf{A}, \mathbf{X}^h) = \frac{-1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} y_{mt}^h \ln(p_t(\mathbf{x}_m^h)) + (1 - y_{mt}^h)(1 - \ln(p_t(\mathbf{x}_m^h))) \tag{3.9}$$

Finally, we use standard L2-regularization $\frac{1}{2} \sum_{t=1}^{T} ||\theta_t||_2^2$ for the Hidden-Task parameters regularizer $\mathcal{R}(\mathbf{A})$.

3.5.2 Training the AU-Emotions Tasks Function

One of the key points in HTL and SHTL is how to obtain the Visible Tasks parameters Φ before training. In our case, we need to obtain a set of linear classifiers $\mathbf{E} = [\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_K}]$ mapping Action Unit activations to an output for each facial expression. For this purpose, we exploit the empirical results reported in [Gosselin et al., 1995, Scherer and Ellgring, 2007]. In these psychological studies, a set of actors were recorded while they interpreted situations involving the six universal basic emotions defined by Ekman. Then, AU annotations were obtained for each video according to the Facial Action Coding System and Action Unit frequencies for each emotion were

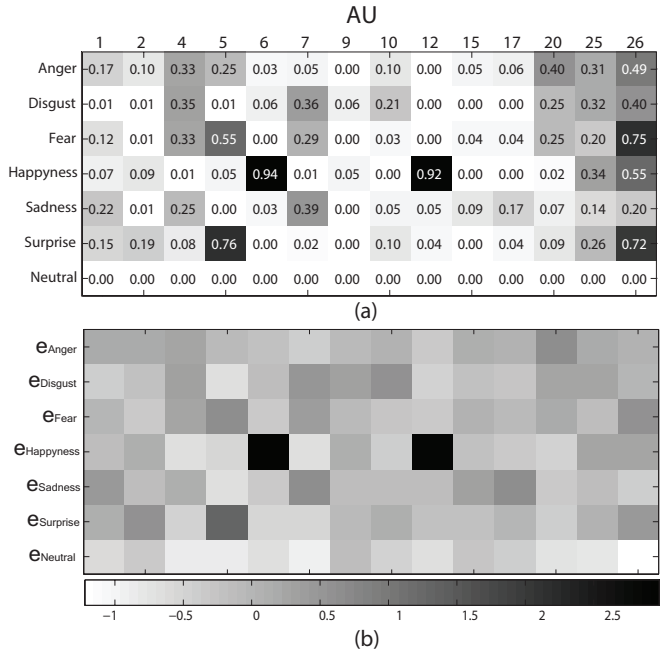

Figure 3.2: (a) Action Unit activation probability for each emotion obtained in [Gosselin et al., 1995]. In Action Unit 20, we have used the results obtained in [Scherer and Ellgring, 2007] for Anger and Fear emotions [1]. (b) Trained linear classifiers $\mathbf{E}$ mapping AU activations to emotions. See text for details.

computed (see Fig. 3.2(a)). More details can be found in the original publications.

We use these empirical results in order to train the Visible-Task classifiers $\mathbf{E}$ as follows. For each emotion, we generate a large number of random samples $\mathbb{R} \in [0,1]^T$ assuming that the probability of an AU activation follows a Bernoulli distribution according to its mean frequency in Fig. 3.2. For each sample dimension, we assign a random value between 0 and 0.5 if the AU is activated and between 0.5 and 1 otherwise. Intuitively, these samples are vectors simulating possible Action Unit activations for each type of emotion according to Eq.

3.6. Finally, we train a linear multiclass-SVM using the generated samples in order to obtain the classifiers $[\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_K}]$. Obtained coefficients for each $\mathbf{e_k}$ are shown in Fig. 3.2(b).

3.5.3 Optimization

According to Eq. 3.4, we need to solve:

$$\min_{\mathbf{A}} (1-\alpha)\mathcal{L}^h(\mathbf{A}, \mathbf{X}^h) + \alpha\mathcal{L}^v(\mathbf{A}, \mathbf{X}^v) + \beta\mathcal{R}(\mathbf{A}) \tag{3.10}$$

in order to obtain the set of optimal Action Unit classifiers $\mathbf{A}$. For this purpose, we follow a gradient-descent approach. Concretely, we use the L-BFGS Quasi-Newton method [Byrd et al., 1994] (see also Appendix A.1) which provides a higher-convergence rate than first order gradient-descent approaches and approximates the Hessian matrix with a low-rank compact form. The gradient of $\mathcal{R}(\mathbf{A})$, $\mathcal{L}^v(\mathbf{A}, \mathbf{X}^v)$ and $\mathcal{L}^h(\mathbf{A}, \mathbf{X}^h)$ w.r.t each vector $\mathbf{a}_t$ are:

$$\nabla\mathcal{L}^v = \frac{-1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}^v(e_k^{(t)} - \sum_{s=1}^{K} p_{nk}e_s^{(t)})p_{nt}(1-p_{nt})\mathbf{x}_n^v$$

$$\nabla\mathcal{R}(\mathbf{A}) = \mathbf{a}_t \ , \ \nabla\mathcal{L}^h = \frac{-1}{MT}\sum_{n=1}^{M}(y_{mt}^v - p_{mt})\mathbf{x}_m^h. \tag{3.11}$$

For shorter notation, we use $p_{nk} = p_k(\mathbf{x}_n^v)$ and $p_{mt} = p_t(\mathbf{x}_m^h)$. $e_k^{(t)}$ is a scalar corresponding to the dimension t of the vector $\mathbf{e}_k$

3.6 Experiments

In Sec. 3.6.1 and Sec. 3.6.2 we describe the different datasets and facial features used in our experiments. In the following sections, we discuss the different experiments and obtained results evaluating the proposed HTL and SHTL frameworks for Action Unit recognition.

[1] As reported in [Scherer and Ellgring, 2007], we observed that AU20 is also present in some Anger and Fear expression images. However, it is not reflected by the empirical results obtained in [Gosselin et al., 1995]

3.6.1 Databases

Action Unit Databases: We have used four different Action Unit databases widely used in the literature: the Extended Cohn-Kanade (CK+) [Lucey et al., 2010], the GEMEP-FERA [Valstar et al., 2012], the UNBC-McMaster Shoulder Pain Expression [Lucey et al., 2011] and the DISFA [Mavadati et al., 2013] datasets. CK+ contains 593 sequences of different subjects performing posed Action Units from the neutral face to the AU appex. Same as [Chu et al., 2013], we use the first frame as a negative sample and the last third frames as positive ones. The GEMEP-FERA data set contains 87 recordings of 7 different actors simulating a situation eliciting a concrete emotion. The UNBC database contains a set of 200 videos of 25 different patients undergoing shoulder pain. These patients were recorded while doing different types of arm movements. Finally, the DISFA dataset contains 27 videos of different subjects watching Youtube videos choosen in order to elicit different types of emotions. AU annotations are provided for each frame. Note that these four data-sets include posed, acted and spontaneous facial behavior. In our experiments, we have considered the recognition of Action Units 1,2,4,5,6,7,9,10,12,15,17,20,25 and 26 which include the 7 most frequent lower and upper AUs over the four datasets.

Facial expression data: In order to obtain a large number of variated facial expression images, we have collected samples from different datasets annotated with the 6 universal emotions (anger, disgust, happiness, sadness, fear and surprise) plus the neutral face. From the Bosphorous Database [Savran et al., 2008], we have used a set of 752 frontal face images from 105 different subjects. From the Radboud Faces [Langner et al., 2010] Database, we have obtained 469 frontal face images from 67 subjects. Finally, with a similar process as followed in the FER2013 Challenge [Goodfellow et al., 2013], we have automatically collected thousands of images from Google and Bing search engines [2]. For this purpose, we used a set of 70 composed

[2]We have considered to collect our own database because the provided images

queries such as "*sad man*","*disgusted woman*" or "*happy face*". Then, images which did not correspond to their emotion query were filtered by a non-expert annotator. Overall, we have collected 3437 facial expression images with a large variety of subjects, illuminations and other factors. In order to test labels reliability, an additional coder repeats the same process in 300 images for each facial expression (2100 images in total). The observed inter-coder agreement was 0.89 with a Cohen's Kappa coefficient of 0.78. Finally, we have augmented the number of samples by flipping each image around the vertical axis.

3.6.2 Facial features

As we have explained in Sec. 3.5.1, we consider a sample $\mathbf{x}$ as a facial-descriptor obtained from a given face image. Before extracting it, we follow a face-alignment process. Firstly, we automatically detect 49 facial-landmarks with the Supervised Descent method [Xuehan-Xiong and De la Torre, 2013]. Secondly, we compute an affine transformation aligning the obtained points with a mean shape. Finally, we apply the transformation to the image and crop the face region (see Fig. 3.3(a)-(b)). From the obtained aligned face, we extract two facial-descriptors from the upper and lower half parts of the face similar to [Chu et al., 2013]. The use of two different features from both parts is motivated by the fact that different Action Units are localized in concrete face areas such as eyes, eyebrows, mouth, etc... Therefore, it is convenient that AU classifiers use one of these descriptors depending on the localization of its corresponding AU. Concretely, we extract a set of SIFT descriptors from local patches centered in a subset of the landmarks (see Fig. 3.3(c)-(d)). Features for each part are concatenated in order to form the final lower and upper facial-descriptors.

---

in [Goodfellow et al., 2013] have a low resolution (48x48) and the annotations are very noisy. It will be made available upon request for the research community

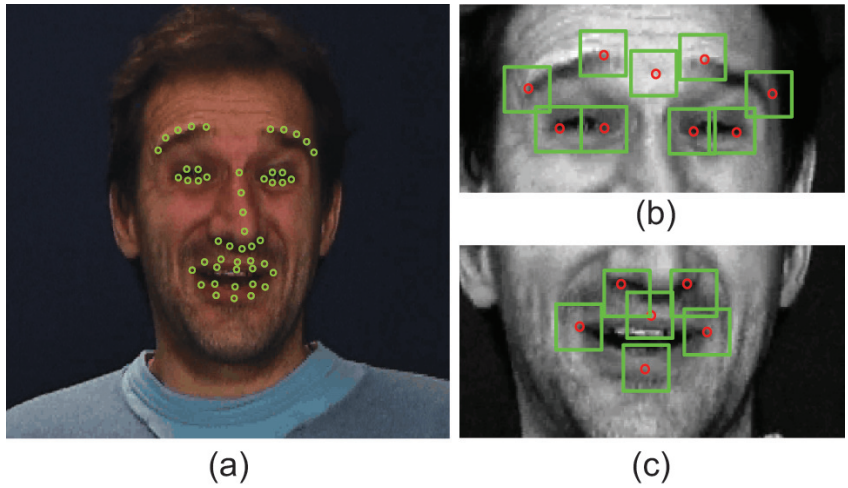

Figure 3.3: Facial-descriptors extracted for the upper and lower part of the face. (a) Original image with the set of 49 landmarks points obtained with [Xuehan-Xiong and De la Torre, 2013]. (c,d) Aligned face image and local patches used to extract the SIFT features composing the lower and upper facial descriptors.

3.6.3 Cross-Databases experiments

We evaluate how HTL and SHTL can be used to improve the generalization ability of AU classifiers by providing additional facial expression samples during training. For this purpose, we have designed a set of cross-database experiments where one Action Unit dataset is used for training and one for testing. In contrast to most works which train and test on the same data-set, a cross-database validation provides more information about how AU classifiers generalize to new subjects and other factors.

Under this setting, we compare the performance of HTL and SHTL with standard Single-Task-Learning (STL). Remember that we refer to STL when only Action Unit training data is used. On the other hand, HTL uses only samples from the Facial Expression dataset and SHTL uses both. As explained in Sec. 3.4.2, these

three approaches are generalized by the proposed SHTL framework by changing the α value in Eq. 3.10. We use α=0 for STL, α=1 for HTL and α=0.5 for SHTL. As a baseline, we also evaluate the performance of a linear SVM classifier trained independently for each AU. Note that SVM can also be considered a Single-Task-Learning approach with a different loss function than our STL. The regularization parameter β has been obtained by cross-validation over the training set. Table 3.1 shows the obtained average AUC and F1-score for the considered set of 14 Action Units[3]. Detailed results for each independent AU are shown in Figures 3.2, 3.3 and 3.4.

		AUC				F1			
Test	Train	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
CK+	UNBC	75.7	78.2	**81.7**	78.3	40.2	43.4	**49.2**	47.2
	FERA	76.6	75.5	**83.4**	80.6	41.6	38.2	**54.7**	51.6
	DISFA	83.4	84.3	**86.1**	83.7	52.8	54.8	**60.1**	56.7
UNBC	CK+	68.2	68.4	**69.7**	**69.7**	**16.9**	**16.9**	15.8	15.6
	FERA	63.8	65.2	**70.0**	69.7	12.9	13.6	**15.7**	15.6
	DISFA	67.1	67.4	**69.2**	68.8	16.3	16.2	**18.0**	16.4
FERA	CK+	70.8	70.8	**72.4**	68.0	43.1	41.3	**44.7**	40.9
	UNBC	67.5	69.4	**71.5**	70.0	42.2	40.5	42.7	**45.5**
	DISFA	70.4	71.3	**72.4**	68.9	44.2	44.3	**45.0**	39.7
DISFA	CK+	71.7	72.6	**76.0**	74.4	30.8	33.5	**39.1**	36.1
	UNBC	69.7	70.3	74.0	**76.7**	32.4	35.7	43.5	**45.4**
	FERA	68.6	70.3	**75.6**	74.4	25.2	25.5	**38.5**	36.1
	Avg.	71.1	72.0	**75.2**	73.6	33.2	33.7	**38.9**	37.3

Table 3.1: Average AU recognition performance obtained with SVM, STL, SHTL and HTL in the set of twelve cross-database experiments. Colors illustrate the different approaches ordered according to their performance.

[3]Only AUs available in the training dataset are used to compute results. HTL and SHTL can learn AU classifiers even when no AU samples are provided in the training set. However, for a fair comparison with STL and SVM, we do not consider these cases to evaluate performance. This explains HTL performance differences on the same test set.

	Test: CK+																							
	Train: UNBC								Train: FERA								Train: DISFA							
	AUC				F1				AUC				F1				AUC				F1			
AU	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
1	-	-	-	-	-	-	-	-	80.7	79.7	**84.4**	83.9	54.6	53.9	**62.8**	61.9	87.5	88.4	**88.9**	83.9	60.1	59.3	**70.0**	61.9
2	-	-	-	-	-	-	-	-	**86.5**	86.5	79.9	74.2	**58.3**	57.7	47.6	45.1	88.3	**89.7**	80.4	74.2	**54.0**	52.2	50.1	45.1
4	76.9	76.9	88.2	**92.4**	45.1	46.8	71.9	**81.6**	74.8	69.6	**92.5**	92.4	58.0	54.6	80.2	**81.6**	89.4	89.4	92.2	**92.4**	75.2	75.1	80.2	**81.6**
5	-	-	-	-	-	-	-	-	85.2	85.3	**93.4**	93.0	55.1	55.1	**76.6**	74.8	90.5	89.9	92.7	**93.0**	48.5	60.9	74.0	**74.8**
6	66.3	81.2	**87.5**	85.6	32.5	51.9	**62.9**	58.7	72.9	71.9	85.2	**85.6**	43.4	41.4	57.7	**58.7**	77.2	77.0	**86.6**	85.6	47.9	47.0	**60.4**	58.7
7	77.9	78.1	**78.5**	60.2	44.3	**44.6**	44.4	32.7	**81.6**	81.5	71.7	60.2	**53.4**	52.9	44.9	32.7	-	-	-	-	-	-	-	-
9	85.9	85.7	90.0	**90.3**	34.8	32.2	46.4	**48.7**	83.3	80.7	**92.3**	90.3	24.0	8.7	**53.5**	48.7	87.1	91.1	**93.0**	90.3	46.1	**53.9**	52.1	48.7
10	56.7	61.2	**67.4**	63.7	8.5	7.3	**11.0**	10.2	68.5	67.7	**73.2**	63.7	10.8	9.8	**14.9**	10.2	-	-	-	-	-	-	-	-
12	94.6	95.5	**97.2**	97.2	76.8	80.1	**85.4**	85.2	92.0	92.0	96.7	**97.2**	72.7	73.1	84.6	**85.2**	95.6	95.9	**97.2**	97.2	82.8	82.9	**85.9**	85.2
15	-	-	-	-	-	-	-	-	69.8	52.3	**89.4**	86.3	29.6	9.2	**54.0**	48.6	85.9	85.4	**88.8**	86.3	36.9	37.9	**50.3**	48.6
17	-	-	-	-	-	-	-	-	65.8	78.8	**88.5**	86.7	24.8	33.7	**70.0**	67.5	81.1	82.5	**88.6**	86.7	63.1	65.7	**73.2**	67.5
20	71.5	**74.4**	68.8	63.3	30.6	**31.0**	22.1	16.0	56.1	59.4	60.5	**63.3**	18.4	**20.1**	19.3	16.0	66.4	**68.5**	64.9	63.3	20.1	**22.5**	21.9	16.0
25	88.4	**90.0**	89.1	79.6	70.1	**77.8**	77.5	70.1	**87.8**	85.5	86.9	79.6	70.0	55.4	**76.3**	70.1	88.3	**90.7**	88.7	79.6	79.0	79.4	**79.4**	70.1
26	62.8	60.8	68.6	**72.2**	18.8	19.0	20.9	**22.1**	67.8	66.2	**73.2**	72.2	9.0	8.6	**23.8**	22.1	62.9	63.4	71.4	**72.2**	20.4	21.2	**23.0**	22.1
Avg.	75.7	78.2	**81.7**	78.3	40.2	43.4	**49.2**	47.2	76.6	75.5	**83.4**	80.6	41.6	38.2	**54.7**	51.6	83.4	84.3	**86.1**	83.7	52.8	54.8	**60.1**	56.7

	Test: FERA																							
	Train: CK+								Train: UNBC								Train: DISFA							
	AUC				F1				AUC				F1				AUC				F1			
AU	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
1	71.4	**74.4**	71.3	61.6	51.5	**54.2**	51.6	41.7	-	-	-	-	-	-	-	-	75.7	**76.8**	72.0	61.6	57.6	**59.4**	52.6	41.7
2	58.1	58.4	**74.2**	71.9	38.8	40.0	**50.3**	36.0	-	-	-	-	-	-	-	-	58.3	60.9	66.9	**71.9**	45.4	**46.5**	46.1	36.0
4	75.2	74.1	**76.0**	72.4	49.7	48.3	**50.6**	44.5	65.3	64.5	65.0	**72.4**	14.9	14.8	30.9	**44.5**	71.9	71.6	**75.2**	72.4	44.2	41.9	**47.9**	44.5
5	81.6	81.7	86.1	**86.2**	48.4	50.9	**58.6**	58.4	-	-	-	-	-	-	-	-	81.2	82.2	**87.3**	86.2	37.3	39.6	**61.4**	58.4
6	**78.9**	78.6	74.5	73.2	**67.5**	65.3	57.2	51.7	67.9	72.9	71.3	**73.2**	50.2	56.2	**58.2**	51.7	77.8	**78.4**	74.7	73.2	**55.2**	54.6	52.7	51.7
7	77.1	**77.7**	74.0	66.8	46.6	42.1	55.6	**60.3**	69.5	**71.6**	71.2	66.8	56.3	60.0	54.7	**60.3**	-	-	-	-	-	-	-	-
9	75.9	78.7	**86.6**	85.7	32.4	24.3	**46.6**	37.4	72.2	69.9	81.9	**85.7**	13.4	11.4	20.9	**37.4**	78.6	80.7	**86.6**	85.7	35.8	33.8	**39.2**	37.4
10	61.9	61.5	**63.1**	58.6	30.9	28.6	33.9	**35.6**	62.1	**65.7**	64.4	58.6	**53.9**	20.6	37.8	35.6	-	-	-	-	-	-	-	-
12	74.7	74.1	**77.1**	76.2	41.0	**42.1**	38.1	40.7	**78.3**	78.2	78.0	76.2	46.6	**48.2**	39.1	40.7	76.7	76.6	**78.1**	76.2	35.9	35.3	**41.2**	40.7
15	**65.7**	61.0	**52.0**	43.3	**21.2**	11.5	11.6	17.1	-	-	-	-	-	-	-	-	54.3	**54.7**	52.6	43.3	30.2	**30.9**	28.9	17.1
17	**71.8**	68.5	65.2	59.4	**38.1**	31.1	14.4	10.1	-	-	-	-	-	-	-	-	68.8	**69.8**	62.2	59.4	34.3	**35.7**	21.7	10.1
20	**63.5**	63.5	60.5	55.9	**19.8**	19.3	19.6	15.7	55.4	**59.9**	57.4	55.9	17.1	**19.7**	15.5	15.7	53.9	**56.1**	55.1	55.9	16.6	**17.3**	16.4	15.7
25	75.8	**76.9**	74.0	65.3	77.5	**79.9**	73.2	59.5	72.6	**77.4**	76.7	65.3	75.8	**79.8**	64.5	59.5	72.3	76.3	**76.6**	65.3	**79.9**	79.6	66.9	59.5
26	60.1	61.9	**78.8**	76.1	39.8	40.2	**64.8**	64.1	64.4	64.8	**77.2**	76.1	51.7	53.4	62.6	**64.1**	74.9	72.2	**82.1**	76.1	57.8	57.1	**64.4**	64.1
Avg.	70.8	70.8	**72.4**	68.0	43.1	41.3	**44.7**	40.9	67.5	69.4	**71.5**	70.0	42.2	40.5	42.7	**45.5**	70.4	71.3	**72.4**	68.9	44.2	44.3	**45.0**	39.7

Table 3.2: Cross-database experiment results. Top: Individual AU performance testing on the CK+ dataset. Bottom: Individual AU performance testing on the FERA dataset.

	Test: DISFA																							
	Train:CK+								Train:UNBC								Train:FERA							
	AUC				F1				AUC				F1				AUC				F1			
AU	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
1	68.8	71.1	**72.0**	64.7	31.1	32.7	**34.0**	26.2	-	-	-	-	-	-	-	-	62.4	61.3	64.4	**64.7**	16.5	17.1	23.9	**26.2**
2	65.5	64.9	**69.3**	65.5	22.9	**24.5**	20.6	15.8	-	-	-	-	-	-	-	-	65.5	**69.6**	64.7	65.5	13.4	16.1	**16.2**	15.8
4	81.8	82.1	82.1	**83.0**	54.6	58.1	58.9	**59.8**	73.0	73.0	77.2	**83.0**	28.1	31.7	47.0	**59.8**	62.7	71.5	82.6	**83.0**	34.1	42.2	57.4	**59.8**
5	82.9	87.3	86.9	**88.5**	19.2	27.1	**37.7**	36.8	-	-	-	-	-	-	-	-	77.6	76.8	87.3	**88.5**	11.2	10.9	**38.7**	36.8
6	58.2	59.4	**67.2**	65.3	25.6	26.3	**33.0**	32.2	61.5	60.7	**67.3**	65.3	17.3	28.8	32.1	**32.2**	60.5	60.4	**69.4**	65.3	30.3	30.3	**35.2**	32.2
7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	64.5	66.8	73.5	**78.0**	8.3	15.2	20.8	**23.6**	67.1	66.7	71.1	**78.0**	13.7	10.4	22.9	**23.6**	65.9	69.5	77.3	**78.0**	10.5	5.1	**26.1**	23.6
10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	90.9	91.1	**92.9**	92.4	61.7	62.2	**74.5**	74.3	89.2	90.4	**92.7**	92.4	64.9	66.6	72.7	**74.3**	86.3	85.3	**92.5**	92.4	59.9	59.0	72.9	**74.3**
15	66.4	66.1	68.5	**69.2**	13.8	17.5	**28.8**	18.9	-	-	-	-	-	-	-	-	66.3	62.2	68.5	**69.2**	11.5	8.2	**23.4**	18.9
17	70.1	66.9	**70.7**	68.0	25.8	24.1	**26.0**	17.9	-	-	-	-	-	-	-	-	56.2	66.0	**72.5**	68.0	15.4	25.5	**31.3**	17.9
20	62.7	62.8	**65.6**	60.4	11.1	**11.2**	10.7	8.8	49.3	50.6	48.6	**60.4**	7.0	6.0	5.8	**8.8**	55.6	57.8	**61.8**	60.4	0.4	0.4	**9.0**	8.8
25	**86.5**	85.8	82.2	75.1	73.7	**74.2**	69.7	61.3	79.0	82.5	**83.4**	75.1	57.5	66.7	**72.2**	61.3	82.3	81.1	**82.4**	75.1	60.7	56.0	**70.5**	61.3
26	61.9	66.5	80.9	**82.5**	21.4	29.2	55.1	**58.0**	68.5	68.3	77.4	**82.5**	38.6	39.9	52.2	**58.0**	82.1	82.0	**83.5**	82.5	39.0	35.2	57.8	**58.0**
Avg.	71.7	72.6	**76.0**	74.4	30.8	33.5	**39.1**	36.1	69.7	70.3	74.0	**76.7**	32.4	35.7	43.5	**45.4**	68.6	70.3	**75.6**	74.4	25.2	25.5	**38.5**	36.1

	Test: UNBC																							
	Train: CK+								Train: FERA								Train: DISFA							
	AUC				F1				AUC				F1				AUC				F1			
AU	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	78.8	78.0	81.3	**86.3**	**9.4**	7.7	7.6	8.3	60.0	73.6	83.4	**86.3**	6.1	**13.6**	8.7	8.3	82.3	81.7	81.9	**86.3**	**11.4**	11.2	7.1	8.3
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	63.9	67.3	**68.3**	63.3	26.6	27.6	28.9	**29.5**	60.1	58.2	**64.8**	63.3	21.6	21.6	26.9	**29.5**	65.3	65.3	**66.3**	63.3	27.1	26.9	**32.3**	29.5
7	68.1	**68.6**	66.5	63.2	19.4	**20.1**	19.9	17.2	58.4	56.9	57.8	**63.2**	15.6	13.9	16.5	**17.2**	-	-	-	-	-	-	-	-
9	70.5	72.4	78.0	**83.8**	**9.8**	7.0	5.1	5.3	63.9	69.0	81.0	**83.8**	2.6	2.5	5.0	**5.3**	74.1	76.8	82.9	**83.8**	5.9	6.9	**7.3**	5.3
10	61.8	60.3	69.2	**82.7**	3.1	2.3	4.6	**8.5**	73.6	68.2	**83.7**	82.7	3.8	2.8	5.4	**8.5**	-	-	-	-	-	-	-	-
12	81.7	82.0	**83.8**	81.8	43.7	44.1	**44.6**	42.6	78.7	78.2	**82.1**	81.8	37.8	37.4	**43.7**	42.6	81.6	81.2	**82.5**	81.8	**46.0**	44.5	43.2	42.6
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20	**60.8**	59.0	44.1	31.1	5.1	**5.7**	2.1	1.5	39.8	**40.7**	30.6	31.1	0.0	0.0	1.3	**1.5**	35.0	**35.5**	32.5	31.1	1.6	1.4	**1.6**	1.5
25	82.8	**83.5**	80.6	69.3	26.6	**29.6**	19.0	12.4	82.9	**83.6**	80.3	69.3	25.7	**27.6**	22.2	12.4	80.3	79.8	**81.2**	69.3	14.7	14.7	**21.6**	12.4
26	45.6	44.5	55.8	**66.1**	8.2	8.3	10.0	**15.0**	57.1	58.4	**66.5**	66.1	3.1	2.8	12.0	**15.0**	51.0	51.7	57.0	**66.1**	7.6	7.5	13.0	**15.0**
Avg.	68.2	68.4	**69.7**	**69.7**	**16.9**	**16.9**	15.8	15.6	63.8	65.2	**70.0**	69.7	12.9	13.6	**15.7**	15.6	67.1	67.4	**69.2**	68.8	16.3	16.2	**18.0**	16.4

Table 3.3: Cross-database experiment results. Top: Individual AU performance testing on the DISFA dataset. AUs 7 and 10 are not available in this dataset. Bottom: Individual AU performance testing on the UNBC dataset. AUs 1, 2, 5,15 and 17 are not available in this dataset.

