Imperial College London
Department of Computing

Robust Subspace Learning for Static and Dynamic Affect and Behaviour Modelling

Christos Georgakis

June, 2017

Supervised by Prof. Maja Pantic

Submitted in part fulfilment of the requirements for the degree of PhD in Computing and the Diploma of Imperial College London. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

Acknowledgements

I would like to thank my supervisor Prof. Maja Pantic for her invaluable guidance over the past years and especially for always being there to support my enthusiasm in exploring new research avenues. I would like to also thank Dr. Yannis Panagakis for his daily supervision, but, most importantly, for helping me shape a pragmatic view of what research signifies. Cheers mate, to our long discussions, real and surreal (kleo), and for being a great companion in robustly dealing with the randomness of London! I would also like to thank Dr. Stavros Petridis for his help and patience during the early stages of my Ph.D. A big thanks goes also to Dr. Stefanos Zafeiriou, for being such a torrential and knowledgeable interlocutor in any kind of discussion, from the most casual to the most philosophical. I also feel extremely obliged to Ms. Teresa Ng and Dr. Amani El-Kholy for being always eager to take care of my administrative tasks. I would also like to thank my examiners Prof. Daniel Rueckert and Dr. Cordelia Schmid for providing constructive suggestions that improved my thesis.

I would also like to say a big thanks to all my friends from the i-BUG group for the fantastic time we spent together in the labs and at QA. I couldn’t have asked for a more interesting and fun mixture of people in a research environment, that’s for sure! I would like to express my special gratitude to my mates, the ‘chat members’, Nondas, Simos, Stefanos, and Thanos. Thanks guys for putting up with my enthusiasm, loud voice, inexplicable voice messages, my random singing, my time-consuming visits to your offices and cigarette breaks, and overall for sharing moments that are and will always be beautiful stories to recount. Nondas thanks for being always there to ‘look’ at the flag together, Simos thanks for being the partner in crime for the best PnR duo ever, Stefanaki thanks for the invaluable diorakia and the operations ’Covent Garden’, Thano thanks for your culinary wonders and for putting us in order whenever necessary. A big shout out also to the Britons James (no scratches) and Patrick (pal), the incredible negotiation master Trigeorge (’this is ridiculous’), the wise Cris Sag, the hombrero Ogi, and my office mates, the 100m runner Greg and the classy Jean. I would like to also thank all the members of the Kourelou band (check their website, they are good) Kostas, Nicos, Odysseas, Pavlos (zeis) and Spiros a.k.a. Bolovinho, for all the incredible gigs in London and summers in Amorgos we spent together, the musical escapes that balanced out the Ph.D. overload! A special mention goes to my best friend Kostas Tsipras here, the unbelievable lad, the rising Greek e-commerce star, the solid bass-guitar tempo keeper, the mary connoisseur, the techno enthusiast, for his endless support.

Finally, I would like to express my deepest gratitude to my brother Angelos and my parents Giorgos and Stathoula. Thank you for being such a loving family and an endless source of strength for anything I set my mind on to achieve.

“*...The total passion for the total height...*”

— Ayn Rand, *The Fountainhead*

Abstract

Machine analysis of human affect and behavior in naturalistic contexts has witnessed a growing attention in the last decade from various disciplines ranging from social and cognitive sciences to machine learning and computer vision. Endowing machines with the ability to seamlessly detect, analyze, model, predict as well as simulate and synthesize manifestations of internal emotional and behavioral states in real-world data is deemed essential for the deployment of next-generation, emotionally- and socially-competent human-centered interfaces. In this thesis, we are primarily motivated by the problem of modeling, recognizing and predicting spontaneous expressions of non-verbal human affect and behavior manifested through either low-level facial attributes in static images or high-level semantic events in image sequences. Both visual data and annotations of naturalistic affect and behavior naturally contain noisy measurements of unbounded magnitude at random locations, commonly referred to as 'outliers'. We present here machine learning methods that are robust to such gross, sparse noise. First, we deal with static analysis of face images, viewing the latter as a superposition of mutually-incoherent, low-complexity components corresponding to facial attributes, such as facial identity, expressions and activation of atomic facial muscle actions. We develop a robust, discriminant dictionary learning framework to extract these components from grossly corrupted training data and combine it with sparse representation to recognize the associated attributes. We demonstrate that our framework can jointly address interrelated classification tasks such as face and facial expression recognition. Inspired by the well-documented importance of the temporal aspect in perceiving affect and behavior, we direct the bulk of our research efforts into continuous-time modeling of dimensional affect and social behavior. Having identified a gap in the literature which is the lack of data containing annotations of social attitudes in continuous time and scale, we first curate a new audio-visual database of multi-party conversations from political debates annotated frame-by-frame in terms of real-valued conflict intensity and use it to conduct the first study on continuous-time conflict intensity estimation. Our experimental findings corroborate previous evidence indicating the inability of existing classifiers in capturing the hidden temporal structures of affective and behavioral displays. We present here a novel dynamic behavior analysis framework which models temporal dynamics in an explicit way, based on the natural assumption that continuous-time annotations of smoothly-varying affect or behavior can be viewed as outputs of a low-complexity *linear dynamical system* when behavioral cues (features) act as system inputs. A novel robust structured rank minimization framework is proposed to estimate the system parameters in the presence of gross corruptions and partially missing data. Experiments on prediction of dimensional conflict and affect as well as multi-object tracking from detection validate the effectiveness of our predictive framework and demonstrate that for the first time that complex human behavior and affect can be learned and predicted based on small training sets of person(s)-specific observations.

History of My Research

Static and dynamic analysis of human non-verbal affect and behavior has primarily motivated our research efforts, the outcomes of which are presented in detail in this thesis. This application domain became the principal area of investigation for my studies after I had explored a different problem which can be subsumed in the category of visual speech biometrics and paralinguistics. Inspired by recent findings suggesting the importance of visual cues in the perception of naturalistic non-verbal behaviors, in the beginning of my Ph.D. studies I focused on foreign accent recognition based exclusively on visual features. We approached this task as a binary classification problem of discriminating native from non-native speech from visual speech episodes captured by mobile devices. Specifically, we investigated to what extent temporal visual speech dynamics related to foreign accent can be modeled and identified when the audio stream is missing or noisy, the speech content unknown and the visual stream acquired under unconstrained conditions. We discovered that by using visual features encapsulating complementary appearance information and sequential classifiers capable of capturing temporal dependencies, one can accurately distinguish native from non-native speech for subjects unseen in the training phase. Overall, our efforts led to the development of the first automated frameworks for visual-only discrimination between native and non-native English speech, which can be exploited to ameliorate the performance of accent-sensitive speech recognizers as well as design alternative, unobtrusive biometric systems. However, in all three distinct approaches that we devised to address this problem, the accent-related information is not explicitly modeled, decoupled and extracted from the visual sensory information, but rather left to feature descriptors and classifiers to implicitly infer. In other words, the modeling approaches employed do not offer interpretability as to what accent traits are captured by both the representation learning and classification stages when trained with visual speech episodes of native and non-native speech. This realization urged me to view this task as a special case of a more generic face analysis problem broken down to interrelated face analysis tasks and seek more interpretable optimization problems based on recent advances on convex optimization and sparse representation learning.

Accent is a soft biometric trait characterizing the speaking style of individuals belonging to a particular language group and, as such, can be represented by a single discrete label for utterances corresponding to a given subject. As such, if one perceives the speech content in a visual speech sequence as ‘ *content*’ of the visual sensory information, one can categorize as belonging to the ‘ *style*’ of speech, along with subject-specific factors related to articulation, eloquence, emotions, expressiveness, openness, to mention but a few. Prompted by this view of visual verbal and non-verbal behavior, in the course of studies I started investigating learning algorithms that can break this high-dimensional information of emotionally- and socially-colored human behavior represented as a single noisy training data matrix into multiple structured class-specific additive matrix components corresponding to different attributes. This conceptual approach to visual human behavior analysis made me depart from the specific problem of accent classification and see the bigger picture, which can be summarized in the question “ *can we devise a model that would utilize label information and suitable norms in the richest canal of human communication, that is, the human face, to to discover components related to different attributes such as identity and facial expression?*”. Since we first focused on static face analysis, it became evident that accent could not be of use in this direction, since it is intrinsically related to dynamics of speech. The facial attributes that were addressed by the method presented in Chapter 3 were identity, facial expression and activation of facial muscle actions. When later I set as ultimate goal of my Ph.D. studies the development of a machine learning method that can explicitly capture temporal dynamics to encode affective and behavioral displays at a finer granularity, I turned the focus of attention towards continuous-time, dimensional affect and behavior analysis. However, since accent cannot be described in terms of continuous-time, real-valued measurements, the problem of accent classification does not lend itself to the investigation of dynamic, subtle affect and behavior and thus was not included in that study. Currently, having investigated both static and dynamic affect and behavior, I certainly believe that accent can be approached by means of robust subspace learning methods as well. However, the soft presence of accent-related information compared to other factors related to subject-specific biometric traits and affective manifestations, renders accent modeling in time a demanding in terms of machine learning effort task that, to my opinion, should be approached holistically in synergy with complementary face analysis tasks related to visual speech. To this end, one could investigate component analysis approaches that depart from a single matrix decomposition and take the form of multi-linear or tensor decompositions that would tackle decoupling of multiple sources of variation in time, one of them being accent. The challenge would be to investigate suitable decomposition formulations and suitable structure-inducing norms so that would facilitate the extraction and explicit modeling of accent-related dynamics from the multi-attribute facial information. Should the reader of this thesis wish to learn more about the part of my research focusing on accent classification, he/she is directed to the following list of publications that stemmed from it.

[1] **C. Georgakis**, S. Petridis, M. Pantic. Discrimination Between Native and Non-Native Speech Using Visual Features Only. IEEE Transactions on Cybernetics (T-CYB), 46(12): pp. 2758–2771, December 2016.

[2] **C. Georgakis**, S. Petridis, M. Pantic. Discriminating Native from Non-Native Speech Using Fusion of Visual Cues. In ACM International Conference on Multimedia (ACM MM), Orlando, Florida, USA, pp. 1177–1180, November 2014.

[3] **C. Georgakis**, S. Petridis, M. Pantic. Visual-only Discrimination between Native and Non-Native Speech. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy, pp. 4828–4832, May 2014.

This thesis is dedicated to yiayia Spyridoula.

Contents

Chapter 1

Introduction

Machine analysis of human affect and behavior stands in the forefront of Artificial Intelligence advancements and human-centered computing developments in our days. Affective computing [179] is a research area that has been active for over twenty years now and has matured to such an extent that the term *machine intelligence* necessarily subsumes *emotional intelligence* of the designed intelligent interfaces. Affective states play a fundamental role in human-human communication in that they motivate human actions and enrich human interactions. Thus, proactive and human-centered interfaces must have the ability to detect, analyze, model, and predict manifestations of internal emotional states in multi-sensory data as well as seamlessly interact with the user by simulating and synthesizing affective displays [165]. On the other hand, human–human communication is always socially situated, with social interactions being not a mere transmission of beliefs and opinions but a part of a larger social interplay [225]. Social signals are omnipresent in social interactions by either contributing to the adjustment of relations between agents (human and artificial) or revealing information about the agents such as turn taking, agreement, politeness, empathy, friendliness, (dis)agreement, conflict [160]. Hence, endowing machines with *social intelligence*, that is, the ability to sense and recognize the user's social signals and behaviors as well as respond and adapt to these signals, is essential for the deployment of next-generation interfaces [161]. Smart devices and robots such as virtual assistants for smart homes or health services, that can not just listen and talk, but also perceive the emotional and social facets in human conversation and respond in a polite, unintrusive, or persuasive manner, are not a far-fetched fiction anymore, but rather a realistic potential outgrowth of modern affective computing and social signal processing research. Despite the tremendous progress made in that direction, there are still numerous challenges in the journey towards cognitive systems that are emotionally and socially intelligent and aware when faced with continuous-time, naturalistic, spontaneous observations of human affect and behavior captured under unconstrained, real-world settings.

Affective Computing [179, 159] focuses on sensing, detecting, interpreting or deliberately influencing human affective states as well as devising appropriate means to handle this affective information in order to enhance current human-computer interaction designs [164]. The problem of modeling human affective behavior computationally has both great practical importance as well as theoretical interest and, as such, has propelled research efforts in a variety of disciplines such as computer science, psychology, neuroscience, and linguistics. As a matter of fact, the construction of any automatic affect recognizer highly depends on our understanding of the nature of affect, which consists of appropriately describing an affective state and defining its association with the human communicative signals (e.g., tone of voice, facial expressions, body gestures) through which is manifested [255].

Numerous approaches to human affect analysis have been proposed in the recent years within the fast-growing fields of machine learning and computer vision, the majority of which are based on non-verbal, audio and/or visual displays [255, 163]. While there has been evidence indicating the correlation of affective states with specific audio signals [100], the visual channel carrying facial expressions and body gestures seems to be most important in the human judgment of behavioral cues [7]. The human face is arguably our preeminent multi-signal input-output communicative system by means of which we communicate and perceive somebody's affective state and intentions on the basis of the portrayed facial expression [104]. Pantic [159] highlights that the facial expression modality should be the core representation of any automatic affect recognition system, while multi-modal frameworks should preferably include also body gestures and acoustic prosodic features. Facial expression recognition has been mainly approached by machine learning frameworks that recognize basic prototypical expression or facial muscle actions in static face images captured under constrained, laboratory settings [166]. Recent efforts in the field are directed towards training recognizers capable of distinguishing among naturalistic facial expressions from spontaneous data, as opposed to deliberate and often exaggerated expressive data [163]. Another emerging trend in facial expression analysis advocates the use of temporal classifiers for facial affective state modeling, prompted by theoretical evidence and experimental findings that stress the importance of the temporal aspect in the human and machine perception of affect [83].

Social signal processing (SSP) [174, 160, 228, 161] is a relatively new research and technological domain that aims at understanding and modeling human social signals and interactions as well as providing computers with similar abilities in human-computer interaction scenarios. Social signals and social behaviors are the expression of one's attitude towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioral cues including gaze exchange, blinks, smiles, head nods, crossed arms, laughter, expressive prosody, and similar [228]. Social signals typically last for a short time (from milliseconds, e.g., turn-taking, to minutes, e.g., mirroring), compared to social behaviors that last longer (from seconds, e.g., agreement, to minutes, e.g., politeness, to hours or days, e.g., empathy) and are distinguished from other internal states in that they explicitly or implicitly confer social intentions of the subjects involved in a social interaction. Similarly to affective displays, social signals are high-level semantic events whose mapping to an internal state or specific meaning is not always uniquely defined, but is rather highly dependent on factors such as content, culture and temporal interval [161]. Despite their inherent subtlety and ambiguity, social signals have been shown to be evident enough that they can be distinguishable by computer-based systems involving audio-visual sensors and machine learning techniques [174]. Consequently, automated analysis of social signals can tremendously facilitate research in social and cognitive sciences by reducing the effort and improving the quality of studies on social phenomena. At the same time, accurately modeling socially-relevant nuances of human non-verbal behavior can not only equip cognitive systems with social awareness but also drive the deployment of more natural, flexible interfaces that will synthesize socially-appropriate signals (e.g., politeness, empathy, interest) to adorn the communication and thus nurture rapport with humans.

The technological advances witnessed in the SSP area are numerous in the past years and have been largely boosted by the progress made in machine analysis of relevant behavioral cues such as blinks, smiles, head nods, laughter, and similar [36, 53, 103, 127, 165, 255, 83]. From the modeling standpoint, more sophisticated data collection techniques have been adopted (e.g., crowdsourcing for obtaining multiple annotations [108], statistical fusion for merging annotations [146]), while more efficient machine learning classifiers have been utilized (e.g., [63, 28, 108]). In terms of applications, research on modeling high-level social phenomena is still at an early stage and limited to a few attempts that have targeted social dominance [93], engagement and hot-spots [242], mimicry [22], personality traits [177], roles in meetings [20], and political stances [226], among others. Another line of research in this domain has approached the problem of recognizing social attitudes, which are defined as positive or negative evaluations of a person or a group of people and include cognitive elements like beliefs, opinions, and social emotions [161]. The omnipresence of social attitudes such as agreement, disagreement and conflict, in everyday social life and their importance in shaping our perception of social interactions has motivated a number of automated approaches to (dis)agreement and conflict [27, 25, 106, 107, 228]. Overall, most of the existing approaches to SSP still approach modeling of high-level social signals and behaviors in a similar fashion to low-level semantic events such as the occurrence of an ironic smile or a hand wave. There is a growing belief that a more principled approach should be followed for the development of next-generation socially-competent systems. The main challenges for these systems will be the ability to perceive the evolution of the multimodal social phenomena by taking into account contextual and temporal information as well as model and distinguish grammars of prototypic persons behaviors [228, 161].

In this Ph.D. thesis, we focus on the problem of modeling manifestations of human non-verbal affect and behavior in still images or video sequences captured under unconstrained, real-world conditions. Both static and dynamic behavior analysis is investigated, with the former targeting face analysis tasks in still images and the latter addressing analysis of dimensional affect or social behavior in continuous time. While the human face acts as the primary sensory input for our work, the developed models can accommodate information coming from other modalities as well. The machine learning frameworks proposed in this thesis share a common denominator in terms of modeling in that they build on *robust subspace learning* to obtain low-dimensional representations of the observed still or sequential facial images that are (possibly) corrupted by sparse errors of unbounded magnitude, commonly referred to as 'outliers' [115]. The term *outlier* refers to observations that violate the assumed statistical model for the data, while a *robust* estimation method is one that can tolerate some percentage of outlying measurements without having the solution arbitrarily skewed [52]. These undesirable artifacts occur frequently in real-world visual data captured 'in-the-wild' and can be due to various factors including occlusion (e.g., by sunglasses, hair, wrapping or hand), illumination (e.g., self-shadowing and specularities), image noise (e.g., scanning of archived data) or computer vision pre-processing errors (e.g., incorrect facial point tracking and image registration), to mention but a few. When the supervised learning paradigm is employed, i.e., human assessments of behavior or affect are utilized in the training phase, outliers can also corrupt these ground truth annotations which may be unreliable mainly due to annotator subjectivity, adversarial annotators or ineffective annotation fusion techniques (e.g., simply considering the mean of multiple annotations) [146]. In the presence of grossly corrupted outlying data in the feature and/or annotation domain, classical subspace learning approaches based on least squares estimation techniques lead to solutions that can be arbitrarily biased and hence are deemed unrealistic [90].

Let us further clarify this point by providing examples of applications where robustness to gross but sparse noise is quintessential in order for a machine learning model to achieve high performance on real-world data. For instance, consider the problem of finding a low-dimensional representation of human faces, that are captured under non-uniform illumination, exhibit various facial expressions and are possibly occluded (e.g., sunglasses, scarves), with the purpose of building a face recognizer. A non-robust subspace estimation method such as the Principal Component Analysis [98] will include the modes of variation owing to the illumination-, expression- and occlusion-related outliers in the derived representation, which will be highly undesirable for the target application. On another example, consider the problem of finding hidden structures explaining the temporal dynamics of a social attitude such as *conflict* among two or more interactants in a naturalistic conversation. If the learning algorithm is not robust to gross noise occurring due to the feature extraction (e.g., inaccurate frontalization of the images) or the annotation process (e.g., large meaningless spikes in the ground truth annotation due to inattentive or spammer annotators), it will end up fitting these noise-related fluctuations as being part of the underlying data generating process (e.g., an auto-regressive model), thus leading to high-complexity latent structures that will not necessarily correspond to the semantics of the observed behavior.

This thesis focuses on developing models that are robust to statistical outliers and constitute the main building block of learning frameworks that can yield accurate low-dimensional embeddings (e.g., appearance, shape, motion, temporal dynamics) of high-dimensional observed data with realistic amounts of unmodeled noise like those encountered in the aforementioned applications. Emphasis is placed on designing methods that can derive robust, discriminant, low-complexity components from facial imagery (possibly) corrupted by gross, sparse outliers in order to recognize facial attributes in static images as well as model and predict the temporal evolution of affective and behavioral attributes as a function of visual cues.

The main questions that this thesis attempts to answer, with respect to human non-verbal affect and behavior, are as follows.

Question 1. Can we jointly extract discriminant low-complexity components associated with facial attributes, such as facial identity, expressions and activation of AUs, from grossly corrupted facial images by means of a single supervised learning algorithm? Can we jointly address classification of these interrelated attributes in a unified classification framework as opposed to treating them independently?

Question 2. Can we use dimensional rather than categorical descriptions to describe subtle, spontaneous manifestations of social attitudes, such as interpersonal conflict, in continuous time, in the same way that we model dimensional affect? Which cues are most suitable to encode expressions of conflict in naturalistic conversations? Are existing off-the-shelf classifiers/regressors able to accurately capture the latent structures associated with the temporal dynamics of real-valued conflict intensity?

Question 3. Can we explicitly model the dynamics of continuous-time, dimensional characterizations of human affect and behavior? Can we learn the functional mapping of sequential observations of behavioral cues to real-valued annotations of affect and behavior by means of a hidden process generating the latter as outputs when the former are viewed as inputs? How accurately can we estimate the complexity, or equivalently, the memory and the other parameters of this underlying process in the presence of sparse, non-Gaussian noise and missing features and/or annotations?

Question 4. Can we predict future values of dimensional affect or behavior manifested in a video sequence based on a few amount of past observations as opposed to relying on large comprehensive training datasets? How is the performance of the predictive framework affected by varying the amount of observations in a sequence used for training and testing?

Motivated by these questions and having identified that there is large room for improvement over existing automated approaches to human affect and behavior analysis in terms of both efficiency and interpretability in real-world scenarios, we develop new machine learning methodologies that have solid theoretical foundations and clear conceptual interpretation so that we can accurately describe spontaneous expressions of affective and social states. Notably, the machine learning methods proposed herein provide powerful modeling platforms that can generalize to automatic analysis of various facial attributes or affective and behavioral displays ‘in-the-wild’ given relevant annotations.

As mentioned above, the learning frameworks proposed in this thesis build on *robust subspace learning* to derive robust, discriminant, low-complexity components from grossly corrupted still or sequential facial images. The proposed models consider – but are not limited to – two main structures for the derived components, namely *low-rank* and *sparsity*. Minimizing the *rank* of a data matrix translates into uncovering linear relationships in noisy data and has been extensively applied for dimensionality reduction and manifold learning [68]. The growing emergence of rank minimization-based methods in computer vision and machine learning (e.g., [122, 123, 168, 155, 124, 144]), has been largely boosted by the the work in [69] which has proposed an efficient convex relaxation to the original intractable rank minimization problem. On the other hand, *sparsity* has traditionally been a fundamental concept in robust statistics [90], where sparsity-promoting norms have been used to equip statistical models with robustness to noise that does not follow the Gaussian assumption but, instead, is better characterized by heavier-than-Gaussian-tailed distributions [90]. The natural sparsity criterion, which dictates to use the ℓ_0-norm to minimize the number of non-zero parameters, leads to an intractable problem. Recent advances in robust compressive sensing [60] advocate the use of ℓ_1-norm, which is the closest convex approximation of the ℓ_0-norm. These findings have given rise to a variety of parsimonious models for recognition [170, 243] that combine the merits of sparse representation, dictionary learning and convex optimization. Finally, the notions of low-rank and sparsity are jointly encountered in various recently proposed computer vision and machine learning methods. The latter are mainly inspired by the seminal work of Candès et al. [32] and are formulated upon the assumption that the observed data can be reconstructed by separable low-rank/sparse underlying components, such as background/foreground streams for video surveillance [29], expression-less/expressive faces for micro-expression recognition in video sequences [234], or background topics/keywords for latent semantic indexing in documents [139].

In what follows, we describe in more technical depth the *contributions* of the work presented in this thesis with respect to the affective and behavioral phenomena investigated, that is, (i) static face analysis (ii) data collection and experimental study on conflict intensity estimation, and (iii) dynamic analysis of dimensional affect and behavior.

The first application domain that we focus on is *static face analysis* in (possibly) grossly corrupted still face images captured under unconstrained conditions including varying illumination, facial expression and heavy contiguous occlusion (e.g., sunglasses, scarf). Face images convey rich information which can be perceived as a superposition of low-complexity components associated with attributes, such as facial identity, expressions and activation of atomic facial muscle actions that correspond to all visually discernible facial movements and are commonly referred to as facial Action Units (AUs) [61]. For instance, low-rank components characterizing neutral, expression-less, facial images are associated with identity, while sparse components capturing non-rigid deformations occurring in certain face regions reveal expressions and AU activations. In Chapter 3, we introduce the Discriminant Incoherent Component Analysis (DICA) to extract low-complexity components corresponding to facial attributes, which are mutually incoherent among different classes (e.g., identity, expression, AU activation) from training data, even in the presence of gross sparse errors. From the feature extraction standpoint, the DICA acts as a discriminant dictionary learning method, since it utilizes label information and structure-inducing norms on the facial aspects in a suitable optimization problem to learn an ensemble of class-specific incoherent facial components. At the classification stage, the DICA lends itself to *sparsity-based recognition* [243] ; an unseen (test) image is expressed as a group-sparse linear combination of the extracted components, where the non-zero coefficients reveal the class(es) of the respective facial attribute(s) that it belongs to.

It becomes evident that the DICA provides an efficient machine learning platform where interrelated classification tasks, such as face recognition and facial expression recognition, can be jointly addressed. This property, albeit highly desirable in modern real-time cognitive applications, has been scarcely addressed in the machine learning community which still approaches face analysis tasks in isolation. Aside from joint face and facial expression recognition, the generalizability and effectiveness of the DICA is demonstrated by conducting experiments on face recognition for varying percentages of corrupted images in the training set, subject-independent expression recognition under varying illumination conditions during training, as well as facial action unit detection. Overall, the DICA constitutes a robust learning framework that can generalize to classification of any number or type of labeled attributes that manifest themselves in the visual stream through specific structures, associated with mutually incoherent modes of variation.

Inspired by the growing, solidly-grounded and well-documented belief in the affective computing [83] and social signal processing literature [161] suggesting that the temporal aspect is crucial for both human and machine perception of spontaneous affective and social behaviors, we directed a large portion of our research efforts towards dynamic, continuous-time modeling of affect and behavior. It is straightforward to realize that temporal modeling of human non-verbal affect and behavior cannot be approached through *categorical* descriptions, that is, non-verbal expressions in terms of basic emotion categories (e.g., happiness, sadness, fear) or discrete social states (e.g., agreement/disagreement, conflict/non-conflict), respectively. Instead, modeling transitions between moderate and naturalistic affective and behavioral displays necessitates the use of *dimensional* descriptions, where affective and social states are characterized in terms of latent dimensions taking real values as a function of time. For dynamic affect modeling, two dimensions have been shown to be sufficient for capturing most of the affective variability: *valence* and *arousal* (V-A) [113], signifying respectively, how positive/negative and active/inactive an emotional state is [113]. While various works and datasets have been proposed in the recent years for dimensional affect recognition (see [83] for a comprehensive survey), analogous continuous-scale characterizations of social behaviors such as interest, politeness, flirting, (dis)agreement, and conflict, are rarely adopted by social signal processing methodologies. This is to a large extent attributed to the lack of annotated data, which mainly stems from the difficulty in obtaining *continuous-time* and *dimensional* annotations of social behavior [24]. As a matter of fact, this task requires suitable real-time annotation tools (e.g., FeelTrace [48]), multiple annotators to reduce the annotator subjectivity effects as well as an efficient technique for merging the multiple annotations in a single ground truth annotation [146].