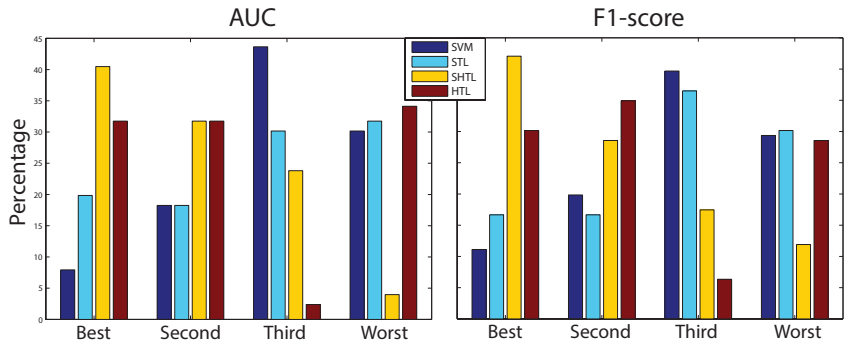

Figure 3.4: Overall, our cross-database experiments include 126 Action Unit detection sub-problems. In order to summarize the presented results, we show the percentage of times where SVM, STL, SHTL and HTL achieves the best,second, third and worst performance across the cited subproblems.

HTL vs STL and SVM: Comparing HTL to STL and SVM, we can observe that HTL achieve comparable or better performance in terms of average AUC and F1 for most of the cross-database experiments. It could seem surprising because HTL does not use any Action Unit annotation during training. However, it confirms our hypothesis that the limited training data of current AU datasets can decrease the quality of learned models. In contrast, HTL uses richer facial expression data which increases its generalization ability over different datasets. Additionally, notice that STL and SVM achieves similar average performance. This can be explained because both are Single-Task-Learning approaches which only use the Action Unit data for training.

SHTL vs STL and HTL: Comparing SHTL with the other approaches, we can observe that SHTL achieves superior performance in most cases. These can be explained because SHTL is able to combine information from the AU and Facial Expression training samples. Analyzing the performance for each AU independently, the results show some variations depending on each experiment. However,

SHTL generally outperforms either HTL or STL. Again, it shows the advantages of using SHTL in order to combine both AU and facial expression training data information.

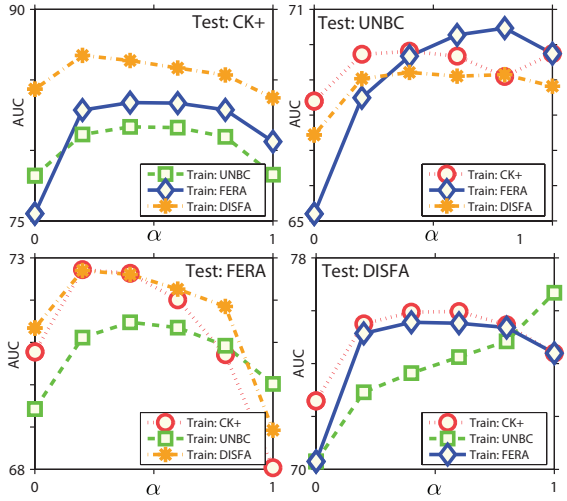

Figure 3.5: Average AU recognition performance in the cross-database experiments varying the α parameter in the range between 0 and 1. See text for details.

Evaluating the effect of α parameter: Previously, we have fixed the α parameter of SHTL to 0.5. This provides a balanced trade-off between Hidden (Action Units) and Visible-Task (Facial Expressions) losses. However, different values for α are also possible. In order to evaluate the impact of the α parameter, we have run the same set of experiments fixing it to different values in the range between 0 to 1. As Figure 3.5 shows, optimal performance is generally obtained with α between 0 and 1 which combines information from AU and Facial Expression databases (SHTL).

We have shown that by using HTL and SHTL, the use of additional training data labelled with prototypical facial expressions improves the generalization ability of learned AU classifiers. Note that we are using simple linear classifiers and standard facial-features.

However, these frameworks are flexible enough to be used with any kind of facial-descriptors or base classifiers.

3.6.4 Single-database experiments

Although cross-database experiments are useful to evaluate the generalization ability of learned models, it is reasonable to ask how SHTL and HTL performs in Action Unit data which have been obtained in similar conditions. In this experiment, we evaluate the previously used methods with a leave-one-subject strategy over the same dataset. Note that this setting is similar to the commonly used in the literature. In this case, for SHTL we have set $\alpha = 0.25$ in order to give more importance to the Hidden-Task loss (Action Unit data). Moreover, for SVM, STL and SHTL we have optimized the classification threshold using the Action Unit training samples during cross-validation. [4]

	AUC				F1			
Train	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
CK+	90.6	91.2	**91.7**	80.6	68.5	68.6	**68.9**	51.7
UNBC	75.3	78.2	**78.8**	69.7	22.7	21.3	**27.1**	15.6
FERA	66.7	66.9	**73.4**	68.0	46.8	48.7	**51.9**	40.9
DISFA	79.6	81.2	**81.5**	74.4	37.6	40.5	**42.9**	36.1
Avg.	78.0	79.4	**81.3**	73.2	43.9	44.8	**47.7**	36.1

Table 3.4: Average Action Unit recognition performance obtained with SVM,STL,SHTL and HTL in single-dataset experiments. Colors illustrate the different approaches ordered according to their performance in each experiment.

Figure 3.4 shows the obtained results. Under this setting, HTL achieves the worst performance. However, it was expected since the

[4]Worst results were observed optimizing the threshold in cross-database experiments.

problem of generalizing to data taken in different conditions is mitigated in this case. SHTL achieves slightly better AUC than STL and SVM in all the cases and a more significant improvement in terms of the F1-score. Therefore, even when data is taken in similar conditions, the use of additional facial expression samples is beneficial. One of the main factors that could explain SHTL improvement is that current Action Unit databases are limited in terms of subject variability. Therefore, SHTL can learn more generic AU classifiers by using training samples from additional subjects present in the facial expressions data. One point that supports that conclusion is that SHTL obtains a significant improvement over the FERA dataset which is the most limited in terms of subjects. In contrast, this improvement is less significative in the CK+ dataset which has the larger number of subjects.

3.6.5 Comparison with related work: Tranductive Learning

In this experiment, we compare SHTL with state-of-the-art transductive learning approaches for AU recognition: STM [Chu et al., 2013], TPT [Sangineto et al., 2014] and SVTPT [Zen et al., 2014]. As we have discussed in Sec. 3.3, these methods use unlabelled data during training in order to learn personalized models for each test subject. In contrast, SHTL is trained with additional facial expressions data which increases its generalization ability to new subjects. We have used similar features and followed the same experimental-setup in order to compare our results with the reported in the cited works. We have retrained the classifiers $\mathbf{e}_k$ (Sec. 3.5.2) using only the subset of 8 AUs evaluated in STM. They also include the 6 AUs used in TPT and SVTPT works. Again, the α parameter of SHTL has been set to 0.25.

Table 3.5 shows the obtained results. As the reader can observe, SHTL achieves competitive performance compared with transductive learning approaches. Concretely, SHTL obtains better AUC in

all cases and similar F1-score over the CK+ dataset. Only STM significantly outperforms the F1-score of SHTL in the FERA dataset. However, it is worth mentioning that Transductive Learning models need to be trained for each subject during testing and requires sufficient samples to correctly estimate the test distribution. In contrast, SHTL just needs to learn a single generic classifier by using the additional facial expression data. Therefore, SHTL is more useful in real applications where training Action Unit classifiers for each subject during testing is not feasible (e.g. online detection of Action Units in video streams).

		SHTL	STM	TPT	SVTPT
AUC	FERA	**76.2**	74.5	-	-
	CK+ (8 AUs)	**93.4**	91.3	-	-
	CK+ (6 AUs)	**93.9**	90.1	91.3	92.7
F1	FERA	55.9	**59.9**	-	-
	CK+ (8 AUs)	**76.5**	**76.6**	-	-
	CK+ (6 AUs)	78.8	74.8	76.8	**79.1**

Table 3.5: SHTL performance and results reported by state-of-the-art transductive Learning approaches for Action Unit recognition on CK+ and FERA datasets.

3.7 Summary

In this Chapter, we have investigated how additional training data annotated with universal facial expressions can improve the generalization ability of Action Unit classifiers. For this purpose, we have proposed the Hidden and Semi-Hidden Task Learning frameworks able to learn a set of Hidden-Tasks (Action Units) when training data is limited or even not available. To address this weakly-supervised setting, these frameworks are able to exploit prior knowledge about the relation between these Hidden-Tasks and a set of Visible-Tasks

(Facial Expressions). Exhaustive experiments have shown that HTL and SHTL improve the generalization ability of Action Unit classifiers by using training data from a large facial expression database. Surprisingly, HTL generally achieves better performance than standard Single-Task Learning in cross-database experiments without using any Action Unit annotation. Moreover, we have also shown the advantages of combining AU and Facial Expressions data information with SHTL. Despite that most existing works on AU recognition have focused on proposing facial features or supervised classification methods, our results suggest that the limitation of training data in AU recognition is an important factor which has been largely overlooked. The proposed HTL and SHTL frameworks can address this problem from a weakly-supervised perspective by using additional training data annotated with facial expression labels which are much easier to obtain. Finally, we consider that HTL and SHTL are general purpose frameworks which could be also useful in other problems where the lack of annotated training data is a challenge. As a future work, it would be interesting to study how to adapt the Visible Task Layer during training by using only the pre-trained parameters as a prior. It could allow SHTL to correct possible inaccuracies of the empirical studies relating Facial Expressions with Action Unit occurrences.

Chapter 4

MULTI-INSTANCE DYNAMIC ORDINAL RANDOM FIELDS FOR WEAKLY-SUPERVISED EXPRESSION INTENSITY ESTIMATION

4.1 Introduction and Motivation

In Chapters 2 and 3, we have presented different weakly-supervised approaches in the context of Discrete Expression Recognition. However, as mentioned in Chapter 1, another important problem in Facial Behavior Analysis is Expression Intensity Estimation. For this reason, in this Chapter we focus on Action Unit (AU) [Mavadati et al., 2013] and Pain Intensity estimation [Lucey et al., 2011]. Both problems can be naturally posed as Dynamical Ordinal Regression problems, where the goal is to predict a value in an ordinal scale for each instant of a sequence. In AU intensity estimation, the objective is to

predict the activation level (in a six-point ordinal scale) of different facial actions at each frame of a video. Similarly, in Pain Intensity estimation we aim to measure an ordinal value representing the level of pain felt by a recorded subject (see Fig. 4.1).

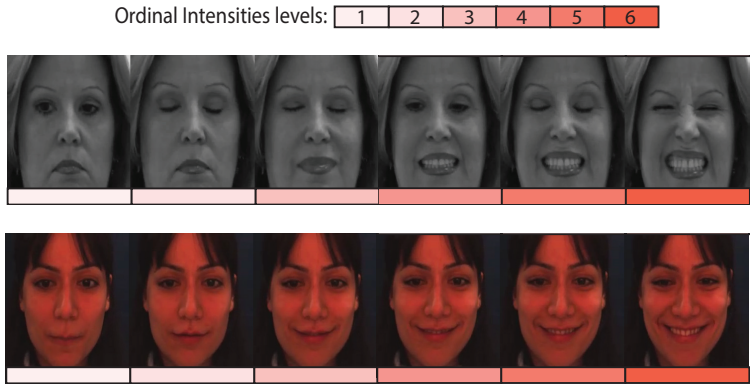

Figure 4.1: Illustration of the Pain and Action Unit intensity problems addressed in this Chapter. Left: Sequence showing different pain levels (coded in an ordinal scale from 1 to 6). Right: Example of different intensities for Action Unit 12 (Lip-Corner Puller) also represented in an ordinal scale.

The vast majority of proposed approaches to address these problems have followed the supervised learning paradigm [Rudovic et al., 2015, Kaltwang et al., 2016, Rudovic et al., 2013], i.e, models are learned using manually annotated intensity levels for each frame in a set of training sequences. Despite the efforts in the field, performance of current approaches following this strategy can still be considered far from optimal. Our hypothesis is that the main reason for this low performance is the limited data used to train supervised models. As previously discussed, annotation in Facial Behavior Analysis is usually an expensive and time-consuming task and, compared to the discrete case, the labelling process of expression intensities is even more tedious. As a consequence, current datasets for Expression Intensity Estimation are suboptimal in terms of size/variability and,

therefore, the use of this limited data for training supervised models can decrease their performance in unseen test samples.

One potential solution to overcome this limitation could be to annotate larger training sets. However, this strategy is not feasible given the cost of the annotation process. In contrast, our explored solution consists of using the weakly-supervised paradigm instead of the fully-supervised one. As a reminder, weakly-supervised approaches aim to learn models using annotations which only provide partial information (weak-labels) about the task that needs to be solved. These weak-labels are much easier to obtain than those for fully-supervised learning, thus allowing us to use larger datasets minimizing the annotation effort. For example, in Pain Intensity estimation, it is much easier to obtain a label for the whole sequence in terms of the maximum pain intensity felt by the recorded subject (e.g. using patients self-reports or external observers). Similarly, annotating Facial Action Unit intensities requires a huge effort by expert coders. In contrast, segmenting sequences according to the increasing or decreasing evolution of AU intensities (i.e, onset and appex segments) is less time-consuming. In this Chapter, we consider these two weakly-supervised settings for Action Unit and Pain Intensity estimation respectively. Our motivation is that models able to learn only from these "cheaper" annotations would allow to leverage larger training sets and thus potentially build more effective models.

Similar to Chapter 2, we address these two Facial Expression Intensity estimation problems by using the Multiple Instance Learning framework. As previously introduced, in traditional Single-Instance-Learning (SIL), the fully supervised setting is assumed with the goal to learn a model from a set of feature vectors (instances) each being annotated in terms of target label y. By contrast, in MIL, the weak supervision is assumed, thus, the training set is formed by bags (sets of instances), and only labels at bag-level are provided. In order to learn a model from this weak-information, MIL assumes that there exists an underlying relation between the label of a bag (e.g., video) and the labels of its constituent instances (e.g., image frames). For

instance, in the standard Multi-Instance-Classification (MIC) [Maron and Lozano-Pérez, 1998] introduced in Chapter 2, labels are considered binary variables $y \in \{-1, 1\}$ and negative bags are assumed to contain only instances with an associated negative label. In contrast, positive bags must contain at least one positive instance. Another example of MIL assumption is related to the Multi-Instance-Regression (MIR) problem [Ray and Page, 2001], where $y \in R$ is a real-valued variable and the maximum instance-label within the bag is assumed to be equal to y.

4.2 Contributions

In order to apply the MIL framework for weakly-supervised Facial Expression Intensity Estimation, in this Chapter we focus on a novel MIL problem that we refer to as Multi-Instance Dynamic Ordinal Regression (MI-DOR). In this case, bags are structured as dynamic sequences of instances with temporal dependencies. Moreover, instance labels (i.e., expression intensities) are considered ordinal variables which can take values in a set of L ordered discrete categories $\{0 \prec ... \prec l \prec L\}$. To address MI-DORF, we propose the Multi-Instance Dynamic Ordinal Random Fields (MI-DORF). To build this framework, we use the notion of Hidden Conditional Ordinal Random Fields (HCORF) [Kim and Pavlovic, 2010a]. Similar to HCORF, MI-DORF is an undirected graphical model where observation labels are modelled as a linear-chain of ordinal latent variables. However, the energy function employed in MI-DORF is designed to explicitly incorporate the weak-relation between latent instance labels and observable sequence weak-labels. For this purpose, we employ high-order potentials modelling different Multiple Instance Learning assumptions. Our main contributions are summarized as follows:

- To the best of our knowledge, no previous works have explored Multi-Instance Dynamic Ordinal Regression. The proposed MI-DORF framework is specifically designed for this type of tasks

by taking into account different assumptions about the weak-relation between instances and sequence labels.

- As far as we know, the proposed framework is the first MIL approach that imposes ordinal structure on instance labels. The proposed method also incorporates dynamic information that is important when modeling temporal structure in instances within the bags (i.e., image sequences). While modeling the temporal structure has been attempted in [Wu et al., 2015, Liu et al., 2016], there are virtually no works that account for both ordinal and temporal data structures within MIL framework.

- We introduce an efficient inference method for the proposed MI-DORF framework which has a similar computational complexity as the forward-backward algorithm [Barber, 2012] (see also Appendix A.3) used in other Latent-Dynamic Models (e.g HCORF). This is despite the fact that we employ high-order potentials modelling the different Multi-Instance assumptions.

- We also introduce a learning and inference approach for the Partially-Observed MI-DOR scenario, where we want to take advantage of a limited number of instance annotations during training. This is useful in cases where this privileged information can be provided together with the weak-labels at sequence-level.

We demonstrate the performance of the proposed framework on weakly-supervised Pain and Action Unit Intensity estimation. We show the superior performance of our method compared with alternative approaches applicable to these scenarios. Our results suggest that the proposed framework can be employed to reduce the annotation effort in Expression Intensity Estimation problems.

4.3 Related Work

Multiple-Instance Learning: In Chapter 2, we divided existing MIL methods between Single-Concept or Multi-Concept approaches. However, in this review we differentiate between the bag-based or instance-based paradigms [Amores, 2013]. Both categorizations are similar but, according to our opinion, the latter is better to clarify the key concepts and contributions presented in this particular Chapter.

In the bag-based methods, a feature vector representation for each bag is first extracted. Then, these representations are used to train standard Single-Instance methods, used to estimate the bag labels. This representation is usually computed by using different types of similarity metrics between training instances. Examples following this paradigm include Multi-Instance Kernel [Gärtner et al., 2002], MILES [Chen et al., 2006] or MI-Graph [Zhou et al., 2009]. The main limitation of these approaches is that the learned models can only make predictions at the bag-level. However, these methods cannot work in the weakly-supervised setting, where the goal is to predict instance-labels (e.g., frame-level intensities) from a bag (e.g., a video). In contrast, instance-based methods directly learn a model which operates at the instance level. For this, MIL assumptions are incorporated by considering instance-labels as latent variables. Using this strategy, traditional supervised models are adapted to incorporate MIL assumptions. Examples of methods following this approach include Multi-Instance Support Vector Machines [Andrews et al., 2003] (MI-SVM), MILBoost [Zhang et al., 2005b], MI Gaussian Processes [Kim and Torre, 2010] or Multi-Instance Logistic Regression [Hsu et al., 2014]. In this Chapter, we follow the instance-based paradigm by treating instance-labels as ordinal latent states in a Latent-Dynamic Model. In particular, we follow a similar idea to that in the Multi-Instance Discriminative Markov Networks [Hajimirsadeghi et al., 2013], where the energy function of a Markov Network is designed to explicitly model weak-relations between bag and instance labels. However, in contrast to the works described above, the

presented MI-DORF framework accounts for the ordinal structure in instance labels, while also accounting for their dynamics.

Latent-Dynamic Models: Popular methods for sequence classification are Latent-Dynamic Models such as Hidden Conditional Random Fields (HCRFs) [Quattoni et al., 2007] or Hidden-Markov-Models (HMMs) [Rabiner and Juang, 1986] (see also Appendix A). These methods are variants of Dynamic Bayesian Networks (DBNs) where a set of latent states are used to model the conditional distribution of observations given the sequence label. In these approaches, dynamic information is modelled by incorporating probabilistic dependence between time-consecutive latent states. MI-DORF builds upon the HCORF framework [Kim and Pavlovic, 2010a] which considers latent states as ordinal variables. However, HCORF follows the SIL paradigm, where the main goal is to predict sequence labels and latent variables are only used to increase the expressive power of the model. In contrast, the energy function of MI-DORF is defined to explicitly encode Multi-Instance relationships between bag and latent instance labels. Note also that more recent works (e.g., [Wu et al., 2015], [Liu et al., 2016]) extended HMMs/HCRFs, respectively, for Multi Instance Classification. The reported results in these works suggested that modeling dynamics in MIL can be beneficial when bag-instances exhibit temporal structure. However, these methods limit their consideration to the case where instance labels are binary and, therefore, are unable to solve MI-DOR problems.

As has been introduced in Sec. 4.2, we also extend MI-DORF to the partially-observed setting, where labels for a small subset of instances are available during training. This scenario has been previously explored using Latent-dynamical models such as Conditional Random Fields [Li et al., 2009] and their extensions (HCRF [Chang et al., 2009]). Although the instance labels are incorporated in these approaches, they can be considered suboptimal for MI-DOR problems, where weak-labels at sequence level need to be taken into account according to the Multi-Instance assumptions.

Non-supervised Facial Behavior Analysis: Research on au-

tomatic Facial Behavior Analysis has usually focused on the fully-supervised setting, however, we can find some exceptions in the literature. In the context of Action Unit detection, Zhou et. al [Zhou et al., 2010] proposed Aligned Cluster Analysis for the unsupervised segmentation and clustering of facial events in videos. Their experiments showed that the obtained clusters were coherent with Action Unit manual annotations. We find another example in [Tax et al., 2010], where Multiple Instance Classification was used to find key frames representing Action Unit activations in sequences. Different from these cited approaches which focus on binary detection, we address weakly-supervised Action Unit intensity estimation. For that purpose, the proposed MI-DORF model is able to learn from segments which are labelled according to the increasing or decreasing evolution of AU intensities (see Sec. 4.1). A similar problem has been recently addressed by Zhao et al. [Zhao et al., 2016b]. Concretely, Ordinal Support Vector Ordinal Regression (OSVR) has been proposed to estimate facial expression intensities using only onset and appex segments during training. However, OSVR presents some limitations in this context. Firstly, it models instance (frame) labels as continuous variables, which implicitly assumes an uniform distribution between the distances of the different ordinal levels. Note that this is a suboptimal modelling of ordinal variables since each label can have a different extent. Secondly, OSVR poses MI-DOR as a ranking problem, where only constraints about the order of the instance intensities is considered. This implies that the scale of predicted values does not necessary match with the ground-truth. In contrast, MI-DORF models instance labels as ordinal variables, thus allowing to predict better labels scaling by determining a priory the number of ordinal levels. Finally, OSVR is a static approach and temporal correlations are not modelled as in MI-DORF.

In the context of weakly-supervised pain detection, MIL approaches have been previously applied by considering that a weak-label is provided for a sequence (indicating the absence or presence of pain). Then, a video is considered as a bag and image frames as instances.

Sikka et al. [Sikka et al., 2013] proposed to extract a Bag-of-Words representation from video segments and treat them as bag-instances. Then, MILBoosting [Zhang et al., 2005b] was applied to predict sequence-labels under the MIC assumption. Following the bag-based paradigm, [Ruiz et al., 2014] developed the Regularized Multi-Concept MIL method capable of discovering different discriminative pain expressions within an image sequence. More recently, [Wu et al., 2015] proposed MI Hidden Markov Models, an adaptation of standard HMM to the MIL problem. The limitation of these approaches is that they focus on the binary detection problem (i.e, pain intensity levels are binarized), and, thus, are unable to consider different intensity levels of pain. This is successfully attained by the proposed MI-DORF.

4.4 Multi-Instance Dynamic Ordinal Regression

In this section, we formalize the MI-DOR problem and two of its particular instances addressed in this Chapter: MaxMI-DOR and RelMI-DOR. We apply these two settings to Pain and Action Unit intensity estimation respectively (see Fig. 4.2). In MI-DOR problems we are provided with a training set $\mathcal{T} = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2)\ , ..., (\mathbf{X}_N, y_N)\}$ formed by pairs of structured-inputs $X \in \mathcal{X}$ and labels y. Specifically, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ are temporal sequences of T observations $\mathbf{x} \in R^d$ in a d-dimensional space [1]. Given the training-set $\mathcal{T}$, the goal is to learn a model $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{H}$ mapping sequences $\mathbf{X}$ to an structured-output $\mathbf{h} \in \mathcal{H}$. Concretely, $\mathbf{h} = \{h_1, h_2, ..., h_T\}$ is a sequence of variables $h_t \in \{0 \prec ... \prec l \prec L\}$ assigning one ordinal value for each observation $\mathbf{x_t}$. In order to learn the model $\mathcal{F}$ from $\mathcal{T}$, it is necessary to incorporate prior knowledge defining the weak-relation between labels y and latent ordinal states $\mathbf{h}$. In MaxMI-DOR, we assume that bag-labels $y \in \{0 \prec ... \prec l \prec L\}$ are also ordinal variables

[1] Total number of observations T can vary across different sequences

and that the maximum value in $\mathbf{h}_n$ must be equal to the label y_n:

$$y_n = \max_h(\mathbf{h}_n) \quad \forall\, (\mathbf{X}_n, y_n) \in \mathcal{T} \tag{4.1}$$

On the other hand, in RelativeMI-DOR the sequence label is a categorical variable taking four possible values $y \in \{\uparrow, \downarrow, \emptyset, \updownarrow\}$. Intuitively, each label indicates the type of evolution within latent labels $\mathbf{h}$. Concretely, in sequences labelled with $y = \uparrow$, there must be an increasing ordinal level transition in, at least, one instant t. Moreover, no decreasing transitions are allowed within the sequence. The opposite occurs in sequences labelled as $y = \downarrow$. In the case of $y = \updownarrow$ the sequence is assumed to contain decreasing and increasing transitions. Finally, when $y = \emptyset$ all the ordinal values in $\mathbf{h}$ should be equal (monotone sequence). Formally, these constraints can be defined as:

$$\forall\, (\mathbf{X}_n, y_n) \begin{cases} y_n = \uparrow & \text{iff} \quad (\exists t\ h_t < h_{t+1}) \wedge (\forall t\ h_t \leq h_{t+1}) \\ y_n = \downarrow & \text{iff} \quad (\exists t\ h_t > h_{t+1}) \wedge (\forall t\ h_t \geq h_{t+1}) \\ y_n = \emptyset & \text{iff} \quad (\forall t\ h_t = h_{t+1}) \\ y_n = \updownarrow & \text{otherwise} \end{cases} \tag{4.2}$$

Note that the definition of these MI-DOR problems differs from standard supervised sequence classification with latent variables. In that case, the main goal is to learn a model $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ mapping $\mathbf{X}$ to sequence labels y.