To fill the aforementioned gap in the availability of data for studying social phenomena in continuous time and scale, we introduce a new database suitable for the investigation of a social attitude, namely *conflict*, in naturalistic conversations. *Conflict* is used to label a range of human experiences, from disagreement to stress and anger, occurring when involved individuals act on incompatible goals, interests, or intentions over resources or attitudes [5, 99]. With conflict having been recognized as one of the main dimensions along which a dyadic or multi-party social interaction is perceived [116], automatic analysis of conflict can tremendously boost the deployment of technologies targeting social interactions understanding and social skills enhancement. The Conflict Escalation Resolution (CONFER) Database presented in Chapter 4, is a collection of audio-visual recordings of spontaneous interactions from political debates where conflicts naturally arise and is the first of its kind to having been annotated in terms of continuous-time and dimensional conflict intensity by multiple annotators. Furthermore, the CONFER database is accompanied by the first experimental study on continuous-time and dimensional conflict intensity estimation, where comparative evaluation of various features and classifiers for the task at hand offer valuable insights for future research.

Dynamic affect and behavior analysis in the presence of gross, but sparse, noise and incomplete visual data is the second vast application domain with which this Ph.D. thesis deals with. This part of our research adopts the temporal modeling paradigm described above for the investigation of affective and behavioral phenomena, which necessitates their description in continuous time and scale. Hence, each recognition task is posited as a regression problem of estimating real-valued descriptions of affect or behavior on a frame-by-frame basis in test sequential observations 'unseen' in the training phase. Most existing approaches to continuous-time modeling of affect and behavior have relied on off-the-shelf classifiers to capture the statistical regularities in the evolution of the relevant cues in time. Representative examples include Hidden Markov Models (HMMs) [43] for video-based facial expression recognition, Dynamic Bayesian Networks (DBN) for complex mental state recognition [63], Hidden Conditional Random Fields (HCRF) for (dis)agreement detection [27, 28], Long-Short Term Memory (LSTM) Neural Networks for continuous prediction of dimensional affect [145], and regression-based approaches for continuous emotion and depression recognition or pain estimation [147, 218, 102]. Despite their merits, these methods rely on large sets of training data to learn a large number of parameters, they are not all suited for regression tasks, while their performance becomes brittle in the presence of gross non-Gaussian noise and incomplete data, which is abundant in real-world (visual) data, as we saw above (see Section 2.2.1). Another fundamental limitation shared by these methods lies on the fact that they do not explicitly model the joint temporal evolution of affective or behavioral characterizations and related cues within a systematic dynamic framework learned from the data. As such, they can be applied neither to learn prototypic manifestations of affect and behavior from data nor to measure the similarity between different behavior and affective displays.

In this thesis, we provide a remedy to the aforementioned problems by proposing a novel framework for dynamic affect and behavior analysis "in-the-wild" in Chapter 5 which models dynamics in an explicit way. This time, robust subspace learning is utilized to robustly learn a low-complexity approximation of the latent auto-regressive process explaining the temporal dependencies of the sequential observations. Specifically, the modeling assumption here is that continuous-time annotations characterizing the temporal evolution of relevant behavior or affect are generated as outputs of a low-complexity *linear dynamical system* when behavioral cues (features) act as system inputs. Having learned this dynamical system from the training data, unknown real-valued descriptions of affect or behavior (system outputs) can be predicted by applying the system equations for the respective features (system inputs). The core of the proposed system learning method is a novel structured rank minimization algorithm for linearly (Hankel)-structured data matrices, which is used to estimate the most crucial parameter, that is, the latent *order* or *memory* of the system. As opposed to existing structured rank minimization methods, the proposed method can handle both (partially) missing data and grossly corrupted observations. In the same time, by utilizing efficient approximations of the rank function and the sparsity-promoting ℓ_0-norm, it provides an estimate of the system order that is close to the true (unknown) order. The other parameters of the system can subsequently be learned by solving a system of linear equations.

Overall, the predictive framework proposed in Chapter 5 is the first machine learning approach to dynamic analysis of dimensional affect and behavior in which annotations and features act as outputs and inputs, respectively, of a low-order linear dynamical system that models the latent temporal structure. In terms of applications, the generalizability of the proposed framework is demonstrated by conducting experiments on three distinct dynamic behavior analysis tasks, namely (i) *conflict intensity prediction*, (ii) *prediction of valence and arousal*, and (iii) *tracklet matching*, that is, multi-(object/person) tracking from detection. The last experiment, in which our method is assigned the task of capturing dynamics related to motion trajectories of multiple objects/people in a heavily occluded visual scenario, serves to highlight that our method can also operate as an unsupervised learning algorithm. Most importantly, the framework we propose departs from a practice commonly adopted in behavioral and affective computing, that is, to train machine learning algorithms by employing large sets of training data that comprehensively cover different subjects, contexts, interaction scenarios and recording conditions. Specifically, we demonstrate for the first time that complex human behavior and affect, manifested by a single person or group of interactants, can be learned and predicted based on a small amount of person(s)-specific observations, amounting to a duration of just a few seconds.

This Ph.D. thesis is structured in a way that facilitates the inspection of the contributions of our work with respect to (i) static (joint) modeling and classification of facial attributes, (ii) data collection and experimental study on conflict intensity estimation in continuous time, and (iii) dynamic modeling and prediction of continuous-time human affect and behavior. The rest of this thesis is structured as follows.

Chapter 2. A review of existing machine learning approaches to *static* face analysis tasks as well as *dynamic* behavior and affect analysis, which are the core application domains with which this thesis deals. For both domains, the focus of this overview is placed on previous works that methodologically stand closer to the models proposed in this thesis. Databases of dyadic or multi-party interactions suitable for automatic analysis of social attitudes are also outlined in a separate section. In every section of this chapter, previous works are related and contrasted to the proposed methods and the contributions of the latter are described in detail.

Chapter 3. The proposal of the *Discriminant Incoherent Component Analysis (DICA)* and its experimental evaluation on static face analysis tasks.

Chapter 4. The release of the *Conflict Escalation Resolution (CONFER) Database* and the first experimental study on continuous-time estimation of real-valued conflict intensity in naturalistic conversations.

Chapter 5. The proposal of the framework for *Dynamic Behavior Analysis via Structured Rank Minimization* and its experimental evaluation on continuous-time prediction of dimensional affect and behavior as well as multi-object tracking by detection.

Chapter 6. A summary and discussion of the outcomes of this thesis accompanied by insights for future work.

The work presented in this thesis has resulted in the following list of publications.

[**1**] **C. Georgakis**, Y. Panagakis, and M. Pantic. Discriminant Incoherent Component Analysis. *IEEE Transactions on Image Processing (T-IP)*, 25(5): pp. 2021–2034, May 2016. (Chapter 3)

[**2**] **C. Georgakis**, Y. Panagakis, S. Zafeiriou, and M. Pantic. The Conflict Escalation Resolution (CONFER) Database. *Image and Vision Computing (IMAVIS), Special Issue on Multimodal Sentiment Analysis and Mining in the Wild Images*, 2017 (accepted for publication). (Chapter 4)

[**3**] **C. Georgakis**, Y. Panagakis, and M. Pantic. Dynamic Behavior Analysis via Structured Rank Minimization. *International Journal of Computer Vision (IJCV), Special Issue on Looking at People*, 2017 (accepted for publication). (Chapter 5)

Chapter 2

Related Work

Contents

In this chapter, we provide an outline of existing machine learning approaches to *static* face analysis tasks as well as *dynamic* behavior and affect analysis, which are the core application domains with which this thesis deals. For both domains, the focus of this overview is placed on previous works that methodologically stand closer to the models proposed in this thesis. For static face analysis, approaches that build upon *subspace learning* and *dictionary learning* for face recognition, facial expression recognition and action unit (AU) detection in static face imagery are outlined. For dynamic behavior and affect analysis, approaches that employ *linear dynamical system learning via structured rank minimization* to learn and predict the temporal dependencies in the evolution of *continuous-time and dimensional* behavior and affect are reviewed. Finally, in the last section of this chapter we provide an outline of existing databases containing audio and/or visual data from dyadic or multi-party social interactions and, as such, are suitable for the investigation of social attitudes such as (dis)agreement and conflict escalation/resolution.

2.1 Static Face and Facial Expression Recognition

Face analysis has been an active research topic over the last thirty years. Human face is a rich source of information consisting of several components which are related to attributes associated with facial identity, emotional expression and activation of atomic facial muscle actions named Action Units (AUs). These components are characterized by specific structures which can assist the semantic interpretation of content in the visual stream. For instance, facial expressions manifest themselves through *sparse* non-rigid deformations occurring in certain face regions [162, 163], while images depicting the neutral face of the same person are expected to be highly correlated and thus drawn from a *low-rank* subspace. Consequently, the extraction of such features of low-complexity (i.e., exhibiting low-rank or sparse structure) is essential for accurate face and expression recognition.

Machine learning systems for static face analysis are usually composed of a four-step framework. In the first step, the face is located in the image region (*face detection*). In the second step, a set of fiducial facial points is used to register the face image, i.e., to remove head pose variation by globally aligning the face images to a frontal reference 'mean face'. In the third step, a collection of measurements corresponding to descriptive shape- and/or appearance-based facial features is obtained from each image (*feature extraction*). Finally, In the fourth step, a classifier is trained to assign to each probe (test) image a label associated with a person's identity, basic emotion (e.g., happiness, surprise, fear) or presence/absence of AUs (*classification*).

Face recognition has been a classical topic of research within the image analysis, computer vision and pattern recognition communities for over 30 years, with diverse applications ranging from surveillance, biometrics and law enforcement to context-aware multimedia environments, computer entertainment and online image search, to mention but a few. In the first comprehensive survey on classic face recognition algorithms in [259], machine-based face recognition is defined as follows: "*given still or video images of a scene (probe set), identify or verify one or more persons in the scene using a stored database of faces (gallery set)*". On the other hand, *facial expressions* have been described at mainly two different levels, following either a "message judgement" or a "sign judgement" approach [167, 163, 162]. The message judgement approach models the face as a single entity and classifies observed facial expressions in terms of a set of universal prototypic emotions [84] (e.g., in terms of six basic emotions proposed by Ekman [61]). The sign judgement approach classifies observed facial expressions following a componential perception model, that is, in terms of facial muscle activations (AUs) that produced the observed expression [167, 163]. These atomic facial actions correspond to all visually discernible facial movements and can be measured according to the facial action coding system (FACS) [61]. Examples of facial action units and combinations of them are illustrated in Fig. 2.1.

In this section, existing approaches to machine analysis of facial identity and expression in

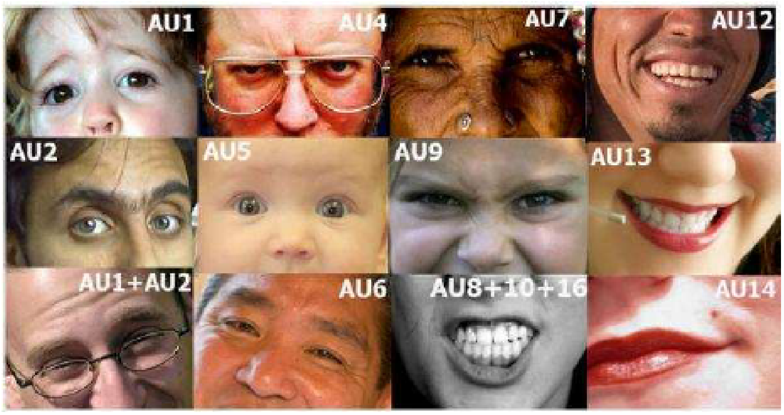


Figure 2.1: Examples of facial action units and their combinations (figure from [164]). 'AU1': Inner Brow Raiser, 'AU2': Outer Brow Raiser, 'AU4': Brow Lowerer, 'AU5': Upper Lid Raiser, 'AU6': Cheek Raiser, 'AU7': Lid Tightener, 'AU8': Lips Toward Each Other, 'AU9': Nose Wrinkler, 'AU10': Upper Lip Raiser, 'AU12': Lip Corner Puller, 'AU13': Cheek Puffer, 'AU14': Dimpler, 'AU16': Lower Lip Depressor

static face imagery are briefly outlined. In particular, attention of this overview is directed towards subspace and dictionary learning methods that bear a greater degree of similarity in terms of methodology with the method proposed in Chapter 3. Robust methods that fall in each of these two categories are outlined separately.

2.1.1 Subspace Learning Methods

Subspace learning methods for static face representation extract discriminative features from the whole face region by means of linear or non-linear projections based on the assumption that the high-dimensional observed faces live in a low-dimensional space. These methods place emphasis on the dimensionality reduction and feature extraction sub-tasks of machine face analysis. The derived representations are usually combined with generic classifiers, such as Nearest Neighbor (NN), Support Vector Machines (SVM) and Bayesian classifiers, to recognize facial identity and expression in 'unseen' test data.

Popular *linear* subspace learning methods that fall in this category include Principal Component Analysis (PCA) [98], Linear Discriminant Analysis (LDA) [16] and Locality Preserving Projections (LPP) [152]. PCA extracts mutually orthogonal basis functions that capture the directions of maximum variance in the face data, and has been one of the most popular dimensionality reduction techniques for face analysis. LDA is a supervised learning method that searches for the project axes on which the within-class scatter of data points is minimized while the between-class scatter is maximized. data points of different classes are far from each other while requiring data points of the same class to be close to each other. Unlike these two methods, LPP has the advantage of capturing the non-linear local structure of the image samples by means of an adjacency graph. Eigenfaces [214], Fisherfaces [16], and Laplacianfaces [88] have utilized PCA, LDA and LPP, respectively, along with a NN classifier for face recognition. Shan et al. [199] have conducted a systematic study on facial expression recognition with linear subspace learning methods.

Non-linear subspace-based methods have been also applied to encode the non-linearity and higher-order statistics of facial attributes. In this category fall *kernel-based extensions* of the aforementioned linear methods such as Kernel PCA and Kernel LDA [250]. These methods operate by non-linearly mapping the face data to a high-dimensional feature space, where the face manifold is linearized and simplified, and subsequently applying the desired linear projection for feature extraction. Although the kernel-based extensions have been shown to outperform their linear counterparts on various tasks, they can be rather problematic since (i) they require tuning of many design parameters, (ii) they often lead to overfitting, and (iii) they are computationally expensive [129]. Another family of non-linear subspace-based dimensionality reduction techniques includes methods such as Locally Linear Embedding [193, 186] and Isomap [13] as well as extensions of them [87, 184]. Common drawbacks of these methods is that it is unclear how to properly select their hyperparameters. Interestingly, Yan et al. [245] have recently shown that several linear and non-linear dimensionality reduction algorithms (e.g., PCA, LDA, ISOMAP, LLE) can be unified within a common framework called graph-embedding, which utilizes a graph similarity matrix to enforce the desired statistical or geometric properties of the data.

Despite their merits and proven efficiency for face recognition [85] and facial expression recognition [199] on benchmarks collected (mostly) under controlled conditions, the aforementioned holistic subspace-based approaches have been shown to be extremely susceptible to even slight local variations due to misalignment, pose, illumination or minor occlusions (e.g., eye blinks, slightly open mouths) [85, 230]. *Local* approaches that extract appearance descriptors such as Local Binary Patterns (LBP) and Gabor wavelets from local face sub-regions furnish a certain degree of stability against small local variations. These descriptors have been extensively employed for facial expression recognition and AU detection [192, 132, 198, 95], while their enhanced invariance properties have been also exploited by the face recognition community [247, 200]. However, these approaches are still susceptible to large, non-localized, non-uniform variations in the data, while they naturally inherit the drawbacks of the respective descriptors (e.g., sensitivity to design parameters for LBP and high dimensionality and computation load for Gabor features).

Robust Approaches. A fundamental limitation shared by the aforementioned subspace learning techniques (both linear and non-linear) is that they are not robust to gross, non-Gaussian noise, commonly referred to as 'outliers'. rely on least-squares (ℓ_2-norm) minimization whose solution can be arbitrarily skewed from the desired solution in the presence of statistically outlying measurements [52, 55, 111, 86, 90].

One of the earliest robust extensions to PCA is the Robust Subspace Learning (RSL) framework introduced within the computer vision community by De la Torre et al. in [52], who replace the least-square metric, that is, the ℓ_2-norm with a robust energy function. Other studies, namely R_1-PCA [55] and PCA-L_1 [111], have replaced the ℓ_2-norm with the ℓ_1-norm, while HQ-PCA employs an information theory-inspired metric to enhance robustness of PCA against outliers. While all the aforementioned robust extensions to PCA operate on the pixel intensity domain, the Image Gradient Orientations (IGO) subspace learning method proposed by Tzimiropoulos et al. [216] operates on the domain of gradient orientations. Also, the ℓ_2-norm based linear correlation of pixel intensities is replaced in the proposed objective function with a cosine-based distance measure that can robustly measure visual similarity in the presence of outliers.

The most successful and widely used robust extension to PCA comes from the compressed sensing literature and is Robust Principal Component Analysis (RPCA) by Candés et al. [32]. Motivated by recent advances in rank minimization and convex optimization [69], the authors in [32] introduce a model which represents the data as superposition of a low-rank matrix and a sparse matrix accounting for outliers of arbitrarily large magnitude. The low-rank and sparsity constraints are enforced by means of the nuclear norm [69] and ℓ_1-norm [60], respectively, thus leading to a a convex program called Principal Component Pursuit. The authors also show that the proposed model is theoretically guaranteed to exactly recover both low-rank and sparse components under some suitable incoherence assumptions. The latter dictate that the singular vectors of the low-rank component be reasonably spread-out, that is, not spiky, and the sparsity pattern be uniform to ensure identifiability of the proposed model.

In the context of face recognition, RPCA has been widely employed, mostly as a pre-processing step to derive 'clean' dictionaries from grossly corrupted training data, which are subsequently utilized for Sparse Representation Classification (SRC) [243] (see section 2.1.2). Recent works on robust facial expression recognition have also employed RPCA for occlusion region reconstruction [133] RPCA has been also used for robust facial expression recognition, mainly as pre-processing step for occlusion removal [96] or occlusion region reconstruction [133]. Recently, Wang et al. [234] have utilized the sparse error component extracted by RPCA to encode subtle facial motion and yield a more localized and identity-free representation for recognition of facial micro-expressions. Finally, a few works have combined RPCA and SRC for facial expression recognition and facial action unit detection and intensity estimation (see section 2.1.2).

2.1.2 Dictionary-based methods

Most of the aforementioned subspace-based methods employ a dimensionality reduction approach that primarily targets the reconstruction and denoising ability of the desired feature space rather than the discriminative power of it from a purely classification standpoint. Conforming to the growing belief that the feature dimensionality and the classifier is what really matters for face recognition [243], more recent works focus more on the classification sub-task of the face analysis pipeline. These approaches represent a probe (test) face image with respect to a single or class-specific *dictionaries*, and classify it building on the assumption that faces belonging to the same subject reside in a low-dimensional linear subspace. Representative works that are subsumed under this category are the *Linear Regression-based Classification (LRC)* [142] and the *Sparse Representation-based Classification (SRC)* [243] frameworks.

In the LRC [142], face recognition is cast as a linear regression problem, where each test sample is represented as a linear combination of training images of each class. Classification is performed by following a *Nearest Subspace* (NS) approach, i.e., the test image is assigned to the subject class that achieves the minimum reconstruction error in the least-squares (ℓ_2-norm) sense. Despite its solid theoretical foundation and computational efficiency, the LRC is not equipped with robustness to illumination, random pixel corruption and its performance is brittle in the presence of the *small sample set* problem.

The SRC [243] has been a face recognition breakthrough in recent years and has largely boosted the research of *sparsity-based recognition*. Unlike the LRC, the SRC method represents the test sample as a sparse linear combination of all the training samples, thus employing a single overcomplete dictionary rather than class-specific dictionaries. Motivated by recent advances in compressed sensing and sparse representation [60], the SRC finds this sparse representation via ℓ_1-minimization. Classification is then performed by using the minimum coding error in the least-squares sense as the decision rule, similarly to the LRC. The SRC algorithm is summarized in Algorithm 1.

Although it has been initially designed for face recognition, the SRC has been shown to be efficient also for recognition of emotional expression and detection of facial action

Algorithm 1 The SRC Algorithm

```
Input: Data: training set $\mathbf{X} = [\mathbf{X_1 X_2} \dots \mathbf{X_{n_c}}] \in \mathbb{R}^{d \times N}$, where $n_c$ is the number of classes and $\mathbf{X_i}$ is the dataset of the $i^{\text{th}}$ class, query image $\mathbf{y} \in \mathbb{R}^{N \times 1}$.
Parameters: $\lambda_{Lasso}$.
Normalize each column of $\mathbf{X}$ to unit $\ell_2$-norm.
Find the sparse representation $\hat{\boldsymbol{\alpha}}$ via $\ell_1$-regularized minimization
$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2 + \lambda_{Lasso}\|\boldsymbol{\alpha}\|_1$.
for $i = 1 : n_c$ do
    Compute the coding residuals $err(i) = \|\mathbf{y} - \mathbf{X_i}\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}}\|$,
    where $\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}}$ is the coding coefficient vector associated with class $i$.
end for
Output: identity($\mathbf{y}$) $= i^* \leftarrow \arg\min_{i \in \{1,2,\dots,n_c\}} err(i)$.
```

units [254, 183, 132, 251, 114, 212]. Zafeiriou and Petrou [254] argue that the application of the SRC is not directly applicable for facial expression recognition as the resulting sparse representation is not meaningful for the task. To alleviate the facial identity effect in the result, they use a dictionary of difference images, that is, they subtract the neutral image from each fully expressive image, for SRC-based expression recognition. Ptucha et al. [183] utilize LPP dimensionality reduction and a statistical mixture model to enhance the discriminative power of SRC classification for facial expression recognition. Mahoor et al. [132] recognize AU combinations in facial images via SRC by constructing a dictionary of mean Gabor features of AU combinations and random features. Ying et al. [251], classify expressions based on the fusion of two SRC-based classifiers, one operating on raw intensities and the other on LBP features. In [114], video-based expression recognition on the CK+ dataset is addressed by sparse representation with respect to spatially localized facial motion flow dictionaries. Recently, Taheri et al. [212] have proposed an efficient dictionary-based approach for facial expression analysis by decomposing expressions in terms of AUs. In particular, the proposed structure-preserving sparse coding framework uses an AU dictionary, constructed by incorporating domain experts' knowledge of AUs or directly learned from the data, for SRC-based AU detection and expression recognition.

Robust approaches. Overall, the *Sparse Representation-based Classification (SRC)* is expected to work well in real-world settings only under two conditions, namely, the availability of (i) a sufficiently large number of samples per class and (ii) a well-controlled, aligned and uncorrupted training set comprehensively covering a wide range of imaging conditions (e.g., in illumination, expression). The Extended SRC (SRC) [243] is proposed to tackle the former problem of *small sample set* by applying an auxiliary intraclass variant dictionary computed from a sufficient number of generic faces. On the other hand, many extensions or alternatives have been proposed to robustify the standard SRC scheme [248, 237, 260, 64, 249, 230, 154, 97], that is, to offer robustness to gross, non-Gaussian outliers including cast shadows, occlusion, or disguise, to name but a few. Most of these works employ robust features, such as those reported in Section 2.1.1, as opposed to pixel intensities to construct the dictionary that is subsequently used for SRC classification. A representative example of this family is the work of Yang et al. [248], who utilize local Gabor features to represent both the dictionary of gallery faces and a learned occlusion dictionary. CP Wei et al. [237] utilize RPCA [32] to pre-process the training images of each class so that the resulting low-rank sub-dictionaries are more discriminative and devoid of gross, non-Gaussian corruptions. However, it is argued that the dictionary derived by this RPCA+SRC scheme might have reduced discriminating ability for the face recognition task, given that the class-specific low-rank components might share spatially correlated features (e.g., locations of the eyes, nose, etc.). To alleviate this problem, the authors propose to add a regularization term to the objective function of RPCA in order to enforce *structural incoherence* among the low-rank matrices of different classes. Experimental results show that the proposed scheme, called *Low-Rank Matrix Recovery with Structural Incoherence (LRSI)*, is more robust to severe illumination variations, contiguous occlusion, and random pixel noise corruptions, than both the standard SRC and RPCA+SRC methods as well as other classifiers. The assets of the combination of robust subspace learning and sparse representation have been also exploited for facial expression analysis. In [140], RPCA is utilized to decompose expression from facial identity and subsequently the intensity of multiple AUs is jointly estimated from a regression model learned through dictionary learning and sparse representation. Finally, the authors in [260] combine the merits of RPCA and LDA by adding discriminant (label) information in the low-rank matrix recovery scheme through Fisher discriminant regularization. The class-specific bases learned from the proposed *Fisher Discrimination based Low Rank matrix recovery (FDLR)* are subsequently fed into the SRC classification scheme.

Other robust alternatives to SRC directly modify the optimization problem used to find the sparse representation (e.g., [64, 249, 230]). Nonetheless, none of these approaches is robust to contiguous occlusion. A partial remedy to this problem is furnished by the work in [154], where the occluded part of the test image is represented as a sparse linear combination of prototype occlusion atoms from a learned occlusion dictionary. Another efficient alternative to SRC is the recently proposed Sparse- and Dense-hybrid Representation (SDR) framework [97]. The proposed method represents a test image by a sparse combination of a class-specific dictionary and a dense combination of a common intra-class variation dictionary plus a term accounting for gross, non-Gaussian corruptions. Theoretically, the SDR is designed to alleviate both the *small sample set* and *corrupted training set* limitations of SRC, while experimentally it is shown to outperform other robust SRC-based approaches in face recognition.

Face and facial expression recognition, despite being two intertwined tasks within the context of face analysis, have hitherto been targeted jointly by just a few works. Vasilescu and Terzopoulos [224] employ an extension of Singular Value Decomposition (SVD) to tensors to uncover subspaces generating different faces, expressions, viewpoints, and illuminations. Another SVD-based work is [233], where the proposed Higher-Order SVD is used to learn the mapping between persons and expressions, which is subsequently utilized to perform facial expression decomposition. Recently, Taheri et al. [211] have adopted a Dictionary-based Component Separation (DCS) algorithm for joint face and expression recognition. The main assumption of the proposed framework is that an expressive face can be represented as a superposition of a neutral, expression-less face with an expression component. First, RPCA [32] is used to construct two data-driven dictionaries, one from low-rank (neutral) components and the other from sparse (expressive) components, and K-SVD [1] is used to refine the dictionaries. A test face image is then decomposed into a neutral component and an expression component, with the former having sparse representation in the neutral dictionary and the latter being sparsely represented using the expression dictionary. The separated components are sparsely decomposed using dictionaries whose grouping structures are enforced into the sparse decomposition results. The updated sparse codes of the neutral and expression component are then used for face and expression recognition, respectively.