4.5 Max-Multi-Instance Dynamic Odinal Random Fields (MaxMI-DORF)

In this section, we present the proposed Max-Multi-Instance Dynamic Odinal Random Fields to solve the MaxMI-DOR problem described in Sec. 4.4.

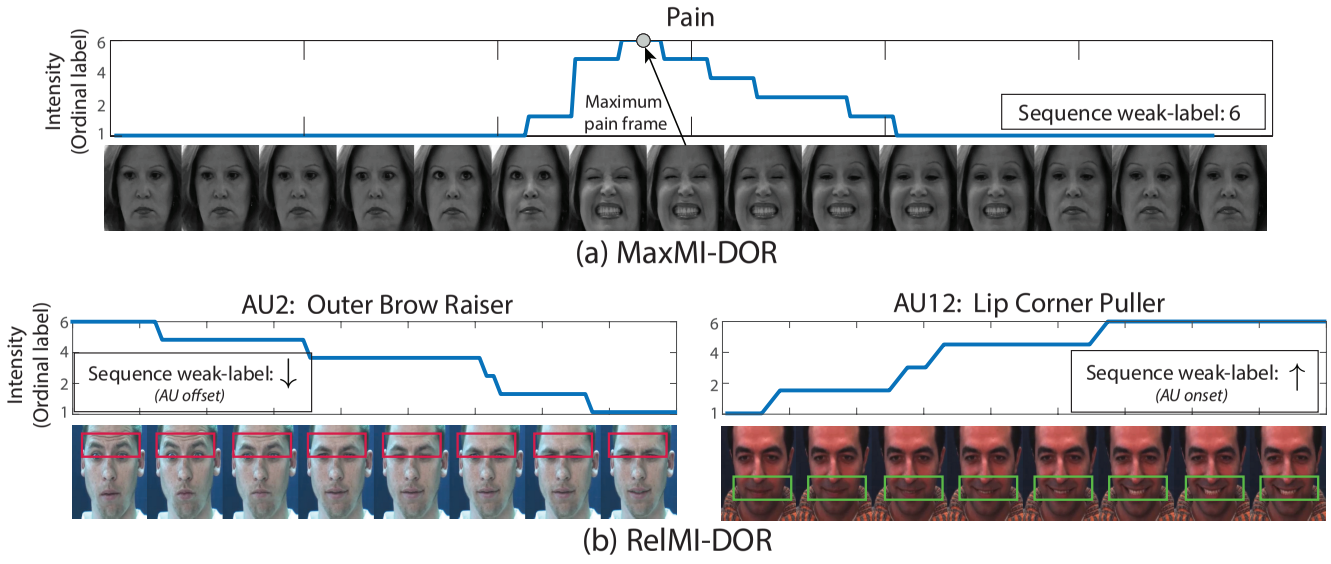

Figure 4.2: Illustration of the MaxMI-DOR (a) and RelMI-DOR (b) problems applied to Pain and Action Unit intensity estimation respectively. In MaxMI-DOR, only a weak-label indicating the maximum level of pain in the sequence is provided during training. In contrast, in RelMI-DOR the video label indicates the increasing or decreasing evolution of the AU intensity within the sequence (onset or offset segments). By only using these weak-labels at sequence-level during training, the goal is to train a model able to predict the expression intensity for each frame of the sequence (blue line).

4.5.1 Model Definition

Similar to Hidden Conditional Ordinal Random Fields, MaxMI-DORF defines the conditional probability of labels y given observations $\mathbf{X}$ using a Gibbs distribution as:

$$P(y|\mathbf{X};\theta) = \sum_{\mathbf{h}} P(y,\mathbf{h}|\mathbf{X};\theta) = \frac{\sum_h e^{-\Psi(\mathbf{X},\mathbf{h},y;\theta)}}{\sum_{y'}\sum_h e^{-\Psi(\mathbf{X},\mathbf{h},y';\theta)}}, \tag{4.3}$$

where θ is the set of the model parameters. As defined in Eq. 4.4, the energy function Ψ defining the Gibbs distribution is composed of the sum of three different types of potentials:

$$\begin{aligned}\Psi(\mathbf{X},\mathbf{h},y;\theta) = \sum_{t=1}^{T}\Psi^N(\mathbf{x}_t,h_t;\theta^N) + \sum_{t=1}^{T-1}\Psi^E(h_t,h_{t+1};\theta^E) \\ + \Psi^M(\mathbf{h},y,\theta^M),\end{aligned} \tag{4.4}$$

The factor graph representation of MaxMI-DORF is shown in Fig. 4.3(a).

MaxMI-DORF: Ordinal node potentials

The node potentials $\Psi^N(\mathbf{x},h;\theta^N)$ aim to capture the compatibility between a given observation $\mathbf{x}_t$ and the latent ordinal value h_t. Similar to HCORF, it is defined using the ordered probit model [Winkelmann and Boes, 2006] (see also Appendix A.2):

$$\Psi^N(\mathbf{x},h=l) = \log\left[\Phi\left(\frac{b_l - \mathbf{fi}^T\mathbf{x}}{\sigma}\right) - \Phi\left(\frac{b_{(l-1)} - \mathbf{fi}^T\mathbf{x}}{\sigma}\right)\right], \tag{4.5}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function (CDF), and $\theta^N = \{\beta, \mathbf{b}, \sigma\}$ is the set of potential parameters. Specifically, the vector $\beta \in R^d$ projects observations $\mathbf{x}$ onto an ordinal line divided

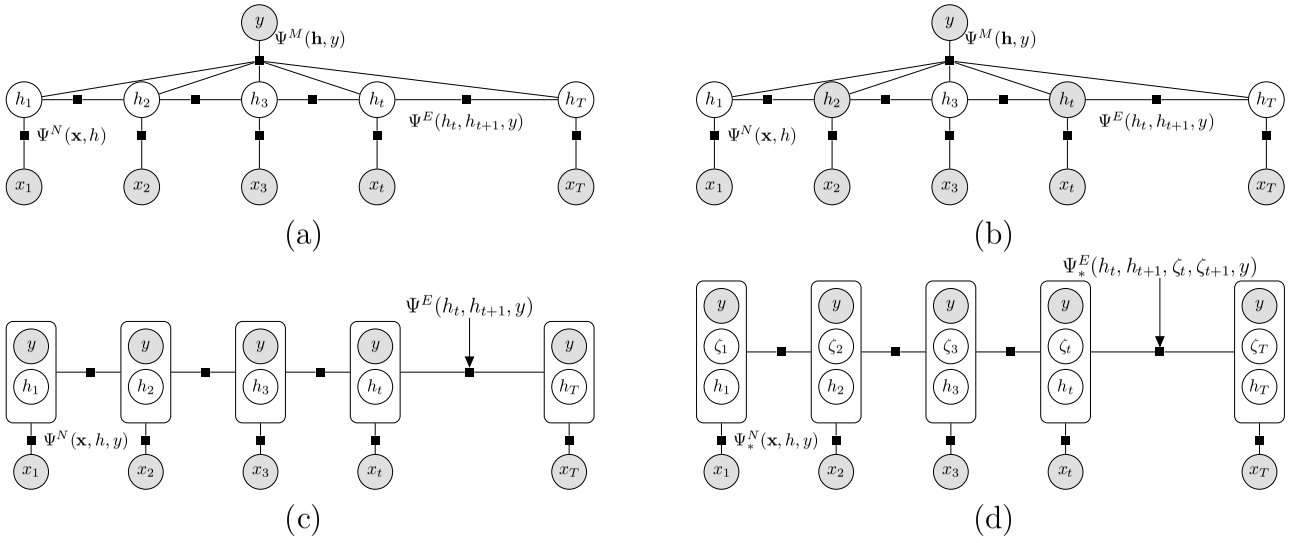

Figure 4.3: (a) Factor graph representation of the proposed MI-DORF framework. Node potentials Ψ^N model the compatibility between a given observation $\mathbf{x}_t$ and a latent ordinal value h_t . Edge potentials Ψ^E take into account the transition between consecutive latent ordinal states h_t and h_{t+1}. Finally, the high-order potential Ψ^M models Multi-Instance assumptions relating all the latent ordinal states $\mathbf{h}_t$ with the bag-label y. (b) Factor graph representation of the Semi-Supervised MI-DORF model, where some instance labels $\mathbf{h}$ are also observable during training. (c) Factor graph of standard Latent-Dynamical models such as HCRF or HCORF. Linear-chain connectivity between latent states $\mathbf{h}$ is preserved, thus allowing efficient inference mechanisms using the forward-backward algorithm (see Appendix A.3) (d) Equivalent model to MI-DORF defined using the auxiliary variables ζ_t for each latent ordinal state. The use of these auxiliary variables and the redefinition of node and edge potentials allows to perform efficient inference by removing the high-order dependency introduced by the potential Ψ^M (see Sec. 4.5.3 and 4.6.2).

by a set of cut-off points $b_0 = -\infty \leq \cdots \leq b_L = \infty$. Every pair of contiguous cut-off points divide the projection values into different bins corresponding to the different ordinal states $l = 1, ..., L$. The difference between the two CDFs provides the probability of the latent state l given the observation $\mathbf{x}$, where σ is the standard deviation of a Gaussian noise contaminating the ideal model (see [Kim and Pavlovic, 2010a] for more details). In our case, we fix $\sigma = 1$, to avoid model over-parametrization.

MaxMI-DORF: Edge potentials

The edge potential $\Psi^E(h_t, h_{t+1}; \theta^E)$ models temporal information regarding compatibilities between consecutive latent ordinal states as:

$$\Psi^E(h_t = l, h_{t+1} = l'; \theta^E) = f(\mathbf{W}_{l,l'}), \tag{4.6}$$

where $\theta^E = \mathbf{W}^{L \times L}$ represents a real-valued transition matrix, as in standard HCRF, and f is a non-linear function defined as:

$$f(s) = -\log(1 + exp(-s)), \tag{4.7}$$

The motivation of using f is to maintain the same range between the values of node and edge potentials. Concretely, f bounds the value of Ψ^E between $[0, -\infty]$ as in the case of the previously defined node potentials.

MaxMI-DORF: Multi-Instance-Ordinal potential

In order to model the MaxMI-DOR assumption (see Eq. 4.1), we define a high-order potential $\Psi^M(\mathbf{h}, y; \theta^M)$ involving label y and all the sequence latent variables $\mathbf{h}$ as:

$$\Psi^M(\mathbf{h}, y; \theta^M) = \begin{cases} w \sum_{t=1}^{T} \mathbf{I}(h_t == y) & \text{iff } \max(\mathbf{h}) = y \\ -\infty & \text{otherwise} \end{cases}, \tag{4.8}$$

where $\mathbf{I}$ is the indicator function, and $\theta^M = w$. Note that when the maximum value within $\mathbf{h}$ is not equal to y, the energy function is equal to $-\infty$ and, thus, the probability $P(y|\mathbf{X};\theta)$ drops to 0. On the other hand, if the MaxMI-DOR assumption is fulfilled, the summation $w\sum_{t=1}^{T}\mathbf{I}(h_t == y)$ increases the energy proportionally to w and the number of latent states $\mathbf{h} \in h_t$ that are equal to y. This is convenient since, in sequences annotated with a particular label, it is more likely to find many latent ordinal states with such ordinal level. Therefore, the defined potential does not only model the MaxMI-DOR assumption but also provides mechanisms to learn how important is the proportion of latent states $\mathbf{h}$ that are equal to the label. Eq. 4.8 is a special case of cardinality potentials [Gupta et al., 2007] also employed in binary Multi-Instance Classification [Hajimirsadeghi et al., 2013].

4.5.2 MaxMI-DORF: Learning

Given a training set $\mathcal{T}$, we learn the model parameters θ by minimizing the regularized log-likelihood:

$$\min \quad \sum_{i=1}^{N} \log P(y|\mathbf{X};\theta) + \mathcal{R}(\theta), \tag{4.9}$$

where the regularization function $\mathcal{R}(\theta)$ over the model parameters is defined as:

$$\mathcal{R}(\theta) = \alpha(||\beta||_2^2 + ||\mathbf{W}||_F^2) \tag{4.10}$$

and α is set via a validation procedure. The objective function in Eq.4.9 is differentiable and standard gradient descent methods can be applied for optimization. To this end, we use the L-BFGS Quasi-Newton method [Byrd et al., 1994] (see also Appendix A.1). The gradient evaluation involves marginal probabilities $p(h_t|\mathbf{X})$ and $p(h_t, h_{t+1}|\mathbf{X})$ which can be efficiently computed using the proposed algorithm in Sec. 4.5.3.

4.5.3 MaxMI-DORF: Inference

The evaluation of the conditional probability $P(y|\mathbf{X};\theta)$ in Eq.4.3 requires computing $\sum_h e^{-\Psi(\mathbf{X},\mathbf{h},y;\theta)}$ for each label y. Given the exponential number of possible latent states $\mathbf{h} \in \mathcal{H}$, efficient inference algorithms need to be used. In the case of Latent-Dynamic Models such as HCRF/HCORF, the forward-backward algorithm [Barber, 2012] can be applied (see also Appendix A.3). This is because the pair-wise linear-chain connectivity between latent states $\mathbf{h}$. However, in the case of MaxMI-DORF, the inclusion of the cardinality potential $\Psi^M(\mathbf{h}, y; \theta^M)$ introduces a high-order dependence between the label y and all the latent states in $\mathbf{h}$. Inference methods with cardinality potentials have been previously proposed in [Gupta et al., 2007, Tarlow et al., 2012]. However, these algorithms only consider the case where latent variables are independent and, therefore, they can not be applied in our case. For these reasons, we propose an specific inference method. The idea behind it is to apply the standard forward-backward algorithm by converting the energy function defined in Eq. 4.4 into an equivalent one preserving the linear-chain connectivity between latent states $\mathbf{h}$.

To this end, we introduce a new set of auxiliary variables $\boldsymbol{\zeta} = \{\zeta_1, \zeta_2, ..., \zeta_T\}$, where each $\zeta_t \in \{0, 1\}$ takes a binary value denoting whether the sub-sequence $\mathbf{h}_{1:t}$ contains at least one ordinal state h equal to y. Now we define an alternative MaxMI-DORF energy function Ψ_* as:

$$\begin{aligned}\Psi_*(\mathbf{X}, \mathbf{h}, \boldsymbol{\zeta}, y; \theta) = &\sum_{t=1}^{T} \Psi_*^N(\mathbf{x}_t, h_t, \zeta_t, y; \theta^N) \\ &+ \sum_{t=1}^{T-1} \Psi_*^E(h_t, h_{t+1}, \zeta_t, \zeta_{t+1}, y; \theta^E),\end{aligned} \tag{4.11}$$

where the new node pontentials Ψ_*^N and edge potentials Ψ_*^E are

given by:

$$\Psi_*^N = \begin{cases} \Psi^N(\mathbf{x}_t, h_t; \theta^N) + w\mathbf{I}(h_t = y) \text{ iff } h_t <= y \\ -\infty \quad \text{otherwise} \end{cases} , \tag{4.12}$$

$$\Psi_*^E = \begin{cases} \mathbf{W}_{h_t, h_{(t+1)}} & \text{iff} \quad \zeta_t = 0 \wedge \zeta_{t+1} = 0 \wedge h_{t+1} \neq y \\ \mathbf{W}_{h_t, h_{(t+1)}} & \text{iff} \quad \zeta_t = 0 \wedge \zeta_{t+1} = 1 \wedge h_{t+1} = y \\ \mathbf{W}_{h_t, h^{(t+1)}} & \text{iff} \quad \zeta_t = 1 \wedge \zeta_{t+1} = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{4.13}$$

Note that Eq. 4.11 does not include the MI potential and, thus, the high-order dependence between the label y and latent ordinal-states $\mathbf{h}$ is removed. The graphical representation of MI-DORF with the redefined energy function is illustrated in Fig.4.3(b). In order to show the equivalence between energies in Eqs. 4.4 and 4.11, we explain how the original Multi-Instance-Ordinal potential Ψ^M is incorporated into the new edge and temporal potentials. Firstly, note that Ψ^N now also takes into account the proportion of ordinal variables h_t that are equal to the sequence label. Moreover, it enforces $\mathbf{h}$ not to contain any h_t greater than y, thus aligning the bag and (max) instance labels. However, the original Multi-Instance-Ordinal potential also constrained $\mathbf{h}$ to contain at least one h_t with the same ordinal value than y. This is achieved by using the set of auxiliary variables ζ_t and the re-defined edge potential Ψ^E. In this case, transitions between latent ordinal states are modelled but also between auxiliary variables ζ_t. Specifically, when the ordinal state in h_{t+1} is equal to y, the sub-sequence $\mathbf{h}_{1:t+1}$ fulfills the MaxMI-DOR assumption and, thus, ζ_{t+1} is forced to be 1. By defining the special cases at the beginning and the end of the sequence ($t = 1$ and $t = T$):

$$\Psi_*^N(\mathbf{x}_1, h_1, \zeta_1, y) = \begin{cases} \Psi_*^N & \text{iff } \zeta_1 = 0 \wedge l_1 < y \\ \Psi_*^N & \text{iff } \zeta_1 = 1 \wedge l_1 = y \\ -\infty & \text{otherwise} \end{cases} , \tag{4.14}$$

$$\Psi_*^N(\mathbf{x}_T, h_T, \zeta_T, y) = \begin{cases} \Psi_*^N(\text{ iff } \zeta_T = 1 \wedge h_T <= y \\ -\infty \quad \text{otherwise} \end{cases} \tag{4.15}$$

we can see that the energy is $-\infty$ when the MaxMI-DOR assumption is not fulfilled. Otherwise, it has the same value than the one defined in Eq.4.4 since no additional information is given. The advantage of using this equivalent energy function is that the standard forward-backward algorithm can be applied to efficiently compute the conditional probability:

$$P(y|\mathbf{X};\theta) = \frac{\sum_{\mathbf{h}} \sum_{\boldsymbol{\zeta}} e^{-\Psi_*(\mathbf{X},\mathbf{h},\boldsymbol{\zeta},y;\theta)}}{\sum_{y'} \sum_{\mathbf{h}} \sum_{\boldsymbol{\zeta}} e^{-\Psi_*(\mathbf{X},\mathbf{h},\boldsymbol{\zeta},y';\theta)}}, \tag{4.16}$$

The proposed procedure has a computational complexity of $\mathcal{O}(T \cdot (2L)^2)$ compared with $\mathcal{O}(T \cdot L^2)$ using standard forward-backward in traditional linear-chain latent dynamical models. Since typically $L << T$, this can be considered a similar complexity in practice. The presented algorithm can also be applied to compute the marginal probabilities $p(h_t|\mathbf{X})$ and $p(h_t, h_{t+1}|\mathbf{X})$. This probabilities are used during training for gradient evaluation and during testing to predict ordinal labels at the instance and bag level.

4.6 Relative-Multi-Instance DORF (RelMI-DORF)

In this section, we present the proposed Relative-Multi-Instance Dynamic Odinal Random Fields to solve the RelMI-DOR problem described in Sec. 4.4.

4.6.1 RelMI-DORF: Model Definition

In RelMI-DORF, ordinal and node potentials are specified as in MaxMi-DORF. However, the Multi-Instance potential $\Psi^M(\mathbf{h}, y)$ it

is now defined as shown in Eq. 4.17. In this case, the potential models the RelMI-DOR assumption, i.e, the weak-relation between the sequence label y and the evolution of latent instance labels $\mathbf{h}$ (see Eq. 4.2).

$$\Psi^M = \begin{cases} 0 \text{ iff } (\exists\, t h_t < h_{t+1}) \wedge (\forall t\, h_t \leq h_{t+1}) \wedge y = \uparrow \\ 0 \text{ iff } (\exists t\, h_t > h_{t+1}) \wedge (\forall t\, h_t \geq h_{t+1}) \wedge y = \downarrow \\ 0 \text{ iff } (\exists t\, h_t > h_{t+1}) \wedge (\exists t\, h_t < h_{t+1}) \wedge y = \updownarrow \\ 0 \text{ iff } (\forall t\, h_t = h_{t+1}) \wedge y = \emptyset \\ -\infty \text{ otherwise} \end{cases} \tag{4.17}$$

Learning in RelMI-DORF can be performed following the same procedure described in Sec. 4.5.2. However, inference requires a special treatment which is described in the following section.

4.6.2 RelMI-DORF: Inference

Similar to the case of MaxMI-DORF, the high-order potential $\Psi^N(\mathbf{h}, y)$ in RelMI-DORF prevents to perform inference using the standard forward-backward procedure. For this purpose, we follow a similar strategy than the one described in Sec. 4.5.3. However, in this case, auxiliary variables ζ_t are defined according to the possible sequence labels in RelMI-DOR. Concretely, $\zeta_t \in \{\uparrow, \downarrow, \emptyset, \updownarrow\}$ indicates the label of the subsequence $\mathbf{h}_{1:t}$ according to the definitions given in Eq. 4.2. The equivalent energy function incorporating this auxiliary variables $\boldsymbol{\zeta}$ can be obtained by redefining the original edge potentials as:

$$\Psi_*^E = \begin{cases} \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \emptyset \wedge \zeta_{t+1} = \emptyset \wedge h_t = h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \emptyset \wedge \zeta_{t+1} = \uparrow \wedge h_t < h_{t+1} \\ \mathbf{W}_{h_t,h^{(t+1)}} & \text{iff } \zeta_t = \emptyset \wedge \zeta_{t+1} = \downarrow \wedge h_t > h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \uparrow \wedge \zeta_{t+1} = \uparrow \wedge h_t \leq h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \uparrow \wedge \zeta_{t+1} = \updownarrow \wedge h_t > h_{t+1} \\ \mathbf{W}_{h_t,h^{(t+1)}} & \text{iff } \zeta_t = \downarrow \wedge \zeta_{t+1} = \downarrow \wedge h_t \geq h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \downarrow \wedge \zeta_{t+1} = \updownarrow \wedge h_t < h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \updownarrow \wedge \zeta_{t+1} = \updownarrow \\ -\infty & \text{otherwise} \end{cases} \tag{4.18}$$

Again, defining the special cases for node potentials at the beginning and ending of the sequence:

$$\Psi_*^N(\mathbf{x}_1, h_1, \zeta_1, y) = \begin{cases} \Psi^N(\mathbf{x}_1, h_1, y) \text{ iff } \zeta_1 = \emptyset \\ -\infty \text{ otherwise} \end{cases}, \tag{4.19}$$

$$\Psi_*^N(\mathbf{x}_T, h_T, \zeta_T, y) = \begin{cases} \Psi^N(\mathbf{x}_T, h_T, y) \text{ iff } \zeta_T = y \\ -\infty \text{ otherwise} \end{cases}, \tag{4.20}$$

it can be shown that the energy function becomes $-\infty$ when the sequence level is not coherent with the evolution of latent instance labels $\mathbf{h}$ (according to RelMI-DOR assumption). Otherwise, it takes the same value than the energy function defined by the original potentials. In this case, computational complexity is $\mathcal{O}(T \cdot (4L)^2)$, which is still linear in terms of the number of instances T.

4.7 Partially-Observed MI-DOR (PoMI-DOR

Although labels at sequence-level are easier to collect, in some applications is feasible to annotate a small subset of the sequence's

instances. In this case, we are interested in learning the model by using weak-labels y but also incorporating the information of these additional annotations. We refer to this problem as Partially-Observed Multi-Instance Dynamic Ordinal Regression (PoMI-DOR). In this case, the training set is formed by triples $\mathcal{T} = \{(\mathbf{X}_1, y_1, \mathbf{h}_1^a) , (\mathbf{X}_2, y_2, \mathbf{h}_2^a) , ..., (\mathbf{X}_N, y_N, \mathbf{h}_N^a)\}$, where $\mathbf{h}_n^a$ contains ground-truth annotations for a subset of sequence instances. Formally, the set $\mathbf{h}_n = \{\mathbf{h}_n^a \cup \mathbf{h}_n^u\}$, where $\mathbf{h}_n^u$ is the subset of ordinal labels corresponding to non annotated instances. Note that when the set $\mathbf{h}_u$ is empty, the problem becomes standard Supervised Dynamic Ordinal Regression. Under this setting, we extend MI-DORF to learn a model maximizing the log-likelihood function of the conditional probability:

$$P(y, \mathbf{h}_a|\mathbf{X}; \theta) = \frac{\sum_{\mathbf{h}^u} e^{-\Psi(\mathbf{X}, \mathbf{h}^u, \mathbf{h}^a, y; \theta)}}{\sum_{y'} \sum_{\mathbf{h}^u} \sum_{\mathbf{h}^a} e^{-\Psi(\mathbf{X}, \mathbf{h}^u, \mathbf{h}^a, y'; \theta)}}, \tag{4.21}$$

for all the sequences in the training set. Note that in this case, the knowledge provided by annotated instances $\mathbf{h}_n^a$ is incorporated into the likelihood function. In order to learn a PoMI-DORF model , the same algorithms presented in Secs. 4.5 and 4.6 can be applied. However, during inference we need to take into account annotations $\mathbf{h}_n^a$ for each sequence. This can be easily achieved by redefining the original node potentials in RelMI-DORF and MaxMI-DORF as:

$$\Psi^N(\mathbf{x}, h_t) = \begin{cases} -\infty \ \text{ iff } \ (h_t \in \mathbf{h}^a) \wedge (h_t^a \neq h_t) \\ \Psi^N(\mathbf{x}_t, h_t) \ \text{ otherwise} \end{cases} , \tag{4.22}$$

Intuitively, observed instance labels $\mathbf{h}^a$ are treated as hard evidences which make the energy function to take a value of $-\infty$ when $\mathbf{h}$ is not consistent with them. This strategy has been previously followed in order to learn Conditional Random Fields [Li et al., 2009] under the partially-observed setting.