2.1.3 Connection to our work

The Discriminant Incoherent Component Analysis (DICA) that we propose in Chapter 3 for face analysis is related to the methods reviewed in Sections 2.1.1 and 2.1.2. In what follows, we relate and contrast the framework of the DICA with the existing approaches by highlighting similarities and differences in the optimization problems and methodologies adopted for these methods.

The fundamental constraint of the majority of the above mentioned methods in for face analysis is that the training data are often assumed to be noise-free. That is, they are collected under well controlled conditions in terms of illumination and pose variations and they do not contain occlusions or disguise. Consequently, the aforementioned methods are not applicable in practical scenarios when both training and test data are contaminated by gross non-Gaussian noise and corruptions (e.g., occlusions and disguise). Instead, the DICA proposed in Chapter 3 of this thesis decomposes training facial images into a superposition of class-specific structured and mutually incoherent components accounting for identity, emotional expression or AUs in the presence of gross but sparse non-Gaussian corruptions. In this light, we stress that the strength of the morphological decomposition of expressive images yielded by the DICA goes beyond a simplistic combination of robust subspace learning-based preprocessing of the data or direct robustification of the SRC framework. As opposed to the DICA, which is a supervised learning algorithm, the robust subspace learning approaches outlined in Sections 2.1.1, such as RPCA [32], ℓ_1-norm based PCA variants [55, 111, 86] or IGO-PCA [216], completely neglect discriminant information in the recovery of the low-dimensional embeddings. Hence, simply applying them for gross corruption removal prior to SRC classification is not deemed suitable for classification due to the reduced discriminative power of the derived representation, as is the case with the RPCA+SRC scheme [237]. The same limitation is shared by model-based robustified alternatives to SRC such as [64, 249, 154, 97] which also completely disregard label information in the construction of the occlusion-robust dictionaries. The only frameworks that employ discriminant information to guide the outlier removal in the dictionary learning stage are the the LRSI [237] and FDLR [260] frameworks. While the LRSI promotes mutual incoherence among the class-specific dictionaries similarly to the DICA, the FDLR utilizes Fisher discriminant regularization which, by construction, cannot guarantee mutual orthogonality of the derived dictionaries. However, both LRSI and FDLR are computationally expensive, since they require solving as many optimization problems as the number of classes, as opposed to the DICA that solves a single optimization problem for the whole training dictionary. Aside from the aforementioned advantages, the DICA is not restricted to a RPCA-like decomposition, but instead it can generalize to the extraction of low-complexity class-specific components of any desirable structure (e.g., low-rank, sparsity, total variation) by enforcing suitable structure-inducing norms for the respective components.

Another major drawback of the aforementioned machine learning techniques to face analysis is that, unlike the DICA, they cannot accomplish joint modeling and classification of multiple facial attributes (e.g., identity, expression, AUs) within a multi-label setting. The Dictionary-based Component Separation (DCS) algorithm proposed by Taheri et al. [211] is the only work in the literature that is designed explicitly to address joint face and expression recognition by learning one neutral and one expression dictionary, respectively. However, the DCS framework relies on multiple runs of RPCA and the iterative K-SVD algorithm for the extraction of the initial dictionaries as well as a label-driven post-processing of the sparse representation codes through an additional group Lasso ℓ_1-norm minimization problem. Instead, the proposed DICA performs joint supervised learning of class-specific, noise-free and mutually incoherent structured components by means of a single matrix decomposition framework.

Finally, despite the widespread use of Convolutional Neural Networks (CNNs) for representation learning for large-scale object detection and recognition [110], there has been no consensus in the literature as to how CNNs can be interpretably and efficiently employed for multi-label classification. Recently, Ghosh et al. [77] propose a multi-label softmax classification loss to tailor CNN training to this scenario and apply it for action unit (AU) detection in facial images. However, their method requires large training effort in augmenting the dataset and making it balanced with respect to multiple labels, while their regularization is still based on ℓ_2-norm minimization which cannot handle sparse corruptions. Most importantly, while sparse and group-sparse constraints on the weights of deep neural networks have been recently employed [194], there has been no previous work proposing deep CNN architectures that could impose such norms directly on the domain of learned representations guided by categorical labels, with that remaining an interesting research direction to be pursued.

Overall, the advantages of the proposed DICA over the existing approaches to static face analysis are as follows.

- The DICA jointly learns low-complexity structures (e.g., low-rank, sparsity) associated with facial attributes from a single matrix decomposition of the training data.

- The derived components that are discriminant (label-driven) and mutually incoherent among different classes.

- The DICA is robust to gross, sparse non-Gaussian noise and corruptions (e.g., occlusions and disguise).

- The DICA lends itself to Sparse Representation Classification (SRC) of 'unseen' test samples in terms of any number or type of interrelated labelled attributes (e.g., identity, expression, AUs) in a multi-label classification setting (e.g. joint face and expression recognition or multiple AU detection).

2.2 Dynamic Behavior and Affect Modeling

Traditionally, research in behavior and affect analysis has focused on recognizing behavioral cues such as smiles, head nods, and laughter [53, 103, 127], pre-defined posed hand gestures (e.g., hand-waving, hand-clapping) [59, 151] or discrete, basic emotional states (e.g., happiness, sadness) [166, 43, 121] mainly from posed data acquired in laboratory settings. However, these models are deemed unrealistic as they are unable to capture the temporal evolution of non-basic, possibly atypical, behaviors and subtle affective states exhibited by humans in naturalistic social settings. In order to accommodate such behaviors and subtle expressions, *continuous-time* and *dimensional* descriptions of human behavior and affect have been recently employed [83, 81, 161, 229]. These approaches are outlined below in Section 2.2.1, with particular focus on previous works that address valence-arousal estimation and estimation of the intensity of social attitudes (e.g., (dis)agreement, conflict).

While dimensional representation and continuous-time prediction of affective states is gaining increasing popularity in automatic affect analysis [83], this approach has not been followed to an equal extent in social signal processing and behavior analysis [161]. Overall, it is considerably hard to obtain *continuous-time* and *dimensional* annotations of human affect and social behavior [24], while for such tasks it is essential to employ multiple annotators and an efficient technique for merging the multiple annotations in a single ground truth annotation [146]. However, the main reason inhibiting the deployment of *continuous* and *dimensional* approaches to behavior and affect analysis is the lack machine learning techniques that are able to sufficiently address the problem, i.e., regression models capable of capturing both discriminative latent structure and temporal dependencies in the data [24]. This problem is addressed by the method proposed in Chapter 5 in this thesis, in which the temporal dependencies characterizing the evolution of dynamic behavior and affect are explicitly modeled through a Linear Dynamical System (LDS) representation based on structured rank minimization. An overview of related structured rank minimization approaches to system learning and their applications in dynamic behavior and affect modeling is provided in Section 2.2.2.

2.2.1 Dimensional and Continuous Representations of Affect and Social Behavior

In this section, we provide a separate overview of existing works that have employed dimensional and continuous-time representations of affect and social behaviors, respectively.

Affect Analysis. Within the affective computing literature, we have witnessed a major shift from categorical descriptions of facial emotion, that is, affective non-verbal expressions in terms of basic emotion categories, towards dimensional descriptions of affect, where affective states are characterized in terms of latent dimensions that are related to each other in a systematic manner [83, 81]. Two dimensions have been shown to be sufficient for capturing most of the affective variability: valence and arousal (V-A), signifying respectively, how positive/negative and active/inactive an emotional state is [113]. This model allows the representation of emotion intensity on a continuous scale as well as similarity and contrast

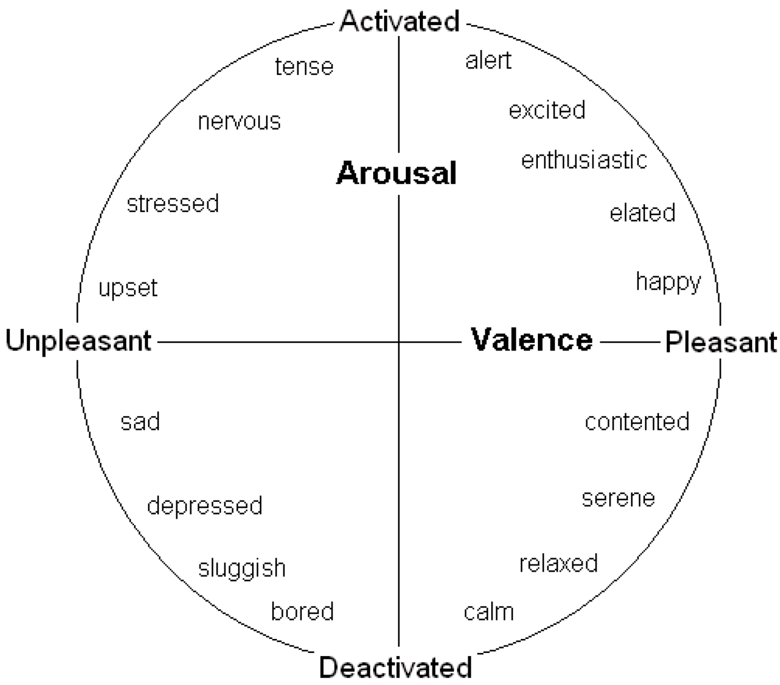


Figure 2.2: The Valence-Arousal Dimensional Model for Emotion (figure from [213]).

between various emotion categories, while it facilitates the analysis of emotion transitions between moderate and naturalistic affective states. The Valence-Arousal dimensional model is graphically illustrated in Fig. 2.2. The survey paper of Gunes and Schuller [83] provides a comprehensive review of existing methodologies that employ the V-A affect representation, categorized with respect to the modality employed. Most of the existing automated approaches to Valence-Arousal (V-A) analysis have been limited to the use of audio cues only [83]. Although the relation of affective dimensions (mostly arousal) to certain acoustic features has been better documented as compared to visual cues, yet there has been evidence that also visual signals (e.g., facial expressions, head shakes, nods) are informative of the V-A dimensions [47, 164]. Such findings have motivated the exploitation of visual features, such as facial expression cues and shoulder movements, in either isolation or combination with audio features, for dimensional affect analysis. Representative examples of this line of research are the works of [81, 147] and [101].

Most of the traditional automated approaches to dimensional affect analysis have compromised to solving a two-class or four-class classification problem, i.e., binary classification with respect to each dimension or classification into the quadrants of the 2D valence-arousal space [83]. In recent years, *continuous-time* estimation of dimensional affect has attracted a lot of interest in the machine learning community. This trend is motivated by recent evidence in psychology [6] and behavioral computing [163, 255] indicating that capturing temporal dynamics and micro-patterns is essential for the human and machine perception of an affective display, especially when it comes to spontaneous and subtle expressions. For instance, spontaneous (Duchenne) smiles are smaller in amplitude, longer in total duration, and slower in onset and offset time than posed smiles (e.g., a polite smile) [62]. Classifiers commonly employed for continuous-time estimation of dimensional affect are Support Vector Regression [205, 82, 145], Relevance Vector Machines (RVM) [147, 101], Long-Short Term Memory (LSTM) Neural Networks [145] as well as Conditional Random Fields [141] and Support Vector Machines [46] on quantized emotion labels [241]. The superior performance yielded by LSTMs over SVR [145, 241] and CRF over SVM [241] on naturalistic expression benchmarks provide strong evidence that temporal classifiers capable of encoding long-range temporal dependencies are more suitable for continuous-time modeling of affect dimensions than frame-based classifiers or regressors. It is worth noting that all the aforementioned works treat valence and arousal independently, which is rather an unorthodox approach given that these two affect dimensions have shown to exhibit high correlation [164]. One exception is the work in [147], in which Output-Associative (OA) RVM are used to model cross-dimensional output dependencies subsequent to a initial layer of regressors. Recently, an extension of the traditional CRF to the case of continuous (real-valued) output, called Continuous Conditional Random Fields (CCRF), is proposed in [14] and shown to outperform SVR on dimensional affect recognition on a subset of the SEMAINE Dataset [138]. A context-aware variant is also proposed in the same work to exploit the non-orthogonality of emotion dimensions. However, none of these models includes latent variables which are deemed essential for capturing fine-grain dynamics in the evolution of affect manifestations. Overall, researchers in the field have not reached consensus on which classifier is better suited for analysis of continuous affective dimensions [83].

Social Behavior Analysis. Social Signal Processing (SSP) is an emerging technological domain that aims providing computers with the ability to sense and understand human social signals [174]. In spite of recent advances in social signal processing [160, 228, 161] and machine analysis of relevant behavioral cues such as blinks, smiles, head nods, laughter, and similar [36, 53, 103, 127, 165, 255, 83], the research in machine analysis and understanding of more complex human social behaviors such as interest, politeness, flirting, (dis)agreement, and conflict escalation/resolution is still limited [27, 25, 106, 107, 228]. As stated eloquently in [161], the journey towards artificial social intelligence and socially-aware computing is still long.

While the suitability of the dimensional characterizations of affect (e.g., valence-arousal) has been extensively demonstrated by experimental studies, appropriate mappings that would render it applicable for recognition of social emotions (e.g., empathy, envy, admiration, compassion) or social attitudes (e.g., conflict, agreement/disagreement) in real-world social settings like patient-doctor discussions, talk-shows, job interviews has not been investigated yet. On the other hand, analogous one-dimensional continuous-scale characterizations of social behaviors such as interest, politeness, flirting, (dis)agreement, and conflict, are rarely adopted by social signal processing methodologies. Most of the existing automated approaches to social behavior analysis target analysis of social behaviors within a classification framework in which either pre-segmented sequences are assigned to a single label (e.g., agreement/disagreement [27, 26] or conflict/non-conflict [106, 107]) or individual frames of the sequence are labelled as quantized levels of behavior (e.g., agreement or conflict [108]). As mentioned above, the non-frequent use of dimensional representation for social signals and behaviors analysis is partially due to the lack of annotated data and suitable annotation tools and, on the other hand, due to the lack of efficient regressors for the tasks at hand [24]. Only a handful of works depart from this practice by using real-valued annotations such as [148, 155], [155] and [102, 101] that deal with estimation of interest, conflict and pain, respectively, on a continuous scale.

Despite evidence suggesting that temporal dynamics of social behavioural cues (i.e., their timing, co-occurrence, speed, etc.) are crucial for the interpretation of the observed social behaviour [228], *continuous-time* modeling of social signals and behaviors is much less investigated within the social signal processing community, as compared to continuous-time analysis of affect. This directly stems from the fact that dimensional characterizations of social signals, that would facilitate the modeling of temporal dynamics of the social signals and related behaviors, are scarcely adopted in the social signal processing frameworks. Machine analysis of social signals such as social dominance [93], codes (e.g., acceptance and blame) [23] and personality traits [177], has mainly addressed the relevant tasks within a SVM-based classification framework. Social emotions such as empathy, envy and admiration, have been mainly recognized based on existing affective computing methodologies based on the emotional expressions of a single subject rather than around the dynamics of the emotional feedback exchange between two subjects [161]. On the other hand, automatic analysis of social attitudes such as conflict, which is defined as a high level of disagreement, has been approached as a binary or multi-class sequence classification problem on pre-segmented conflict/non-conflict episodes by means of audio features and static classifiers such as SVM [106, 107] and Gaussian Processes [108] (see Section 2.3).

Only recently, temporal classifiers have been employed within the social signal processing community to target analysis of social attitudes, mainly for (dis)agreement detection. [26, 108]. For instance, El Kaliouby and Robinson [63] employ Hidden Markov Models (HMM) in sliding windows to detect behavioral cues such as head motion, facial action units and mouth actions and subsequently Dynamic Bayesian Networks (DBN) [63] to recognize complex mental states (6 classes: agreeing, concentrating, disagreeing, interested, thinking and unsure). Bousmalis et al. [27] employ HMM, Support Vector Machines (SVM) and Hidden Conditional Random Fields (HCRF) or (dis)agreement recognition. In this study, various non-verbal audiovisual cues such as head motions, hand and shoulders movements and auditory features (fundamental frequency and energy) are employed and it is shown that HCRF outperform SVM and HMMs. In [28], a non-parametric variant of HCRF, termed Infinite Hidden Conditional Random Fields (IHCRFs) is proposed which perform equally well with HCRF. IHCRF have the advantage that they can learn an appropriate latent structure of the model without specifying a priori the appropriate number of hidden clusters of cues. Although IHCRF have been shown to be less prone to overfitting than the standard HCRF, they still require the tuning of many parameters and they do not scale well to large datasets, since computationally intensive inference is needed due to the large number of hidden states involved. Overall, while these approaches employ classification schemes that can capture statistical regularities of social behavioral cues in time, they still approach the automatic analysis of the social phenomena within a binary or multi-class classification framework, i.e., by assigning a discrete or quantized label to the entire test sequence. This limitation is mainly due to the HCRF model and its variants having been designed for sequence classification rather than continuous-time regression. One model that overcomes this problem is the one proposed in [105]. However, it is unable to capture latent structure which is essential for modeling complex behaviors.

Limitations. Despite their merits, dynamics classifiers such as LSTMs, CRF, HCRF and IHCRF used by the aforementioned approaches to dimensional affect and behavior analysis, come with a number of limitations, namely (i) they rely on large sets of training data to learn a large number of parameters (e.g. LSTMs), (ii) they involve a large number of hyperparameters and are prone to overfitting (e.g. CRF and HCRF), (iii) they do not scale well to large datasets as they involve computationally intensive inference techniques (e.g., IHCRF). Notably, the CRF and HCRF models cannot cope with continuous-time, frame-by-frame regression, as they are designed to tackle sequence classification tasks. Also, these methods do not model in an explicit way the joint temporal evolution of affective or behavioral characterizations and related features within a systematic dynamic framework learned from the data (e.g., an auto-regressive process). Hence, these models can be applied neither to learn prototypic manifestations of affect and behavior from data nor to measure the similarity between different behavior and affective displays. Such properties are extremely useful for several applications such as video indexing, skimming and summarization, that deal with human behavior and affect data. Most importantly, all these approaches are fragile in the presence of gross, non-Gaussian noise and incomplete data, which are abundant in real-world data.

In the following section, we provide a concise review of methods that explicitly model temporal dependencies in sequential data through learning a linear dynamical system generating the observations. These methods rely on rank minimization of structured matrices constructed from the data to estimate the system order, that is, the model complexity. The system parameters can easily be learned at a later stage by solving a system of linear equations. The predictive framework for dynamic behavior and affect analysis proposed in the Chapter 5 in this thesis has as its main building block a system learning algorithm, which employs a novel structured rank minimization method that is robust to grossly corrupted and incomplete data.

2.2.2 Linear Dynamical System Learning via Structured Rank Minimization

Dynamical systems are able to compactly model the temporal evolution of time-varying data. While the dynamic model can be considered known in some applications (e.g., Brownian dynamics in motion models), it is in general unknown and, hence, should be learned from the available sequential data.

Recent advances in systems theory [221, 70] have provided us with tools that enable us to reliably uncover linear temporal dependencies in observed sequential data, under the assumption that the latter have been generated by a Linear Time Invariant (LTI) system of low complexity, i.e., low order. Specifically, it has been shown that the rank of a matrix constructed from noiseless sequential observations according to a specific linear matrix structure equals the order of a state-space model used for the realization of the data as a LTI system [221] (see Section 5.2.2). The linear matrix structure used in this context is the Hankel structure which enforces constant entries along the skew diagonals [70]. On the other hand, the order of the system is the most crucial parameter for a linear dynamical system, since it captures its memory and is a measure of its complexity. Having estimated the order of the underlying LTI system generating the observations, the system parameters can then be easily obtained by solving a system of linear equations [221].

However, real-world data are inexact and thus Hankel matrices constructed from them are full-rank. Hence, a structured matrix rank minimization – henceforth called *structured rank* *minimization* – needs first to be solved in order to estimate the unknown system order and learn the temporal dependencies in the sequential data. A (Hankel)-structured rank minimization problem seeks a matrix which is as close as possible, in the least square sense, to the observed data and the rank of its associated Hankel matrix is minimal (see Chapter 5 for a detailed problem formulation and definitions). Since the original problem is combinatorial due to the discrete nature of the rank function, several approximations have been proposed in the literature. Fazel et al. [70] propose a convex approximation by employing the nuclear norm, which is the convex surrogate of the rank function [69]. Non-linear approximations based on the variational form of the nuclear norm have been also developed [203, 252]. Furthermore, to estimate the rank of an incomplete Hankel matrix (i.e., in the presence of missing data), the models in [134, 54, 11] have been also proposed. Representative structured rank minimization models along with the optimization problems that they solve are listed in Table 5.1. Detailed discussion regarding the limitations of these models, as compared to the proposed structured rank minimization model in this thesis, is provided in Chapter 5.

As shown in Section 2.1, minimizing the rank of a data matrix translates into uncovering linear relationships in noisy data and, as such, it is often used to derive discriminative low-dimensional representations for semantic analysis of static human behavior. For sequential data, a low-rank approximation of a Hankel matrix is intrinsically related to the assumption that the observed data are a trajectory of a low-complexity Linear Time-Invariant (LTI) system, as mentioned above. This well-established connection between LTI system learning and (Hankel)-structured matrix rank minimization has been utilized extensively in the fields of system analysis and control theory for *system identification and realization* and in finance for *time-series analysis and forecasting* [70].

In the last decade, linear dynamical system learning via structured rank minimization has been exploited to address computer vision problems. The modeling assumption on which these approaches are built is that smoothly-varying dynamic behavior phenomena can be postulated to be trajectories of a LTI system. Some works employ the dynamics-revealing Hankel matrices constructed from observations as features termed *hankelets* [118, 182, 181], or use the Singular Value Decomposition (SVD) and rank of Hankel matrices for feature extraction or event detection [117, 10, 209, 19]. Other approaches resort to low-rank Hankel matrix approximation and completion to learn the dynamics-related information and predict missing observations [56, 58, 57, 12, 54]. Applications of the aforementioned works include tracklet matching [56, 57, 54], multi-camera tracking [10], activity recognition [117, 118, 19, 182], emotion recognition [181], video inpainting [58], causality detection [12], and anomaly detection [209]. However, none of these works has approached learning of behavior dynamics based on continuous-time and dimensional annotations of behavior or affect and corresponding visual features. This is accomplished by the proposed predictive framework for dynamic behavior and affect analysis presented in Chapter 5.

2.2.3 Connection to our work

The dynamic behavior analysis framework that we propose in Chapter 5 performs dimensional and continuous behavior and affect prediction by learning a linear dynamical system generating the sequential observations via structured rank minimization. In what follows, we relate and contrast the proposed framework with the existing approaches to dimensional and continuous behavior and affect analysis reviewed in Section 2.2.1 and 2.2.2.

As mentioned above, despite recent advances in affect computing and social signal processing, the majority of existing methods to affect and behavior modeling employ categorical representations and/or rely on static, frame-based classifiers. Although this is partially due to the lack of annotated data and annotation tools, the lack of dimensional and continuous models is to a large extent attributed to the non-existence of machine learning regressors that can efficiently capture the latent temporal dependencies in the observed data. Two off-the-shelf neural and bayesian networks variants widely used for affect and social behavior modeling are LSTMs and HCRFs, respectively. The main limitation of LSTMs is that they require large amounts of training data to learn their parameters, while HCRFs are not suitable for continuous-time regression (see Section 2.2.1).

While LSTMs have been shown to be highly efficient for a variety of sequential learning tasks such as speech recognition and machine translation in the last five years [210] , yet they do not come without downsides. LSTM training involves multiple non-linear mappings in the flow of gradients among memory units which prohibits the interpretability of which is the ‘memory’ of the latent temporal process and which long-term temporal dependencies are actually being captured. The latter are not explicitly modeled and thus LSTMs cannot represent the learned temporal dependencies in a systemic formulation that would allow reproducibility and intra-sequence comparison of dynamics. This problem is alleviated by our method in Chapter 5, which explicitly encodes them by means of a generative auto-regressive model which can be exploited to represent, cluster, compare and contrast different displays of affect and behavior based on the learned dynamics. The recently proposed Variational Recurrent Neural Networks (VRNNs) [42] partially overcome the aforementioned drawback of LSTMs, by representing the hidden variable within a bayesian rather than deterministic framework. Specifically, in each time step variational autoencoders [109] are used for representation learning, while the recurrent units are controlled by a latent variable so that they are allowed to take an infinite number of states, unlike fixed state space models like HMMs [43]. While this variational bayes setup allows for a probabilistic representation of the temporal dynamics in terms of posterior distributions, this model is highly non-linear, involves computationally expensive inference and learning, requires a lot of training data and, most importantly, it has yet to be seen how this model would perform in high-dimensional data involving large amounts of unmodeled noise.

On the contrary, our method proposed in Chapter 5 requires tuning of a single parameter, it can take both convex and non-convex objective functions, does not impose any modeling assumptions (e.g., form of the priors) other than the auto-regressive assumption, and it can perform well when trained with small amounts of (possibly) corrupted data. The proposed predictive framework presented in Chapter 5 goes beyond an implicit encoding of the latent temporal dependencies in the observations. Rather, dynamics are modeled in an explicit way. Specifically, the modeling assumption here is that continuous-time annotations characterizing the temporal evolution of relevant behavior or affect are considered as outputs of a linear dynamical system, while features describing behavioral cues are deemed system inputs. Existing linear dynamical system learning approaches reviewed in Section 2.2.2 have mainly employed the dynamics-revealing properties of rank-deficient Hankel matrices constructed from data targeting primarily event detection and other computer vision applications based on matrix completion (e.g., tracklet matching, video inpainting). None of these works have formally modeled the temporal patterns in dimensional characterizations of behavior and affect through actually learning the linear dynamical system that generates these sequential data as its outputs.

Our work is the first machine learning approach to dynamic behavior and affect analysis in which annotations and features act as outputs and inputs, respectively, of a low-order linear dynamical system that models the latent temporal structure. By explicitly learning systems accounting for manifestations of dynamic phenomena in visual data, our method lends itself to various video analysis and computer vision applications that require measuring similarity between the portrayed actions or events, such as measuring behavior similarity, human action classification, video skimming and summarization, among others. The proposed framework can efficiently learn the system generating the data by employing a novel *ℓ_q-norm regularized (Hankel) structured Schatten-p norm minimization* problem solved by an efficient first-order algorithm. As opposed to existing structured rank minimization methods, the proposed method can handle both (partially) missing data and grossly corrupted observations.

Gross, non-Gaussian noise is frequent in real-world (visual) data due to pixel corruptions, partial image texture occlusions or feature extraction failure (e.g., incorrect object localization, tracking errors), while human assessments of behavior or affect may be unreliable mainly due to annotator subjectivity or adversarial annotators [146]. Most importantly, the proposed dynamic behavior and affect analysis framework departs from a practice commonly adopted in behavioral and affective computing, that is, to train machine learning algorithms by employing large sets of training data that comprehensively cover different subjects, contexts, interaction scenarios and recording conditions. Specifically, we demonstrate for the first time that complex human behavior and affect, manifested by a single person or group of interactants, can be learned and predicted based on a small amount of person(s)-specific observations, amounting to a duration of just a few seconds.