4.8 Experiments

In this section, we describe the experiments performed in order to evaluate the presented MaxMI-DORF and RelMI-DORF methods. For each case, we perform experiments over synthetic and real data. For MaxMI-DOR we consider the problem of weakly-supervised Pain Intensity estimation, where sequence-labels correspond to the maximum pain felt by the subject. On the other hand, for RelMI-DOR we test our approach in Action Unit intensity prediction, where we assume that only onset and apex labels for video segments are available during training.

4.8.1 Compared methods

The presented framework is designed to address Multiple Instance Learning problems when bags are structured as temporal sequences of instances with ordinal labels. Given that this has not been attempted before, we compare them with alternative methods that can be also used in these problems but present some limitations: either ignore the MIL assumptions (Single-Instance), do not model dynamic information (Static) or do not take into account the ordinal nature of instance labels.

Single-Instance Ordinal Regression (SIL-OR): MaxMI-DOR can be posed as a supervised learning problem with noisy labels. The main assumption is that the majority of instances will have the same label than their bag. In order to test this assumption, we train standard Ordinal Regression [Winkelmann and Boes, 2006] at instance-level by setting all their labels to the same value as their corresponding bag. This baseline can be considered an Static-SIL approach to solve the MaxMI-DOR problem.

Static Multi-Instance Ordinal Regression (MI-OR): Again for MaxMI-DOR, we have implemented this Static Multi-Instance approach. This method is inspired by MI-SVM [Andrews et al., 2003], where instance labels are considered latent variables and are itera-

tively optimized during training. To initialize the parameters of the ordinal regressor, we follow the same procedure as described above in SIL-OR. Then, ordinal values for each instance are predicted and modified so that the MaxMI-DOR assumption is fulfilled for each bag. Note that if all the predictions within a bag are lower than its label, the instance with the maximum value are set to the bag-label. On the other hand, all the predictions greater than the bag-label are decreased to this value. With this modified labels, Ordinal Regression is applied again and this procedure is applied iteratively until convergence.

Multi-Instance-Regression (MIR): As discussed in Sec. 4.1, the MaxMI-DOR problem is closely related with Multiple-Instance-Regression. Several methods have been proposed in the literature for MIR when instance labels are real-valued variables. In order to evaluate the performance of this strategy, we have implemented a similar method as used in [Hsu et al., 2014]. Specifically, a linear regressor at the instance-level is trained by optimizing a loss function over the bag-labels. This loss models the MIR assumption by using a soft-max function which approximates the maximum instance label within a bag predicted by the linear regressor. Note that a similar approach is also applied in Multi-Instance Logistic Regression [Ray and Craven, 2014]. In these works, a logistic loss is used because instance labels take values between 0 and 1. However, we use a squared-error loss to account for the different ordinal levels.

MaxMI-DRF: This approach is similar to the proposed MaxMI-DORF. However, MaxMI-DRF ignores the ordinal nature of labels and models them as categorical variables. For this purpose, we replace the MaxMI-DORF node potentials by a multinomial logistic regression model [2]. Inference is performed by using the same algorithm described in Sec. 4.5.3.

[2]The potential with the Multinomial Logistic Regession model is defined as $\log(\frac{\exp(\beta_l^T x)}{\sum_{l' \in L} \exp(\beta_{l'}^T x)})$. Where all $\mathbf{fi_l}$ defines a linear projection for each possible ordinal value l [Walecki et al., 2015]

RelMI-DRF: Similar to MaxMI-DRF, this method is equivalent to RelMI-DORF but modelling instance labels as categorical variables.

Latent-Dynamic Models (HCRF/HCORF): In MaxMI-DOR and Rel-MIDOR a label at sequence-level is provided during training. Therefore, it is possible to apply existing Latent-Dynamic Models such as HCRF [Quattoni et al., 2007] or HCORF [Kim and Pavlovic, 2010a] for both problems. Despite these two methods model dynamics and incorporate the information provided by sequence-labels, they do not explicitly take into account the Multi-Instance assumptions.

Ordinal Support Vector Regression (OSVR): This method presented in [Zhao et al., 2016c] can be applied for RelMI-DOR. However,it is an Static approach that do not consider dynamic information. Moreover, it models instance labels as continuous variables instead of ordinal.

Methods for Partially-Observable MI-DOR: In our experiments, we evaluate Max-MIDORF and Rel-MIDORF when some instance labels are also available during training (see Sec. 4.7). In order to compare their performance under this setting, we evaluate the partially-observed extensions of CRF [Li et al., 2009] and HCRF [Chang et al., 2009]. Ordinal versions of these two approaches has been also implemented.

Methods for Supervised Dynamical Ordinal Regression: To fully evaluate the performance of methods trained using only weak-labels, we compare the previous described methods with two fully-supervised models for sequence classification CRF [Lafferty et al., 2001] and CORF [Kim and Pavlovic, 2010b]. These approaches are learned with complete information (i.e, instance labels for all the training instances).

4.8.2 Evaluation and Metrics

In order to evaluate the performance of the different methods, we report results in terms of instance-labels predictions. Note that in

the MIL literature, results are usually reported at bag-level. However, in MI-DOR problems, the only goal is to predict instance labels (pain or AU intensities) inside the bag (video). Given the ordinal nature of the labels, we use Pearson's Correlation (CORR), Mean-Average-Error (MAE) and Intra-Class-Correlation (ICC) as evaluation metrics. In all our experiments, we use a portion of training sequences as a validation set. This set is used to optimize the different regularization parameters for all the methods using standard grid-search strategy. Specifically, for MaxMI-DOR and RelMI-DOR problems, we optimize these parameters according to the at sequence level in the validation set. In the case of partially-observed problems, we consider instance-level predictions performance for the subset of available instance labels.

4.8.3 MaxMI-DOR and RelMI-DOR: Synthetic Data

Synthetic Data generation:

Given that no standard benchmarks are available for MI-DOR problems, we have generated synthetic data. In order to create sequences for MaxMI-DOR, we firstly sample a sequence of ordinal values using a random transition matrix representing change probabilities between temporally-consecutive ordinal levels. Secondly, we generate random parameters of an Ordinal Regressor as defined in Eq. 4.5. This regressor is used to compute the probabilities for each ordinal level in a set of feature-vectors randomly sampled from a Gaussian distribution. Thirdly, the corresponding sequence observation for each latent state in the sequence is randomly chosen between the sampled feature vectors according to the obtained probability for each ordinal value. Finally, the sequence-label is set to the maximum ordinal state within the sequence following the MaxMI-DOR assumption and Gaussian noise ($\sigma = 0.25$) is added to the feature vectors. Fig. 4.4(a-c) illustrates this procedure. For RelMI-DOR, we follow a similar strategy

to generate the synthetic sequences. However, the transition matrix is forced to contain a probability of 0 for decreasing transitions in case the sequence label is $y = \uparrow$ and for increasing transitions if $y = \downarrow$. For testing, we create unsegmented sequences (with increasing and decreasing transitions) by concatenating two segments generated following the previous procedure.

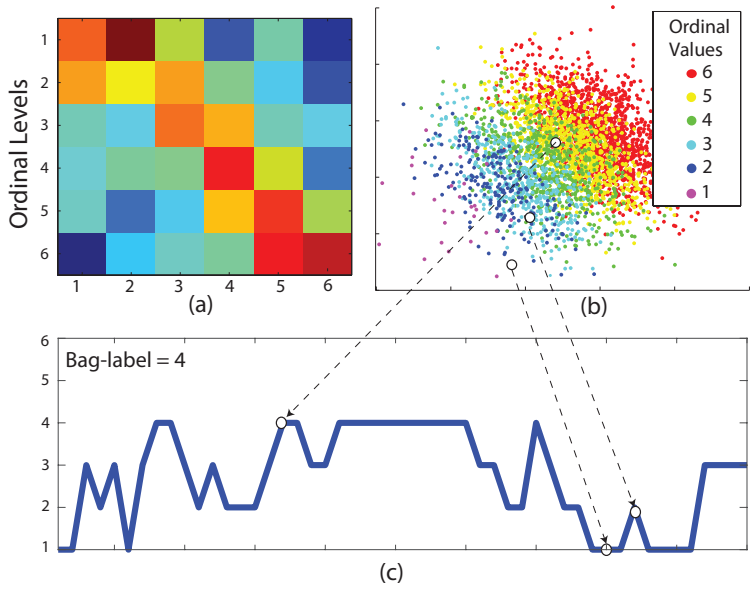

Figure 4.4: Description of the procedure used to generate synthetic sequences. (a) A random matrix modelling transition probabilities between consecutive latent ordinal values. (b) Ordinal levels assigned to the random feature vectors according to the ordinal regressor. (c) Example of a sequence of ordinal values obtained using the generated transition matrix. The feature vector representing each observation is randomly chosen between the samples in (b) according to the probability for each ordinal level.

Experimental setup and results

Following the strategy described above, we have generated ten different data sets for RelMI-DOR and MaxMI-DOR by varying the ordinal regressor parameters and transition matrix. Specifically, each dataset is composed of 100 sequences for training, 150 for testing and 50 for validation. The sequences have a variable length between 50 and 75 instances in MaxMI-DOR and between 15 and 25 in RelMI-DOR. The dimensionality of the feature vectors was set to 10 and the number of ordinal values to 6. For partially-observed MI-DOR, we have randomly choose one instance per sequence of which its label is also used during training. Table 4.1 and 4.2 shows the results computed as the average performance over the ten datasets for MaxMI-DOR and RelMI-DOR respectively. We also report results for fully-supervised CRF and CORF trained considering all the instance labels.

MaxMI-DOR discussion

In the MaxMI-DOR problem, SIL methods (SIL-OR, HCRF and HCORF) obtain lower performance than their corresponding MIL versions (MI-OR, MaxMI-DRF and MaxMI-DORF) in all the evaluated metrics. This is expected since SIL approaches ignore the Multi-Instance assumption. Moreover, HCORF and MaxMI-DORF obtain better performance compared to HCRF and MaxMI-DRF. This is because the former model instance labels as nominal variables, thus, ignoring their ordinal nature. Finally, note that MI-DORF outperforms the static methods MI-OR and MIR. Although these approaches use the Multi-Instance assumption and incorporate the label ordering, they do not take into account temporal information. In contrast, MaxMI-DORF is able to model the dynamics of latent ordinal states and use this information to make better predictions when sequence observations are noisy. As Fig. 4.5(a) shows, MI-OR predictions tend to be less smooth because dynamic information is not taken into account. In contrast, MaxMI-DORF better estimate the actual ordinal

levels by modelling transition probabilities between consecutive ordinal levels. Looking into the results achieved by the different methods in the PoMI-DOR setting, we can derive the following conclusions. Firstly, HCORF and HCRF improve their performance by taking into account the additional information provided by instance labels. However, we can observe that, under this setting, CRF and CORF obtain lower results than HCORF and HCRF. This is because the latter are able to use the sequence-label information together with the provided by labelled instances. Secondly, observe that MaxMI-DRF and MaxMI-DORF still achieve better performance than methods that do not consider the MIL assumption (CORF, CRF, HCRF and HCORF). This shows the importance of explicitly incorporate the MaxMI-DOR assumption in the model even though instance labels can be available during training. Finally, note that MaxMI-DORF obtains again the best performance, even close to fully-supervised CRF and CORF. This suggest that the need of annotated instances is highly-reduced if the weak-information provided by sequence labels is properly used following the MIL assumption.

Table 4.1: Results on Synthtic Data (MaxMI-DOR)

Setting	Method	CORR ↑	MAE ↓	ICC ↑
MaxMI-DOR	**SI-OR**	0.79	1.31	0.46
	MI-OR	0.82	0.62	0.70
	HCRF [Quattoni et al., 2007]	0.05	1.99	0.05
	HCORF [Kim and Pavlovic, 2010a]	0.73	0.74	0.65
	MIR [Hsu et al., 2014]	0.79	0.65	0.69
	MaxMI-DRF	0.77	0.77	0.71
	MaxMI-DORF	**0.86**	**0.41**	**0.85**
PoMaxMI-DOR (1 sample/seq.)	**CRF [Li et al., 2009]**	0.74	0.63	0.74
	CORF [Li et al., 2009]*	0.84	0.46	0.83
	HCRF [Chang et al., 2009]	0.79	0.57	0.78
	HCORF [Chang et al., 2009]*	0.86	0.42	0.85
	MaxMI-DRF	0.82	0.52	0.81
	MaxMI-DORF	**0.87**	**0.38**	**0.87**
Supervised DOR	**CRF [Lafferty et al., 2001]**	0.88	0.35	0.88
	CORF [Kim and Pavlovic, 2010b]	**0.89**	**0.35**	**0.88**

(*)Indicates an originally nominal method that we have extended to deal with ordinal labels.

Table 4.2: Results on Synthtic Data (RelMI-DOR)

Setting	Method	CORR ↑	MAE ↓	ICC ↑
RelMI-DOR	**HCRF [Quattoni et al., 2007]**	0.36	1.82	0.32
	HCORF [Kim and Pavlovic, 2010a]	0.85	1.32	0.80
	OSVR [Zhao et al., 2016c]	0.87	3.51	0.10
	RelMI-DRF	0.77	1.36	0.49
	RelMI-DORF	**0.89**	**0.74**	**0.84**
PoRelMI-DOR (1 sample/seq.)	**CRF [Li et al., 2009]**	0.82	0.64	0.81
	CORF [Li et al., 2009]*	0.89	0.43	0.89
	HCRF [Chang et al., 2009]	0.83	0.60	0.83
	HCORF [Chang et al., 2009]*	0.89	0.44	0.88
	OSVR [Zhao et al., 2016c]	0.87	0.61	0.85
	RelMI-DRF	0.88	0.49	0.87
	RelMI-DORF	**0.92**	**0.36**	**0.91**
Supervised DOR	**CRF [Lafferty et al., 2001]**	0.93	0.31	0.93
	CORF [Kim and Pavlovic, 2010b]	**0.93**	**0.29**	**0.93**

(*)Indicates an originally nominal method that we have extended to deal with ordinal labels.

RelMI-DOR discussion

In the RelMI-DOR problem, we observe similar results as in MaxMI-DOR. Firstly, note that non-ordinal approaches (HCRF and RelMI-DRF) obtain the worst performance in most cases. Secondly, RelMI-DORF obtains better performance than HCORF by explicitly modelling the Multi-Instance-Assumption. Finally, OSVR achieves a competitive performance in terms of correlation compared with RelMI-DORF. However, it obtains poor results in terms of MAE and ICC. As discussed in Sec. 4.3, OSVR considers labels as continuous variables and does not explicitly model the RelMI-DOR assumption. Instead, it only ranks the instance labels within the sequence. Therefore, it fails to properly estimate the actual scale of the predicted values.

When some instance labels are provided (PoRel-MIDOR), all the methods improve their performance by exploiting this additional information. However, the improvement in terms of MAE and ICC is much higher than for correlation. This is because in the RelMI-DOR problem, sequence labels only provide information about the evolution of instance labels within the sequence. Therefore, models can achieve a good performance predicting sequence-labels even though

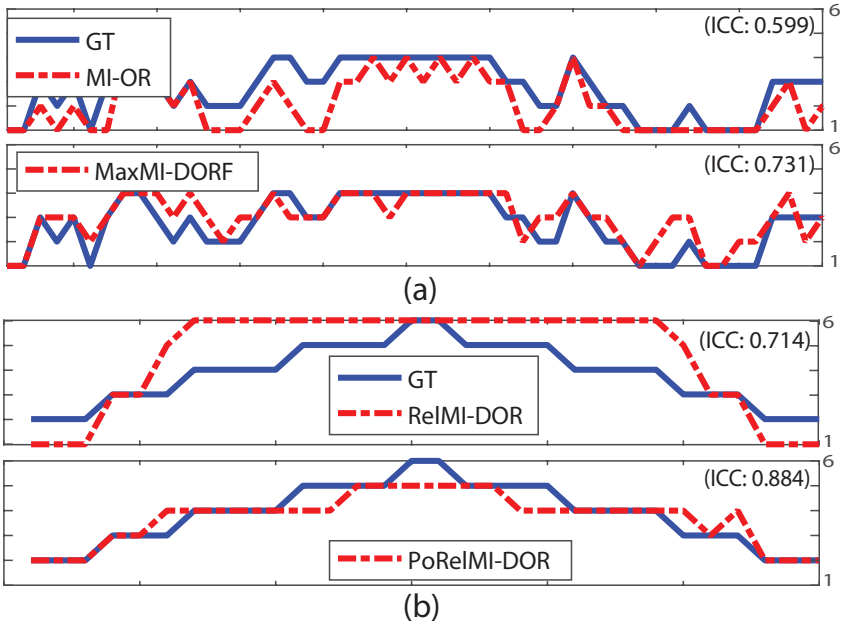

Figure 4.5: (a) Examples of instance-level predictions in a sequence for MI-OR and MaxMI-DORF. (b) Examples of instance-level predictions in a sequence for RelMI-DORF in the case of non-observed and partially-observed instance labels during training.

the ordinal levels are not accurate. In contrast, when some instance labels are incorporated during training, a better estimate of the ordinal levels can be achieved. This is illustrated in Fig. 4.5(b) where RelMI-DORF predictions are more accurate when it is trained using instance labels. Finally, note that RelMI-DORF under the PoRelMI-DOR setting achieves again competitive performance compared to fully-supervised CRF and CORF. This supports our hypothesis that Dynamic Ordinal Regression problems can be addressed using the proposed MIL framework.

4.8.4 MaxMI-DOR: Weakly-supervised pain intensity estimation

In this experiment, we test the performance of MaxMI-DORF for weakly-supervised Pain Intensity estimation. As detailed in Sec. 4.1, our main motivation is that pain intensity annotation is very time consuming. However, the maximum pain felt by a subject during a sequence is much easier to obtain.

UNBC Dataset

We use the UNBC Shoulder-Pain Database [Lucey et al., 2011] which contains recordings of different subjects performing active and passive arm movements during rehabilitation sessions. In this dataset, pain intensities at each frame are given in terms of the PSPI scale [Prkachin, 1992]. This ordinal scale ranges from 0 to 15. Given the imbalance between low and high pain intensity levels, we follow the same strategy as [Rudovic et al., 2015]. Specifically, pain labels are grouped into 5 ordinal levels as: 0(0),1(1),2(2),3(3),4-5(4),6-15(5). These frame-by-frame pain annotations are considered the instance labels in MaxMI-DOR. On the other hand, bag (video) labels are extracted as the maximum pain level within each sequence. In order to extract facial-descriptors at each video frame representing the bag instances, we compute a geometry-based facial-descriptor as follows. Firstly, we obtain a set of 49 landmark facial-points with the method described in [Xuehan-Xiong and De la Torre, 2013]. Then, the obtained points locations are aligned with a mean-shape using Procrustes Analysis. Finally, we generate the facial descriptor by concatenating the x and y coordinates of the aligned points.

Experimental setup and results

Similar to the experiment with synthetic data (Sec. 4.8.4), we consider two scenarios for weakly-supervised pain intensity estimation. The first one is the MaxMI-DOR setting, where only bag labels are

used. Apart from the baselines described in Sec. 4.8.1, in this scenario we also evaluate the performance of the approach presented in [Sikka et al., 2013] which considers pain levels as binary variables. For this purpose, we use the MILBoosting [Zhang et al., 2005b] method employed in the cited work and considered videos with a pain label greater than 0 as positive. Given that MI-Classification methods are only able to make binary predictions, we use the output probability as indicator of intensity levels , i.e., the output probability is normalized between 0 and 5. In the second scenario (Partially-Observed MaxMI-DOR setting), we randomly select different percentages of annotated frames inside each sequence. This simulates that the time required to annotate the dataset has been significantly reduced by only labelling a small subset of the frames. Concretely, we consider the 5% and 10% of annotated frames in each sequence. Under these different experimental setups, we perform Leave-One-Subject-Out Cross Validation where, in each cycle, we use 15 subjects for training, 1 for testing and 9 for validation. In order to reduce computational complexity and redundant information between temporal consecutive frames, we have down-sampled the sequences using a time-step of 0.25 seconds. Table 4.3 shows the results obtained by the evaluated methods following the described procedure. Results for fully-supervised CRF and CORF are also reported.

Discussion

By looking into the results in the MaxMI-DOR setting, we can derive the following conclusions. Firstly, SI approaches (SI-OR, HCORF and HCRF) obtain worse performance than MI-OR and MIR. Specially, HCORF and HCRF obtain poor results. This is because pain events are typically very sparse in these sequences and most frames have intensity level 0 (neutral). Therefore, the use of the MIL assumption has a critical importance in this problem in order to correctly locate pain frames. Secondly, MIR and MI-OR obtain better results than MaxMI-DRF. This can be explained because the latter

Table 4.3: Results on the UNBC Database

Setting	Method	CORR ↑	MAE ↓	ICC ↑
MaxMI-DOR	**SI-OR**	0.22	2.20	0.08
	MI-OR	0.29	0.84	0.27
	MILBoost [Zhang et al., 2005b]	0.23	2.38	0.09
	HCRF [Quattoni et al., 2007]	0.09	1.73	0.05
	HCORF[Kim and Pavlovic, 2010a]	0.06	1.23	0.05
	MIR [Hsu et al., 2014]	0.32	1.03	0.25
	MaxMI-DRF	0.16	1.96	0.08
	MaxMI-DORF	**0.36**	**0.71**	**0.34**
PoMaxMI-DOR (5% of data)	**CRF [Li et al., 2009]**	0.31	0.66	0.30
	CORF [Li et al., 2009]*	0.39	0.58	0.38
	HCRF [Chang et al., 2009]	0.32	0.76	0.29
	HCORF [Chang et al., 2009]*	0.38	0.68	0.36
	MaxMI-DRF	0.32	0.72	0.30
	MaxMI-DORF	**0.43**	**0.52**	**0.42**
PoMaxMI-DOR (10% of data)	**CRF [Li et al., 2009]**	0.29	0.65	0.28
	CORF [Li et al., 2009]*	0.44	0.55	0.43
	HCRF [Chang et al., 2009]	0.34	0.63	0.32
	HCORF [Chang et al., 2009]*	0.45	0.58	0.44
	MaxMI-DRF	0.34	0.55	0.34
	MaxMI-DORF	**0.46**	**0.51**	**0.46**
Supervised DOR	**CRF [Lafferty et al., 2001]**	0.45	**0.50**	0.44
	CORF [Kim and Pavlovic, 2010b]	**0.48**	0.56	**0.48**

(*)Indicates an originally nominal method that we have extended to deal with ordinal labels.

consider pain levels as nominal variables and is ignorant of the ordering information of the different pain intensities. Finally, MILBoost trained with binary labels also obtains low performance compared to the MI-OR and MIR. This suggest that current approaches posing weakly-supervised pain detection as a MI-Classification problem are unable to predict accurately the target pain intensities. By contrast, MaxMI-DORF obtains the best performance across all the evaluated metrics. We attribute this to the fact that it models the MIL assumption with ordinal variables. Moreover, the improvement of MaxMI-DORF compared to static approaches, such as MI-OR and MIR, suggests that modelling dynamic information is beneficial in this task.

In the Partially-observed setting, all the methods improve their performance by considering the additional information provided by

labelled instances. However, note that approaches modelling the ordinal structure of labels (CORF, HCORF and MaxMI-DORF) still outperform nominal methods (CRF, HCRF and MaxMI-DRF) under this setting. Moreover, MaxMI-DORF also achieves the best performance with 5% and 10% of labeled frames. Despite the other approaches also consider instance labels, MaxMI-DORF better exploits sequence labels information by explicitly modelling the MIL assumption. It is worth to mention that considering only 10% of annotated frames, MaxMI-DORF obtain competitive performance against fully-supervised approaches. Concretely, it outperforms CRF in terms of ICC/CORR and CORF in terms of MAE. This suggest that the effort needed to annotate pain intensity databases, could be highly-reduced using the proposed weakly-supervised framework. In order to give more insights into this issue, Fig. 4.6(b) shows the performance in terms of ICC as the percentage of annotated frames increases. As we can observe, MaxMI-DORF outperforms other methods with 0%, 5% and 10% of annotated frames. When this percentage increases to 25%, the performance of partially-observed CORF, HCORF and MaxMI-DORF is comparable to the one achieved by fully-supervised CORF. However, note that labelling 25% of samples does not suppose a significant reduction of the annotation time in a real scenario.

Finally, in Fig. 4.6(b) we show qualitative examples comparing predictions of the best evaluated methods under the different settings. When only bag-labels are used for training, MI-OR predictions are less accurate than the ones obtained by MaxMI-DORF. Moreover, MaxMI-DORF estimates better the actual pain levels in the partially-observed setting, where a small subset of instance labels are used. These predictions are more accurate than the ones obtained with partially-observed HCORF which does not take into account the MIL assumption. This is reflected by the ICC depicted in the sequences, showing that the proposed MaxMI-DORF method outperforms the competing approaches on target data.

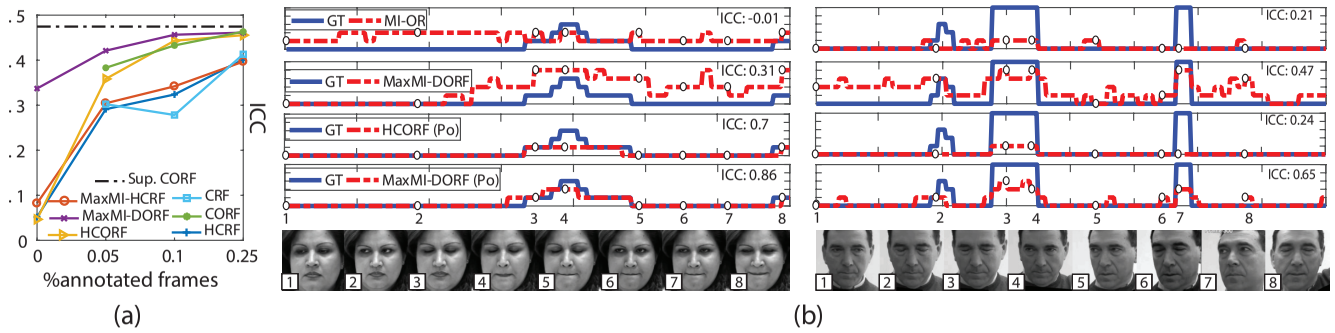

Figure 4.6: (a) ICC achieved in the UNBC dataset considering different percentages of labelled instances in the training set. Black line shows the performance of a fully-supervised CORF trained with all the instance labels. (b) Visualization of the pain intensity predictions in different sequences of the UNBC dataset. From top to bottom: MI-OR and MaxMI-DORF without using instance labels. Partially-observed HCORF and MaxMI-DORF using 10% of annotated frames.

4.8.5 RelMI-DOR: Weakly-supervised AU intensity estimation

In this section, we test the performance of RelMI-DORF for weakly-supervised Action Unit intensity estimation. Similarly to pain intensity, AU labelling requires a huge effort for expert coders. However, segmenting videos according to the increasing or decreasing evolution of AU intensities (i.e. onset and offset sequences) is less time-consuming.

DISFA Dataset

We employ the DISFA Database [Mavadati et al., 2013], which is a popular benchmark for AU intensity estimation. It contains naturalistic data consisting on 27 annotated sequences of different subjects watching videos eliciting different types of emotions. Specifically frame-by-frame AU intensities are provided for 12 AUs (1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, 26) in a six-point ordinal scale (*neutral*<A<B<C<D<E). As far as we know, this is the largest available dataset in terms of the number of Action Units annotated. Although the UNBC dataset also provides AU intensity annotations for 11 AUs, we found that the number of onset and appex events for each of them is very limited. Therefore, we discard it for this experiments. To the best of our knowledge, no previous works have evaluated DISFA under the weakly-supervised setting. The described AU annotations represent the instance labels in our RelMI-DOR problem. As previously discussed, bags are considered onset and apex sequences where the intensity of a given AU is monotone increasing ($y = \uparrow$) or decreasing ($y = \downarrow$). These segments have been automatically extracted with an exhaustive search over the whole video using the ground-truth intensity labels at frame-level. The intervals corresponding to the lowest and highest intensity levels in each segment are cut so that they do not last for more than 0.25 seconds. This procedure simulates that a given annotator has only labelled onset and offset segments instead of specific AU intensities for all the frames. The number of

extracted segments for each AU following this strategy is shown in Table 4.4. To compute the facial descriptors at each frame, we use the same procedure described in Sec. 4.8.4. However, for upper-face AUs (1,2,4,5,6 and 9) only the locations of landmark points corresponding to the eyes and eyebrows are used. Similarly, we only use the points extracted from the mouth and nose for lower-face AUs (12, 15, 17, 20, 25 and 26).

Experimental setup and results

Using the training segments for each AU, we evaluate the different methods using a subject-independent 5-fold cross validation. Specifically, 3 folds are used for training and 1 for testing and validation purposes. During testing, the trained models are evaluated on the original non-segmented videos. The motivation is that, in a real scenario, onset and apex segmentation is not known for testing sequences. We also consider the partially-observed setting, where labels for 5% and 10% of frames are available during training (PoRelMI-DOR). Table 4.5 shows the performance obtained by the evaluated methods computed as the average for all the considered AUs. Specific results in terms of ICC for independent AUs are shown in Table 4.4.

Discussion

When instance labels are not used during training (RelMI-DOR setting), we can observe that HCRF and HCORF obtain poor results compared to OSVR and RelMI-DORF. This can be explained because the former methods explicitly model the increasing/decreasing intensity constraints provided by sequence weak-labels. Moreover, the low results obtained by RelMI-DRF compared to RelMI-DORF suggest that modelling intensities as nominal variables is suboptimal in this scenario. Also note that OSVR obtains worse results in terms of ICC and MAE compared to RelMI-DORF. Given that performances in terms of CORR are more similar, it shows the limitation of OSVR to predict the actual scale of instance ordinal labels. Considering the

Table 4.4: Results (ICC) for independent AUs in the DISFA Database. In parentheses, number of onset and apex segments extracted

Setting	Method	AU1 (342)	AU2 (230)	AU4 (572)	AU5 (216)	AU6 (364)	AU9 (159)	AU12 (642)	AU15 (210)	AU17 (575)	AU20 (199)	AU25 (800)	AU26 (723)	AVG
RelMI-DOR	**HCRF [Quattoni et al., 2007]**	0.06	0.03	0.11	0.01	0.03	0.02	0.14	0.01	0.06	0.01	0.45	**0.24**	0.10
	HCORF [Kim and Pavlovic, 2010a]	0.02	0.01	0.05	0.08	0.02	0.01	0.04	0.01	0.01	0.00	0.06	0.01	0.03
	OSVR [Zhao et al., 2016c]	0.10	0.13	0.21	0.04	0.16	0.09	0.40	**0.09**	0.04	0.04	0.37	0.17	0.15
	RMI-HCRF	0.02	0.04	0.10	0.03	0.12	0.01	0.30	0.04	-0.02	0.02	0.40	0.22	0.11
	RMI-DORF	**0.34**	**0.30**	**0.27**	**0.17**	**0.30**	**0.10**	**0.60**	0.07	**0.08**	**0.04**	**0.70**	0.21	**0.26**
PoRelMI-DOR (5% of frames)	**CRF [Li et al., 2009]**	0.24	0.33	0.18	0.17	0.40	0.07	0.71	0.14	0.13	0.08	0.85	0.21	0.29
	CORF [Li et al., 2009]*	0.20	0.39	0.21	0.26	0.41	0.10	0.77	0.14	**0.15**	**0.11**	0.80	0.32	0.32
	HCRF [Chang et al., 2009]	0.26	0.35	0.18	0.17	0.42	0.08	0.72	0.10	0.13	0.08	**0.86**	0.23	0.30
	HCORF [Chang et al., 2009]*	0.24	0.34	0.25	**0.30**	0.40	0.10	**0.78**	0.15	**0.15**	**0.11**	0.81	0.35	0.33
	OSVR [Zhao et al., 2016c]	0.15	0.20	**0.30**	0.16	0.34	0.11	0.73	0.16	0.09	**0.09**	0.78	**0.37**	0.29
	RMI-HCRF	0.12	0.39	0.04	0.18	**0.51**	0.10	0.24	0.17	0.06	0.09	0.25	0.14	0.19
	RMI-DORF	**0.38**	**0.47**	0.28	0.29	0.44	**0.11**	**0.78**	**0.18**	**0.15**	**0.11**	0.78	0.35	**0.36**
PoRelMI-DOR (10% of frames)	**CRF [Li et al., 2009]**	0.27	0.44	0.21	0.19	0.46	0.06	0.72	**0.22**	0.16	0.07	**0.84**	0.23	0.32
	CORF [Li et al., 2009]*	0.26	0.45	0.28	0.32	0.39	0.11	0.76	0.17	0.09	0.09	0.78	0.31	0.33
	HCRF [Chang et al., 2009]	0.36	0.46	0.20	0.24	0.40	0.08	0.73	0.26	0.12	0.08	0.84	0.29	0.34
	HCORF [Chang et al., 2009]*	0.25	0.44	0.26	0.35	0.42	0.11	0.77	0.20	0.16	0.09	0.78	0.32	0.35
	OSVR [Zhao et al., 2016c]	0.15	0.22	0.29	0.17	0.34	**0.13**	0.74	0.17	0.10	0.09	0.77	**0.37**	0.29
	RMI-HCRF	0.28	0.44	0.24	0.21	**0.49**	0.08	0.71	0.20	0.14	0.12	0.72	0.22	0.32
	RMI-DORF	**0.39**	**0.50**	**0.29**	**0.39**	0.44	0.12	**0.78**	0.21	**0.17**	**0.11**	0.81	0.32	**0.38**
Supervised DOR	**CRF [Lafferty et al., 2001]**	0.33	0.44	0.26	0.33	**0.51**	0.08	0.74	**0.24**	**0.14**	**0.11**	**0.84**	0.24	0.35
	CORF [Kim and Pavlovic, 2010b]	**0.40**	**0.47**	**0.28**	**0.35**	0.45	**0.11**	**0.78**	0.20	**0.14**	0.09	0.81	**0.32**	**0.37**

(*)Indicates an originally nominal method that we have extended to deal with ordinal labels.

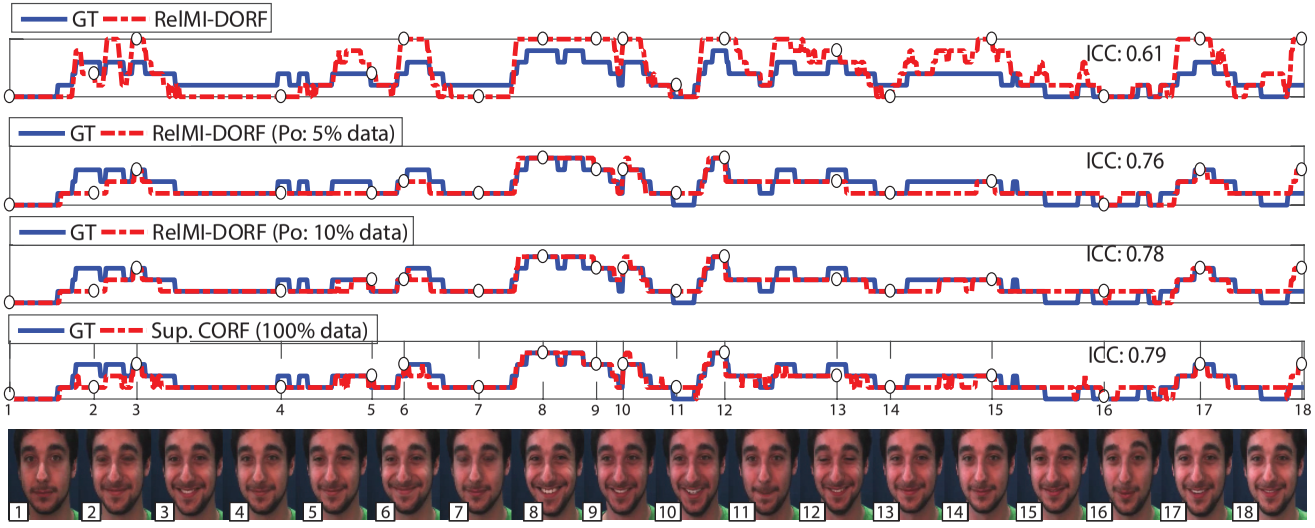

Figure 4.7: Visualization of AU12 (Lip-Corner puller) intensity predictions in a subsequence of the DISFA dataset. From up to bottom: RelMI-DORF without using instance labels and with 5% and 10% of annotated frames. Supervised CORF using all the frame labels during training. Intensity estimation for RelMI-DORF tends to be more accurate as more instance labels are considered during training. Using only a 10% of annotated frames, RelMI-DORF achieves similar accuracy than a fully-supervised CORF.

Table 4.5: Average performance across AUs on the DISFA dataset.

Setting	Method	CORR ↑	MAE ↓	ICC ↑
RelMI-DOR	HCRF [Quattoni et al., 2007]	0.21	2.04	0.10
	HCORF [Kim and Pavlovic, 2010a]	0.26	3.49	0.03
	OSVR [Zhao et al., 2016c]	0.35	1.38	0.15
	RelMI-DRF	0.19	1.70	0.11
	RelMI-DORF	**0.40**	**1.13**	**0.26**
PoRelMI-DOR (5% frames)	CRF [Li et al., 2009]	0.33	0.55	0.29
	CORF [Li et al., 2009]*	0.37	0.57	0.32
	HCRF [Chang et al., 2009]	0.34	0.59	0.30
	HCORF [Chang et al., 2009]*	0.38	0.62	0.33
	OSVR [Zhao et al., 2016c]	0.36	0.81	0.29
	RelMI-DRF	0.23	0.64	0.19
	RelMI-DORF	**0.40**	**0.51**	**0.36**
PoRelMI-DOR (10% frames)	CRF [Li et al., 2009]	0.36	0.50	0.32
	CORF [Li et al., 2009]*	0.39	0.56	0.33
	HCRF [Chang et al., 2009]	0.38	0.57	0.34
	HCORF [Chang et al., 2009]*	0.40	0.59	0.35
	OSVR [Zhao et al., 2016c]	0.37	0.80	0.29
	RelMI-DRF	0.36	0.50	0.32
	RelMI-DORF	**0.42**	**0.48**	**0.38**
Supervised DOR	CRF [Lafferty et al., 2001]	0.39	**0.44**	0.35
	CORF [Kim and Pavlovic, 2010b]	**0.41**	0.50	**0.37**

(*)Indicates an originally nominal method that we have extended to deal with ordinal labels.

results for independent AUs, we observe that RelMI-DORF achieves the best performance for most cases. Note however, that results for some particular AUs (9,15,17, 20) is low for all the methods. We attribute this to the fact that, the activation of these AUs is typically more subtle and high-intensity levels are scarce.

By looking into the results in the partially-observed setting, we can derive the following conclusions. Firstly, all the methods improve their average performance as the percentage of instance labels increases. However, this improvement is more significant for ICC and MAE. This shows that, when instance labels are not available during training, the tendency of intensity levels can be captured. However, accurate predictions of particular ordinal labels requires the additional information provided by frame-by-frame annotations. To illustrate this, in Fig. 4.7 we show AU12 predictions attained by RelMI-DORF using different percentages of annotated frames. Sec-

ondly, note that approaches modelling the ordinal structure of labels usually achieves better performance than nominal methods in terms of ICC and CORR. In contrast, CRF and HCRF obtain lower MAE than CORF and HCORF. This can be explained because the majority of sequence frames has AU intensity level of 0 (neutral). As a consequence, CRF and HCRF tends to assign most of the frames to this level, thus minimizing the absolute error. In contrast, ordinal methods are more robust to imbalanced intensity levels and capture better changes in AU intensities. Finally, note that the proposed RelMI-DORF method obtains the best average performance considering 5% and 10% of annotated frames. Regarding specific AUs, RelMI-DORF obtains better results for most cases and competitive performance against the best method otherwise. Finally, note that RelMI-DORF results with 10% of annotated frames is comparable to the achieved by the fully-supervised approaches CRF and CORF. Specifically, only supervised CRF outperforms RelMI-DORF in terms of average MAE. The slightly worse results of supervised CORF compared with RelMI-DORF suggest that considering intensity annotations for all the frames may cause overfitting and decrease performance on unseen test sequences. This can be seen more clearly by looking at the results of independent AUs, where RelMI-DORF obtain slightly better performance than fully-supervised CORF in some cases. In conclusion, the presented results support our hypothesis that it is possible to use the proposed RelMI-DORF model in order to reduce the annotation effort required for AU intensity estimation.

4.9 Summary

In this Chapter, we have addressed Facial Expression Intensity Estimation from a weakly-supervised perspective. For this purpose, we have presented the MI-DORF framework for the novel task of Multi-Instance Dynamic-Ordinal Regression. To the best of our knowledge,

this is the first MIL approach that imposes an ordinal structure on instance labels, and also attains dynamic modeling within bag instances. By considering different weak-relations between instance and bag labels, we have developed two variants of this framework: RelMI-DORF and MaxMI-DORF. Moreover, we have extended the proposed framework for Partially-Observed MI-DOR problems, where a subset of instance labels are also available during training. Although the presented MI-DORF framework has many potential applications in multiple domains, we have focused on Action Unit and Pain Intensity estimation problems, where we have demonstrated the ability of the proposed models to reduce the annotation effort in these tasks. Following this idea, in future work we plan to explore the combination of the proposed model with Deep Learning approaches. Coupling Probabilistic Models with Convolutional Neural Networks has been recently addressed in the field [Walecki et al., 2016]. However, the cited approach followed a supervised learning strategy during training. In contrast, coupling Deep Networks with the presented MI-DORF model would provide a principled way to train this powerful models using larger datasets, which could be easily annotated using only weak-labels at video level. This issue is discussed more in detail in Chapter 6.

Chapter 5

FUSION OF VALENCE AND AROUSAL ANNOTATIONS THROUGH DYNAMIC SUBJECTIVE ORDINAL MODELLING

5.1 Introduction and Motivation

As discussed in Chapter 1, facial expressions are accepted to carry important information about human emotions [Vinciarelli et al., 2009] and, thus, its automatic understanding has a wide range of potential applications in the context of Affect Analysis. In this scenario, a lot of research has focused on building computer vision systems able to map facial behavior to a representation of human affect [Nicolaou et al., 2011]. One of the most popular representations for this purpose is the Valence-Arousal (V-A) space [Russell, 1980]. Formalized by Russell through the Circumplex model of affect, Valence and Arousal have

been identified as the underlying dimensions of human emotion. Valence refers to how pleasant or unpleasant is an affective state while Arousal indicates the activation or deactivation level. These two dimensions have been consistently identified in experiments across various modalities [Cliff and Young, 1968, Green and Cliff, 1975, Osgood et al., 1975], which supports their validity. On the downside, Valence and Arousal are abstract dimensions whose exact meaning, apart from subjective, is not common knowledge (e.g. as opposed to the 6 universal emotions).

An essential issue when training and validating computer vision systems based on the V-A representation, is how to obtain ground-truth annotations from collected data (e.g, videos of human interactions). This is typically addressed based on manual annotations from expert human observers. However, labels obtained in this way are evidently subjective and have been shown to suffer from large inter-observer variations [Devillers et al., 2006, Nicolaou et al., 2014, Sukno et al., 2016, Yannakakis and Martínez, 2015a]. Subjectivity is unavoidable as it is inherent to affect annotations, regardless of using V-A or any other representation. However, it has been shown that the consistency of subjective annotations can be considerably improved if these are performed based on discrete (instead of continuous) labels and maintaining their ordinal relations [Metallinou and Narayanan, 2013, Yannakakis and Martínez, 2015a]. The reasons behind this finding seem related to the invalid assumptions underlying the use of both continuous and non-ordinal (nominal) labels. For instance, nominal labels assume the same degree of confusion between neighbouring and far away labels, while continuous methods assume a linear relation between the true labels and the subjective perception of annotators [Yannakakis and Martínez, 2015b].

Even if following the above recommendations, it is not possible to rely on individual annotations to obtain reliable ground truth. Therefore, consensus from pools of observers are usually preferred (see Fig. 5.1). Nevertheless, such consensus is not straightforward to obtain and the problem of fusing annotations from multiple ob-

servers has attracted considerable attention [Artaechevarria et al., 2009, Wang et al., 2013a, Langerak et al., 2010, Warfield et al., 2004, Warfield et al., 2008, Landman et al., 2012, Asman and Landman, 2012, Commowick and Warfield, 2010, Zhou et al., 2014, Metrikov et al., 2015, Lakshminarayanan and Teh, 2013, Commowick et al., 2012, Asman and Landman, 2011]. However, a vast majority of algorithms treat annotated labels simply as nominal classes, e.g. without taking into account their ordinal relations. As explained above, this has been found suboptimal for affective annotations, both theoretically [Jamieson et al., 2004, Yannakakis and Martínez, 2015b] and empirically [Yannakakis and Hallam, 2011, Yannakakis and Martínez, 2015a].

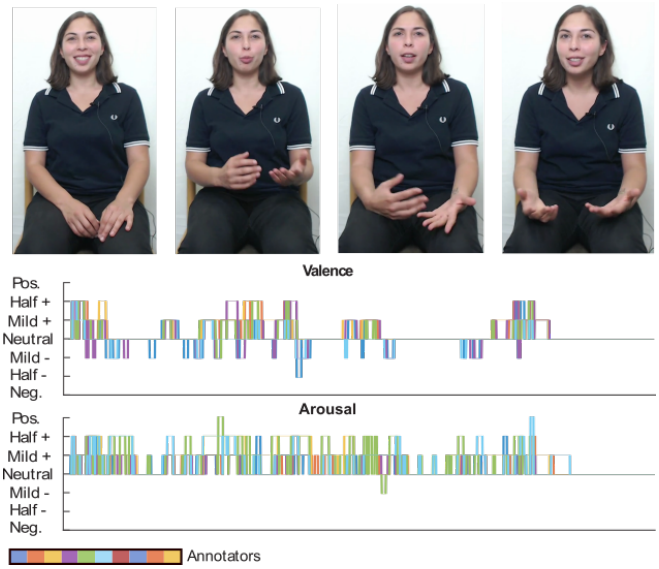

Figure 5.1: Example of a video sequence annotated by a set of annotators according to the Arousal and Valence dimensions (represented in an ordinal scale)

5.2 Contributions

In this Chapter, we present a novel probabilistic framework to address the fusion of V-A annotations from multiple human observers. Note that this task can be considered a weakly-supervised learning problem since the goal is to estimate a consensus (not annotated task) from a set of subjective annotations (weak-labels) given by a set of observers. Our main contributions are summarized as follows:

- The proposed method explicitly considers the ordinal structure in V-A labels and models the gap between the labeling scale and the subjective perception of each annotator. Fig. 5.2 illustrates this concept. While the fusion of ordinal labels has been investigated in other domains [Zhou et al., 2014], to the best of our knowledge, this is the first time it is applied to V-A annotations.

- In contrast to previous methods in annotation fusion, the presented framework is able to exploit dynamic information present in temporal sequences of annotations. Despite the fact that this information is irrelevant in other applications, it is of special importance in the context of V-A label fusion, where data typically consist of annotated videos of human interactions.

- In our experiments over synthetic and real data annotated with ordinal V-A labels, we show the superior performance of the presented method with respect to alternative approaches that ignore either the ordinal structure of labels or the dynamic information.

5.3 Related Work

Valence and Arousal annotations: The use of dimensional approaches to represent emotions has increasingly gained popularity

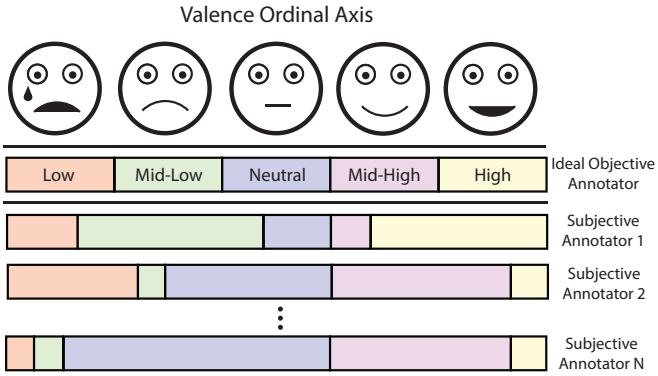

Figure 5.2: Illustration of the ordinal subjective assumption to fuse Valence and Arousal annotations. While the *objective* distance between consecutive ordinal labels is hypothetically uniform, each observer has his/her own perception of both the position and extent of them. As stated in [Jamieson et al., 2004], there is no justification for the assumption that subjective annotations follow a linear scale (e.g. the perceived distance between pleasant and neutral not necessarily matches the one between neutral and unpleasant). Thus, the only assumption we make in the proposed model is that the order of perceived labels is maintained across annotators, not their distances.

in the context of affective computing and related fields [Gunes and Schuller, 2013]. Several possible dimensions have been proposed to represent affect, with Valence and Arousal emerging as the most popular ones [Koelstra et al., 2012, McKeown et al., 2012, Sukno et al., 2016, Baveye et al., 2015]. However, in spite of their widespread use, generating reliable annotations in V-A space has proven challenging, not only in terms of achieving consensus but even in the way to address data annotation.

Based on the original definition by Russell [Russell, 1980], several authors directly targeted annotations in V-A space in the continuous domain [Koelstra et al., 2012, McKeown et al., 2012, Ringeval et al., 2013]. This means that each (human) observer is asked to rate a video

(either in segments or in continuous time) with points in $\mathbb{R}^2$, typically ranging between -1 and 1 in each axis and constrained to be within the unit circle [Cowie et al., 2000]. Such annotations have proven very variable, producing large dispersions between annotators, even when trying to map a single specific emotion into V-A space [Robinson and Baltrušaitis, 2015]. Some methods for the fusion of multiple continuous annotations have been proposed [Nicolaou et al., 2014, Gupta et al., 2017], but their ability to model annotator's subjectivity is limited. Indeed, while fusion in continuous space might seem attractive from an ordinal perspective, it actually implies assuming a linear relationship between the ground truth and the observer's annotation. This is implicit in the work by Nicolaou et al. [Nicolaou et al., 2014] who, moreover, provides only an estimate of how annotations change over time but loses their actual scale, which makes it impossible to apply their method to obtain a properly scaled consensus. In a very recent work, Gupta et al. [Gupta et al., 2017] make the linear mapping between observations and ground truth explicit. Linearity is indeed a key assumption to derive their equations within an EM framework. Extending the mapping to consider non-linear relations would allow a better modeling of annotator's subjectivity, but this possibility has so far not been explored.

On the other hand, some alternatives to the continuous representation of V-A have been recently explored. For example, Baveye et al. [Baveye et al., 2015] provide a ranking between events based on pairwise comparisons from a large number of observers. Thus, they do not provide actual labels but an ordering of events in V-A axes. This strategy is motivated by the argument that human observers are better at producing relative (e.g. pair-wise) comparisons than absolute ratings. Yannakakis et al. [Yannakakis and Martínez, 2015a] provide experimental evidence to support this theory. They compared the V-A annotations performed in continuous space against annotations produced in a ordinal discretized V-A space and found the latter to be clearly more consistent. Interestingly, they also showed that if annotations are performed in continuous V-A space and then dis-

cretized, their consensus also improves but not as much as in the case of annotating directly in the discretized ordinal space. Recent efforts in producing V-A labellings have followed this direction [Sukno et al., 2016].

Given the aforementioned advantages of using an ordinal scale to represent Valence and Arousal dimensions, in this Chapter we focus on the problem of fusing annotations which are given according to discrete but ordered categories. As far as we know, this is the first time that this problem is explored in the context of V-A affect annotations.

Discrete label fusion: Fusion of discrete annotations is a necessity for several fields and it is especially important when the true labels (*ground truth*) are unknown. In such cases, we wish to estimate the ground truth by merging the estimates (annotations) from a number of observers. The basic intuition is that given a sufficient number of observers, we should be able to extract a consensus from their annotations that is reasonably close to the ground truth. Straightforward solutions to label fusion include averaging, majority voting and extensions such as weighting [Artaechevarria et al., 2009, Wang et al., 2013a] or iterative outlier removal [Langerak et al., 2010].

A more principled solution consists on adopting a probabilistic framework that jointly estimates the ground truth and the annotators' subjective perception of labels. The latter is done by means of modelling a conditional probability which indicates, for each observer, what is the likelihood of annotating/perceiving a value given a fixed ground-truth label (*annotator's perception model*). Several approaches, have followed this line, among which STAPLE (Simultaneous truth and performance level estimation) is the most popular [Warfield et al., 2004]. STAPLE employs an expectation maximization (EM) algorithm to iteratively estimate the ground truth and perception model parameters, such that the probability of the observed annotations is maximized. STAPLE has been successfully used in several applications and numerous extensions to the framework have also been proposed. Among the most notable ones we

can cite variants for handling partial observations [Landman et al., 2012], variations or instabilities [Asman and Landman, 2012, Commowick and Warfield, 2010, Commowick et al., 2012] in annotator performances and variable difficulty throughout the annotation task [Asman and Landman, 2011].