In terms of applications, the generalizability of the proposed framework is demonstrated by conducting experiments on 3 distinct dynamic behavior analysis tasks, namely (i) *conflict intensity prediction*, (ii) *prediction of valence and arousal*, and (iii) *tracklet matching*. In the first two tasks our method is assigned the task of learning an input/output dynamical system accounting for the sequential observations of features and annotations, while in the third task it acts as an unsupervised learning algorithm in distinguishing motion trajectories corresponding to different objects/persons. The visual modality, which is consistently disregarded by many existing approaches to the first two tasks, is employed in all three tasks. It is also worth noting that the presented experiments on conflict intensity prediction constitute the first approach to continuous-time and dimensional intepersonal conflict analysis.

Overall, the contributions of the dynamic behavior and affect analysis framework presented in Chapter 5 over existing approaches are as follows.

- The proposed framework explicitly learns a linear dynamical system generating the sequential observations of human affect or behavior
- Our system learning framework is the first that models. continuous-time characterization of behavior or affect as the output of a linear time-invariant system when behavioral cues act as the input. As such, it provides a generic learning method for continuous-time and dimensional modeling of behavior or affect.
- The main building block of the proposed framework is a novel structured rank minimization algorithm that can robustly learn the underlying dynamics from grossly corrupted and/or (partially) missing data.

- The proposed framework can accurately predict future behavior and affect based on just a few seconds of past person(s)-specific observations.
- Our method can be easily extended to learn prototypic behavior or affect patterns as well as directly compare behavioral or affective displays.

2.3 Databases of Dyadic or Multi-party Interactions Suitable for Automatic Analysis of Social Attitudes

Social signals and social behaviors are the expression of one's attitude towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioral cues including facial expressions, body postures, gestures, and vocal outbursts [228]. Social signals typically last for a short time (from milliseconds, e.g., turn-taking, to minutes, e.g., mirroring), compared to social behaviors that last longer (from seconds, e.g., agreement, to minutes, e.g., politeness, to hours or days, e.g., empathy) and are expressed as temporal patterns of non-verbal behavioral cues [161]. Since humans are predominantly social beings, the importance of social signals in everyday life situations is self-evident. Human social interactions are omnipresent in multimedia data (e.g., television programs, movies, etc.) and thus the automatic analysis and understanding of human social signals and social behaviors from audio-visual recordings is a cornerstone in the deployment of content-based multimedia indexing and retrieval, machine-mediated communication, state-of-the-art human-computer interfaces, to mention but a few.

Social attitudes can be defined as positive or negative evaluations of a person or a group of people and include cognitive elements like beliefs, opinions, and social emotions [161]. Agreement and disagreement are related to social attitudes; *agreement* between two persons usually entails alliance and mutually positive attitude. On the other hand, *disagreement* typically implies mutually negative attitude. Finally, *conflict* describes a high level of disagreement, or "escalation of disagreement", where at least one of the involved interlocutors feels emotionally offended. A range of human experiences, from disagreement to stress and anger, occurring when involved individuals act on incompatible goals, interests, or actions, can be labeled as conflict.

Despite the increasing popularity of the social signal processing domain [160, 228, 161], the research in machine analysis and understanding of social attitudes such as (dis)agreement, and conflict escalation/resolution is still limited. As highlighted above, this can be partially attributed to an overall lack of suitable annotated data that could be used to train the machine learning detectors for recognition of relevant phenomena [25, 161]. Bousmalis et al. [26] provide a

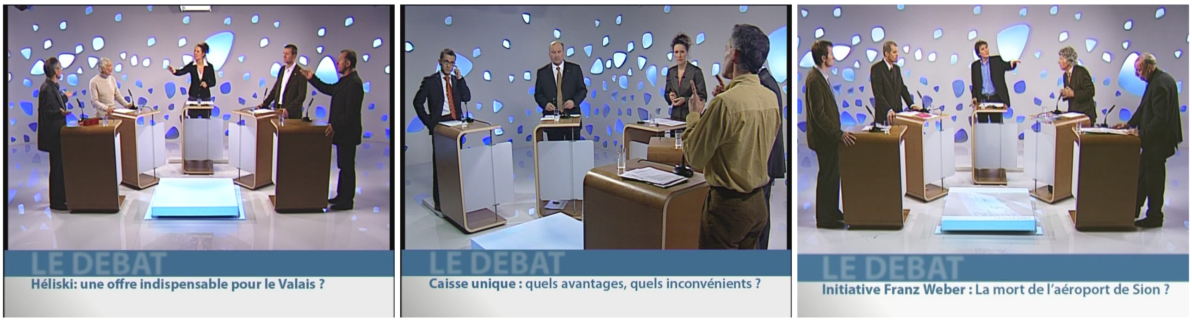

LE DEBAT
Héliski: une offre indispensable pour le Valais ?
LE DEBAT
Caisse unique : quels avantages, quels inconvénients ?
LE DEBAT
Initiative Franz Weber : La mort de l'aéroport de Sion ?

Figure 2.3: Characteristic frames from three episodes of the Canal9 Corpus [227].

survey of databases that have been released for automatic analysis of (dis)agreement, along with a review of related cues and tools. Existing databases that lend themselves to (dis)agreement and conflict analysis are mainly based on two distinct recording setups, that is, *televised political debates* and *group meetings*. Political debates offer an interesting platform for the analysis of social attitudes since they contain real-world competitive multi-party conversations where participants do not act in a simulated context, but rather participate in an event that has a major impact on their real life (for example, in terms of results at the elections) [106]. Consequently, even if some constraints are imposed by the debate format, the participants have real motivations leading to spontaneous disagreement and conflict. The most well-known database of this family is the Canal9 Corpus [227] a collection of 43 hours and 10 minutes of audio-visual recordings from 70 real televised debates(in French) on Canal 9, a Swiss television network. Characteristic frames from the Canal9 Corpus are illustrated in Fig.2.3. On the other hand, databases of recorded group meetings, since they include social interactions where (dis)agreement and conflict frequently arise. Representative examples of group meeting datasets are the AMI [137] and AMIDA [33] corpora which portray group meetings based on role playing for the design of new remote control. AMI and AMIDA are equipped with a rich set of annotations including transcriptions of the meetings, dialogue act and topic segmentation and labeling as well as head and hand gestures, among others. Characteristic frames from the AMI Corpus are illustrated in Fig.2.4. In what follows, we provide an overview of existing datasets that are have been or can be utilized for automatic analysis of conflict, since conflict is extensively investigated in this thesis both through experimental studies and a newly released database.

2.3.1 Databases for conflict analysis.

The only existing database that have been released primarily to serve research on machine analysis of *conflict* is the SSPNet Conflict Corpus [196], which consists of 1430 clips of 30 seconds extracted from the Canal9 Corpus [227] – a collection of audio-visual recordings from 45 political debates aired on the Swiss TV (in French) – corresponding to 138 subjects in total. Each clip of the database has been annotated in terms of a single continuous conflict score in the range $[-10, +10]$ for the purposes of the sequence-level binary classification and regression tasks of the Conflict Sub-Challenge included in the Interspeech 2013 Computational Paralinguistics Challenge [196]. Pesarin et al. [175] have manually segmented 13 debates from the SSPNet Conflict Corpus, with a total duration of 6 h and 27 min, into conflictual and non-conflictual intervals for conflict detection. Recently, Kim et al. [108] have relied on Mechanical Turk crowdsourcing to have the corpus annotated in terms of continuous (real-valued) conflict intensity, using two separate questionnaires, one for the *physical* layer and the other for the *inferential* layer of the conversation. However, the annotations of both works mentioned above constitute a sequence-level rather than a frame-by-frame characterization of conflict.

Audio-visual recordings of political debates have been recently utilized for research on detection of the similar behavioral phenomenon of (dis)agreement [27] (see [26]) for a survey). The latter has been also investigated by means of experiments performed on meeting corpora such as the AMI Corpus [137] and the ICSI Corpus [92]. Other databases, albeit not annotated in terms of (dis)agreement or conflict, that contain multiple instances of the latter behaviors as well as other social behaviors (e.g., interest, politeness, mimicry, flirting), social signals (e.g., social dominance, engagement, hot-spots, acceptance, blame) and personality traits (e.g., emotional stability, extraversion, conscientiousness) and thus could be used to develop relevant automated frameworks include [33, 91, 178, 76], the Green Persuasive Dataset and the newly released SEWA Database. It is worth mentioning that the SEWA Database is the largest and the richest DB of human conversational and emotional behaviour that has been released so far.

Finally, naturalistic datasets that capture human-computer or human-human elicited emotionally colored interaction such as the SAL [160] and SEMAINE [160] datasets or the Belfast Induced Natural Emotion Database (BINED) [207], respectively, contain naturalistic data that could be useful in training robust tools for detecting behavioral cues associated with (dis)agreement or conflict.

[1] The Green Persuasive Database can be found online at `http://sspnet.eu/2009/12/green-persuasive-database/`.

[2] The SEWA Database is available online at `http://db.sewaproject.eu/`.

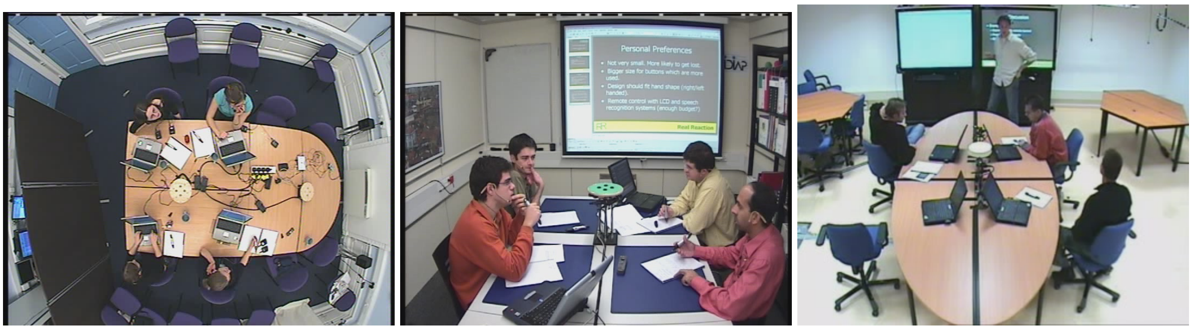


Figure 2.4: Characteristic frames from three episodes of the AMI Corpus [137].

Table 2.1: Summary of the databases that have or could be used for automatic analysis of conflict as well as other social behaviors and signals.

Database	**# Subjects**	**Duration**	**Audio-visual?**	**Synchronous?**	**Cultural Background**
Canal 9 [227]	190	43 h 10 min	✔	✔	Swiss (French-speaking)
SSPNet Conflict Corpus [196] (subset of Canal 9)	138	11 h 55 min	✔	✔	Swiss (French-speaking)
[27] (subset of Canal 9)	28	?	✔	✔	Swiss (French-speaking)
AMI [137]	213	100 h	✔	✔	Mostly non-native English speakers
AMIDA [33]	?	10 h	✔	✔	Mostly non-native English speakers
ICSI [92]	53	75 h	✗	–	28 native English speakers (mostly American), the rest non-native (12 German)
Green Persuasive[1]	16	?	✔	✔	?
Wolf [91]	36	7h	✔	✔	Mostly non-native English speakers
Mission Survival [178]	44	6h	✔	✔	Canadian
MAHNOB Mimicry [22]	60	11h	✔	✔	Spanish, French, Greek, English, Dutch, Portuguese, Romanian
SEWA[2]	398	34h 35 min	✔	✔	British, German, Hungarian, Greek, Serbian, Chinese
[76]	208	62 h 48 min	✔	✔	77% Caucasian, 8% Afr. American, 5% Asian or Pac. Islander, 5% Latino(a), 1% Native American, 4% Other
SAL [160]	?	10 h	✔	✔	Mostly native English speakers
SEMAINE [160]	150	80 h	✔	✔	Mostly native English speakers
BELFAST [207]	256	13 h 5 mins	✔	✔	Northern Irish, Peruvian

Table 2.1 provides a concise summary of the existing databases that have already or could be used for automatic analysis of conflict and similar social signals and phenomena. From Table 2.1 and the overview above, it becomes evident that there is a lack of data for automatic analysis of social attitudes such as conflict and (dis)agreement. While most of the existing databases of dyadic or multi-party interactions can be exploited for these tasks, they often do not provide annotations of any social signals whatsoever. Thus, promoting this field necessitates significant effort in the development and release of new comprehensive datasets, captured under naturalistic settings and containing specialized annotations in terms of social attitudes, signals or behaviors. As far as automatic conflict analysis is concerned, a remedy to this problem is provided by the newly released Conflict Escalation Resolution (CONFER) Database presented in Chapter 4. In what follows, we relate and contrast the CONFER Database with the above outlined databases for conflict analysis.

2.3.2 Connection to our work

The CONFER Database presented in Chapter 4, is a collection of audio-visual recordings of naturalistic interactions from political debates where conflicts naturally arise. The database contains approximately 142 minutes of recordings in Greek language, split over 120 non-overlapping episodes of spontaneous conversations that involve two or three interactants. The audio-visual episodes have been filmed in real-world “in-the-wild” conditions involving a wide range of views, amenable lighting conditions, spontaneous and overlapping speech, and abrupt head and body movements or occlusions. Most importantly, all episodes have been annotated by 10 experts, in terms of continuous conflict intensity. Along with the audio-visual episodes and the annotations, audio and visual features (facial tracking points and local appearance descriptors) are also provided with the database, thus facilitating further experimental studies. Please see Chapter 4 for a comprehensive description of the database.

The main advantage offered by the CONFER Database over existing databases is that it is annotated in terms of *continuous* and *dimensional* conflict intensity. Every episode is annotated on a frame-by-frame basis in terms of real-valued conflict intensity on a continuous scale by 10 experts. This renders the CONFER Database the first of its kind to provide dimensional and continuous-time characterizations of conflict intensity. As a matter of fact, all existing databases for (dis)agreement and conflict analysis come with binary or discretized labels characterizing pre-segmented episodes of agreement/disagreement or conflict/non-conflict. Furthermore, in each episode of the CONFER Database the interactants participating in the debate are visible in every frame of the video stream, thus rendering continuous-time visual processing of all parties feasible. This is in contrast to other datasets, such as the Canal9 corpus of political debates and its “derivatives” [196] and [27], where not all participants are visible at all times and also extreme camera angles prohibit automatic visual analysis (e.g., facial expression analysis) at times. Finally, the CONFER Database is accompanied by the first experimental study on dimensional and continuous conflict intensity estimation. In this study, various audio and visual features and fusion of them as well as classifiers are examined in subject-independent experiments for the task at hand.

Overall, the CONFER Database fills a significant gap in the availability of data for the investigation of social phenomena by providing for the first time continuous-time annotations of conflict intensity, represented in a dimensional rather than a categorical approach. As such, it is expected to promote research in the development of machine learning methodologies that model the temporal evolution of social signals and behaviors in naturalistic settings as a real-valued function of time.

Chapter 3

Discriminant Incoherent Component Analysis

Contents

The first application domain that we focus on in this thesis is *static face analysis* in (possibly) grossly corrupted still face images captured under unconstrained conditions including varying illumination, facial expression and heavy contiguous occlusion (e.g., sunglasses, scarf). From the modeling standpoint, we view information conveyed by face images as a superposition of low-complexity components associated with attributes, such as facial identity, expressions and activation of action units. Motivated by the growing need for modern emotionally-aware interfaces to recognize these attributes in a joint fashion, we build on discriminant dictionary learning and sparsity-based recognition to develop a novel learning method that can jointly approach these interrelated classification tasks. In terms of applications, emphasis is placed on face recognition and facial expression, addressed jointly or in isolation under varying types and levels of data corruption, as well as action unit detection.

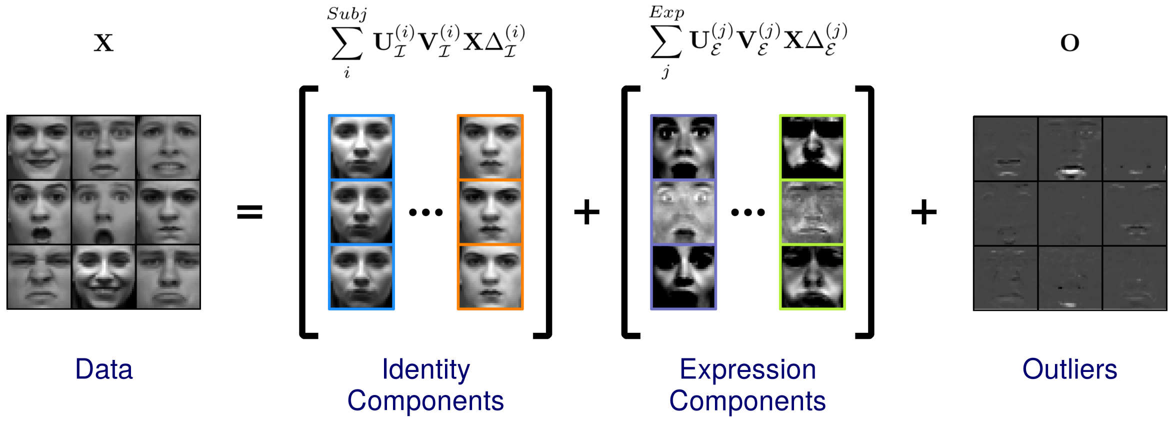

$\mathbf{X}$
$\sum_{i}^{Subj} \mathbf{U}_{\mathcal{I}}^{(i)} \mathbf{V}_{\mathcal{I}}^{(i)} \mathbf{X} \Delta_{\mathcal{I}}^{(i)}$
$\sum_{j}^{Exp} \mathbf{U}_{\mathcal{E}}^{(j)} \mathbf{V}_{\mathcal{E}}^{(j)} \mathbf{X} \Delta_{\mathcal{E}}^{(j)}$
$\mathbf{O}$
=
+
+
Data
Identity Components
Expression Components
Outliers

Figure 3.1: The proposed Discriminant Incoherent Component Analysis (DICA), as applied to the multi-label setting of joint face and expression recognition. The data matrix $\mathbf{X}$ containing expressive face images is expressed as a superposition of identity- and expression-specific mutually incoherent components, under the assumption of possible gross errors (outliers).

3.1 Introduction

Human face is a rich source of information consisting of several components which are related to attributes associated with facial identity, emotional expression and activation of action units (AUs). These components are characterized by specific structures which can assist the semantic interpretation of content in the visual stream. For instance, facial expressions manifest themselves through *sparse* non-rigid deformations occurring in certain face regions [162, 163], while images depicting the neutral face of the same person are expected to be highly correlated and thus drawn from a *low-rank* subspace. Consequently, the extraction of such features of low-complexity (i.e., exhibiting low-rank or sparse structure) is essential for accurate face and expression recognition.

As we saw in Section 2.1, a fundamental constraint of existing methods for face analysis is that the training data is often assumed to be noise-free. That is, they are collected under well controlled conditions in terms of illumination and pose variations and they do not contain occlusions or disguise. Consequently, they are not applicable in practical scenarios when both training and test data are contaminated by gross non-Gaussian noise and corruptions (e.g., occlusions and disguise). Moreover, the majority of these works approach the tasks of face and expression recognition separately rather than within a joint framework, despite them being two intertwined tasks.

To alleviate the aforementioned drawbacks and motivated by recent advances in robust subspace learning [155, 168, 187, 156, 157, 206], in this chapter we propose the Discriminant Incoherent Component Analysis (DICA) in order to decompose training facial images into a superposition of class-specific structured and mutually incoherent components accounting for identity, emotional expression or AUs in the presence of gross but sparse non-Gaussian corruptions. In other words, we model expressive faces as expressionless faces capturing the identity, superimposed by sparse images of non-rigid deformations corresponding to facial expressions, plus sparse components corresponding sparse errors of large magnitude, which cannot be explained by labels. To learn such a decomposition, we impose low-rank constraints on the components capturing the face's identity and sparsity constraints to those related to expressions. The proposed model can be also used to recover more localized sparse components related to AUs. Having found an ensemble of class-specific incoherent components, a test image is expressed as a group-sparse linear combination of these components with non-zero coefficients corresponding to the identity and expression class that the test sample belongs to.

Below, we list the main contributions of the generic supervised learning framework based on the DICA that is presented in this chapter.

1. The DICA provides a generic method to decompose data into class-specific structured and incoherent components, and a sparse matrix accounting for outliers.

2. An efficient Alternating-Directions Method of Multipliers (ADMM)-based algorithm is presented that can solve suitable optimization problems for the DICA, according to the desirable component structure.

3. A dictionary-based classification framework is proposed, according to which a test sample is collaboratively represented via class-specific components extracted by the DICA.

Overall, the discriminative representation furnished by the DICA proves efficient for static face analysis tasks. The performance of the DICA is assessed by conducting experiments on joint face and expression recognition, face recognition under varying percentages of training data corruption, subject-independent expression recognition under varying illumination conditions during training, and facial action unit detection, using 4 datasets. The proposed method outperforms the methods that is compared to in all the aforementioned tasks.

The remainder of this chapter is as follows. In Section 3.2, the DICA and its algorithmic framework are detailed. A dictionary-based framework for classification via the DICA is described in section 3.3. The performance is assessed experimentally on both synthetic and real-world data in Section 3.4. Section 3.5 concludes the chapter and gives insight for future research directions.

Notations. Matrices (vectors) are denoted by uppercase (lowercase) boldface letters , e.g., $\mathbf{A}$, $\mathbf{B}$, ($\mathbf{a}$, $\mathbf{b}$). $\mathbf{I}$ denotes the identity matrix of compatible dimensions. The ith element of vector $\mathbf{x}$ is denoted as x_i, while the ith column of matrix $\mathbf{X}$ is denoted as $\mathbf{x_i}$. For the set of real numbers, the symbol $\mathbb{R}$ is used. We refer to a set of N real matrices of varying dimensions as $\{\mathbf{X}^{(n)} \in \mathbb{R}^{p_n \times q_n}\}_{n=1}^{N}$. Regarding vector norms, $\|\mathbf{x}\|= \sqrt{\sum_i x_i^2}$ denotes the Euclidean norm. Regarding matrix norms, $\|\mathbf{X}\|_*$ denotes the nuclear norm, which equals the sum of singular values, while $\|\mathbf{X}\|$ denotes the spectral norm, which equals the largest singular value. $\|\mathbf{X}\|_1= \sum_i \sum_j |x_{ij}|$ is the element-wise matrix ℓ_1-norm, and $\|\mathbf{X}\|_F= \sqrt{\sum_i \sum_j x_{ij}^2} = \sqrt{\mathrm{tr}(\mathbf{X}^T\mathbf{X})}$ is the Frobenius norm, with $\mathrm{tr}(\cdot)$ denoting the trace of a square matrix. Finally, $\lambda_{\max}[\mathbf{X}]$ denotes the largest eigenvalue of a square matrix $\mathbf{X}$.

3.2 Discriminant Incoherent Component Analysis

In this section, the DICA is described along with its solver.

3.2.1 Problem Statement

The goal of the DICA is to robustly learn components from training samples that 1) are discriminant and exhibit low-complexity structures (e.g., low-rank or sparsity) associated with facial attributes, 2) are mutually incoherent among different classes, and 3) facilitate the classification of test samples by means of sparse representation.

Let $\mathbf{x} \in \mathbb{R}^d$ be a vectorized expressive face image and $\boldsymbol{l} \in \{0,1\}^{n_c}$ the label vector associated with it, whose non-zero elements are those corresponding to the identity and expression class it belongs to (n_c denotes the total number of classes). We seek to decompose $\mathbf{x}$ as a sum of n_c class-specific components $\mathbf{y}^{(i)} \in \mathbb{R}^d$, capturing the discriminant characteristics of each class. Thus, $\mathbf{x}$ is expressed as

$$\mathbf{x} = \sum_{i=1}^{n_c} \mathbf{y}^{(i)} \tag{3.1}$$

We assume that each class-specific component $\mathbf{y}^{(i)}$ lies in a linear orthonormal subspace spanned by $\mathbf{U}^{(i)} \in \mathbb{R}^{d \times m^{(i)}}$, and $\mathbf{V}^{(i)} \in \mathbb{R}^{m^{(i)} \times d}$ denotes the projection matrix that embeds $\mathbf{x}$ onto the $m^{(i)}$-dimensional space, while also preserving the structure (e.g., low-rank or sparsity) related to the class-specific attribute. Therefore, $\mathbf{y}^{(i)}$ is written as

$$\mathbf{y}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{x}\,, \tag{3.2}$$

Following [261] and [155], the above mentioned formulation enables us to impose a specific structure on the projection spaces $\mathbf{V}^{(i)}$, by minimizing a suitable structure-inducing norm

$\|\mathbf{V}^{(i)}\|_{(\cdot)}$; this is either the nuclear norm [68] which imposes low-rank on the projection spaces corresponding to facial identities, or the ℓ_1-norm [60] which enables to learn sparse projections for facial expressions or AUs. By incorporating (3.2) into (3.1), $\mathbf{x}$ is written as

$$\mathbf{x} = \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{x}\,, \tag{3.3}$$

Clearly, to perfectly disentangle the class-specific components $\mathbf{y}^{(i)}$ (i.e., to ensure the identifiability of (3.1)), the column spaces that they are stemming from should be mutually incoherent, that is ${\mathbf{U}^{(i)}}^T\mathbf{U}^{(j)} = \mathbf{0}$ for $i \neq j$. We observe that Equation (3.3), combined with the mutual incoherence property ${\mathbf{U}^{(i)}}^T\mathbf{U}^{(j)} = \mathbf{0}$ for $i \neq j$, entails ${\mathbf{U}^{(i)}}^T \simeq \mathbf{V}^{(i)}$ for $i = 1, 2, \ldots, n_c$. In other words, matrices ${\mathbf{U}^{(i)}}^T$ and $\mathbf{V}^{(i)}$ are proportional for every class i. This further entails that ${\mathbf{U}^{(i)}}^T\mathbf{U}^{(j)} = \mathbf{0}$ is equivalent to $\mathbf{V}^{(i)}{\mathbf{V}^{(j)}}^T = \mathbf{0}$ for $i \neq j$.

To account also for the possible presence of facial aspects that cannot be explained by labels, including outliers and gross corruptions, we include the additive term $\mathbf{o} \in \mathbb{R}^d$ in the decomposition (3.3), which is written as

$$\mathbf{x} = \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{x} + \mathbf{o}\,, \tag{3.4}$$

Having found the decomposition (3.4), the representation vector $[(\mathbf{V}^{(1)}\mathbf{x})^T, (\mathbf{V}^{(2)}\mathbf{x})^T, \cdots, (\mathbf{V}^{(n_c)}\mathbf{x})^T]^T$ is expected to be group-sparse, with non-zero elements corresponding to the class(es) the sample $\mathbf{x}$ belongs to.

The DICA learns the reconstruction matrices $\{\mathbf{U}^{(i)}\}_{i=1}^{n_c}$ and projection matrices $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$ by employing the training matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$ which contains in its columns the vectorized training face images, with d being the dimensionality of each image and N the number of training observations. Let us denote by $\mathbf{X}_{\mathcal{S}^{(i)}} \in \mathbb{R}^{d\times N}$ the column-sparse matrix whose non-zero columns are the columns of $\mathbf{X}$ with label i. Therefore, with the set $\mathcal{W} = \{\{\mathbf{U}^{(i)} \in \mathbb{R}^{d\times m^{(i)}}\}_{i=1}^{n_c}, \{\mathbf{V}^{(i)} \in \mathbb{R}^{m^{(i)}\times d}\}_{i=1}^{n_c}, \mathbf{O} \in \mathbb{R}^{d\times N}\}$ containing all the unknown variables, the DICA solves

$$\begin{aligned} &\operatorname*{arg\,min}_{\mathcal{W}} \lambda^{(i)} \sum_{i=1}^{n_c} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta \sum_{i\neq j} \|\mathbf{V}^{(i)}{\mathbf{V}^{(j)}}^T\|_F^2 + \lambda_1\|\mathbf{O}\|_1\,, \\ &\text{s.t.} \quad i) \quad \mathbf{X} = \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} + \mathbf{O}\,, \\ &\qquad\; ii) \quad {\mathbf{U}^{(i)}}^T\mathbf{U}^{(i)} = \mathbf{I}\,, \quad i = 1, 2, \ldots, n_c\,, \end{aligned} \tag{3.5}$$

where the structure-inducing norm $\|\mathbf{V}^{(i)}\|_{(\cdot)}$ is either the nuclear norm for face-specific projections or the ℓ_1-norm for expression-specific and AU-specific projections. The term

Algorithm 2 ADMM solver for the DICA (3.5)

```
Input: Data: $\mathbf{X} \in \mathbb{R}^{d \times N}$. Parameters: $\lambda^{(i)}$, $\eta$, $\lambda_1$, and $\{m^{(i)}\}_{i=1}^{n_c}$.
Normalize each column of $\mathbf{X}$ to unit $\ell_2$-norm.
Initialize: Set $\{\{\mathbf{U}^{(i)}[0]\}, \{\mathbf{V}^{(i)}[0]\}\}_{i=1}^{n_c}$, $\mathbf{O}[0]$, $\mathbf{Y}[0]$ to zero matrices. Set $\mu[0] = 1/\|\mathbf{X}\|$, $\rho = 1.1$, $\mu_{\max} = 10^{10}$.
while not converged do
  for $i = 1 : n_c$ do
    Calculate $L = 1.02\lambda_{\max}\left[\mu[t]\mathbf{X}_{\mathcal{S}^{(i)}}\mathbf{X}_{\mathcal{S}^{(i)}}^T + 2\eta\sum_{j \neq i}\mathbf{V}^{(j)}[t]^T\mathbf{V}^{(j)}[t]\right]$.
    if $\mathbf{V}^{(i)}$ is associated with nuclear norm then
      $\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{D}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - L^{-1}\nabla f(\mathbf{V}^{(i)}[t])\right]$.[1]
    else if $\mathbf{V}^{(i)}$ is associated with $\ell_1$-norm then
      $\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{S}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - L^{-1}\nabla f(\mathbf{V}^{(i)}[t])\right]$.
    end if
    $\mathbf{U}^{(i)}[t+1] \leftarrow \mathcal{P}\left[\left(\mathbf{X} - \sum_{j \neq i}\mathbf{U}^{(j)}[t]\mathbf{V}^{(j)}[t+1]\mathbf{X}_{\mathcal{S}^{(j)}} - \mathbf{O}[t] + \mu[t]^{-1}\mathbf{Y}[t]\right)\left(\mathbf{V}^{(i)}[t+1]\mathbf{X}_{\mathcal{S}^{(i)}}^T\right)\right]$.
  end for
  $\mathbf{O}[t+1] \leftarrow \mathcal{S}_{\lambda_1/\mu[t]}\left[\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}[t+1]\mathbf{V}^{(i)}[t+1]\mathbf{X}_{\mathcal{S}^{(i)}} + \mu[t]^{-1}\mathbf{Y}[t]\right]$.
  Update the Lagrange multiplier by $\mathbf{Y}[t+1] \leftarrow \mathbf{Y}[t] + \mu[t]\left(\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}[t+1]\mathbf{V}^{(i)}[t+1]\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}[t+1]\right)$.
  Update $\mu$ by $\mu[t+1] = \min(\rho \cdot \mu[t], \mu_{\max})$.
end while
Output: $\{\mathbf{U}^{(i)} \in \mathbb{R}^{d \times m^{(i)}}, \mathbf{V}^{(i)} \in \mathbb{R}^{m^{(i)} \times d}\}_{i=1}^{n_c}$, $\mathbf{O} \in \mathbb{R}^{d \times N}$.
```

$\sum_{i \neq j}\|\mathbf{V}^{(i)}\mathbf{V}^{(j)^T}\|_F^2$ induces mutual incoherence among the projection spaces and $\mathbf{O} \in \mathbb{R}^{d \times N}$ denotes the outlier matrix accounting for components that cannot be explained by the summand containing the class-specific reconstructions. The positive parameters $\lambda^{(i)}$, η, and λ_1 control the norm imposed on $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$, the mutual incoherence for all component pairs, and the sparsity of outliers $\mathbf{O}$, respectively.

In Fig. 3.1, one can see how the proposed DICA is applied to the multi-label scenario of joint face and expression recognition. In that case, each training image is characterized by two labels, one for identity and the other for expression. The data matrix $\mathbf{X}$, containing the vectorized training images, is accordingly represented as a superposition of discriminant and mutually incoherent class-specific components (low-rank for identity and sparse for expression), plus an outlier matrix $\mathbf{O}$ accounting for unbounded sparse errors.

3.2.2 Alternating-Direction Method-Based Algorithm

The Alternating-Directions Method of Multipliers (ADMM) [18] is employed hereby to solve (3.5). The (partial) augmented Lagrangian function for (3.5) is defined as:

$$\begin{aligned}\mathcal{L}(\mathcal{W}, \mathbf{Y}, \mu) = {} & \lambda^{(i)} \sum_{i=1}^{n_c} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta \sum_{i \neq j} \|\mathbf{V}^{(i)} \mathbf{V}^{(j)^T}\|_F^2 \\ & + \lambda_1 \|\mathbf{O}\|_1 + \text{tr}\left(\mathbf{Y}^T\left(\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)} \mathbf{V}^{(i)} \mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\right)\right) \\ & + \frac{\mu}{2} \|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)} \mathbf{V}^{(i)} \mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\|_F^2 ,\end{aligned} \tag{3.6}$$

where μ is a positive parameter and $\mathbf{Y} \in \mathbb{R}^{d \times N}$ is the Lagrange multiplier related to the linear constraint in (3.5).

At each iteration, (3.6) is minimized with respect to each variable in $\mathcal{W}$ in an alternating fashion and, subsequently, the Lagrange multiplier $\mathbf{Y}$ and parameter μ are updated. The iteration index is denoted herein by t. The notation $\mathcal{L}(\mathbf{U}^{(i)}, \mathbf{Y}[t], \mu[t])$ is used to denote the solution stage in which all other variables but $\mathbf{U}^{(i)}$ are kept fixed, and similarly for the other unknown variables. Thus, given the variables $\mathcal{W}[t]$, the Lagrange multiplier $\mathbf{Y}[t]$ and the parameter $\mu[t]$ at iteration t, the updates of ADMM are calculated as follows.

Update the primal variables:

$$\begin{aligned}\mathbf{U}^{(i)}[t+1] & = \operatorname*{arg\,min}_{\mathbf{U}^{(i)}} \mathcal{L}(\mathbf{U}^{(i)}, \mathbf{Y}[t], \mu[t]) \\ & \quad \text{s.t.} \quad \mathbf{U}^{(i)^T} \mathbf{U}^{(i)} = \mathbf{I}, \quad i = 1, 2, \ldots, n_c \\ & = \operatorname*{arg\,min}_{\mathbf{U}^{(i)}} \frac{\mu[t]}{2} \|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)} \mathbf{V}^{(i)} \mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1} \mathbf{Y}[t]\|_F^2 \\ & \quad \text{s.t.} \quad \mathbf{U}^{(i)^T} \mathbf{U}^{(i)} = \mathbf{I}, \quad i = 1, 2, \ldots, n_c\end{aligned} \tag{3.7}$$

$$\begin{aligned}\mathbf{V}^{(i)}[t+1] & = \operatorname*{arg\,min}_{\mathbf{V}^{(i)}} \mathcal{L}(\mathbf{V}^{(i)}, \mathbf{Y}[t], \mu[t]) \\ & = \operatorname*{arg\,min}_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta \sum_{i \neq j} \|\mathbf{V}^{(i)} \mathbf{V}^{(j)^T}\|_F^2 \\ & + \frac{\mu[t]}{2} \|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)} \mathbf{V}^{(i)} \mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1} \mathbf{Y}[t]\|_F^2 \\ & = \operatorname*{arg\,min}_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + f(\mathbf{V}^{(i)}), \quad i = 1, 2, \ldots, n_c\end{aligned} \tag{3.8}$$

Algorithm 3 Framework for face/expression recognition.

```
Input: Data: training set $\mathbf{X} \in \mathbb{R}^{d\times N}$, query image $\mathbf{y} \in \mathbb{R}^{N\times 1}$. Parameters: $\lambda_{Lasso}$.
Normalize each column of $\mathbf{X}$ to unit $\ell_2$-norm.
Compute low-rank matrices $\{\mathbf{A}^{(i)}\}_{i=1}^{n_c}$ by performing RPCA [32] on each class-specific sub-matrix $\mathbf{X}^{(i)}$.
Initialize: For each subspace $i \in \{1,2,\dots,n_c\}$, set $\mathbf{U}^{(i)}[0] = \mathbf{M}^{(i)}$, and $\mathbf{V}^{(i)}[0] = \mathbf{M}^{(i)^T}$, where $\mathbf{A}^{(i)} = \mathbf{M}^{(i)}\mathbf{\Sigma}\mathbf{N}^{(i)^T}$ is the skinny SVD of $\mathbf{A}^{(i)}$.
Calculate $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$ according to Algorithm 2, using the nuclear- ($\ell_1$-) norm in Problem (3.5) for face (expression) recognition.
Form dictionary $\mathbf{D} = \left[\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \dots, \mathbf{D}^{(n_c)}\right]$, with $\mathbf{D}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}$ , $i \in \{1,2,\dots,n_c\}$.
Normalize each column of $\mathbf{D}$ to unit $\ell_2$-norm.
Perform SRC: $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|^2 + \lambda_{Lasso}\|\boldsymbol{\alpha}\|_1$.
for $i = 1 : n_c$ do
    $err(i) = \|\mathbf{y} - \mathbf{D}\delta^{(i)}(\hat{\boldsymbol{\alpha}})\|$.
end for
$i^* \leftarrow \arg\min_{i\in\{1,2,\dots,n_c\}} err(i)$.
Output: subject (expression) label $i^*$.
```

$$
\begin{aligned}
\mathbf{O}[t+1] &= \arg\min_{\mathbf{O}} \mathcal{L}(\mathbf{O}, \mathbf{Y}[t], \mu[t]) \\
&= \arg\min_{\mathbf{O}} \lambda_1\|\mathbf{O}\|_1 \\
&+ \frac{\mu[t]}{2}\|\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1}\mathbf{Y}[t]\|_F^2
\end{aligned}
\tag{3.9}
$$

Update the Lagrange Multiplier:

$$
\mathbf{Y}[t+1] = \mathbf{Y}[t] + \mu[t]\left(\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\right) \tag{3.10}
$$

Equations (3.7)-(3.9) are solved by means of the operators and Lemmas that are introduced next. We begin by defining the shrinkage operator [32] as $\mathcal{S}_\tau[a] = \operatorname{sgn}(a)\max(|a|-\tau, 0)$, whose matrix version is obtained by applying it element-wise. Also, if $\mathbf{A} = \mathbf{M}\mathbf{\Sigma}\mathbf{N}^T$ denotes the SVD of a matrix $\mathbf{A}$, the singular value thresholding operator (SVT) is defined as in [31]: $\mathcal{D}_\tau[\mathbf{A}] = \mathbf{M}\mathcal{S}_\tau[\mathbf{\Sigma}]\mathbf{N}^T$. Based again on the SVD of $\mathbf{A}$, the Procrustes operator is defined as

$\mathcal{P}[\mathbf{A}] = \mathbf{M}\mathbf{N}^T$ and solves the problem in the following Lemma.

Lemma 1 [261]: *The constrained minimization problem:*

$$\underset{\mathbf{B}}{\arg\min}\, \|\mathbf{A} - \mathbf{B}\|_F^2 \quad \text{s.t.} \quad \mathbf{B}^T\mathbf{B} = \mathbf{I} \tag{3.11}$$

has a closed-form solution given by $\mathbf{P} = \mathcal{P}[\mathbf{A}]$.

The solution of (3.8) is presented in detail in the Appendix and is based on the SVT (shrinkage) operator when the nuclear- (ℓ_1-) norm is employed for the component $\mathbf{V}^{(i)}$. Moreover, the minimizer of (3.9) is based on the shrinkage operator. Finally, (3.7) is solved as in Lemma 1. The ADMM-based solver of (3.5) is wrapped up in Algorithm 2. For all experiments presented herein, Algorithm 2 is terminated when $\|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\|_F / \|\mathbf{X}\|_F < 10^{-7}$, or when 1000 iterations are reached.

Computational Complexity and Convergence In the case where the nuclear norm is enforced on $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$, the cost of each iteration in Algorithm 2 is mainly associated with the calculation of the SVT operator in Step 7. Hence, each iteration has a complexity equal to that of SVD, i.e., $\mathcal{O}(\max(d^2N, dN^2))$. In the case where the ℓ_1-norm is used, the shrinkage operator becomes the most time-consuming calculation, thus entailing linear complexity $\mathcal{O}(dN)$. As far as convergence of Algorithm 2 is concerned, the convergence of the ADMM to local minima has not been proved for the cases where the latter is adopted to solve non-convex problems [18, 173]. A systematic convergence proof does not fall within the scope of this thesis, yet for proof of the weak convergence of Algorithm 2 one can follow the approach in [124]. Nonetheless, the experiments in Section 3.4 serve as a testament to the guaranteed convergence of Algorithm 2.

3.3 DICA-based Classification

In this section, a dictionary-based framework built upon the DICA (3.5) is proposed. This can be tailored accordingly to cope with either a single- or a multi-label scenario. Herein, the framework is presented for the problems of face and expression recognition, viewed either as separate single-label tasks or jointly within a multi-label setting. For the multi-label scenario,

[1] f is the smooth differentiable part of the minimizer (3.8).

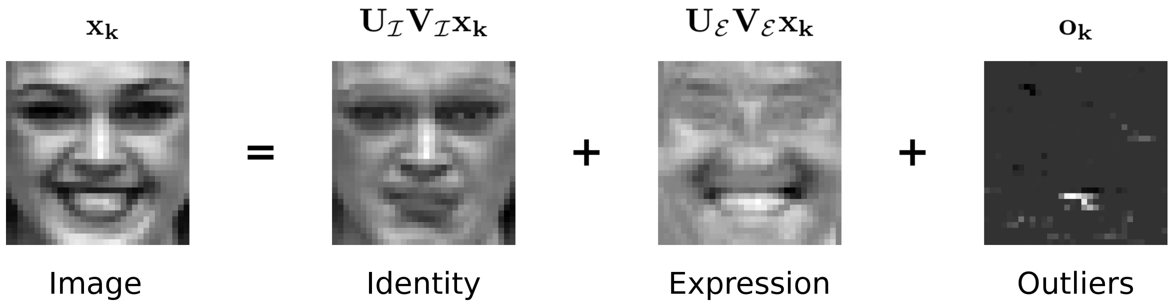

$\mathbf{x_k}$
$\mathbf{U}_{\mathcal{I}}\mathbf{V}_{\mathcal{I}}\mathbf{x_k}$
$\mathbf{U}_{\mathcal{E}}\mathbf{V}_{\mathcal{E}}\mathbf{x_k}$
$\mathbf{o_k}$
=
+
+
Image
Identity
Expression
Outliers

Figure 3.2: Decomposition of an expressive image from the CK+ Dataset into an identity component, an expression component and a sparse error term accounting for outliers, as produced by the DICA.

an extension of our framework, which can deal with the facial action unit detection task, is also described.

3.3.1 Single-Label Case: Face/Expression Recognition

Suppose each column $\mathbf{x_n}$ of our training data matrix $\mathbf{X} \in \mathbb{R}^{d \times N}$ represents a vectorized image, with subject (expression) label $i \in \{1, 2, \dots, n_c\}$, where n_c equals the number of subjects (expressions). Let us also denote by $\mathbf{X}^{(i)} \in \mathbb{R}^{d \times n^{(i)}}$ the matrix that is composed of the $n^{(i)}$ columns of $\mathbf{X}$ that are associated with the subject (expression) label i.

First, for face (expression) recognition, the nuclear- (ℓ_1-) norm is chosen for $\mathbf{V}^{(i)}$ in the DICA, as the goal here is to uncover low-rank (sparse) components. Second, RPCA [32] is performed on each $\mathbf{X}^{(i)}$ for warm initialization of $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ in (3.5). Specifically, each basis $\mathbf{U}^{(i)}$ and component $\mathbf{V}^{(i)}$ is initialized as $\mathbf{U}^{(i)} = \mathbf{M}^{(i)}$ and $\mathbf{V}^{(i)} = \mathbf{M}^{(i)^T}$, respectively, where $\mathbf{A}^{(i)}$ denotes the low-rank matrix yielded by RPCA for subject (expression) i and $\mathbf{A}^{(i)} = \mathbf{M}^{(i)}\mathbf{\Sigma}\mathbf{N}^{(i)^T}$ denotes its skinny SVD. Note that setting $\mathbf{V}^{(i)} = \mathbf{M}^{(i)^T} = \mathbf{U}^{(i)^T}$ is an intuitive choice, considering that $\mathbf{V}^{(i)}$ and $\mathbf{U}^{(i)^T}$ are proportional to each other, as shown in Section 3.2.1. Choosing an initial estimate that is close to the optimum sought can markedly speed up the convergence of a non-convex optimization problem like the DICA [18]. RPCA has been proved efficient in recovering low-complexity facial components, while also being robust to gross errors in the data [211]. This motivates its choice for the initialization step, while its positive impact on the convergence speed was corroborated by preliminary experiments. Third, Problem (3.5) is solved according to Algorithm 2.

Following a SRC-like approach, the class-specific reconstruction images $\{\mathbf{D}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}\}_{i=1}^{n_c}$ are concatenated to construct the dictionary $\mathbf{D}$. Then, for each query image $\mathbf{y} \in \mathbb{R}^{d \times 1}$ a vector $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{N \times 1}$ is sought so that $\mathbf{y}$ is represented as a sparse linear combination of the dictionary atoms, i.e., $\mathbf{y} = \mathbf{D}\hat{\boldsymbol{\alpha}}$. The sparse coefficient vector $\hat{\boldsymbol{\alpha}}$ is obtained by solving

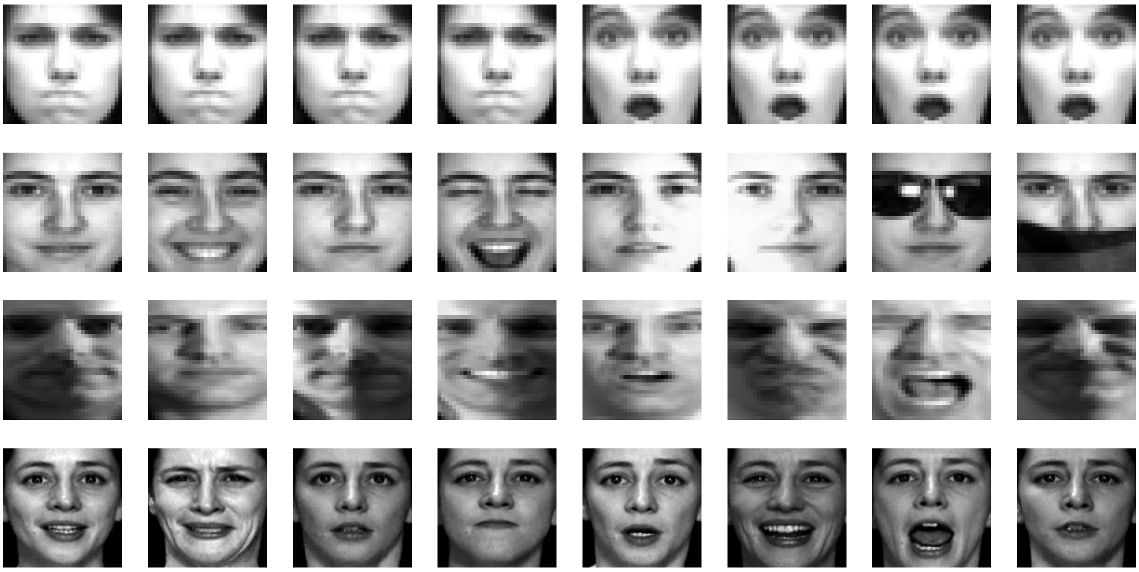


Figure 3.3: Example registered images from each of the 4 datasets used. From top to bottom: CK+ [130], AR [135], CMU Multi-PIE [80], GEMEP-FERA [219].

the Lasso minimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|^2 + \lambda_{Lasso}\|\boldsymbol{\alpha}\|_1 \tag{3.12}$$

Finally, the subject (expression) label i^* is estimated as that accounting for the minimum class-specific reconstruction error of $\mathbf{y}$, i.e.,

$$i^* = \arg\min_{i\in\{1,2,\ldots,n_c\}} \|\mathbf{y} - \mathbf{D}\delta^{(i)}(\hat{\boldsymbol{\alpha}})\|\,, \tag{3.13}$$

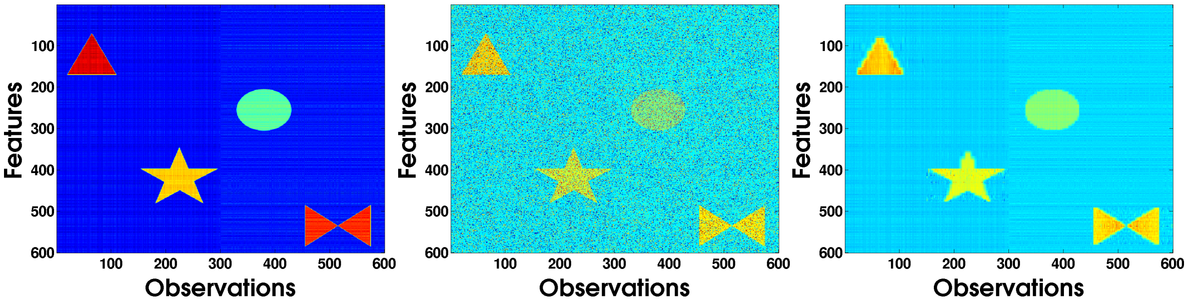
where $\hat{\boldsymbol{\alpha}}$ is the solution of (3.12), and $\{\delta^{(i)}(\cdot) : \mathbb{R}^{N\times 1} \mapsto \mathbb{R}^{N\times 1}\}_{i=1}^{n_c}$ are class-specific selector operators calculated as

$$\delta^{(i)}(q_n) = \begin{cases} q_n, & \text{if } n \in \mathcal{S}^{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

The proposed single-label framework is summarized in Algorithm 3 for face/expression recognition.

3.3.2 Multi-Label Case: Joint Face and Expression Recognition & Action Unit Detection

The framework described in the previous section is extended to the multi-label case, where each observation is associated with multiple labels w.r.t. different attributes. Two face analysis



(a) Synthetic Data (b) Data corrupted with sparse noise (c) Reconstruction by the DICA

Figure 3.4: Illustration of corrupted synthetic data reconstruction, as produced by the DICA. Each 600×150 subset of the data matrix (where the first dimension is the feature space and the second dimension is the ambient space) is a superposition of one of the two low-rank components (depicted as 600×300 blue striped backgrounds in (a)) and one of the four block-sparse components, which form a shape of filled triangle, asterisk, circle and butterfly, respectively. (a) Original synthetic data, (b) Synthetic data of (a) contaminated with additive sparse noise, (c) Low-Rank/Sparse Reconstruction of the corrupted signal as produced by the DICA.

tasks that fall in this multi-label case are (a) joint face and expression recognition, and (b) facial action unit (AU) detection. In this section, we choose to present the DICA-based classification framework tailored to the aforementioned tasks, on which our experimental validation in Section 3.4 is based.

Joint Face and Expression Recognition First, the DICA (3.5) is solved for the total number of classes $n_c = n_s + n_e$, with n_s (n_e) being the number of subjects (expressions). Similarly to the single-label case, for the subject- (expression-)specific components $i \in \{1, 2, \dots, n_s\}$ ($i \in \{n_s + 1, n_s + 2, \dots, n_s + n_e\}$) the nuclear- ($\ell_1$-) norm is enforced on the corresponding $\mathbf{V}^{(i)}$. Second, the derived identity-related reconstruction images are used to form the identity dictionary $\mathbf{D}_{\mathcal{I}}$, while the expression-related reconstruction images are used to form the expression dictionary $\mathbf{D}_{\mathcal{E}}$. The final dictionary consists of the concatenation of $\mathbf{D}_{\mathcal{I}}$ and $\mathbf{D}_{\mathcal{E}}$ as $\mathbf{D} = \begin{bmatrix} \mathbf{D}_{\mathcal{I}} & \mathbf{D}_{\mathcal{E}} \end{bmatrix}$.