When focusing on ordinal affective annotations, an issue of particular importance is the fact that labels are not simply unrelated categories but, instead, they naturally follow a relative ordering. For example, in the Valence axis, *pleasant* is further from *unpleasant* than from *neutral*. Ignoring the ordinal nature of labels has been found suboptimal, not only within affective computing but for subjective annotations in general [Yannakakis and Martínez, 2015b]. However, this issue has been largely overlooked in the label fusion literature. Among recent efforts to incorporate ordinal constraints we find the methods from [Zhou et al., 2014], based on entropy optimization within a mini-max framework, [Metrikov et al., 2015], based on latent trait models and [Lakshminarayanan and Teh, 2013], based on Bayesian inference. The Ordinal Min-Max Entropy method [Zhou et al., 2014] incorporates ordinal constraints on the annotators' perception model by means of an auxiliary variable that converts multi-label comparisons into a binary problem and optimize the conditional entropy jointly across all possible binary splittings. More closely related to our approach, [Metrikov et al., 2015] and [Lakshminarayanan and Teh, 2013] use Gaussian priors to model the probabilistic labeling of each annotator conditional to the true (but unknown) labels. However, strictly speaking, the approach from Metrikov et al. cannot ensure that ordinal constraints hold, since the Gaussian models for each label are completely independent of each other. Lakshminarayanan et al. resolve this by a mapping strategy based on predefined thresholds that naturally follow the desired ordering of labels, but this limits the flexibility of the approach to model annotator differences.

In contrast to the previously described works, our method employs an ordered probit model [Agresti, 2010] (see Appendix A.2)

to explicitly incorporate ordinal constraints in the annotators' perception model. This approach is both flexible enough to account for annotator-specific differences in perception while still ensures that ordinality is strictly fulfilled. Moreover, the presented framework is also able to incorporate dynamic information, useful when dealing with temporal sequences of annotations. To the best of our knowledge, temporal modelling has not been considered before in the context of ordinal annotation fusion.

5.4 Problem definition

In the following, we formally describe the annotation fusion problem addressed. We assume that a training set of N annotated sequences $\mathbb{D} = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, ..., \mathbf{D}^{(N)}\}$ is provided. Concretely, each $\mathbf{D^{(i)}}$ is an $A \times T$ matrix containing a set of T annotations for a total of A observers, where T is the number of items per sequence [1]. From now on, we will refer to the the label assigned by annotator a to the item t as $D_{at}^{(i)}$. Moreover, we consider the scenario where $D_{at}^{(i)} \in \{0 \prec ... \prec l \prec L\}$ is an ordinal variable taking L possible values.

Similar to previous works on label fusion [Warfield et al., 2004], we aim to learn a parametric model maximizing the log-likelihood over the training set $\mathbb{D}$ as:

$$\underset{\theta}{arg\,max}\ \ L(\mathbb{D} \mid \theta) = \log(p(\mathbb{D} \mid \theta)) \tag{5.1}$$

where θ is the set of model parameters. For this purpose, we define for each sequence a new set of latent variables $\mathbf{g^{(i)}} = \{g_1^{(i)}, g_2^{(i)}, ..., g_T^{(i)}\}$ representing the ground-truth ordinal labels for each item. Given $\mathbf{g^{(i)}}$, the log-likelihood can be expressed as:

[1] For notation simplicity, we use the same number of annotators A and T for each sequence. However, the proposed methods can also handle cases where they vary across sequences

$$L(\mathbb{D} \mid \theta) = \sum_{i=1}^{N} \log \Big(\sum_{\mathbf{g}} \prod_{a} p(\mathbf{D}_a^{(i)} \mid \mathbf{g}; \theta) p(\mathbf{g}) \Big) \tag{5.2}$$

by assuming conditional independence between observer annotations for each sequence and marginalizing over all the possible latent ground-truth labellings $\mathbf{g}$.

Given the parameters θ that maximize Eq. 5.2, the probability of a given ground-truth labelling $\mathbf{g}^{(i)}$ for a given sequence can be obtained from:

$$p(\mathbf{g}^{(i)}|\mathbf{D}^{(i)}, \theta) = \frac{\prod_a p(\mathbf{D}_a^{(i)}|\mathbf{g}^{(i)}; \theta) p(\mathbf{g}^{(i)})}{\sum_{\mathbf{g}} \prod_a p(\mathbf{D}_a^{(i)}|\mathbf{g}; \theta) p(\mathbf{g})} \tag{5.3}$$

where $p(\mathbf{g})$ defines a prior over $\mathbf{g}$.

5.5 Static Ordinal Annotation Fusion

Following, we describe the proposed Static Ordinal Annotation Fusion (SOAF). This model aims to solve the problem defined in Sec. 5.4 by ignoring temporal information. For this purpose, we assume that item ground-truth labels $g_t^{(i)}$ within a sequence are independent. Under such assumption, Eq. 5.2 can be expressed as:

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T} \log \Big(\sum_{l} \prod_{a} p(\mathbf{D}_{at}^{(i)} \mid g_t^{(i)} = l, \theta^a) p(g_t^{(i)} = l) \Big), \tag{5.4}$$

Note that we have defined a set of independent parameters $\theta = \{\theta^1, \theta^2, ..., \theta^A\}$ for each annotator. These parameters model the conditional probability $p(\mathbf{D}_{at}|g_t, \theta^a)$ and, thus, describe the subjective perception of each annotator a for a given latent ground-truth ordinal label (annotator perception model). Under the defined model,

the ground-truth probability for any sequence item can be easily computed as:

$$p(g_t^{(i)} = l|\mathbf{D}^{(i)}, \theta) = \frac{\prod_a p(\mathbf{D}_{at}^{(i)}|g_t^{(i)} = l)}{\sum_{l'} \prod_a p(\mathbf{D}_{at}^{(i)}|g_t^{(i)} = l')} \tag{5.5}$$

by assuming an uniform prior distribution over all $p(g_t^{(i)})$

5.5.1 Ordinal annotator perception model

In order to incorporate the ordinal constraints into the annotator's perception model, we propose to use an ordered probit model [Agresti, 2010] (see Appendix A.2 for more details) to define conditional probabilities $p(\mathbf{D}_{at} = l|g_t = l', \theta^a)$ as:

$$p(\mathbf{D}_{at} = l|g_t = l', \theta^a) = \Phi\Big(\frac{c_l^a - l'}{\sigma^a}\Big) - \Phi\Big(\frac{c_{(l-1)}^a - l'}{\sigma^a}\Big). \tag{5.6}$$

Here, $\Phi(\cdot)$ denotes the normal cumulative distribution function (CDF), and $\theta^a = \{\mathbf{c^a}, \sigma^a\}$ is the set of annotator parameters. Specifically, $\mathbf{c}^a = \{c_0^a = -\infty \leq c_1^a \leq \cdots \leq c_L^a = \infty\}$ are monotonically increasing thresholds dividing a continuous line into L bins corresponding to different ordinal values. The difference between the two CDFs provides the probability of a perceived label l given the ground-truth ordinal value l'. Moreover, $\sigma^a > 0$ is the standard deviation of Gaussian noise modelling uncertainty in the observer annotations (see Fig. 5.3 for an illustration). This model has been previously explored in the context of facial expression intensity estimation [Rudovic et al., 2012].

In order to ensure $\sigma > 0$ and the monotonically increasing constraints of thresholds $\mathbf{c}$, we use a re-parametrization strategy similar to [Kim and Pavlovic, 2010a]. Concretely, we define $c_l = c_1 + \sum_{s=1}^{l-1} \delta_s^2$ and $\sigma = \tau^2$. With this parametrization, the maximization of Eq. 5.4 becomes an unconstrained optimization problem which can be solved as described in the following section.

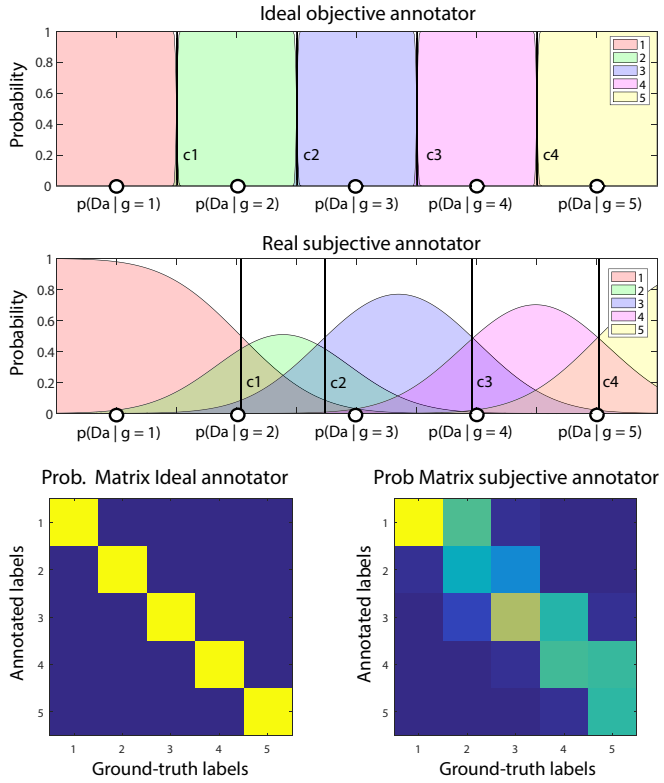

Figure 5.3: Illustration of the employed ordered probit model defining the annotator perception models $p(\mathbf{D}_{at}|g_t, \theta^a)$. Top: Ideal objective annotator perceiving equally-distant ordinal labels with no uncertainty ($\sigma = 0$). Middle: Real annotator where the perception of different labels is non-linear but follows ordinal constraints ($\sigma \approx 0.5$ modelling perception noise). Note that for both annotators, the perceived distance between ordinal values are determined by thresholds $\mathbf{c}$. The monotonically increasing constraints over these thresholds ensure that the likelihood of perceived labels are ordered. Bottom: For both cases, matrices representing the conditional probabilities $p(\mathbf{D}_{at}|g_t, \theta^a)$ for each pair of ground-truth and perceived ordinal labels.

5.5.2 Learning

In order to learn the optimal model parameters θ given a set $\mathbb{D}$, we use standard gradient ascent. Specifically, we employ the LBFGS Quasi-Newton method [Byrd et al., 1994] (see also Appendix A.1), which generally provides a higher-convergence rate than first-order approaches. The derivatives of the log-likelihood function $L(\mathbb{D} \mid \theta)$ w.r.t the annotator parameters θ^a can be expressed as:

$$\frac{\delta L}{\delta \theta^a} = \sum_{i,t,l} \frac{p(g_t^{(i)} = l|\mathbf{D}^{(i)}, \theta)}{p(\mathbf{D}_{at}^{(i)}|g_t^{(i)} = l)} \cdot \frac{\delta p(\mathbf{D}_{at}^{(i)}|g_t^{(i)} = l)}{\delta \theta^a} \tag{5.7}$$

where the gradients $\frac{\delta p(\mathbf{D}_{at}^{(i)}|g_t^{(i)}=l)}{\delta \theta^a}$ with the defined ordinal probit model can be easily computed as detailed in [Kim and Pavlovic, 2010a].

5.6 Dynamic Ordinal Annotation Fusion

One of the main assumptions made in SOAF (Sec 5.5), is that ground-truth latent variables g_t^i for a given sequence are independent. This assumption is suboptimal in temporal sequences (e.g. videos) where ground-truth labels tend to be temporally-correlated. In order to incorporate dynamic information, we extend SOAF to Dynamic Ordinal Annotation Fusion (DOAF). For this purpose, we follow a first-order Markovian assumption where a given g_t is dependent on the previous ground-truth label g_{t-1}. In DOAF, the log-likelihood can be expressed as:

$$L = \sum_{i=1}^{N} \log \Big(\sum_{\mathbf{g}} \Big[p(g_1^{(i)}) p(\mathbf{D}_{:1}^{(i)} \mid g_t^{(1)}, \theta^a) \cdot \prod_{t=2}^{T} p(\mathbf{D}_{:t}^{(i)} \mid g_t^{(i)}, \theta^a) p(g_t^{(i)}|g_{t-1}^{(i)}; \theta^D) \Big] \Big), \tag{5.8}$$

where $p(\mathbf{D}_{:t}^{(i)} \mid g_t^{(i)}; \theta^a) = \prod_a p(\mathbf{D}_{at}^{(i)} \mid g_t^{(i)}; \theta^a)$ and $\theta^D = \{\alpha_{1|1}, \ldots, \alpha_{l|l'}, \ldots, \alpha_{L|L}\}$ is a new set of parameters defining the label transition probabilities $p(g_t|g_{t-1})$. Concretely, we use a soft-max function:

$$p(g_t = l | g_{t-1} = l'; \theta^D) = \frac{e^{\alpha_{l|l'}}}{\sum_s e^{\alpha_{s|l'}}}, \tag{5.9}$$

which ensures the conditional probability constraints $\sum_l p(l|l') = 1$ and $p(l|l') \geq 0$.

Note that DOAF is a special case of a Hidden Markov Model [Rabiner and Juang, 1986] (see also Appendix A.3) where latent states represent the ground-truth labels and emission probabilities are defined by $p(\mathbf{D}_{:t}^{(i)} \mid \theta^a)$ (see Fig. 5.4). Therefore, marginal probabilities $p(g_t^{(i)}|\mathbf{D}^{(i)}, \theta)$ of item latent labels can be computed using the forward-backward algorithm [Barber, 2012] (see also Appendix A.3) employed in HMMs. Similar to SOAF, we assume a uniform prior distribution over initial latent ordinal states $p(g_1)$.

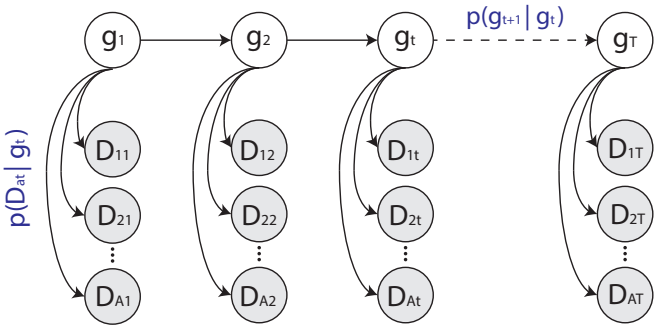

Figure 5.4: Graphical representation of the proposed DOAF model. Given a sequence of T items, independent labels $\mathbf{D}_{at}$ for each annotator a and item t are provided. The consensus label g_t for each item is treated as a latent variable defined by two probabilities: $p(\mathbf{D}_{at}|g_t)$, representing the subjective perception for the annotator a given the provided label and (ii) $p(g_{t+1}|g_t)$, which models temporal correlations between consecutive consensus labels g_t and g_{t+1}

5.6.1 Learning

In DOAF, we use a similar learning procedure than the one described in Sec. 5.5.2 for SOAF. In this case, the gradients of parameters θ^a can be computed as defined in Eq. 5.7. However, marginal probabilities $p(g_t^{(i)}|\mathbf{D}^{(i)}, \theta)$ need to be obtained using the forward-backward procedure. On the other hand, the gradient of transition parameters θ^D can be computed as:

$$\frac{\delta L}{\delta \alpha_{s|s'}} = \sum_{i,t,l,l'} \Big[\mathbb{1}(s' = l') p^{it}_{l,l'} [\mathbb{1}(s = l) - p^{it}_{s|s'}] \Big], \tag{5.10}$$

where $\mathbb{1}(\cdot)$ is an indicator function and

$$p^{it}_{l|l'} = p(g_t^{(i)} = l | g_{t-1}^{(i)} = l'; \theta^D) \tag{5.11}$$

$$p^{it}_{s,s'} = p(g_t^{(i)} = l, g_{t-1}^{(i)} = l' \mid \mathbf{D}^{(i)}; \theta). \tag{5.12}$$

Again, marginal probabilities $p(g_t^{(i)} = l, g_{t-1}^{(i)} = l' \mid \mathbf{D}^{(i)})$ can be computed with the forward-backward procedure.

5.7 Experiments

5.7.1 Evaluation criteria and metrics

In order to compare the proposed SOAF and DOAF frameworks with alternative approaches, we use two different criteria. In the first one, we evaluate the prediction of ground-truth labels. Formally, given the learned annotator perception models represented by $p(\mathbf{D}_{at}|g_t)$, we estimate the most likely ground-truth labels $\mathbf{g}^{(*)}_{pred}$ for a new test sequence $\mathbf{D}^{(*)}$ (see Eq. 5.3). Assuming that we know the real annotations $\mathbf{g}^{(*)}_{real}$ for $\mathbf{D}^{(*)}$, we compare predictions and real labels for each sequence item. For that purpose, we employ standard metrics used in the context of ordinal regression [Rudovic et al., 2015]. Concretely, we use the Pearson's Correlation Coefficient (CORR), Mean Absolute Error (MAE) and the Intra-Class-Correlation Coefficient (ICC).

Despite the main goal of annotation fusion methods is to predict ground-truth labels, evaluation under the aforementioned criterion is generally not feasible in real scenarios, since the actual $\mathbf{g}^{(*)}_{real}$ is not known. In order to compare different methods in the context of V-A annotations, we use an alternative evaluation criterion. It is based on the assumption that, given the perception model $p(\mathbf{D}_{at}|g_t)$, we should be able to predict new annotations for the observer a given a predicted ground truth $\mathbf{g}^{(*)}_{pred}$. Formally, the predicted annotation for a given sequence and observer can be computed as:

$$p(\mathbf{D}^{(*)}_{a}|\mathbf{D}^{(*)}_{\forall a' \neq a}) = \sum_{\mathbf{g}^{(*)}} p(\mathbf{D}^{(*)}_{a}|\mathbf{g}^{(*)}) p(\mathbf{g}^{(*)}|\mathbf{D}^{(*)}_{\forall a' \neq a}) \tag{5.13}$$

where $\mathbf{D}^{(*)}_{\forall a' \neq a}$ refers to the sequence annotations for all the observers except a. Given that we know $\mathbf{D}^{(*)}_{a}$ for a test sequence, the same metrics previously described can be used to evaluate the model's performance. Note that this criteria jointly evaluates the annotators' model and ground truth estimation, since both are needed to estimate the annotator labellings. For instance, even if we had the optimal perception model for a given annotator, it would be impossible to correctly generate his labelling for a given test sequence if the estimated ground-truth was not accurate.

5.7.2 Baselines

In our experiments, we compare the proposed SOAF and DOAF methods with alternative approaches that ignore either the ordinal structure of labels (nominal) or the dynamic information (static). Following, we describe them.

Majority Voting (MV): The ground-truth labelling is predicted with a majority voting strategy. Concretely, the estimated label for a given item is chosen as the majority ordinal level across all the annotators. Given that this approach does not explicitly compute the annotator's perception model $p(\mathbf{D}_{at}|g_t)$, we empirically compute

it from the training annotated sequences and the estimated ground-truth. MV follows an static-nominal assumption.

STAPLE: This method is one of the most popular approaches for fusing annotations with nominal labels. (see Sec. 5.3 and [Warfield et al., 2004]). We have used our own implementation of this method which also follows a static-nominal assumption.

Static Nominal Annotation Fusion (SNAF): This approach is equivalent to the proposed SOAF model but modelling ordinal labels as nominal. Concretely, $p(\mathbf{D}_{at}|g_t)$ for each annotator is defined using a parametrized soft-max function (see Eq. 5.9). It can be easily shown that SNAF maximizes the same log-likelihood function as STAPLE. However, SNAF is trained using gradient-ascent whereas STAPLE uses an Expectation-Maximization algorithm.

Dynamic Nominal Annotation Fusion (DNAF): In this case, DNAF is equivalent to the proposed DOAF model but modelling ordinal labels as nominal similar to SNAF. Therefore, we can consider DNAF a dynamic-nominal approach.

Ordinal Minimax Conditional Entropy (OMME): This method can be considered the state-of-the-art approach for static ordinal annotation fusion (see Sec. 5.3). In our experiments, we use the implementation provided by the authors of the original paper. Similarly to the case of MV, we empirically compute the annotator conditional probabilities $p(\mathbf{D}_{at}|g_t)$.

5.7.3 Synthetic Experiments

To validate the benefits of the proposed framework while fusing ordinal annotations of temporal sequences, we have performed a set of experiments using synthetic data. The use of these data allows to evaluate the performance of different approaches while predicting latent ground-truth labels. As explained in Sec. 5.7.1, this ground-truth is not known in real data and, therefore, it is not feasible to evaluate methods according to this criterion.

Data generation and experimental setup: In order to cre-

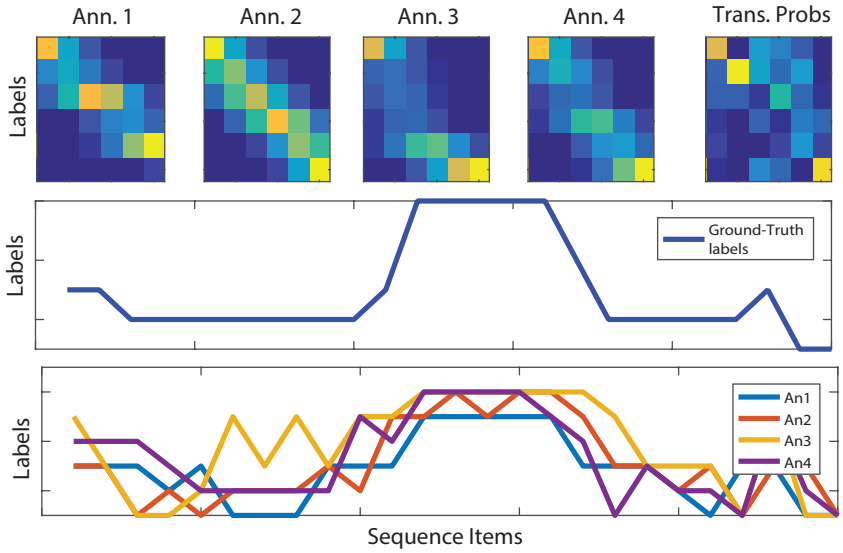

Figure 5.5: Illustration of the process followed to generate synthetic data sequences. From top to bottom: (i) Matrices representing the annotator perception models (A=4) and temporal transition probabilities from a randomly generated DOAF model. (ii) Example of a ground-truth sequence sampled according to the defined transition probabilities. (iii) Randomly generated annotations according to the defined ground-truth sequence and perception models.

ate a synthetic dataset of annotated sequences $\mathbb{D}$, we use the following procedure. Firstly, we generate a DOAF model (see Section 5.6) by randomly defining a set of parameters $\theta = \{\theta^1, ..., \theta^A, \theta^D\}$. The number of ordinal levels and annotators has been set to $L = 6$ and $A = 4$ respectively. Secondly, for each sequence $\mathbf{D}^{(\mathbf{i})}$ we sample the ground-truth labels $\mathbf{g}_{real}^{(*)}$ by using the conditional probabilities $p(g_t|g_{t-1}; \theta^D)$ and a uniform prior distribution over $p(g_1)$. Sequence length has been set to $T = 50$. Finally, for each observer we generate his annotations $\mathbf{D}_{\mathbf{a}}^{(\mathbf{i})}$ by sampling from his perception model $p(\mathbf{D}_{at}|g_t)$ and $\mathbf{g}_{real}^{(*)}$. Figure 5.5 illustrates this process. Using this procedure, we have generated 100 synthetic datasets with 10 and 20 sequences

for training and testing respectively (randomly generating different DOAF parameters for each dataset). Training sequences are used to learn the different model parameters whereas test sequences are used to compute the different metrics described in 5.7.1

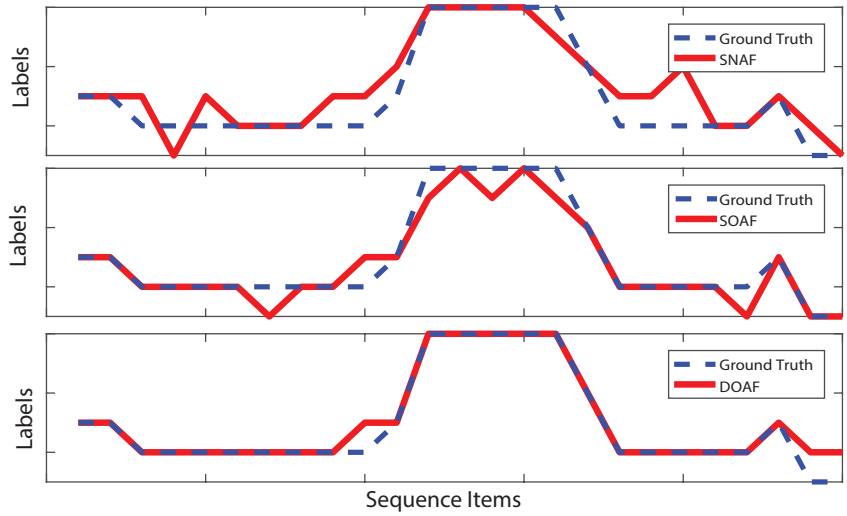

Figure 5.6: Examples of ground-truth predictions in a synthetic sequence for SNAF, SOAF and DOAF. Note that SOAF predicts more accurately the actual latent ordinal levels than SNAF, which models labels as nominal variables. Moreover, SOAF predictions tend to be less temporally smooth than in the DOAF case. This is because the latter incorporates dynamic information which takes into account the conditional dependencies between temporally consecutive items in the sequence.

Results and discussion: Table 5.1 shows the average results over the 100 synthetic datasets. By looking into the results of the compared methods, we can derive the following conclusions. Firstly, MV obtains the worst performance among all metrics. This was expected since it follows a static-nominal approach and does not take into account the perception model for each annotator. Secondly, static-nominal approaches such as SNAF and STAPLE generally per-

Table 5.1: Obtained results results on synthetic generated data

	Ground-truth Predictions			Annotation Predictions		
	CORR	**MAE**	**ICC**	**CORR**	**MAE**	**ICC**
MV	0.859	0.648	0.853	0.674	0.944	0.673
STAPLE	0.869	0.648	0.862	0.737	0.790	0.736
SNAF	0.867	0.653	0.860	0.737	0.792	0.736
DNAF	0.901	0.501	0.895	0.740	0.783	0.739
OMME	0.901	0.516	0.894	0.738	0.790	0.737
SOAF	0.928	0.420	0.923	0.744	0.771	0.743
DOAF	**0.941**	**0.259**	**0.940**	**0.756**	**0.746**	**0.755**

form worse than static-ordinal methods (OMME and SOAF). This shows the importance of taking into account the ordinal structure of labels in this kind of problems. Thirdly, note that the dynamic-nominal approach DNAF obtains slightly better results than all the other static-nominal approaches by considering the dynamic information present in temporal sequences of annotations. Finally, DOAF obtains the best performance in all cases by taking into account both dynamic information and introducing ordinal constraints in the annotator perception models. To illustrate this conclusions, we show in Fig. 5.6 an example of qualitative results obtained by SNAF, SOAF and DOAF in the synthetic test sequence shown in Fig. 5.5.