Subsequently, the SRC algorithm is modified accordingly to solve jointly for the identity and expression coefficient vectors $\hat{\boldsymbol{\alpha}}_{\mathcal{I}}$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{E}}$, respectively:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{\mathcal{I}}, \hat{\boldsymbol{\alpha}}_{\mathcal{E}} &= \underset{\boldsymbol{\alpha}_{\mathcal{I}},\boldsymbol{\alpha}_{\mathcal{E}}}{\arg\min} \frac{1}{2}\|\mathbf{y} - \begin{bmatrix}\mathbf{D}_{\mathcal{I}} & \mathbf{D}_{\mathcal{E}}\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha}_{\mathcal{I}} \\ \boldsymbol{\alpha}_{\mathcal{E}}\end{bmatrix}]\|^2 \\ &+ \frac{\lambda_{Lasso}}{2}\| \begin{bmatrix}\boldsymbol{\alpha}_{\mathcal{I}} \\ \boldsymbol{\alpha}_{\mathcal{E}}\end{bmatrix} \|_1 \\ &= \underset{\boldsymbol{\alpha}_{\mathcal{I}},\boldsymbol{\alpha}_{\mathcal{E}}}{\arg\min} \frac{1}{2}\|\mathbf{y} - \mathbf{D}_{\mathcal{I}}\boldsymbol{\alpha}_{\mathcal{I}} - \mathbf{D}_{\mathcal{E}}\boldsymbol{\alpha}_{\mathcal{E}}\|^2 \\ &+ \frac{\lambda_{Lasso}}{2}\|\boldsymbol{\alpha}_{\mathcal{I}}\|_1 + \frac{\lambda_{Lasso}}{2}\|\boldsymbol{\alpha}_{\mathcal{E}}\|_1\end{aligned} \tag{3.15}$$

Finally, the component separation approach of [211] is followed, where the reconstruction image $\hat{\mathbf{y}}_{\mathcal{I}} = \mathbf{D}_{\mathcal{I}}\hat{\boldsymbol{\alpha}}_{\mathcal{I}}$ based on the identity dictionary $\mathbf{D}_{\mathcal{I}}$ is utilized for face recognition, and, similarly, the reconstruction image $\hat{\mathbf{y}}_{\mathcal{E}} = \mathbf{D}_{\mathcal{E}}\hat{\boldsymbol{\alpha}}_{\mathcal{E}}$ based on the expression dictionary $\mathbf{D}_{\mathcal{E}}$ is utilized for expression recognition, according to the following minimum-residual rules:

$$i^*_{\mathcal{I}} = \underset{i\in\{1,2,\ldots,n_s\}}{\arg\min} \|\hat{\mathbf{y}}_{\mathcal{I}} - \mathbf{D}_{\mathcal{I}}\delta^{(i)}(\hat{\boldsymbol{\alpha}}_{\mathcal{I}})\| \tag{3.16}$$

$$i^*_{\mathcal{E}} = \underset{i\in\{n_s+1,n_s+2,\ldots,n_s+n_e\}}{\arg\min} \|\hat{\mathbf{y}}_{\mathcal{E}} - \mathbf{D}_{\mathcal{E}}\delta^{(i)}(\hat{\boldsymbol{\alpha}}_{\mathcal{E}})\| \tag{3.17}$$

In Fig. 3.2, one can see the decomposition of an expressive image into a identity-related component, an expression-related component and a sparse error term. The identity (expression) component is formed out of the reconstruction of the original image based on the corresponding subject- (expression-)specific subspace. It can be visually verified that indeed the identity (expression) component contains no expression- (subject-)related information, due to its calculation based on images of all training expressions (subjects) and the mutual incoherence property. Finally, the outliers term encodes whatever image features deviate in a non-Gaussian sense from the class-specific decomposition that model (3.5) dictates.

Facial Action Unit Detection The DICA (3.5) is applied for the total of n_c of AU-specific classes, using the ℓ_1-norm to enforce sparse structure on the respective components $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$. Note that a training image with more than one AUs activated can appear multiple times in (3.5), through the corresponding class-specific sub-matrices $\mathbf{X}_{\mathcal{S}^{(i)}}$. Similarly to Algorithm (3), reconstruction images are next used to form class-specific dictionaries $\mathbf{D}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}$, $i \in \{1, 2, \ldots, n_c\}$, each of which is associated only with the respective AU label, regardless of the possible presence of other AUs in the corresponding training images. The final dictionary $\mathbf{D} \in \mathbb{R}^{d\times N}$ is formed out of the concatenation of all class-specific dictionaries $\{\mathbf{D}^{(i)}\}_{i=1}^{n_c}$. Next, for each test set vector $\mathbf{y} \in \mathbb{R}^{d\times 1}$ the sparse coefficient vector $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{N\times 1}$ and the reconstructed test vector $\hat{\mathbf{y}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$ are obtained by solving (3.12).

Classical SRC, formulated as in Equation (3.13), is not directly applicable to the action unit detection task, as the latter necessitates binary classification for each of the AU-specific classes. The sparse similarity voting approach in [191] is adopted herein for classification. Let $\boldsymbol{l_n} \in \{0,1\}^{n_c}$ be the binary label vector associated with the dictionary atom $\mathbf{d_n}$. By construction, only one element of $\boldsymbol{l_n}$ will be non-zero for our framework, i.e., that which corresponds to the AU label of the class-specific dictionary $\mathbf{d_n}$. Let also $\mathbf{L} \in \{0,1\}^{n_c\times N}$ be the label matrix for the whole dictionary, with corresponding label vectors $\boldsymbol{l_n}$ in its columns. Then, the multi-label *confidence* vector $\mathbf{c} \in \mathbb{R}^{n_c}$ for the test sample $\mathbf{y}$, is given by

$$\mathbf{c} = \sum_{n=1}^{N} w_n \boldsymbol{l_n} = \mathbf{L}\mathbf{w}\,, \tag{3.18}$$

where w_n denotes the similarity between the test vector $\mathbf{y}$ and its reconstruction by the n-th dictionary atom, given by

$$w_n = \frac{\hat{\alpha}_n \mathbf{d_n}^T \mathbf{y}}{\|\mathbf{y}\|\|\hat{\mathbf{y}}\|} \tag{3.19}$$

Each element c_i of the label vector $\mathbf{c}$ in (3.18) can be perceived as a *confidence* score with regards to the test sample belonging to the i-th AU class. Finally, binary labels for the test sample with respect to each class are obtained by thresholding each c_i via ROC analysis [67].

3.4 Experiments

Our method is evaluated on four distinct tasks: (a) face recognition, (b) facial expression recognition, (c) joint face and expression recognition, and (d) facial action unit detection. Our dictionary-based framework for joint face and expression recognition is evaluated on CK+ Dataset [130], while experiments on subject-independent facial expression recognition are conducted on both CK+ [130] and CMU Multi-PIE [80] datasets. For face recognition experiments and action unit detection experiments, AR database [135] and GEMER-FERA database [219] is used, respectively.

The proposed method is compared to the approaches of Linear Regression Classifier (LRC) [142], Sparse Representation-based Classification (SRC) [243], as well as Robust Principal Component Analysis and SRC (RPCA+SRC) and Low-Rank Matrix Recovery with Structural Incoherence (LRSI) combined with SRC [237]. For RPCA+SRC, RPCA [32] is applied for each subject and the resulting low-rank (sparse) matrices are used for SRC-based

Table 3.1: Quantitative reconstruction results produced by the DICA on the synthetic data shown in Fig. 3.4. For a given component $\mathbf{X}^{(i)}$, the reconstruction metric used here corresponds to $\|\mathbf{X}^{(i)} - \hat{\mathbf{X}}^{(i)}\|_F / \|\mathbf{X}^{(i)}\|_F$, where $\hat{\mathbf{X}}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}$.

Reconstructions	
Clean Signal	0.369
Error Signal	0.916
Low-Rank Component 1	0.986
Low-Rank Component 2	0.972
Triangle	0.933
Asterisk	0.928
Circle	0.916
Butterfly	0.927
Relative Constraint	$9.9 \cdot 10^{-8}$

face (expression) recognition similarly to [211]. For LRSI, the algorithm in [237] is applied subject-wise for face recognition and expression-wise for expression recognition; the nuclear norm is used for all components. In case of identical experimental protocol, LRSI results correspond to those reported in [237]. Unlike [237], where PCA is used to reduce dimensionality, vectorized images in the pixel domain are used for all experiments, with the exception of AU detection experiments in Section 3.4.5.

Implementation details For both our method and LRSI, the parameter η that controls incoherence is set to the value 10^{-1}, which was proved efficient upon preliminary experiments. For the DICA, various values, different for each task, are examined for the parameter $\lambda^{(i)}$ controlling the norm $\|\mathbf{V}^{(i)}\|_{(\cdot)}$ and the outlier-related parameter λ_1 in Problem (3.5), and the best score achieved is reported each time. For each RPCA+SRC and LRSI optimization problem applied class-wise, the value $\lambda_1 = 1/\sqrt{\max(d, n^{(i)})}$ is used for the parameter associated with the sparse error term, which is an efficient heuristic according to [32].

For the face recognition experiments in Section 3.4.3, the Lasso minimization problem (3.12) for the SRC-based approaches is solved by means of the Homotopy method [246], in order for our results to be comparable to those in [237]. For all SRC-based experiments in Sections 3.4.2, 3.4.4, and 3.4.5, the Efficient Euclidean Projections method [125] is chosen to solve the Lasso problems (3.12) and (3.15), thanks to its fast implementation and robustness to matrix singularities.

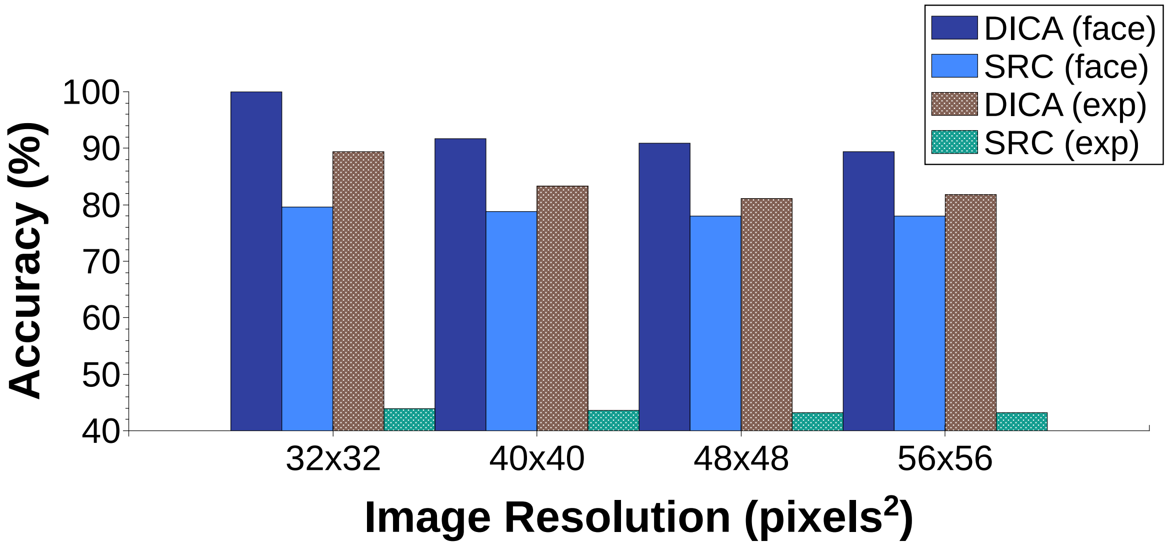


Figure 3.5: Face and expression recognition accuracies (%), as produced by the DICA and SRC for the first fold of the protocol for the CK+ Dataset, varying with the image resolution.

For all experiments with the DICA, the regularization parameter λ_{Lasso} of the Lasso minimization problems (3.12) and (3.15) is examined amongst the values $\{10^{-5}, 5{\cdot}10^{-5}, 10^{-4}, \ldots, 5{\cdot}10^{-1}\}$, and the best result is reported each time. For joint face and expression recognition, recognition accuracies reported correspond to the best average score over the two tasks. For all experiments with the other SRC-based approaches, that is, SRC, RPCA+SRC, and LRSI, λ_{Lasso} is fixed to 10^{-3}.

The DICA is also evaluated by means of experiments with synthetic data in Section 3.4.1. The results of these experiments serve as an important proof of concept since (a) they validate the effectiveness of our method both qualitatively and quantitatively, and (b) they provide evidence that our method can be applied equally well to any labeled data populations, thus serving diverse applications other than face analysis tasks.

3.4.1 Experiment on Synthetic Data

Our method is first evaluated on synthetic data corrupted with sparse, non-Gaussian noise. Each data point is constructed as a superposition of a low-rank and block-sparse component. In more detail, we first create a rank-2 component $\mathbf{X}^{(1)}$ with column space $\mathbf{U}^{(1)} \in \mathbb{R}^{600\times 2}$, based on the first two principal components of a random matrix $\mathbf{A} \in \mathbb{R}^{600\times 300}$. Next, we form a second rank-2 component $\mathbf{X}^{(2)}$ with column space $\mathbf{U}^{(2)} = \mathbf{R}\mathbf{U}^{(1)}$, where $\mathbf{R}$ is a random orthogonal matrix; as a result of this, the two components are mutually incoherent. Subsequently, four block-sparse components $\mathbf{X}^{(i)} \in \mathbb{R}^{600\times 150}$ $(3 \leq i \leq 6)$ are constructed, with their non-zero

Table 3.2: Recognition Rates (%) for Joint Face & Expression Recognition and Subject-Independent Expression Recognition on CK+ Dataset.

Method	Joint Face & Expression Recognition		Subject-Independent Expression Recognition
	Face	Expression	
LRC [142]	86.2	57.7	60.1
SRC [243]	75.4	41.4	53.5
RPCA+SRC [237]	89.6	59.5	70.6
LRSI [237]	92.9	75.5	71.4
DICA	96.7	83.6	75.7

elements corresponding to visually discernible shapes, that is, triangle, asterisk, circle and butterfly, respectively. Those are then added to the low-rank components to form the matrices $\mathbf{Y}_1 = \mathbf{X}^{(1)} + \mathbf{X}^{(3)} + \mathbf{X}^{(4)}$ and $\mathbf{Y}_2 = \mathbf{X}^{(2)} + \mathbf{X}^{(5)} + \mathbf{X}^{(6)}$. Our final clean data matrix $\mathbf{Y}$ is the result of concatenation of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ along the second dimension, and can be seen in Fig. 3.4a.

Subsequently, sparse, non-Gaussian noise is added to the original signal $\mathbf{Y}$ to simulate a more realistic scenario. First, a matrix containing only values in $\{+1, -1\}$ is created as $\mathbf{E} = \text{sgn}(\mathbf{B})$, where $\mathbf{B} \in \mathbb{R}^{600\times600}$ is a random matrix and sgn denotes the sign function. The final error matrix $\mathbf{O}$ is formed by setting to zero those entries of $\mathbf{E}$ whose indices i and j satisfy the rule $\mathcal{N}[i,j] \leq 0.8$, where $\mathcal{N} \in \mathbb{R}^{600\times600}$ is a matrix whose elements follow the Normal distribution. The final corrupted signal $\tilde{\mathbf{Y}} = \mathbf{Y} + \mathbf{O}$ and the low-rank/sparse reconstruction produced by the DICA (3.5) can be seen in Fig. 3.4b and Fig. 3.4c, respectively. It is evident that our method reconstructs accurately all components, both the low-rank components lying in the background and the sparse components appearing as shapes, while, at the same time, isolates the sparse, gross errors. Quantitative results are reported in Table 3.1, in terms of normalized reconstruction error for each component, that is, $\|\mathbf{X}^{(i)} - \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}\|_F/\|\mathbf{X}^{(i)}\|_F$. It is worth noting that all subspace-specific reconstruction errors along with the clean signal reconstruction error $\|\mathbf{Y} - \sum_{i=1}^{6}\mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}}\|_F/\|\mathbf{Y}\|_F$ have low value, corroborating the conclusions drawn for our method from the qualitative inspection of Fig. 3.4.

3.4.2 Joint Face & Expression Recognition on CK+ Dataset

Our method is evaluated on the two-label setting of joint face and expression recognition. CK+ [130] has been widely used for the task of face and posed expression recognition. It contains 123 subjects in a total of 593 sequences, 327 out of which are annotated with respect to the emotion portrayed. As our method does not consider the temporal dimension, only the last 4 frames are used as expressive images for each sequence, as those are close to the apex phase of the expression. The experimental setup is identical to that of [211]. Specifically, a subset of 25 subjects, corresponding to 108 sequences, is used herein that meet the following criteria: (a) there are at least 4 annotated sequences for each of them, and (b) they perform one of the 6 universal emotions[2](*Anger, Disgust, Fear, Happiness, Sadness and Surprise*). The first condition is essential in order for the subjects to appear with a sufficient amount of images in the training set (at least 12 images), and the resulting dictionary to be balanced (for the face recognition part). Example images for a female subject of CK+ can be seen in Fig. 3.3.

To examine how image dimensionality affects accuracy in both face and expression recognition and tune it accordingly, the following experiment is conducted. Specifically, the DICA and SRC are tested on joint face and expression recognition with the image resolution varying through the range 32×32, 40×40, 48×48 and 56×56 pixels. Note that all images have been previously converted to gray scale and aligned based on the location of the eyes. For each subject, 3 sequences are randomly picked to be used for training, leaving the rest for testing. The parameters of the DICA and SRC are optimized separately for each resolution and the best accuracy obtained is reported in Fig. 3.5. The choice of 32×32 pixels for the image size consistently leads to the best performance. This behavior was expected as by using a smaller image size the *curse of dimensionality* is avoided (given that no feature extraction is performed to the aim of dimensionality reduction). It is also worth mentioning that using a smaller resolution for the DICA has the additional benefit of speeding-up the convergence, which increases quadratically with the dimensionality owing to the SVT operator (see Section 3.2.2). Accuracy achieved using the three remaining resolutions does not vary largely. In view of the above, the image size is fixed to 32×32 pixels for all experiments of this section.

For joint face and expression recognition, for each subject, 3 sequences are randomly selected to be used for training, and the remaining sequences are used for testing. This process is repeated 10 times, and the average scores for the face and expression recognition tasks are reported. Leave-one-subject-out expression recognition experiments are also conducted and the average rate over 25 folds is reported. For all experiments, parameters $\lambda^{(i)}$ controlling the nuclear norm of the identity-related $\mathbf{V}^{(i)}$ in Problem (3.5) are set to 1. For joint face and expression recognition, the values for λ_1 and the expression-related $\lambda^{(i)}$ accounting for the best average score over the two tasks were found to be 10^{-2} and 10^{-2}, respectively. For expression recognition, the corresponding values were 10^{-2} and $5 \cdot 10^{-2}$, respectively.

Recognition rates for both tasks are reported in Table 3.2. The merits of the DICA for face and expression recognition are directly evident from Table 3.2: it is the best-performing method for both tasks, yielding face and expression recognition accuracies of 96.7% and 83.6%,

[2] 18 sequences depicting ‘Contempt’ are not included.

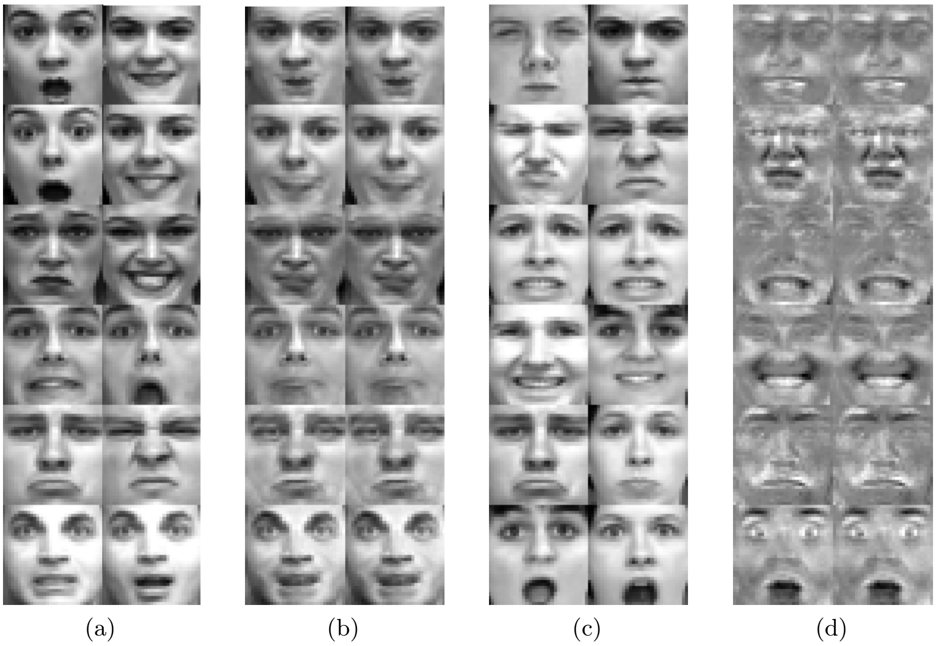


Figure 3.6: Joint Face and Expression Recognition on the CK+ Database: (a) Training images from six subjects showing various expressions, (b) Low-rank reconstruction produced by the DICA for each identity class, (c) Training images from six expression classes (from top to bottom: *Anger, Disgust, Fear, Happiness, Sadness, Surprise*) posed by various subjects, (d) Sparse reconstruction produced by the DICA for each expression class.

respectively[3]. LRSI comes second in performance, by a negative margin of 3.8% and 8.1% for face and expression recognition, respectively. Surprisingly, LRC provides scores close to those obtained by RPCA+SRC, presumably due to the beneficial effect of small training size and the similarity between training and test data populations. It is worth stressing that results of the DICA and RPCA+SRC correspond to the same sparsity parameter $\lambda_{Lasso}/2$ being used for the two dictionaries in (3.15). We believe that by separately optimizing the sparsity parameters for the SRC coefficients of identity and expression classes, that is, $\boldsymbol{\alpha}_{\mathcal{I}}$ and $\boldsymbol{\alpha}_{\mathcal{E}}$, respectively, one can achieve even higher performance.

Our method achieves the best score of 75.7% in the second setup also, where facial expression is recognized on data from subjects unseen in the training phase. LRSI is again the second-best-performing method with 71.4%. SRC performs poorly in this setup too, primarily due to test images being associated with sparse linear combinations of similar faces rather

[3] The recognition scores obtained for the dictionary-based component separation (DCS) algorithm from [211] are 99.1% and 81.6% for joint face and expression recognition, respectively, and 86.8% for subject-independent expression recognition. These results are only to some extent comparable to those reported in Table 3.2, given that the dataset and protocol are identical. However, bear in mind that in [211], K-SVD [1] is also applied to refine the identity and expression dictionaries, which are initially provided by RPCA [32]. For this reason, the corresponding results are not considered in the discussion of this section.

Table 3.3: Recognition Rates (%) for Protocol 1 (Sunglasses) and Protocol 2 (Scarf) with varying percentage of occluded images ($n_o/7$) in the AR Database training set.

Method	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf
	$0\% = 0/7$		$14\% = 1/7$		$29\% = 2/7$		$43\% = 3/7$	
LRC [142]	61.3	59.5	69.2	66.7	72.9	73.3	73.3	73.2
SRC [243]	72.3	71.4	82.4	83.3	88.6	89.3	88.9	90.1
RPCA+SRC [237]	75.4	85.0	81.6	89.4	87.7	90.7	88.8	87.3
LRSI (reported in [237])	73.0	72.8	84.2	82.6	83.7	80.5	83.7	79.6
DICA	85.9	88.3	93.5	94.4	93.4	94.0	93.3	93.1

Table 3.4: Recognition Rates (%) for Protocol 3 (Sunglasses+Scarf) with varying percentage of occluded images ($2n_o/(7 + 2n_o)$) in the AR Database training set.

	Sunglasses+Scarf			
Method	$0\% = 0/(7 + 0)$	$22\% = 2/(7 + 2)$	$36\% = 4/(7 + 4)$	$46\% = 6/(7 + 6)$
LRC [142]	59.9	66.2	69.1	70.3
SRC [243]	71.6	82.1	89.0	90.5
RPCA+SRC [237]	72.5	86.3	90.8	93.1
LRSI (reported in [237])	62.8	80.8	81.8	82.8
DICA	81.8	93.8	95.2	95.4

than similar expressions in the dictionary.

Fig. 3.6 illustrates the low-rank identity-based reconstruction (Fig. 3.6b) and the sparse expression-based reconstruction (Fig. 3.6d), as produced by our method for the joint face and expression recognition experiment on CK+ images, grouped by subject (Fig. 3.6a) and by expression (Fig. 3.6c), respectively. Note that no expression variations are retained in the subject-based reconstruction, while, at the same time, the sparse expression components contain no subject-related information. It is also worth observing that the expression components (Fig. 3.6d) are 'denser' and also account for higher values in the image regions where the action units 'shaping' each corresponding expression lie [220] (e.g., Brow-Lowerer AU4 for 'Anger', or Lip Corner Depressor AU15 for 'Sadness'). Overall, the resulting reconstructions are discriminant for both tasks.

3.4.3 Face Recognition on AR Dataset

For the task of face recognition, the focus of experiments is to investigate methods' performances for varying percentage of face images corrupted due to occlusion in the training set. This is a frequently-occurring scenario in real-world biometrics applications, where noise-free training data is hard to be attained (e.g., due to uncontrolled recording conditions and huge amount of data). To this end, the AR Database [135] is used, which includes a total of 4,000 frontal images for 126 individuals. The face images exhibit variations with respect to expression, illumination and two types of occlusion, that is, sunglasses and scarf (see Fig. 3.3). For each subject, images are taken in two sessions, each one constituent of 13 images: 3 images with sunglasses, 3 with scarves, 4 with different expressions, and the remaining 3 with different illuminations. The latter 7 images, which do not include occlusions, are considered as neutral images for the experiments in this section.

A randomly picked subset of 100 subjects is used for our experiments. Three protocols are tested in an identical way as in [237], corresponding to occlusion in the training images due to (1) sunglasses, (2) scarf, and (3) sunglasses and scarf, respectively. Note that sunglasses account for occlusion of about 20% of the face image, whereas for the scarf scenario this percentage amounts to about 40%.

The three protocols are outlined below:

- **Protocol 1:** For each subject, n_{cl} neutral images and $n_o \in \{0, 1, 2, 3\}$ occluded images (sunglasses) from Session 1 are used for training, where $n_{cl} + n_o = 7$. 7 neutral images and 3 occluded images (sunglasses) from Session 2 are used for testing.
- **Protocol 2:** Same as Protocol 1, with occluded images containing scarf rather than sunglasses.
- **Protocol 3:** For each subject, $n_{cl} = 7$ neutral images, $n_{sg} \in \{0, 1, 2, 3\}$ sunglasses images, and $n_{sc} \in \{0, 1, 2, 3\}$ scarf images, from Session 1 are used for training, where $n_{sg} = n_{sc}$. Here, the amount of training images per subject varies from 7 to 13, as opposed to the first two protocols, in which it is fixed to 7. All 13 images (7 neutral, 3 sunglasses, 3 scarf) from Session 2 are used for testing.

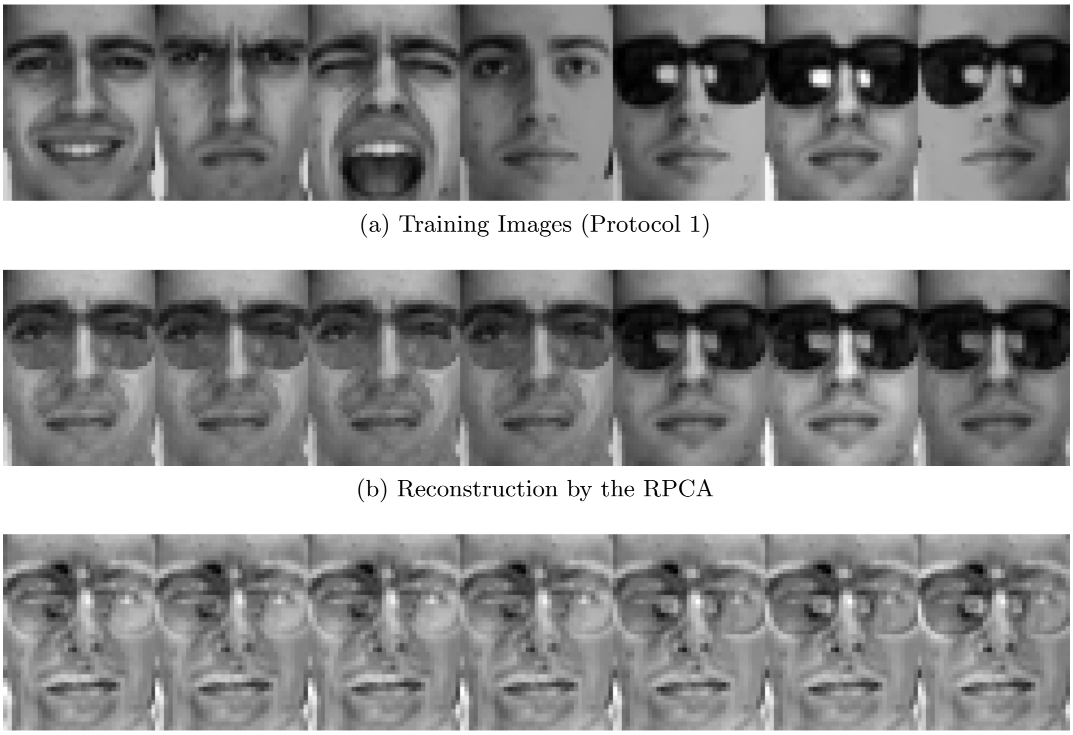
Results are shown in Table 3.3 for Protocols 1 and 2, and in Table 3.4 for Protocol 3. The DICA achieves the most accurate recognition in all scenarios, reaching 95.4% accuracy in Protocol 3 when 46% of training images are corrupted. The value of parameter λ_1 that yielded the best scores for our method was 10. It is worth noting that all methods show a significant increase in performance in all three protocols when at least one occluded image per subject is included in the training set, as compared to the case of 100% clean data. Notably, the performance achieved by the DICA fluctuates less as the percentage of training set corruption increases, as compared to that of the other methods. This is because components produced in the output of the DICA are by definition mutually incoherent, regardless of how many images with similar corruptions in similar face regions across classes are used for training. In Protocol 3, where two different kinds of data corruption are present, RPCA+SRC consistently achieves the second-best accuracy. It is also worth observing that, even for large percentages of training set corruption, SRC performs quite accurately also. This can be attributed to the efficiency of SRC in scenarios where the training and test set distributions are characterized by similar variations [257]. LRSI shows poor performance possibly due to its inability to suppress the effect of occlusion in the generated subspaces. LRC underperforms the rest of the methods in all cases. This can be largely attributed to singularities occurring in the matrix $\mathbf{D}^{\mathbf{T}}\mathbf{D}$, where $\mathbf{D}$ is the dictionary matrix (see [257, 142]).

In Fig. 3.7, the performance of our method and RPCA is comparatively illustrated on an instance of Protocol 1, that is, 7 images of a male subject, 3 of which are occluded by sunglasses. One can observe that both methods successfully remove variations caused by expression or illumination in the derived low-rank reconstruction. Nonetheless, our method succeeds to discard the occlusion in the reconstruction images, as opposed to the RPCA. This is due to the fact that presence of sunglasses in the reconstructed images of all subject classes would clash with the mutual incoherence property, which entails that class-specific components are as close as possible to being orthogonal. The same holds for Protocol 2, where the occlusion due to scarf covers even larger part of the image. Reconstructions yielded by our method for images of the same subject in Protocols 1 and 2 are shown in Fig. 3.8, for the scenario in which the occluded images cover 3/7 of the training set.

3.4.4 Expression Recognition on CMU Multi-PIE Dataset

In Section 3.4.2 we presented expression recognition experiments for the case of different subjects being included in the training and test set. Aiming to evaluate the effectiveness of our method in a scenario where labels from an additional source of variation, such as illumination, are not utilized in our discriminant analysis during training, we perform expression recognition also on the CMU Multi Pose Illumination, and Expression (Multi-PIE) Database [80]. This dataset contains 337 subjects, corresponding to about 750,000 images with 19 illumination variations, 15 different poses, and 6 facial expressions (*Neutral, Smile, Surprise, Disgust, Scream, Squint*). In the current study, only the frontal pose images are considered. For the presented experiments, 50 subjects are randomly selected. For each subject, 5 different illumination conditions are generated (corresponding to pan angles $-30°$, $-15°$, $0°$, $15°$, $30°$) for all 6 expressions, resulting in 30 images per subject. Some characteristic images from Multi-PIE are illustrated in Fig. 3.3.

The same protocol used in Section 3.4.2 is adopted for facial expression recognition. Subject-independent experiments are conducted and the average score over 50 runs is reported. The best values for the sparsity-controlling parameters λ_1 and $\lambda^{(i)}$ for the expression components



(a) Training Images (Protocol 1)

(b) Reconstruction by the RPCA

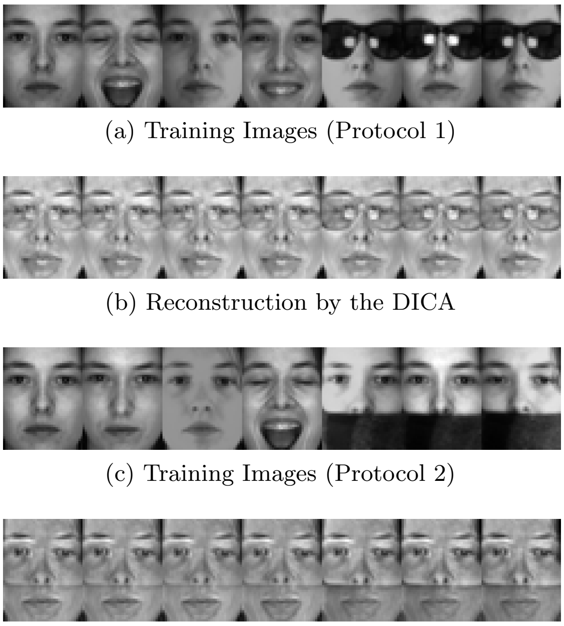
(c) Reconstruction by the DICA

Figure 3.7: Face Recognition on the AR Database: Reconstruction images, as produced by the RPCA (b), and the DICA (c), on all training images of a subject in Protocol 1 (3/7=43% of occluded images (sunglasses)) (a).

were found to be 10 and 1, respectively. Recognition rates are reported in Table 3.5. Here, illumination conditions vary a lot across training images, rendering the task even more challenging. Still, our method achieves the best accuracy of 74.4%, followed by LRSI that achieves 67.3%. RPCA+SRC and SRC perform rather similarly, meaning that RPCA preprocessing fails in this case to uncover the class-specific low-rank manifolds. Note also that LRC shows a surprisingly poor performance. Again, the DICA efficiently decouples expression-related deformations from subject-specific characteristics and other effects, thereby enabling us to construct a much more discriminative expression dictionary.

3.4.5 Facial action unit detection on GEMEP-FERA Dataset

In this section, the efficiency of the DICA in decomposing an expressive image into mutually incoherent sparse components related to AUs is examined. The training subset of the GEMEP-FERA [219] dataset is used for subject-independent action unit detection experiments. It contains 7 subjects depicted in 87 image sequences, which are FACS-labeled on a frame-by-frame basis in terms of AUs. Herein, we use only the images in which at least one out of 8



(a) Training Images (Protocol 1)

(b) Reconstruction by the DICA

(c) Training Images (Protocol 2)

(d) Reconstruction by the DICA

Figure 3.8: Face Recognition on the AR Database: Reconstruction produced by the DICA ((b),(d)), on all training images of a subject in Protocols 1 and 2 (3/7=43% of occluded images - sunglasses in (a) and scarf in (c), respectively).

Table 3.5: Recognition Rates (%) for Subject-Independent Expression Recognition on Multi-PIE Dataset.

Method	Expression Recognition
LRC [142]	18.0
SRC [243]	58.9
RPCA+SRC [237]	60.4
LRSI [237]	67.3
DICA	74.4

action units is activated. The AUs considered are: AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU12 (Lip Corner Puller), AU15 (Lip Corner Depressor), and AU17 (Chin Raiser). Images are converted to gray scale, aligned based on the location of the eyes, and, subsequently, resized to 128× 128 pixels. Characteristic images are shown in Fig. 3.3. Intensities from 22×22 pixel patches around 15 facial points (extracted by the tracker in [190]) are gathered in a single vector for each image. The final feature vector is composed of PCA coefficients corresponding to components that account for 98% of the total variance (374 components in our experiments).

Seven-fold subject-independent cross-validation is performed, so that all images for the 7 subjects are tested. For each fold, a randomly selected subset of the training images, evenly distributed across subjects and AU labels, is used. For the DICA, the action unit detection framework described in Section 3.3.2 is used. Specifically, the rank $m^{(i)}$ of each subspace is set to 5, while the remaining parameters are optimized similarly to the previous experiments. The values of the sparsity-controlling parameters λ_1 and $\lambda^{(i)}$ accounting for the best performance were found to be 0.05 and 1, respectively.

Except for the DICA, LRC and SRC are also examined, while RPCA+SRC and LRSI are not considered, as their design is not adaptable to this task. Multi-Label k-Nearest Neighbours (ML-kNN) [258] ($k = 10$ neighbours) and Rank-SVM [65] (with polynomial kernel of degree 8) are also examined, as they are general-purpose algorithms for multi-label classification. For the DICA, each dictionary atom is associated with a single AU label (see Section 3.3.2), as opposed to other methods, for which the training data retain their initial multi-class labelling. For the dictionary-based methods, namely the DICA, LRC and SRC, ROC ranking [67] is employed to threshold the class-specific *confidence* scores obtained by (3.18) and thus provide multi-class predictions for each test sample. Finally, for all algorithms examined in the experiments of this section, the $F1$ score, defined as $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$, is used as the evaluation metric.

Action unit detection results in terms of $F1$ score, as produced by each method, are reported in Table 3.6 for each action unit along with the average performance over all AU classes. For comparison purposes, we choose to also include in Table 3.6 the results reported in [40] for the same evaluation protocol for Selective Transfer Machine (STM), which is a recently published successful method for AU detection. The DICA achieves similar performance to that of STM[4], while it outperforms all other methods. SRC also achieves high performance, thus validating previous evidence that sparse representation is efficient for the AU detection task [132]. LRC, as well as the baseline methods ML-kNN and Rank-SVM, attain much poorer performance.

3.5 Conclusion

A method for recovering mutually incoherent and structured components in *face imagery*, relying on discriminant information as well as structure-inducing norms on the facial aspects, has been proposed in this chapter. An ADMM-based algorithm that can solve appropriate minimization problems for the DICA, according to the matrix norm imposed, while also being robust to gross outliers through sparsity regularization, has been also proposed. Finally, a dictionary-based framework that combines the DICA with sparse representation to jointly

[4] The difference in average performance over all AUs achieved by the DICA and the STM is not significant, according to a paired t-test at significance level 0.05.

Table 3.6: $F1$ Scores (%) for each action unit and method examined in the action unit detection experiments on the GEMEP-FERA Database training set.

AU	ML-kNN [258]	Rank-SVM [65]	LRC [142]	SRC [243]	STM [40]	DICA
1	53.8	67.0	37.7	60.5	68.1	66.3
2	41.6	46.3	47.2	58.3	65.5	58.9
4	20.3	20.4	61.5	58.3	43.3	55.5
6	62.2	68.2	57.1	63.9	71.6	70.2
7	53.7	61.9	66.2	67.8	66.2	70.6
12	75.8	77.7	75.8	76.9	82.1	78.0
15	28.1	44.8	46.2	30.1	39.3	41.0
17	39.3	38.0	18.6	37.8	35.9	32.0
Avg.	46.9	53.0	51.3	56.7	59.0	59.1

address interrelated classification tasks within multi-label scenarios has been presented. The experimental validation of our method was primarily focused on face analysis tasks. The effectiveness of the DICA was first demonstrated on synthetic data contaminated with sparse, non-Gaussian noise. Next, extensive experiments were conducted on joint face and expression recognition, face recognition for varying percentages of corrupted images in the training set, subject-independent expression recognition under varying illumination conditions during training, as well as facial action unit detection. The DICA outperformed all methods that were used for comparison, in all tasks and experimental scenarios.

Overall, the DICA is a robust learning framework that can generalize to classification of any number or type of labeled attributes that manifest themselves in the visual stream through specific structures, associated with mutually incoherent modes of variation. Generally speaking, it can be subsumed under the category of supervised subspace learning methods for structured component extraction from high-dimensional data. Viewing it from the viewpoint of static face analysis, the DICA serves as a robust convex model that yields class-specific low-dimensional representations whose reconstruction in the pixel intensity domain render dictionary learning and sparse-based recognition of facial attributes more efficient compared to using raw pixel intensities. This was experimentally corroborated above, where the superiority of the DICA over other SRC-like frameworks was evidenced. An additional advantage of the DICA over intensity-based representations and other subspace learning algorithms is its ability to disregard sparse corruptions of large magnitude affecting more than one classes, whose presence in the feature representation would otherwise be largely harmful for the subsequent recognition tasks. For instance, in the scenario where ones wishes to perform face recognition with an uncontrolled training set that includes images with sunglasses for multiple subjects, an algorithm like PCA would include the sunglasses in the derived low-dimensional representations of the training images due to them accounting for large variation. On the contrary, the DICA disregards these shared corruptions since their presence in the derived coefficients would clash with the imposed mutual incoherence property, thus leading to corruption-free, hence more discriminative representations. However, it is worth mentioning that the DICA has been designed to work well only when applied to well-aligned, pose-free facial images. In preliminary experiments with not aligned or badly aligned images showed that pose variability and misalignment errors can be detrimental to the decoupling of structured components performed by the DICA, as this information is undesirably assigned to class-wise components designed to contain only face- or expression- related information. Furthermore, the DICA has been envisioned, designed and developed as a robust matrix decomposition framework that operates on the pixel intensity domain – thus circumventing the necessity of feature extraction – for the extraction of facial components having an known, interpretable structure that can be enforced by an appropriate structure-inducing norm such as low-rank expression-less face components or components of sparse expressions and region-specific, sparse AU activations. Nevertheless, we showed experimentally that the DICA can derive discriminative representations even when not applied on pixel intensities but rather on a different domain like that of PCA coefficients, as was the case in the above presented AU detection experiment. In other words, even the induction of sparsity constraint on the localized, AU-specific PCA features is not directly interpretable to the human eye, we saw that the DICA can achieve state-of-the-art recognition results. This makes us believe that the DICA is versatile in the sense that it can be applied equally or more efficiently on images pre-processed with other handcrafted features like LBP [3] or [51], with the purpose of deriving class-specific components corresponding to local intensity or orientation information, respectively, to mention but a few. In a similar fashion, it would be interesting to explore how the DICA decomposition would perform in representations derived from deep convolutional neural networks (CNNs) such as the well-known VGG-16 model [204]. In this way, the DICA could drive deep learning features derived in an unsupervised way to class-specific mutually incoherent manifolds, either in a unified or an alternating two-step optimization procedure. Another interesting direction that could be explored, in the intersection space of deep learning and component analysis, would be to combine these components in the exactly opposite way. In other words, one could use the DICA as a pre-processing step that would remove sparse corruptions affecting one or multiple classes and other irrelevant for the task variations before going on to discover more fine-grain image properties with CNNs, either in one pass or multiple alternating passes. By means of this combined framework that combines the best of the two worlds, we believe that one could achieve higher recognition performance that just employing CNNs on pixel intensities, since in the former case the bulk of the multi-scale feature detection process carried out by CNNs would be focused on already discriminant and corruption-free images.

Overall, the DICA is an efficient component analysis algorithm that is quite flexible with the respect to the tasks that it can address but also the learning pipelines of which it can form part when applied to face analysis tasks. Having provided a generic learning method that can jointly address recognition of intertwined facial attributes represented by categorical descriptions in still face images, in the subsequent chapters of this thesis we turn our attention to continuous-time modeling of affective and behavioral displays represented by dimensional descriptions.

Chapter 4

The Conflict Escalation Resolution (CONFER) Database

Contents

We highlighted in Section 2.2 that temporal modeling of human non-verbal affect and behavior cannot be approached through *categorical* descriptions, that is, non-verbal expressions in terms of basic emotion categories (e.g., happiness, sadness, fear) or discrete social states (e.g., agreement/disagreement, conflict/non-conflict), respectively. Instead, modeling transitions between moderate and naturalistic affective and behavioral displays necessitates the use of *dimensional* descriptions, where affective and social states are characterized in terms of latent dimensions taking real values as a function of time. We also identified that continuous-scale characterizations of high-level semantic social behaviors such as interest, politeness, flirting, (dis)agreement, and conflict, are rarely adopted by social signal processing methodologies. As made clear in Section 2.3, this gap in the literature is to a large extent attributed to an overall lack of suitable annotated data that could be used to train learning algorithms for recognition of social phenomena at a finer granularity. In this chapter, we provide a remedy to this problem by providing a new database suitable for the investigation of a social attitude, namely conflict, in continuous scale and time.

4.1 Introduction

Conflict is used to label a range of human experiences, from disagreement to stress and anger, occurring when involved individuals act on incompatible goals, interests, or intentions over resources or attitudes [5, 99]. Various research studies in human sciences argue that a "disagreement" does not have to result in a conflict; conflict describes a high level of disagreement, or "escalation of disagreement", where at least one of the involved interlocutors feels emotionally offended. Similarly to other phenomena arising in social interactions [228, 161], conflict is largely manifested by means of non-verbal behavioral cues including facial expressions, body postures, gestures, and head movements, as well as conversational social signals including interruptions, overlapping speech, loudness and other cues associated with turn-organization [44]. Conflict, which has been recognized as one of the main dimensions along which a dyadic or multi-party social interaction is perceived, is usually accompanied by negative effects on communication and social life [116]. Hence, automatic analysis of conflict can be a cornerstone in the deployment of technologies targeting social interactions understanding and social skills enhancement such as content-based multimedia indexing and retrieval, machine-mediated communication, socially intelligent human-computer interfaces, to mention but a few.

Although conflict has been extensively investigated in human sciences, it has not received the same level of attention by the computing community. In spite of recent advances in social signal processing [160, 228, 161] and machine analysis of cues related to social behaviors [36, 53, 103, 127, 165, 255, 83], research on machine analysis of conflict is still limited to just a few works that target automatic conflict detection based on audio features [106, 107, 108] or (dis)agreement detection [27, 25, 71]. As already mentioned, this is primarily due to the lack of suitable databases (for an overview on existing databases that have already been or could be used for automatic analysis of conflict and similar social signals and phenomena in dyadic or multi-party conversations, the reader is directed to Section 2.3). Furthermore, given that interpersonal conflict is a mode of dyadic or multi-party interaction, automatic analysis of conflict is by itself a difficult task in terms of machine learning effort, since it requires the simultaneous analysis of more than one subjects at the same time. Also, the particularities of non-verbal communication due to conflictual conversation pose additional challenges to the related audio signal processing and computer vision tasks. For instance, interruptions and overlapping speech are more frequent when conflict takes place, which can largely affect the accuracy of speaker diarization or subsequent stages of audio feature extraction. When the visual modality is also considered, irregular postures or frequent and intense head and hand movements can lead to increased levels of visual noise pertaining to missing and incomplete

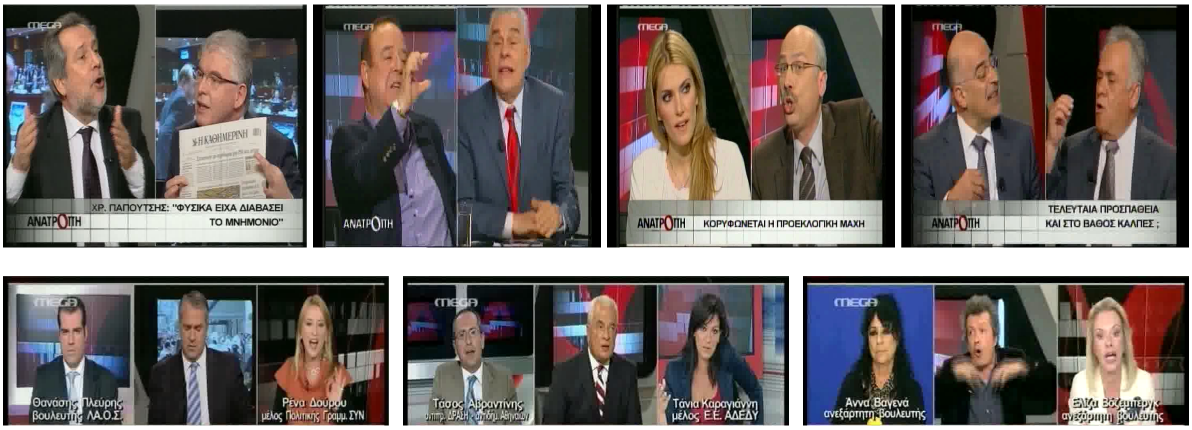

MEGA
Η ΚΑΘΗΜΕΡΙΝΗ
ΧΡ. ΠΑΠΟΥΤΣΗΣ: "ΦΥΣΙΚΑ ΕΙΧΑ ΔΙΑΒΑΣΕΙ ΤΟ ΜΝΗΜΟΝΙΟ"
ΑΝΑΤΡΟΠΗ
ΚΟΡΥΦΩΝΕΤΑΙ Η ΠΡΟΕΚΛΟΓΙΚΗ ΜΑΧΗ
ΤΕΛΕΥΤΑΙΑ ΠΡΟΣΠΑΘΕΙΑ ΚΑΙ ΣΤΟ ΒΑΘΟΣ ΚΑΛΠΕΣ ;
Θανάσης Πλεύρης βουλευτής ΛΑ.Ο.Σ.
Ρένα Δούρου μέλος Πολιτικής Γραμμ. ΣΥΝ
Τάσος Αβραντίνης αντιπρ. ΔΡΑΣΗ - αντιδημ. Αθηναίων
Τάνια Καραγιάννη μέλος Ε.Ε. ΑΔΕΔΥ
Άννα Βαγενά ανεξάρτητη βουλευτής
Έλσα Βόζεμπεργκ ανεξάρτητη βουλευτής

Figure 4.1: Characteristic frames from episodes of the Set *two* (top row) and *three* (bottom row) of the CONFER Database.

data (e.g., partial image texture occlusions) or feature extraction errors (e.g., incorrect object localization, tracking errors).

Overall, previous works on automatic conflict analysis are characterized by the following main limitations.

- They are evaluated on corpora containing conversations that are captured in controlled, simulated conditions or on pre-segmented episodes of conflict/non-conflict.
- They are based exclusively on the audio modality (e.g., prosodic, conversational features), such as the works of Kim et al. [106, 107, 108], who investigated the degree of conflict in broadcasted political debates. The only audio-visual approach to conflict detection that we are aware of is [155], where robust, multi-modal fusion of audio-visual cues is utilized.
- They only deal with conflict detection or conflict escalation/resolution detection. These are approached as classification tasks aiming at estimating a single binary label (conflict/non-conflict) or discrete conflict intensity levels for the entire sequence or segments of it. The only work in the literature – that we are aware of – that has approached conflict in dimensional rather than categorical terms, i.e., as a continuous (real-valued) variable, and conflict intensity estimation as a regression task is [108].

In this chapter, we provide a comprehensive description of the *Conflict Escalation Resolution (CONFER) Database*, a collection of audio-visual recordings of naturalistic interactions from political debates suitable for the investigation of conflict behavior. These recordings have been manually extracted from more than 60 hours of live political debates, televised in Greece between 2011 and 2012. In contrast with other corpora, political debates are real-world competitive multi-party conversations where participants do not act in a simulated context, but participate in an event that has a major impact on their real life (for example, in terms of results at the elections) [106]. Consequently, even if some constraints are imposed by the debate format, the participants have real motivations leading to real conflicts.

From the entire dataset, 120 video excerpts have been extracted from a total of 27 TV broadcasts, with total duration amounting to approximately 142 minutes. The dataset is split into 2 sets, namely *two* and *three*, which consist of recordings containing interactions that involve two or three participants, respectively. All 120 videos have been annotated by 10 experts, in terms of continuous conflict intensity. The CONFER Database has been partially presented at previous works (see [158, 155, 253]), but a complete description of the data and available annotations, has not been reported so far. The database is publicly available for non-commercial use at `http://ibug.doc.ic.ac.uk/resources/confer/`. Along with the audio-visual episodes and the annotations, audio and visual features (facial tracking points and SIFT) are also provided (see Section 4.3).

This work is novel not only in providing a comprehensive description of this database, which is suitable for the investigation of conflict behavior in naturalistic conversations, but also in reporting baseline experiments that could serve as a benchmark for efforts in the field. These experiments primarily aim to overcome the last two of the aforementioned limitations of previous works on automatic conflict analysis, namely by i) examining both audio and visual features as well as fusion of them for the target problem, and ii) addressing *continuous-time* (frame-by-frame) estimation of *continuous-valued* conflict intensity. For each Set of the database, we conduct two baseline experiments in which the efficiency for the problem at hand of various visual (shape- and appearance-based) descriptors and audio features as well as fusion of them, and classifiers, respectively, is examined. A cross-validation experimental scenario is employed in order to assess performance of the baseline predictive frameworks on collectively all audio-visual recordings of the CONFER Database. A challenging experimental protocol is established with all experiments being subject- and session-independent. This is to ensure that the sequences used for testing involve different subjects from different TV broadcasts compared to those used in the training phase.

The remainder of this chapter is as follows. Section 4.2 presents in detail the audio-visual data and conflict intensity annotations included in the CONFER Database. Section 4.3 describes the methodology employed for the baseline experiments on continuous conflict intensity estimation

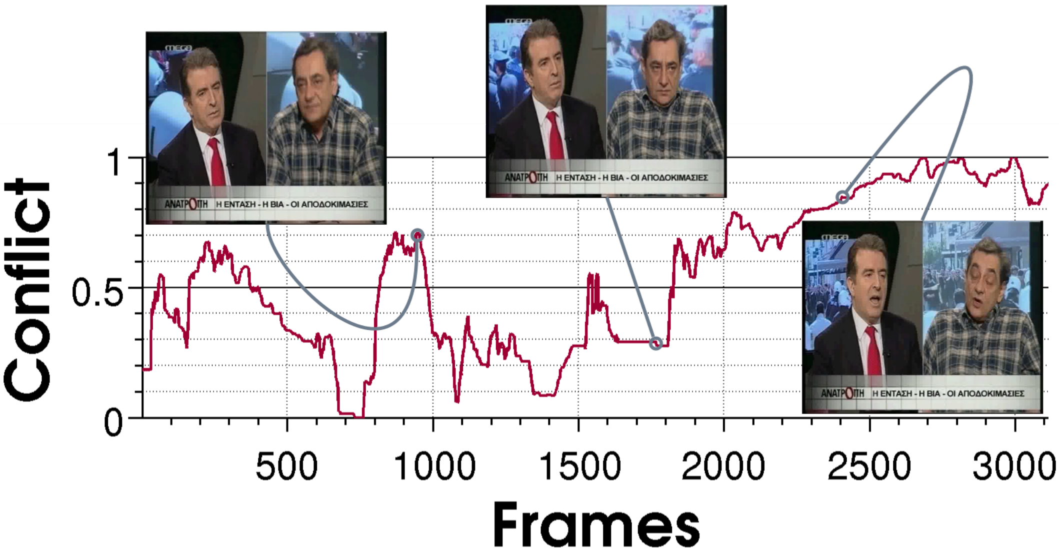


Figure 4.2: Conflict intensity annotations along with three characteristic frames shown for the sequence *20120326˙seq3* from the Set *two* of the CONFER Database.

that are presented in Section 4.4, while Section 4.5 concludes the chapter.