5.7.4 Valence and Arousal annotations fusion

In order to evaluate the proposed method in the context of V-A annotation fusion, we have used the database described in [Sukno et al., 2016]. To our knowledge, this is the largest database providing a set of annotated videos using ordinal V-A labels. Moreover, we have discarded databases with continuous ratings, such as SEMAINE [McKeown et al., 2012], because quantization of continuous annotations does not result in ordinal data. In contrast, ratings from [Sukno et al., 2016] were annotated according to a small set of ordered discrete labels, which can be validated as an ordinal setting [Yannakakis

and Martínez, 2015a]. The rationale is as follows: with only a few labels to choose from, you intuitively compare among them. As the choices increase, this task is more difficult. In the extreme case (continuous), it is impossible for an annotator to keep strict ordinality and subsequent discretization cannot fix that.

Database and experimental setup: The database consists of 64 videos of human interactions -with a total duration of approximately 3.5h- annotated by a maximum of 11 human experts. Valence and Arousal dimensions are labelled on different axes and represented with a set of 7 ordinal labels: {positive, half positive, mild positive, neutral, mild negative half negative, negative}. We have performed an 8-fold cross-validation for both dimensions, where 56 videos have been used for training and 8 for testing. Similar to synthetic experiments, training videos are used to learn the annotators' perception models which are then employed for evaluation on test sequences. In the case of the Arousal dimension, we do not use the lowest negative label since it never appears in the dataset. In order to reduce computational complexity and remove temporal redundancy, all sequences have been sub-sampled to only contain time instances where any annotator reported a change in the affective state.

Table 5.2: Obtained results on Arousal and Valence annotations of human interaction recordings

	Arousal Annotations			**Valence Annotations**		
	CORR	**MAE**	**ICC**	**CORR**	**MAE**	**ICC**
MV	0.308	0.529	0.300	0.483	0.486	0.471
STAPLE	0.343	0.513	0.337	0.496	0.463	0.479
SNAF	0.359	0.506	0.352	0.493	0.462	0.471
DNAF	0.332	0.538	0.330	0.503	0.482	0.497
OMME	0.352	0.516	0.349	0.514	0.481	0.511
SOAF	0.368	0.497	0.354	0.509	0.454	0.457
DOAF	**0.400**	**0.492**	**0.391**	**0.542**	**0.445**	**0.516**

Results and discussion: Given that ground-truth labels $\mathbf{g}_{real}^{(*)}$ for real sequences are unknown, we use the second evaluation criterion described in Sec. 5.7.1. Table 5.2 shows the results obtained for each affective dimension following the previously described cross-validation procedure. We can see that the static-ordinal approaches OMME and SOAF generally obtain better performance than static-nominal methods (SNAF and STAPLE). Indeed, between these four methods, the best performance for any metric is always obtained by either OMME or SOAF. This shows the advantages of considering labels' ordinal structure in this context. Secondly, DOAF outperforms static-ordinal approaches by incorporating dynamic information into the ground-truth labelling estimation. However, note that DNAF does not actually outperforms SNAF. This suggest that the advantage of modelling temporal correlations can only be fully achieved if appropriately considering the ordinality of labels. In conclusion, our results show the benefits of the proposed DOAF model for V-A annotations fusion. Fig. 5.7 shows an illustrative example of the estimated ground-truth labels by DOAF in a test sequence.

5.8 Summary

In this Chapter, we have proposed a novel probabilistic framework for the fusion of ordinal annotations in temporal sequences. This problem can be considered a weakly-supervised task since the goal is to estimate a common ground-truth from a pool of subjective annotations which only provides weak information. Moreover, it is of special importance in the context of affective computing, where collected data is typically formed by videos of human interactions annotated in terms of V-A affective labels. Recent works have shown that the consistency of V-A annotations can be considerably improved if these are performed based on an ordinal scale. Thus, in contrast to previous methods for annotation fusion, our approach explicitly introduces ordinal constraints into the annotators' perception model

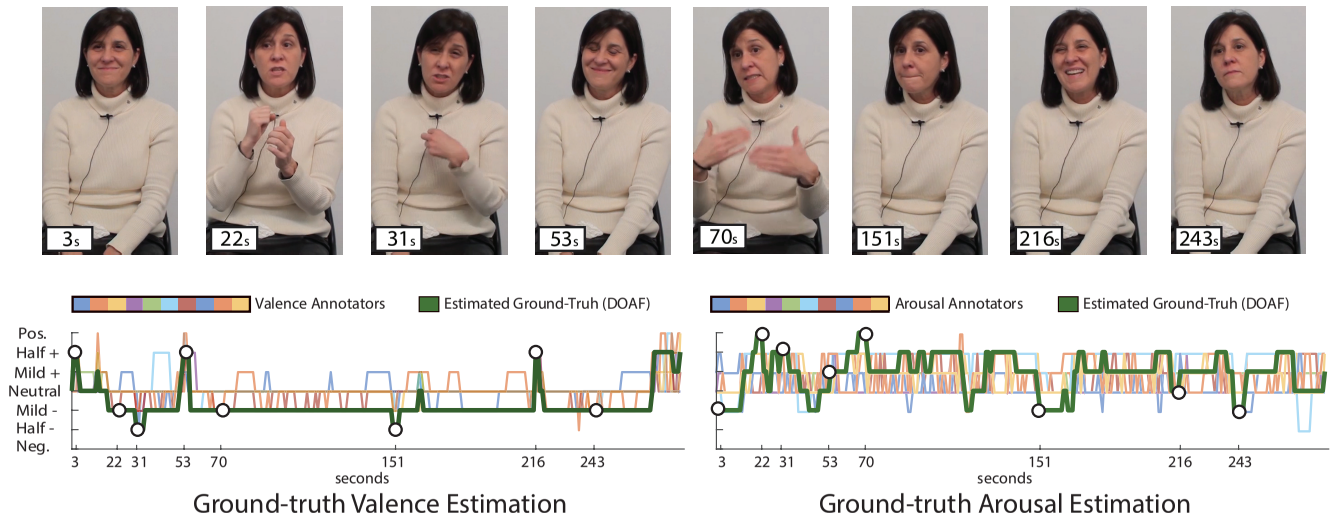

Figure 5.7: Estimated ground-truth from a set of V-A annotations in a test video. Despite the noisy subjective annotations provided by different observers, our method is able to estimate a sequence of ground-truth labels coherent with the non-verbal behavior displayed by the subject.

and incorporates dynamic information useful when dealing with temporal sequences. In our experiments over synthetic and real data, we show that the proposed method outperforms alternative approaches which do not take into account either the ordinal structure of labels or the dynamic information. Future datasets may benefit from the presented framework as it would help to provide more reliable ground-truth to train and validate automatic affect analysis models.

Chapter 6

DISCUSSION AND FUTURE RESEARCH

In **Chapter 1**, we summarized the main motivation of this Thesis with the following question: *Can we address Facial Behavior Analysis problems by changing the fully-supervised paradigm by a weakly-supervised one?* Our interest was justified by the hypothesis that weakly-supervised approaches may provide a potential solution to alleviate the need of annotated data in facial expression analysis problems. According to our opinion, addressing this question is of special importance given the laborious labelling process typically required for collecting databases. Additionally, researchers have largely overlooked this issue as is demonstrated by the fact, that the vast majority of proposed approaches in the field has followed the supervised-learning paradigm. Following this motivation, the goal of the presented Thesis has been to explore different weakly-supervised methods and show their potential applications in this context. In this Chapter, we discuss our main contributions and findings, as well as different research lines that can be addressed in the future.

In **Chapter 2**, we have introduced a novel problem called Facial Behavior Categorization. In this case, the goal is to infer high-level semantic labels from the facial behavior displayed by subjects in video

recordings. This task is implicitly related with the standard Discrete Expression Recognition problem, where models are trained in order to recognize a set of predefined gestures (e.g. Action Units). However, in Facial Behavior Categorization, the weakly-supervised setting is assumed and expressions are not explicitly labelled but need to be automatically learned from the information provided by the high-level labels at video-level. To address this problem, we have proposed a novel Multiple Instance Learning method specially designed for this purpose: Regularized Multi-Concept MIL. In our context, the demonstrated ability of RMC-MIL to successfully address different Facial Behavior Categorization problems is relevant in some aspects. Firstly, annotation of high-level video labels is usually a much easier task than the labelling of specific expressions at frame-basis. Therefore, learning to interpret facial behavior from this weak-information supposes a huge advantage with respect to the fully-supervised approach. For example, in our experiments, we have shown how RMC-MIL is able to discover discriminant facial gestures from annotations related with the emotional response of subjects while watching an advertisement. Note that these labels were obtained from self-reports and they did not require external annotators. Another interesting observation is that, in many applications, the recognition of predefined gestures such as the Action Units is not the final goal. Instead, the main interest consist in inferring higher-level information from long-term facial behavior. Therefore, Facial Behavior Categorization provides a more natural formulation to develop solutions in these scenarios.

Despite of the potential applications of Facial Behavior Categorization, in **Chapter 3** we have addressed a more standard problem in the field: Action Unit recognition. The reason is that the Facial Action Coding System is generally accepted as the most objective manner to describe expressions. For this purpose, we have presented the Hidden and Semi-Hidden Task Learning frameworks in order to improve the performance of AU classifiers by using data annotated according to the six universal facial expressions. Given that their annotation is much simpler than in the case of Action Units,

these weakly-supervised methods allow to train models using larger datasets and, thus, increasing their generalization ability. In our experiments, we have provided empirical evidence of this phenomena by performing a set of exhaustive Cross-Database experiments over several AU datasets. Interestingly, we have demonstrated how Hidden Task Learning is able to outperform the standard supervised approach even when no Action Unit labels are available during learning. According to our opinion, this result confirms our initial hypothesis that limited training data in facial expression analysis has a negative impact on the performance of fully-supervised methods. In contrast, the good results obtained by HTL and SHTL indicate that coupling additional data with weakly-supervised approaches is a potential solution to overcome this challenge.

Apart from Action Unit detection, a relevant problem in Automatic Facial Behavior Analysis is Expression Intensity Estimation. In order to explore the potential of weakly-supervised approaches in this scenario, in **Chapter 4** we have presented Multi-Instance Dynamic Ordinal Random Fields (MI-DORF). Using this framework, expression intensity estimation can be addressed without the need of frame-by-frame annotations. Instead, MI-DORF is trained by using only labels providing weak-information about the evolution of intensity levels within a video sequence. Moreover, we have also extended MI-DORF to Partially-Observed MI-DORF, which allows to use additional annotations of expression intensities for a small subset of video frames. By evaluating this method on Action Unit and Pain Intensity Estimation tasks, we have demonstrated the potential ability of MI-DORF to reduce the annotation effort in these type of problems. Specifically, we have showed that MI-DORF can learn underlying variables that are significantly correlated with frame intensity labels. Even though our results in this case were lower than fully-supervised approaches, our method provides a good trade-off between the annotation effort and the accuracy of intensity predictions. While we do not claim to replace the AU/Pain annotation process using only weak-labels at sequence-level, this setting may be preferable

in some applications. For example, when the focus is on capturing the variation in target facial behaviour rather than obtaining highly accurate frame labels (e.g., for monitoring changes in patient's pain intensity levels). In this case, our approach has clear advantages over the fully supervised methods which require a time-consuming annotation process. On the other hand, the competitive results of Partially-Observed MI-DORF compared to the evaluated fully-supervised approaches, indicate that annotation effort can be highly-reduced when combined with weak-labels. This provides an opportunity to replace the limited-size datasets currently used to solve these problems, by large-scale databases which could be efficiently labelled.

Finally, in **Chapter 5** we have addressed another important challenge derived from the difficult annotation process involved in Facial Behavior Analysis: label reliability. Specifically, we have focused on the fusion of Valence-Arousal ordinal labels from multiple annotators. This problem is very relevant in the context of Affect Analysis given that annotations are inherently subjective and inter-observer agreement is typically low. For this purpose, we have presented a novel probabilistic framework able to estimate a common ground-truth from videos annotated by a set of coders. Different from previous approaches proposed for this purpose, our method explicitly models the subjective perception for each specific annotator as well as the temporal information present in annotated sequences. In our experiments, we have shown the importance of these two issues while fusing Arousal and Valence annotations. We believe that future databases may benefit from the presented framework, since it will allow to obtain more objective ground-truth in order to train reliable models for automatic Affect Analysis.

In conclusion, we consider that the presented research has provided strong evidences for the benefits of weakly-supervised approaches in the context of Automatic Facial Behavior Analysis. In our opinion, **future work** may achieve significant advances in the field by following this research line. For example, we hypothesize that these type of methods will allow to train more powerful and expressive

models by taking advantage of a vast amounts of weakly-annotated data. In this context, it is worth to mention a recent trend exploring Deep Learning in the context of Automatic Facial Behavior Analysis [Tősér et al., 2016, Zhao et al., 2016a]. Although these models have a high modelling power, the performance improvement with respect to traditional shallow methods are not as impressive as the achieved in most other Computer Vision problems. One of the possible reasons explaining this phenomena is the limitation of available training data which complicates the training of such models. This challenge has been recently identified and addressed by incorporating special mechanisms in the learning process of supervised Convolutional Neural Networks [Han et al., 2016, Walecki et al., 2017]. However, we believe that coupling Deep Learning with weakly-supervised approaches, would provide a more principled way to learn this powerful models by increasing the available training data. Also related with recent trends in Deep Learning, another interesting open question is whether weakly-supervised approaches could be combined with Deep Generative models such as Variational Autoencoders [Kingma and Welling, 2013]. Via semi-supervised strategies, these methods have been shown to provide powerful mechanisms to incorporate non-annotated data into the learning process [Kingma et al., 2014]. Therefore, coupling weak-labels with unsupervised learning objectives would provide the opportunity to increase even more the amount of training data and thus, improve the quality of learned models.

As a final remark, we hope that the research presented in this Thesis encourages further investigation exploring weakly-supervised methods in the context of Automatic Facial Behavior Analysis. According to our opinion and the presented results, this approach is a very promising way to achieve significant advances in the field.

Appendix A

TECHNICAL DETAILS

A.1 L-BFGS Quasi Newton method

In order to optimize the parameters of the different proposed models, we use a standard gradient-descent strategy during training. Specifically, for all the methods presented in this Thesis, we use the "Limited-Memory Broyden–Fletcher–Goldfarb–Shanno" algorithm [Byrd et al., 1994] (L-BFGS). Following, we give some technical details about this optimization method.

L-BFGS belongs to the family of Quasi-Newton methods [Dennis and Moré, 1977] which are designed to solve optimization problems of the form:

$$\min_{\theta} f(\theta), \tag{A.1}$$

where f is a twice-differentiable function depending on variables θ. L-BFGS applies an iterative gradient-descent procedure in order to minimize f. Specifically, given variables θ_t at iteration t, f is minimized by updating the variables θ as:

$$\theta_{t+1} = \theta_{t+1} - B_t \cdot \nabla f(\theta_t), \tag{A.2}$$

where $\nabla f(\theta_t)$ is the gradient of f w.r.t θ_t and B_t is an approximation

of the inverse Hessian Matrix. The key difference between L-BFGS and other Quasi-Newton methods is how B_t is computed at each iteration. Given that matrix B_t has size $K \times K$, where K is the number of variables θ, its computation and storage can be very expensive. For this reason, L-BFGS approximates B_t with a low-rank compact form as:

$$B_t = B_0 - \sum_{i=0}^{t-1} a_i a_i^T + \sum_{i=0}^{t-1} b_i b_i^T$$

$$\text{with} \quad a_i = \frac{B_i s_i}{\sqrt{s_i^T B_i s_i}}, \quad b_i = \frac{y_i}{\sqrt{y_i^T s_i}}, \quad B_0 = \lambda_i^{-1}\mathbf{I}, \tag{A.3}$$

where

$$y_i = \nabla f(\theta_{i+1}) - \nabla f(\theta_i), \quad s_i = \theta_{i+1} - \theta_i \quad \text{and} \quad \lambda_i = \frac{s_{i-1}^T y_{i-1}}{||s_{i-1}||_2^2} \tag{A.4}$$

are defined by using the variables θ_i and gradients $\nabla f(\theta_i)$ computed from previous iterations. Compared to other Quasi-Newton methods, this approximation for B_t is much more efficient in terms of space and computational time.

A.2 Ordered Probit Model

In Chapters 4 and 5 we employ the Ordered Probit model as a key component of the proposed methods. In order to clarify the concepts and assumptions related with this model, in this Appendix we provide some technical details. This model has been traditionally employed in Ordinal Regression problems [Agresti, 2010], where the goal is to estimate the likelihood $p(y = l|z)$ for an ordinal variable y given a continuous variable $z \in \mathbb{R}$. Here, l refers to a value belonging to set of possible discrete ordered categories $y \in \{0 \prec 1 \prec ... \prec l \prec L\}$. The symbol "$\prec$" indicates that the different values are not independent, but are related according to an increasing monotonicity constraint. This contrasts to the assumption made in other models such as the

Multinomial Probit [McCulloch and Rossi, 1994], where the discrete categories $y \in \{0, 1, ..., l, L\}$ are assumed to not share any relation.

Given the continuous variable $z \in \mathbb{R}$, the noise-free ordinal likelihood [Chu and Ghahramani, 2005] can be modelled as:

$$p^*(y = l|z) = \begin{cases} 1 & iff \;\; z \in (b_{l-1}, b_l] \\ 0 & otherwise \end{cases} \quad , \tag{A.5}$$

where $b_0 = -\infty \leq \cdots \leq b_L = \infty$ are a set of values dividing the continuous line where z lies into L contiguous intervals (see Fig. A.1).

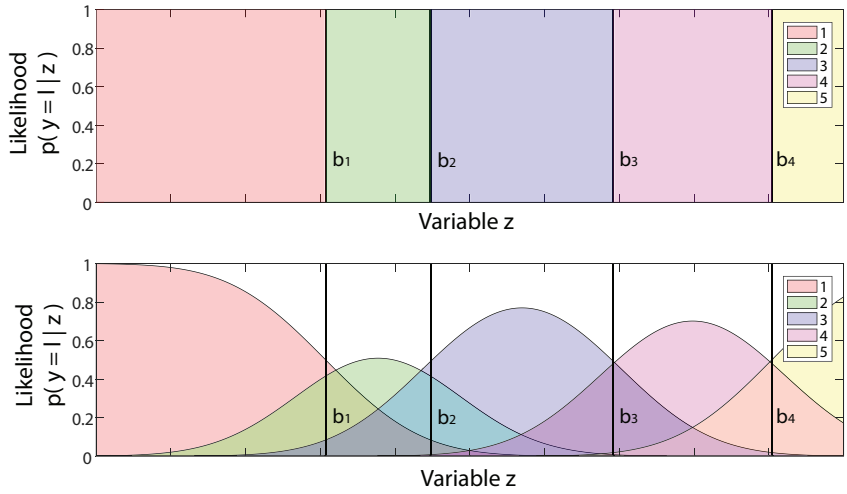

Figure A.1: Illustration of the likelihood $p(y = l|z)$ defined by the Ordered Probit model (example with the number of possible ordinal values $L = 5$. Top: Ideal noise-free case. Bottom: Assuming Gaussian noise contaminating variable z.

Assuming that the variable z is contaminated by a Gaussian noise ϵ with mean 0 and standard deviation σ, the noisy likelihood defined by the Ordered Probit model (see Fig. A.1) can be defined as:

$$p(y = l|z) = \int_{\epsilon \in R} p^*(y = k|z) \cdot \mathcal{N}(\epsilon; 0, \sigma^2) \tag{A.6}$$

$$= \Phi\Big(\frac{b_l - z}{\sigma}\Big) - \Phi\Big(\frac{b_{l-1} - z}{\sigma}\Big), \tag{A.7}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function (CDF).

A.3 The Forward-Backward Algorithm

In Chapters 4 and 5, we propose two Probabilistic Graphical Models (PGMs) in order to address different weakly-supervised facial behavior analysis problems. Our approaches are variants of Dynamic Bayesian Networks where we assume a temporal sequence of T observed variables $\mathbf{x}_{1:T} = \{\mathbf{x}_1\ , \mathbf{x}_2\ , ..., \mathbf{x}_T\}$ and latent variables $\mathbf{h}_{1:T} = \{h_1\ , h_2, ...\ , h_T\}$. Similar to other PGMs such as Hidden Markov Models [Rabiner and Juang, 1986], Conditional Random Fields [Lafferty et al., 2001] or Hidden Conditional Random Fields [Quattoni et al., 2007], each variable $\mathbf{x}_t$ depends only on its corresponding latent variable h_t. Moreover, the conditional dependence between variables h_t can be represented by a linear-chain connectivity (see Fig. A.2).

During the training and testing phases of the presented models, it is necessary to compute the conditional probabilities $p(h_t|\mathbf{x}_{1:T})$ for every latent variable h_t. This is accomplished by using the forward-backward algorithm. In order to clarify the key aspects of this algorithm, we describe its application to the specific case of Hidden Markov Models. However, the same procedure is applied in our proposed methods as detailed in Chapters 4 and 5.

In a HMM (see Fig. A.2(a)), the joint probability of latent variables $\mathbf{h}_{1:T}$ and observations $\mathbf{x}_{1:T}$ is defined as:

$$p(\mathbf{x}_{1:T}, \mathbf{h}_{1:T}) = \prod_{i=1}^{T} p(\mathbf{x}_t|h_t)p(h_t|h_{t-1}) \tag{A.8}$$

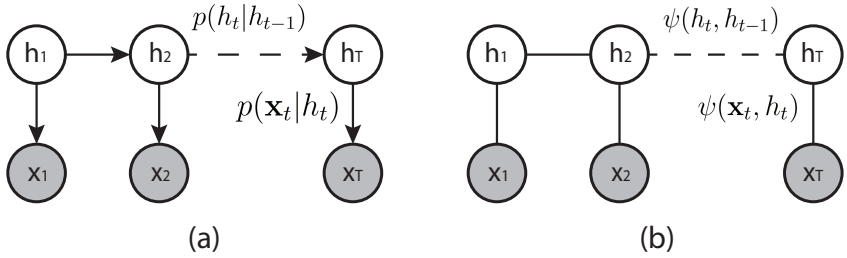

Figure A.2: Graphical representation of Dynamic Bayesian Networks with linear chain connectivity between latent variables. (a) HMM defined by the conditional probabilities $p(\mathbf{x}_t|h_t)$ and $p(h_t|h_{t-1})$. (b) CRF defined by the potentials $\psi(\mathbf{x}_t, h_t)$ and $\psi(h_t, h_{t-1})$. Note that CRF can be understood as an undirected graphical model analogous to HMM.

where $p(h_1|h_0) = p(h_1)$.

Given the above expression, the conditional probability $p(\mathbf{h}_{1:T}|\mathbf{x}_{1:T})$ can be defined as:

$$p(\mathbf{h}_{1:T}|\mathbf{x}_{1:T}) = \frac{p(\mathbf{h}_{1:T}, \mathbf{x}_{1:T})}{p(\mathbf{x}_{1:T})} \propto p(\mathbf{h}_{1:T}, \mathbf{x}_{1:T}), \tag{A.9}$$

Therefore, we can use marginalization in order to compute the probability $p(h_t|\mathbf{x}_{1:T})$ for a given h_t as:

$$p(h_t|\mathbf{x}_{1:T}) \propto \sum_{h_{1:t-1}} \sum_{h_{t+1:T}} \prod_{i=1}^{T} p(\mathbf{x}_i|h_i)p(h_i|h_{i-1}). \tag{A.10}$$

Note that Eq. A.10 requires to sum over a very large number of possible latent variable configurations $\{h_1, ..., h_{t-1}, h_{t+1}, ..., h_T\}$. Thus, a naive force algorithm would have an exponential complexity. In contrast, the forward-backward algorithm is designed to make this computation more efficient. Specifically, note that we can use the associative and commutative properties in order to express Eq. A.10 as the product of two terms $\alpha(h_t)$ and $\beta(h_t)$:

$$p(h_t|\mathbf{x}_{1:T}) \propto \alpha(h_t)\beta(h_t)$$

$$\propto \overbrace{p(\mathbf{x}_t|h_t) \sum_{h_{t-1}} p(h_t|h_{t-1})p(x_{t-1}|h_{t-1}) \sum_{h_{t-2:1}} \prod_{i=1}^{t-2} p(\mathbf{x}_i|h_i)p(h_i|h_{i-1})}^{\alpha(h_t)}$$

$$\times \underbrace{\sum_{h_{t+1}} p(h_{t+1}|h_t)p(\mathbf{x}_{t+1}|h_{t+1}) \sum_{h_{t+2:T}} \prod_{i=t+2}^{T} p(\mathbf{x}_i|h_i)p(h_i|h_{i-1})}_{\beta(h_t)} \quad \text{(A.11)}$$

which can be defined with a recursive formula as:

$$\alpha(h_t) = p(\mathbf{x}_t|h_t) \sum_{h_{t-1}} p(h_t|h_{t-1})\alpha(h_{t-1}) \quad \text{(A.12)}$$

$$\beta(h_t) = \sum_{h_{t+1}} p(h_{t+1}|h_t)p(\mathbf{x}_{t+1}|h_{t+1})\beta(h_{t+1}) \quad \text{(A.13)}$$

with $\alpha(h_1) = p(\mathbf{x}_1|h_1)p(h_1)$ and $\beta(h_T) = 1$. Eqs. A.12 and A.13 can be easily shown from:

$$p(h_t|\mathbf{x}_{1:T}) \propto \alpha(h_t)\beta(h_t)$$

$$\propto p(\mathbf{x}_t|h_t) \sum_{h_{t-1}} p(h_t|h_{t-1}) \overbrace{p(x_{t-1}|h_{t-1}) \sum_{h_{t-2:1}} \prod_{i=1}^{t-2} p(\mathbf{x}_i|h_i)p(h_i|h_{i-1})}^{\alpha(h_{t-1})}$$

$$\times \sum_{h_{t+1}} p(h_{t+1}|h_t)p(\mathbf{x}_{t+1}|h_{t+1}) \underbrace{\sum_{h_{t+2:T}} \prod_{i=t+2}^{T} p(\mathbf{x}_i|h_i)p(h_i|h_{i-1})}_{\beta(h_{t-1})} \quad \text{(A.14)}$$

These recursive definitions are used by the forward-backward algorithm to efficiently compute $\alpha(h_t)$ and $\beta(h_t)$ in linear time with respect to T. Concretely, the algorithm is divided in two steps. In

the forward pass, the value $\alpha(h_t)$ is computed by iterating from h_1 to h_t using the recursive definition in Eq. A.12. Similarly, in the backward step, $\beta(h_t)$ is estimated by iterating in the reverse order from h_T to h_t using Eq. A.13.