4.2 Database

In this section, we provide a comprehensive description of the CONFER Database, a collection of audio-visual recordings of naturalistic interactions from political debates.

Data. The database consists of video excerpts from televised political debates in Greek language. In particular, it contains episodes of conflict escalation and resolution, which have been extracted from more than 60 hours of live political debates aired as a part of the Anatropi Greek TV show[1]. Each debate includes at least two guests discussing under the moderation of the TV host.

From the entire collection of the TV programme broadcasts, 120 non-overlapping episodes of conflict escalation have been manually extracted. These audio-visual excerpts are divided into two Sets, which are balanced in terms of total duration, namely the Set *two* that contains 73 episodes of dyadic interactions, and the Set *three* that contains 47 episodes of interactions among three subjects. Overall, these episodes correspond to a total duration of approximately 142 minutes and to a total number of 54 subjects, 43 male and 11 female. It is worth mentioning that the episodes contain debates that may have more

[1] `http://www.megatv.com/anatropi/`.

than one instances of conflict escalation, yet they always end with conflict resolution. For all recordings, the video stream has been recorded at 25 frames per second, while the sample rate of the audio channel is 22050 Hz. Each video sequence of the dataset has a spatial resolution of 720×576 pixels and has all participants involved in the episode in view. The duration of the episodes varies from 20.2 seconds to 534.0 seconds, having a mean and standard deviation of 71.0 and 70.5 seconds, respectively, as computed for the whole dataset. Characteristic frames from the dataset are depicted in Fig. 4.1.

Due to the spontaneous and competitive nature of the interactions contained in the CONFER Database, various types and levels of noise are incurred in the data. Regarding the audio channel, speaker diarization and speech recognition are rendered difficult since the interlocutors often interrupt or talk over one another, driven by anger or agitation or aiming to dominate the dialogue. In some of the recordings, a third party speaking in the background is involved. Also, in most of the cases speech is emotionally colored and thus often fragmented and disorganized or extremely rapid and even unintelligible.

Regarding the visual stream, camera angles can vary a lot across episodes or even within the same episode, while illumination conditions vary less. Depending on the way the interlocutors are positioned in the studio, the former are often portrayed at large head pose *pan* angles or even in almost-profile view, due to them looking at their interlocutor rather than the camera fixed on them. Moreover, due to the involved parties being engaged in naturalistic competitive conversations, the subjects often perform abrupt and extreme head movements (e.g., head nods, shakes, tilts), body movements (e.g., forward/backward leaning, spinning periodically on their swivel chairs) and gestures (e.g., hand crosses, hand wags). The aforementioned conditions pose obstacles to the computer vision pre-processing tasks, such as face detection, facial point tracking and registration [188, 189], since the latter have to cope with frequent and large out-of-plane head rotations and occlusions [168, 74].

Annotations. The data have been annotated on a frame-by-frame basis in terms of continuous (real-valued) conflict intensity by 10 expert annotators, all of them being native Greek speakers. The annotation task is carried out in real time, i.e., while the annotators are watching each audio-visual excerpt, by employing a joystick-based annotation tool. The tool records the conflict intensity level in the continuous range $[0, 1000]$ at a variable sampling rate, which is approximately 64 samples per second in average. All annotations are subsequently down-sampled to the video frame rate of 25 frames per second. The procedure followed so as to extract a single ground truth annotation sequence from the 10 available annotations for each episode of the CONFER Database is described in detail in Section 4.3.2. Ground truth annotations of conflict intensity are plotted as a function of time for a sequence of the database along with three characteristic frames in Fig. 4.2.

The annotators have been advised to annotate the videos by considering the *physical* (related to the behavior being observed) and the *inferential* (related to the interpretation of the discussion) layer of the conversation [106]. The *physical layer* includes the behavioral cues observed during conflicts and include interruptions, overlapping speech, cues related to turn organization in conversations as well as head nodding, fidgeting and frowning [44]. The *inferential layer* is based on the perception of the competitive processes occurring in conversations where conflict is viewed as a 'mode of interaction' governed by the principle that *"the attainment of the goal by one party precludes its attainment by the others"* [99, 71]. For instance, conflicting goals often lead to attempts of limiting, if not eliminating, the speaking opportunities of others in conversations. In view of the demanding nature of the task of annotating conflict in real time and in terms of both conversational layers, all annotators were initially 'trained' on a small subset ($\sim$10%) of the CONFER Database episodes. In particular, they were instructed to watch these episodes as many times as they considered necessary and retain the annotation that best assessed conflict intensity in terms of both layers. For each of the remaining episodes of the database, the annotators were allowed two plays, and again the most suitable annotation was retained.

4.3 Methodology

In this section, the methodology employed for the baseline experiments conducted on the CONFER Database for audio-visual continuous-time conflict intensity estimation is described.

4.3.1 Sets and Protocol

Two baseline experiments are conducted for each Set of the CONFER Database. A *subject- and session-independent cross-validation* experimental protocol is employed. Specifically, each Set is divided in 5 segments, balanced in terms of duration, containing videos that include different interactants and have been broadcast at different times. In each fold, 3 segments are used for training, one for validation (parameter tuning) and the remaining one for testing, and the average value over all test sequences of each evaluation metric (see Section 4.3.5) is retained. The process is repeated 5 times, until all episodes have been used for testing. Finally, the mean and standard deviation of the metrics, as computed over all 5 folds, are reported.

4.3.2 Annotations

Recent studies on combining multiple annotations of human behavior or affect have provided evidence suggesting that the average of multiple annotations can lie far away from the actual ground truth and thus lead to ill-generalizable models [146]. This is mainly due to the subjectivity of annotators and the variability related to their age and gender or their stress, fatigue, attention or even intention while annotating (e.g., there can be *spammer* annotators that they do not even pay attention during the annotation process). Furthermore, when the task in question is temporal, additional noise in the set of multiple annotations is entailed by the temporal lags in the perception and annotation of the related events.

Motivated by the aforementioned findings, herein we follow a supervised approach to fusing the multiple available annotations. Specifically, Canonical Correlation Analysis (CCA) [8] is employed for each sequence to extract subspaces that are maximally correlated for the set of 10 annotations available and the corresponding audio-visual feature set. For all experiments presented in this paper, the coefficient corresponding to the first component of the CCA-derived annotation subspace is used as the ground truth annotation for each episode. The latter is rescaled in the continuous range $[0, 1]$. Original annotations of conflict intensity from the 10 annotators as well as the CCA-derived annotation for a sequence of the CONFER Database are plotted as a function of time in Fig. 4.3.

4.3.3 Features

The various audio and visual features as well as fusion of them that are used in the experiments of this study are described in what follows.

Audio features. As mentioned above, most of the existing approaches to automatic conflict analysis have relied almost exclusively on audio features [106, 107, 108] such as spectral, prosodic, durational, lexical and turn organization descriptors. A concise review of audio-based approaches to (dis)agreement and conflict detection is provided in [108].

In this work, we employ the openSMILE feature extractor [66] to obtain the ComParE acoustic feature set of 65 low-level descriptors (LLD) (4 energy-related, 55 spectral and 6 voicing-related), which has been successfully applied for automatic recognition of paralinguistic phenomena [238]. The 65 LLD used are summarized in Table 3 in [185]. The audio features extracted for each sequence of the CONFER Database are down-sampled to 25 Hz frequency to match the frame rate of the video stream. Similarly to [145, 238] the audio features of each sequence are z-normalized (each feature component is normalized to mean=0 and standard

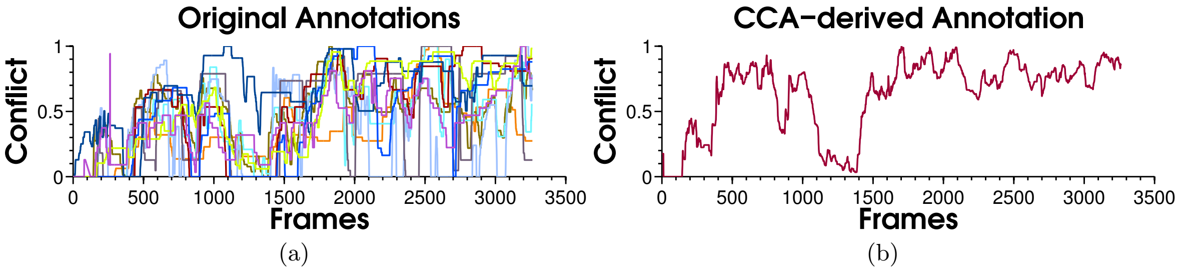


Figure 4.3: Annotations illustrated as a function of time for the sequence *20120206˙seq5* from the Set *three* of the CONFER Database. (a) Original annotations from 10 annotators rescaled in $[0, 1]$, and (b) Ground truth annotations derived by performing CCA on the original annotations and the corresponding features.

deviation=1).

Visual features. In recent years, research in behavioral and affective computing as well as signal processing has gradually shifted from audio-only (or even video-only) systems to audio-visual approaches [255, 83]. As a matter of fact, the latter have been shown to outperform uni-modal frameworks in various related tasks such as continuous interest prediction [148, 155], detection of behavioral mimicry [21], and dimensional and continuous affect prediction [145], to mention but a few. Notably, other challenging problems such as accent classification [75, 72, 73] and pain intensity estimation [101] have been addressed based exclusively on visual features.

Motivated by the aforementioned works and deviating from a common practice in automatic conflict analysis where only audio features are employed (e.g., [106, 107, 108]), in this paper we utilize also visual features for conflict intensity estimation. Our aim is to capture facial behavioral cues that are deemed intrinsically correlated with conflict, such as smiling, blinking, head nodding, flouncing and frowning [44, 25]. Both shape- and appearance-based descriptors are examined. Note that the video stream of each episode from the CONFER Database is spatially cropped at each frame so that a separate video stream is obtained for each one of the participants involved in the conversation. The Menpo project [4] has been employed in this study for all visual feature extraction tasks, which are described as follows.

Facial point tracking: First, 68 fiducial facial points are detected at each frame of each cropped video sequence portraying a single interactant. To this end, we employ the Gauss-Newton Deformable Part Model in [215], which when combined with a person-specific face detector produces very accurate results [39]. The coordinates of 49 facial landmarks are retained for each frame by excluding the facial points that correspond to the face boundaries. Next, the effects of head translation, scale and in-plane rotation are removed by universally aligning the

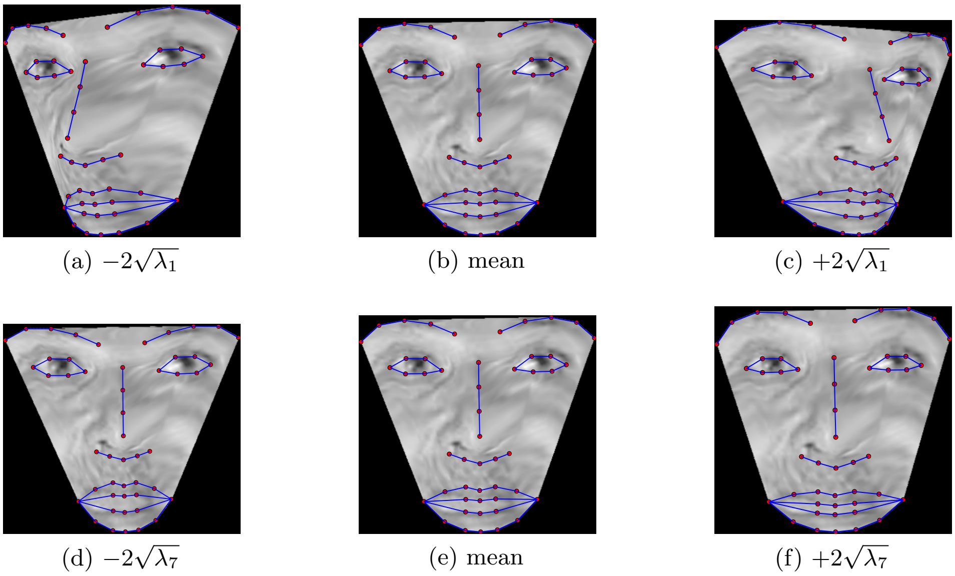


Figure 4.4: Effect on the mean shape ((b), (e)) of varying the 1st ($i = 1$) component (pose-related) and the 7th ($i = 7$) component (expression-related) of the Active Shape Model used herein for shape feature extraction by $-2\sqrt{\lambda_i}$ and $2\sqrt{\lambda_i}$, where λ_i denotes the respective eigenvalue.

tracking points with the 'mean' shape computed over all frames through a 2-D non-reflective similarity transformation.

Shape features: Principal Component Analysis (PCA) [98] is applied on the aligned tracking points to yield a low-dimensional shape descriptor for each frame. In particular, the coordinates of the 49 facial landmarks are projected onto the subspace spanned by the 'eigenshapes' of a pre-trained Active Shape Model (ASM) [45]. The latter has been previously trained on collectively 4 datasets of faces "in-the-wild", and thus its principal components efficiently 'explain' variations of shape corresponding, for instance, to out-of-plane rotations, different face anatomy characteristics and subtle expression-related deformations. For each video frame, 18 coefficients that account for 95% of the total variance are retained for each subject. The final feature vector for each frame is obtained by concatenating the descriptors for all interactants in the episode.

Inspired by [176], we follow a face-anatomy-driven rather than a simply data-driven approach to identifying the most suitable feature representation of facial shape for the problem at hand. To this end, we visually inspect the deformation pattern associated with each component of the ASM. We observe that the first 6 components capture head movements (rigid face motion), while the remaining 12 components capture expression-related deformations (non-rigid face motion). The discriminative power of both pose- and expression-related shape features as well as the combination of them – which we henceforth call *Pose*, *Expression* and *Points*, respectively – is investigated for the target problem. In Fig. 4.4, one can see the mean shape and the effect on it of varying the 1st ($i = 1$) component (pose-related) and the 7th ($i = 7$) component (expression-related) by $-2\sqrt{\lambda_i}$ and $+2\sqrt{\lambda_i}$, where λ_i denotes the variance explained by the respective component. It is evident that the former component is associated with out-of-plane head rotation (*yaw*), whereas the latter component is associated with deformations related to sadness/happiness (*frown*/*smile*).

Appearance features: Previous frameworks targeting biometrics and affective computing tasks such as face recognition [2] and pain intensity estimation [101] have relied on appearance features locally extracted from a pre-defined grid of rectangular regions in face images registered in frontal pose. However, this technique is not suitable for databases including images that portray faces with large head pose angles, as is the case with the CONFER Database, since the 2D registration process unavoidably induces pixel artifacts and texture discontinuities. Furthermore, some researchers are critical of the grid-based feature extraction, suggesting that the sub-regions are not necessarily well aligned with meaningful facial features [89].

Motivated by these findings and other recent works [9, 74, 187], in this study we adopt a hybrid approach to appearance feature extraction. In particular, we first apply the same transformation used for point registration to the pixel intensities of each face image to remove translation, scale and in-plane rotation effects. Subsequently, features are extracted from the intensities lying within rectangular regions (patches) of dimension 20 ×20 pixels centered at each facial point. Facial point tracking and point/image registration results are depicted for each interlocutor in Fig. 4.5 for 2 characteristic frames from a sequence of the CONFER Database.

Two appearance-based descriptors are examined herein, namely *Scale-Invariant Feature Transform (SIFT)* [128] and *Discrete Cosine Transform (DCT)* [180]. SIFT is a rotation- and scale-invariant descriptor that captures local orientation information in images, while DCT is a frequency-based descriptor that projects pixel intensities onto real cosine basis functions. For SIFT, we extract a 4×4 array of 8-bin orientation histograms for each image patch. For DCT, the two-dimensional DCT is employed and the first 128 out of the zig-zag-arranged coefficients, which correspond to the lowest frequencies, are retained, so that the final dimensionality matches that of SIFT. For both descriptors, the features calculated from the total of 49 patches are concatenated into a single vector. For each frame, the final representation is formed by concatenating the feature vectors for all interlocutors (two or three). Finally, dimensionality is reduced in a supervised manner, by applying CCA on the features and corresponding annotations. The CCA coefficients of the feature set corresponding to 95% of the total energy

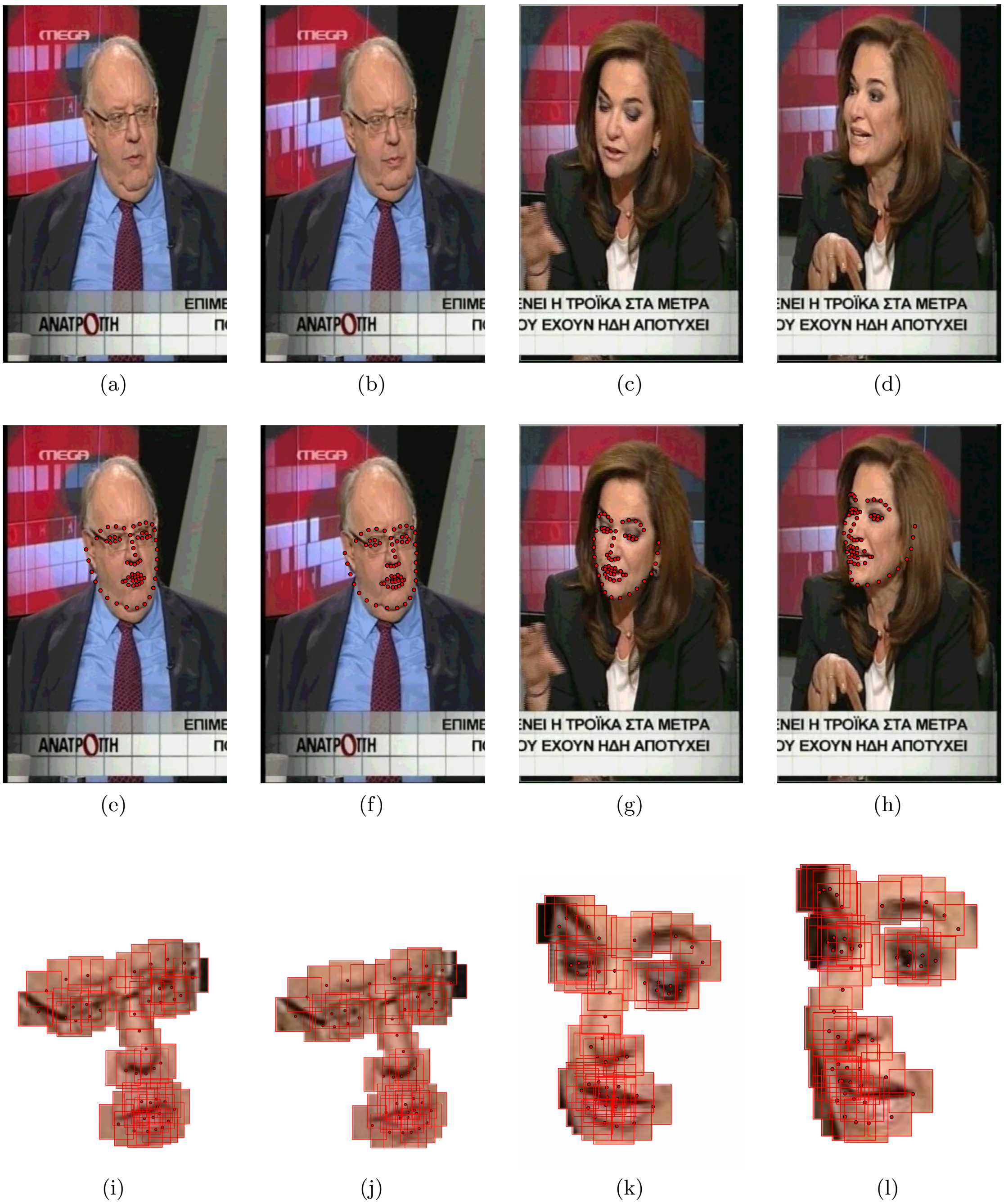


Figure 4.5: Tracking and point/image registration results shown for each subject for 2 characteristic frames (frame 683 and frame 762) of the sequence *20111212˙seq1* from the Set *two* of the CONFER Database. (a)-(d) Original video frames, (e)-(h) Original tracking points superimposed on the original video frames, and (i)-(l) Rectangular patches extracted around the aligned points on the aligned video frames.

are retained, thus resulting in dimensionality of 75 (65) and 63 (56) for the Set *two* (*three*) for SIFT and DCT, respectively. Note that all visual features are ℓ_2-normalized.

Fusion. To investigate which features carry complementary information with regards to manifestations of conflict in conversational and emotional behavior data and thus could help improve performance of conflict analysis tools, in this study we examine also *feature-level fusion*. Both *intra-modality* (Video-Video) and *inter-modality* (Audio-Visual) fusion is investigated. For the former case, we combine the expression-related shape descriptor with each appearance descriptor, that is, *Expression+SIFT* and *Expression+DCT* as well as the *Points* descriptor (the whole shape-based feature vector) with the appearance descriptor that performs best in the first baseline experiment (see Section 4.4.1). This is done on the feature level, that is, by concatenating the respective feature vectors.

For audio-visual (AV) fusion, we follow a more sophisticated approach, motivated by recent evidence which suggests that feature-level AV fusion can be sub-optimal and highly problematic mainly due to (i) the two modalities being recorded at different measurement and temporal scales and (ii) the detrimental effect of increased dimensionality on the classifier's performance [255]. To overcome the aforementioned limitations, we perform CCA to derive linear, maximally correlated components among the audio and visual feature sets. After retaining the components that account for the 95% of energy for each of the sets, the resulting CCA coefficients are concatenated to form the final AV feature representation. Note that for AV fusion, audio features are combined only with the best-performing out of the (single- or multi-feature) visual descriptors examined in the first baseline experiment (see Section 4.4.1).

4.3.4 Classifiers

Four classifiers that have been extensively used for temporal modeling of human behavior and affect are examined, namely *Support Vector Regression (SVR)* [205], *Random Forests for Regression (RF)* [30], *Continuous Conditional Random Fields (CCRF)* [14], and *Long-Short Term Memory (LSTM) Neural Networks* [78]. LIBSVM [34], `scikit-learn` [172], [14], and the CUda RecurREnt Neural Network Toolkit (CURRENNT) [239] are used to train SVR, RF, CCRF and LSTMs, respectively. For each fold of the cross-validation experiments, the validation set is used to optimize the classifiers in terms of Correlation (COR) for SVR, RF and CCRF (see Section 4.3.5), and Root Mean Squared Error (RMSE) for LSTMs[2].

[2] CURRENNT [239] only supports RMSE criterion for the objective function of LSTMs.

SVR [223] is a discriminative regression framework that extends Support Vector Classification (SVC) to the continuous (real-valued) targets, and is one of most commonly used regressors in the fields of affective computing and social signal processing [83, 228] with applications to various tasks such as continuous and dimensional emotion prediction [145], and social signal/behavior (e.g., laughter/conflict) detection/recognition [196], to mention but a few. In this study, linear SVR with ϵ-insensitive loss function is examined, whose parameters are optimized by means of a suitable grid search. In particular, the regularization parameter C is optimized in the set $\{10^{-5}, 10^{-4}, \ldots, 1\}$, the convergence tolerance parameter *tol* in the set $\{10^{-5}, 10^{-4}, \ldots, 10^{-2}\}$, while for the ϵ parameter 50 values logarithmically spaced in the range $[10^{-2}, 1]$ are examined.

RF [30] is an ensemble learning algorithm that combines unpruned Decision Tree learners based on random split selection of feature subspaces. RF have gained popularity in recent years within the computer vision and machine learning communities (e.g., [197, 38]) as they combine the ability to handle large training datasets with computational efficiency and good generalizability. The two most critical parameters in the RF design, that is the number of trees T in the forest and the number of features F selected to split each node, are optimized in the range $T \in \{100, 500, 1000, 2000\}$ and $F \in \{\sqrt{p}, p/3, p/2\}$, respectively, where p denotes the dimensionality of the feature vector.

CCRF [14] is an undirected graphical model-based discriminative framework that extends the traditional Conditional Random Fields (CRF) [112] to the case of continuous (real-valued) output. CCRF have been applied in combination with SVR for the task of continuous and dimensional emotion prediction [14]. Herein, we follow the approach in [14] in using linear SVR (exactly as described above) to learn the *vertex* (static) features of the graphical model and ten *edge* (temporal) features, that is, 5 neighbor $n = \{1, 2, \ldots, 5\}$ and 5 distance similarities $\sigma = \{2^{-6}, 2^{-7}, \ldots, 2^{-11}\}$ (see [14] for details).

LSTMs [79] constitute an extension of the traditional Recurrent Neural Network architecture that is efficient in capturing contextual statistical regularities with large and unknown lags in time-series data. LSTMs have been successfully applied to various behavioral and affective computing tasks such as continuous and dimensional affect prediction [240, 145], visual-only accent classification [72], and audio-visual depression scale prediction [35]. Herein, we use bi-directional LSTMs with 1 hidden layer of 128 memory blocks. The output layer consists of a single node whose sigmoid-function activation is used as the estimate of the conflict intensity. The networks are trained with stochastic gradient descent with a batch size of 5 sequences for a maximum of 1000 epochs. Finally, zero-mean Gaussian noise of variance 01 is added to the features and early stopping is employed to prevent overfitting.

4.3.5 Evaluation Metrics

Performance is measured for each test sequence based on two metrics, namely the *Pearson's Correlation coefficient (COR)* and the *Intra-class Correlation Coefficient (ICC)* [202]. Both metrics are computed for each test sequence, and the average value over all test sequences is retained for each fold. Finally, the mean and standard deviation of each metric over all 5 folds are reported.

The Pearson's Correlation coefficient (COR) is, along with the Mean Squared Error (MSE), the most commonly used evaluation metric in the affective computing literature [83, 145]. We have opted to use COR in this study over MSE since the former can capture linear structural information about how ground truth annotations and predictions vary together through the calculation of the covariance [83]; if the two measurements have a perfect linear relationship, then COR becomes 1 (complete positive relationship) or -1 (complete negative relationship). This property of the correlation is deemed advantageous for the experimental setting of our study that deals with continuous-time (frame-by-frame) estimation of conflict intensity.

The Intra-class Correlation Coefficient (ICC) [202], initially proposed as a metric for rater reliability in behavioral measurements, has been recently applied in providing a measure of 'consistency' or 'agreement' between ground truth annotations of behavioral or affective attributes provided by humans and corresponding predictions yielded by automated approaches (e.g., [101, 217]). It typically expresses the fraction of the total variance across all ratings and subjects (including random error in the 'judgements') 'explained' by the component of variance due to the targets alone [202]. Herein, we employ the coefficient ICC(3,1), which corresponds to the scenario *'Each target is