Once both terms are computed, the conditional probability of h_t given observations $\mathbf{x}_{1:T}$ can be easily computed as:

$$p(h_t|\mathbf{x}_{1:T}) = \frac{\alpha(h_t)\beta(h_t)}{\sum_{h_t} \alpha(h_t)\beta(h_t)} \tag{A.15}$$

The same algorithm can be also employed in order to compute $p(h_t, h_{t-1}|\mathbf{x}_{1:T})$.

Bibliography

[Agresti, 2010] Agresti, A. (2010). *Analysis of ordinal categorical data*. Wiley.

[Ambadar et al., 2005] Ambadar, Z., Schooler, J. W., and Cohn, J. F. (2005). Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions. *Psychological science*.

[Amores, 2013] Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*.

[Andrews et al., 2002] Andrews, S., Tsochantaridis, I., and Hofmann, T. (2002). Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*.

[Andrews et al., 2003] Andrews, S., Tsochantaridis, I., and Hofmann, T. (2003). Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*.

[Argyriou et al., 2007] Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. *Advances in Neural Information Processing Systems*.

[Artaechevarria et al., 2009] Artaechevarria, X., Munoz-Barrutia, A., and Ortiz-de Solórzano, C. (2009). Combination strategies in multi-atlas image segmentation: application to brain mr data. *IEEE Transactions on Medical Imaging*.

[Asman and Landman, 2011] Asman, A. J. and Landman, B. A. (2011). Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (collate). *IEEE Transactions on Medical Imaging*.

[Asman and Landman, 2012] Asman, A. J. and Landman, B. A. (2012). Formulating spatially varying performance in the statistical fusion framework. *IEEE Transactions on Medical Imaging*.

[Azizpour and Laptev, 2012] Azizpour, H. and Laptev, I. (2012). Object detection using strongly-supervised deformable part models. *Proc. European Conf. on Computer Vision*.

[B. Martinez, 2016] B. Martinez, M. V. (2016). Advances, challenges, and opportunities in automatic facial expression recognition. In M. Kawulok, E. Celebi, B. S., editor, *Advances in Face Detection and Facial Image Analysis*. Springer.

[Bahdanau et al., 2016] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*. IEEE.

[Baltrušaitis et al., 2013] Baltrušaitis, T., Banda, N., and Robinson, P. (2013). Dimensional affect recognition using continuous conditional random fields. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Barber, 2012] Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

[Baveye et al., 2015] Baveye, Y., Dellandrea, E., Chamaret, C., and Chen, L. (2015). Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*.

[Birgin et al., 2000] Birgin, E. G., Martínez, J. M., and Raydan, M. (2000). Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*.

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

[Byrd et al., 1994] Byrd, R. H., Nocedal, J., and Schnabel, R. B. (1994). Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*.

[Chang et al., 2009] Chang, K.-Y., Liu, T.-L., and Lai, S.-H. (2009). Learning partially-observed hidden conditional random fields for facial expression recognition. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Chen et al., 2006] Chen, Y., Bi, J., and Wang, J. Z. (2006). Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Chen and Wang, 2004] Chen, Y. and Wang, J. Z. (2004). Image categorization by learning and reasoning with regions. *J. of Machine Learning Research*.

[Chu and Ghahramani, 2005] Chu, W. and Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*.

[Chu et al., 2013] Chu, W.-S., De la Torre, F., and Cohn, J. F. (2013). Selective transfer machine for personalized facial action unit detection. *Proc. Computer Vision and Pattern Recognition*.

[Chu et al., 2017] Chu, W.-S., De la Torre, F., and Cohn, J. F. (2017). Selective transfer machine for personalized facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Cliff and Young, 1968] Cliff, N. and Young, F. W. (1968). On the relation between undimensional judgments and multidimensional scaling. *Organizational Behavior and Human Performance*.

[Cohn et al., 2009] Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. In *Proc. Int. Conf. Affective Computing and Intelligent Interaction and Workshops, 2009*.

[Commowick et al., 2012] Commowick, O., Akhondi-Asl, A., and Warfield, S. K. (2012). Estimating a reference standard segmentation with spatially varying performance parameters: local map staple. *IEEE Transactions on Medical Imaging*.

[Commowick and Warfield, 2010] Commowick, O. and Warfield, S. K. (2010). Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori staple. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer.

[Cootes et al., 2001] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Cowie et al., 2000] Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., and Schröder, M. (2000). 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop on speech and emotion*.

[Darwin, 1998] Darwin, C. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

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

[De la Torre et al., 2011] De la Torre, F., Simon, T., Ambadar, Z., and Cohn, J. F. (2011). Fast-facs: A computer-assisted system to increase speed and reliability of manual facs coding. In *Proc. Affective Computing and Intelligent Interaction*. IEEE.

[Dennis and Moré, 1977] Dennis, Jr, J. E. and Moré, J. J. (1977). Quasi-newton methods, motivation and theory. *SIAM review*.

[Devillers et al., 2006] Devillers, L., Cowie, R., Martin, J., Douglas-Cowie, E., Abrilian, S., and McRorie, M. (2006). Real life emotions in french and english tv video clips: an integrated annotation protocol combining continuous and discrete approaches. In *Proc. Int. Conf. on Language Resources and Evaluation*.

[Ding et al., 2016] Ding, X., Chu, W.-S., De la Torre, F., Cohn, J. F., and Wang, Q. (2016). Cascade of tasks for facial expression analysis. *Image and Vision Computing*.

[Du et al., 2014] Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*.

[Ekman, 1993] Ekman, P. (1993). Facial expression and emotion. *American psychologist*.

[Ekman and Friesen, 1978] Ekman, P. and Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press*.

[Ekman and Friesen, 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*.

[Ekman et al., 1978] Ekman, P., Friesen, W. V., and Hager, J. C. (1978). Facial action coding system (facs). *A technique for the measurement of facial action. Consulting, Palo Alto*.

[Ekman and Rosenberg, 1997] Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[El Meguid and Levine, 2014] El Meguid, M. K. A. and Levine, M. D. (2014). Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers. *IEEE Transactions on Affective Computing*.

[Eleftheriadis et al., 2016] Eleftheriadis, S., Rudovic, O., Deisenroth, M. P., and Pantic, M. (2016). Variational gaussian process auto-encoder for ordinal prediction of facial action units. In *Asian Conference on Computer Vision*, Taipei, Taiwan. Springer.

[Eleftheriadis et al., 2015] Eleftheriadis, S., Rudovic, O., and Pantic, M. (2015). Discriminative shared gaussian processes for multi-view and view-invariant facial expression recognition. *IEEE Transactions on Image Processing*.

[Everingham et al., 2015] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *Int. Journal of Computer Vision*.

[Fasel and Luettin, 2003] Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*.

[Foulds and Frank, 2010] Foulds, J. and Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review*.

[Friesen and Ekman, 1983] Friesen, W. V. and Ekman, P. (1983). Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*.

[Fu et al., 2011] Fu, Z., Robles-Kelly, A., and Zhou, J. (2011). Milis: Multiple instance learning with instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Gärtner et al., 2002] Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. J. (2002). Multi-instance kernels. In *Proc. International Conference on Machine Learning*. ACM.

[Geiger et al., 2012] Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Girard et al., 2015] Girard, J. M., Cohn, J. F., Jeni, L. A., Lucey, S., and la Torre, F. D. (2015). How much training data for facial action unit detection? In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Goodfellow et al., 2013] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *Advances in Neural Information Processing Systems*.

[Gosselin et al., 1995] Gosselin, P., Kirouac, G., and Doré, F. Y. (1995). Components and recognition of facial expression in the communication of emotion by actors. *Journal of personality and social psychology*.

[Gower, 1975] Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

[Green and Cliff, 1975] Green, R. S. and Cliff, N. (1975). Multidimensional comparisons of structures of vocally and facially expressed emotion. *Perception & Psychophysics*.

[Gunes and Schuller, 2013] Gunes, H. and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*.

[Gupta et al., 2017] Gupta, R., Audhkhasi, K., Jacokes, Z., Rozga, A., and Narayanan, S. (2017). Modeling multiple time series annotations based on ground truth inference and distortion. *IEEE Transactions on Affective Computing*.

[Gupta et al., 2007] Gupta, R., Diwan, A. A., and Sarawagi, S. (2007). Efficient inference with cardinality-based clique potentials. In *Proc. International Conference on Machine Learning*. ACM.

[Hajimirsadeghi et al., 2013] Hajimirsadeghi, H., Li, J., Mori, G., Zaki, M., and Sayed, T. (2013). Multiple instance learning by discriminative training of markov networks. In *Uncertainty in Artificial Intelligence*. Elsevier.

[Han et al., 2016] Han, S., Meng, Z., Khan, A.-S., and Tong, Y. (2016). Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in Neural Information Processing Systems*.

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

[Hsu et al., 2014] Hsu, K.-J., Lin, Y.-Y., and Chuang, Y.-Y. (2014). Augmented multiple instance regression for inferring object contours in bounding boxes. *IEEE Transactions on Image Processing*.

[Hurley and Rickard, 2009] Hurley, N. and Rickard, S. (2009). Comparing measures of sparsity. *IEEE Transactions on Information Theory*.

[Jamieson et al., 2004] Jamieson, S. et al. (2004). Likert scales: how to (ab) use them. *Medical education*.

[Jampour et al., 2015] Jampour, M., Mauthner, T., and Bischof, H. (2015). Multi-view facial expressions recognition using local linear regression of sparse codes. In *Proc. Computer Vision Winter Workshop*.

[Jiang et al., 2011] Jiang, B., Valstar, M. F., and Pantic, M. (2011). Action unit detection using sparse appearance descriptors in space-time video volumes. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Kaltwang et al., 2015] Kaltwang, S., Todorovic, S., and Pantic, M. (2015). Latent trees for estimating intensity of facial action units. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Kaltwang et al., 2016] Kaltwang, S., Todorovic, S., and Pantic, M. (2016). Doubly sparse relevance vector machine for continuous facial behavior estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Kim and Pavlovic, 2010a] Kim, M. and Pavlovic, V. (2010a). Hidden conditional ordinal random fields for sequence classification. In *Machine Learning and Knowledge Discovery in Databases*. Springer.

[Kim and Pavlovic, 2010b] Kim, M. and Pavlovic, V. (2010b). Structured output ordinal regression for dynamic facial emotion intensity prediction. In *Proc. European Conf. on Computer Vision*. Springer.

[Kim and Torre, 2010] Kim, M. and Torre, F. (2010). Gaussian processes multiple instance learning. In *Proc. International Conference on Machine Learning*. ACM.

[Kingma et al., 2014] Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Koelstra et al., 2012] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*.

[Kotsia and Pitas, 2007] Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. International Conference on Machine Learning*. ACM.

[Lakshminarayanan and Teh, 2013] Lakshminarayanan, B. and Teh, Y. W. (2013). Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. *arXiv preprint arXiv:1305.0015*.

[Landman et al., 2012] Landman, B. A., Asman, A. J., Scoggins, A. G., Bogovic, J. A., Xing, F., and Prince, J. L. (2012). Robust statistical fusion of image labels. *IEEE Transactions on Medical Imaging*.

[Langerak et al., 2010] Langerak, T. R., van der Heide, U. A., Kotte, A. N., Viergever, M. A., van Vulpen, M., and Pluim, J. P. (2010). Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *IEEE Transactions on Medical Imaging*.

[Langner et al., 2010] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and Emotion*.

[Leistner et al., 2010] Leistner, C., Saffari, A., and Bischof, H. (2010). Miforests: Multiple-instance learning with randomized trees. In *Proc. European Conf. on Computer Vision*. Springer.

[Lewis et al., 2010] Lewis, M., Haviland-Jones, J. M., and Barrett, L. F. (2010). Handbook of emotions. chapter 13.

[Li and Sminchisescu, 2010] Li, F. and Sminchisescu, C. (2010). Convex multiple-instance learning by estimating likelihood ratio. In *Advances in Neural Information Processing Systems*.

[Li et al., 2009] Li, X., Wang, Y.-Y., and Acero, A. (2009). Extracting structured information from user queries with semi-supervised conditional random fields. In *Proc. Conf. Research and development in information retrieval*. ACM.

[Lisetti and Schiano, 2000] Lisetti, C. L. and Schiano, D. J. (2000). Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragmatics & cognition*.

[Littlewort et al., 2007] Littlewort, G. C., Bartlett, M. S., and Lee, K. (2007). Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proc. Int. Conf. Multimodal interfaces*.

[Liu et al., 2016] Liu, J., Chen, C., Zhu, Y., Liu, W., and Metaxas, D. N. (2016). Video classification via weakly supervised sequence modeling. *Computer Vision and Image Understanding*.

[Lucey et al., 2009] Lucey, P., Cohn, J., Lucey, S., Sridharan, S., and Prkachin, K. M. (2009). Automatically detecting action units from

faces of pain: Comparing shape and appearance features. In *Proc. Computer Vision and Pattern Recognition Workshops*. IEEE.

[Lucey et al., 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proc. Computer Vision and Pattern Recognition Workshops*. IEEE.

[Lucey et al., 2011] Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., and Matthews, I. (2011). Painful data: The unbc-mcmaster shoulder pain expression archive database. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Lyons et al., 1998a] Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998a). Coding facial expressions with gabor wavelets. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Lyons et al., 1998b] Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J., and Budynek, J. (1998b). The japanese female facial expression (jaffe) database. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Mahoor et al., 2009] Mahoor, M. H., Cadavid, S., Messinger, D. S., and Cohn, J. F. (2009). A framework for automated measurement of the intensity of non-posed facial action units. In *Proc. Computer Vision and Pattern Recognition Workshops*.

[Maron and Lozano-Pérez, 1998] Maron, O. and Lozano-Pérez, T. (1998). A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*.

[Matthews and Baker, 2004] Matthews, I. and Baker, S. (2004). Active appearance models revisited. *Int. Journal of Computer Vision*.

[Mavadati et al., 2013] Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*.

[McCulloch and Rossi, 1994] McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*.

[McDuff et al., 2013a] McDuff, D., el Kaliouby, R., Demirdjian, D., and Picard, R. (2013a). Predicting online media effectiveness based on smile responses gathered over the internet. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[McDuff et al., 2013b] McDuff, D., Kaliouby, R., Senechalz, T., Amrz, M., Cohn, J., and Picard, R. (2013b). Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected in-the-wild. In *Proc. Computer Vision and Pattern Recognition Workshops*. IEEE.

[McKeown et al., 2012] McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*.

[Metallinou and Narayanan, 2013] Metallinou, A. and Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Metrikov et al., 2015] Metrikov, P., Pavlu, V., and Aslam, J. A. (2015). Aggregation of crowdsourced ordinal assessments and integration with learning to rank: A latent trait model. In *Proc. Int. Conf. Information and Knowledge Management*. ACM.

[Mollahosseini et al., 2016] Mollahosseini, A., Hasani, B., Salvador, M. J., Abdollahi, H., Chan, D., and Mahoor, M. H. (2016). Facial expression recognition from world wild web. In *Proc. Computer Vision and Pattern Recognition Workshops*. IEEE.

[Murphy and Russell, 2002] Murphy, K. P. and Russell, S. (2002). Dynamic bayesian networks: representation, inference and learning.

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

[Nicolaou et al., 2011] Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, pages 92–105.

[Nicolaou et al., 2014] Nicolaou, M. A., Pavlovic, V., and Pantic, M. (2014). Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Osgood et al., 1975] Osgood, C. E., May, W. H., and Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press.

[Pandey and Lazebnik, 2011] Pandey, M. and Lazebnik, S. (2011). Scene recognition and weakly supervised object localization with deformable part-based models. In *Proc. IEEE Int. Conf. on Computer Vision*. IEEE.

[Pantic and Patras, 2006] Pantic, M. and Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*.

[Pantic et al., 2005] Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *Proc. Int. Conf. Multimedia and Expo*. IEEE.

[Pfister et al., 2011] Pfister, T., Li, X., Zhao, G., and Pietikäinen, M. (2011). Recognising spontaneous facial micro-expressions. In *Proc. IEEE Int. Conf. on Computer Vision*. IEEE.

[Ponce-López et al., 2016] Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J., and Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision Workshops*. Springer.

[Prkachin, 1992] Prkachin, K. M. (1992). The consistency of facial expressions of pain: a comparison across modalities. *Pain*.

[Quattoni et al., 2007] Quattoni, A., Wang, S., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Rabiner and Juang, 1986] Rabiner, L. R. and Juang, B.-H. (1986). An introduction to hidden markov models. *ASSP Magazine*.

[Ray and Craven, 2014] Ray, S. and Craven, M. (2014). Supervised versus multiple instance learning: An empirical comparison. In *Proc. International Conference on Machine Learning*. ACM.

[Ray and Page, 2001] Ray, S. and Page, D. (2001). Multiple instance regression. In *Proc. International Conference on Machine Learning*. ACM.

[Raykar et al., 2008] Raykar, V. C., Krishnapuram, B., Bi, J., Dundar, M., and Rao, R. B. (2008). Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proc. International Conference on Machine Learning*. ACM.

[Ringeval et al., 2013] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Robinson and Baltrušaitis, 2015] Robinson, P. and Baltrušaitis, T. (2015). Empirical analysis of continuous affect. In *Proc. Int. Conf. on Affective Computing and Intelligent Interaction*. Springer.

[Rudovic et al., 2012] Rudovic, O., Pavlovic, V., and Pantic, M. (2012). Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Rudovic et al., 2013] Rudovic, O., Pavlovic, V., and Pantic, M. (2013). Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. In *International Symposium on Visual Computing*. Springer.

[Rudovic et al., 2015] Rudovic, O., Pavlovic, V., and Pantic, M. (2015). Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Ruiz et al., 2014] Ruiz, A., Van de Weijer, J., and Binefa, X. (2014). Regularized multi-concept mil for weakly-supervised facial behavior categorization. In *Proc. British Machine Vision Conference*. BMVA Press.

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*.

[Sangineto et al., 2014] Sangineto, E., Zen, G., Ricci, E., and Sebe, N. (2014). We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proc. ACM Multimedia*. ACM.

[Savran et al., 2008] Savran, A., Alyüz, N., Dibeklio\uglu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*. Springer.

[Scherer and Ellgring, 2007] Scherer, K. R. and Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*.

[Schmidt et al., 2009] Schmidt, M. W., Berg, E., Friedlander, M. P., and Murphy, K. P. (2009). Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *Proc. International Conference on Artificial Intelligence and Statistics*.

[Scovanner et al., 2007] Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proc. International Conference on Multimedia*.

[Shan, 2008] Shan, C. (2008). *Inferring facial and body language*. PhD thesis, Queen Mary University of London.

[Shotton et al., 2006] Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2006). Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. European Conf. on Computer Vision*. Springer.

[Sikka et al., 2013] Sikka, K., Dhall, A., and Bartlett, M. (2013). Weakly supervised pain localization using multiple instance learning. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Sukno et al., 2016] Sukno, F., Domínguez, M., Ruiz, A., Schiller, D., Lingenfelser, F., Pragst, L., Kamateri, E., and Vrochidis, S. (2016). A multimodal annotation schema for non-verbal affective analysis in the health-care domain. In *Proc. Int. Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction*. ACM.

[Tarlow et al., 2012] Tarlow, D., Swersky, K., Zemel, R. S., and Adams, R. P. (2012). Fast exact inference for recursive cardinality models. In *Conf. on Uncertainty in Artificial Intelligence*. Elsevier.

[Tax et al., 2010] Tax, D. M., Hendriks, E., Valstar, M. F., and Pantic, M. (2010). The detection of concept frames using clustering multi-instance learning. In *Proc. Int. Conf. on Pattern Recognition*. IEEE.

[Tian, 2004] Tian, Y.-l. (2004). Evaluation of face resolution for expression analysis. In *Proc. Computer Vision and Pattern Recognition Workshops*. IEEE.

[Tősér et al., 2016] Tősér, Z., Jeni, L. A., Lőrincz, A., and Cohn, J. F. (2016). Deep learning for facial action unit detection under large head poses. In *European Conference on Computer Vision Workshops*. Springer.

[Valstar et al., 2007] Valstar, M. F., Gunes, H., and Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. In *Proc. Int. Conf. Multimodal interfaces*.

[Valstar et al., 2012] Valstar, M. F., Mehu, M., Jiang, B., Pantic, M., and Scherer, K. (2012). Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*.

[Valstar and Pantic, 2006] Valstar, M. F. and Pantic, M. (2006). Biologically vs. logic inspired encoding of facial actions and emotions in video. In *Proc. Int. Conf. Multimedia and Expo*. IEEE.

[Valstar and Pantic, 2007] Valstar, M. F. and Pantic, M. (2007). Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *Int. Workshop on Human-Computer Interaction*. Springer.

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

[Velusamy et al., 2011] Velusamy, S., Kannan, H., Anand, B., Sharma, A., and Navathe, B. (2011). A method to infer emotions from facial action units. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE.

[Vezhnevets et al., 2011] Vezhnevets, A., Ferrari, V., and Buhmann, J. M. (2011). Weakly supervised semantic segmentation with a multi-image model. In *Proc. IEEE Int. Conf. on Computer Vision*. IEEE.

[Vezhnevets et al., 2012] Vezhnevets, A., Ferrari, V., and Buhmann, J. M. (2012). Weakly supervised structured output learning for semantic segmentation. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Vinciarelli et al., 2009] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Transactions on Image and Vision Computing*.

[Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *Int. Journal of Computer Vision*, 57(2):137–154.

[Walecki et al., 2016] Walecki, R., Rudovic, O., Pantic, M., and Pavlovic, V. (2016). Copula ordinal regression for joint estimation of facial action unit intensity. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Walecki et al., 2015] Walecki, R., Rudovic, O., Pavlovic, V., and Pantic, M. (2015). Variable-state latent conditional random fields

for facial expression recognition and action unit detection. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Walecki et al., 2017] Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B., and Pantic, M. (2017). Deep structured learning for facial expression intensity estimation. In *Proc. Computer Vision and Pattern Recognition*, Honolulu, Hawaii. IEEE.

[Wang et al., 2013a] Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. (2013a). Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Wang et al., 2012] Wang, Q., Si, L., and Zhang, D. (2012). A discriminative data-dependent mixture-model approach for multiple instance learning in image classification. In *Proc. European Conf. on Computer Vision*. Springer.

[Wang et al., 2013b] Wang, Z., Li, Y., Wang, S., and Ji, Q. (2013b). Capturing global semantic relationships for facial action unit recognition. In *Proc. IEEE Int. Conf. on Computer Vision*. IEEE.

[Warfield et al., 2004] Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*.

[Warfield et al., 2008] Warfield, S. K., Zou, K. H., and Wells, W. M. (2008). Validation of image segmentation by estimating rater bias and variance. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*.

[Warren et al., 2009] Warren, G., Schertler, E., and Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*.

[Whitehill et al., 2009] Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., and Movellan, J. (2009). Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Whitehill and Omlin, 2006] Whitehill, J. and Omlin, C. W. (2006). Haar features for facs au recognition. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

[Winkelmann and Boes, 2006] Winkelmann, R. and Boes, S. (2006). *Analysis of microdata*. Springer Science & Business Media.

[Wu et al., 2015] Wu, C., Wang, S., and Ji, Q. (2015). Multi-instance hidden markov model for facial expression recognition. In *International Conference on Automatic Face and Gesture Recognition*. IEEE.

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

[Yang et al., 2007] Yang, P., Liu, Q., and Metaxas, D. N. (2007). Boosting coded dynamic features for facial action units and facial expression recognition. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Yannakakis and Hallam, 2011] Yannakakis, G. N. and Hallam, J. (2011). Ranking vs. preference: a comparative study of self-reporting. In *Proc. Int. Conf. Affective Computing and Intelligent Interaction*. Springer.

[Yannakakis and Martínez, 2015a] Yannakakis, G. N. and Martínez, H. P. (2015a). Grounding truth via ordinal annotation. In *Proc. Int. Conf. Affective Computing and Intelligent Interaction*, pages 574–580. IEEE.

[Yannakakis and Martínez, 2015b] Yannakakis, G. N. and Martínez, H. P. (2015b). Ratings are overrated! *Frontiers in ICT*.

[Youssif and Asker, 2011] Youssif, A. A. and Asker, W. A. (2011). Automatic facial expression recognition system based on geometric and appearance features. *Computer and Information Science*.

[Zen et al., 2014] Zen, G., Sangineto, E., Ricci, E., and Sebe, N. (2014). Unsupervised domain adaptation for personalized facial emotion recognition. *Proc. Int. Conf. on Multimodal Interaction*.

[Zeng et al., 2009] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Zhang et al., 2005a] Zhang, C., Platt, J. C., and Viola, P. A. (2005a). Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems*.

[Zhang et al., 2005b] Zhang, C., Platt, J. C., and Viola, P. A. (2005b). Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems*.

[Zhao and Pietikainen, 2007] Zhao, G. and Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Zhao et al., 2016a] Zhao, K., Chu, W.-S., and Zhang, H. (2016a). Deep region and multi-label learning for facial action unit detection. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Zhao et al., 2016b] Zhao, R., Gan, Q., Wang, S., and Ji, Q. (2016b). Facial expression intensity estimation using ordinal information. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Zhao et al., 2016c] Zhao, R., Gan, Q., Wang, S., and Ji, Q. (2016c). Facial expression intensity estimation using ordinal information. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Zhong et al., 2012] Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., and Metaxas, D. (2012). Learning active facial patches for expression analysis. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Zhou et al., 2014] Zhou, D., Liu, Q., Platt, J. C., and Meek, C. (2014). Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proc. International Conference on Machine Learning*. ACM.

[Zhou et al., 2010] Zhou, F., De la Torre, F., and Cohn, J. F. (2010). Unsupervised discovery of facial events. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[Zhou et al., 2009] Zhou, Z.-H., Sun, Y.-Y., and Li, Y.-F. (2009). Multi-instance learning by treating instances as non-iid samples. In *Proc. International Conference on Machine Learning*. ACM.

[Zhu and Ramanan, 2012] Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE.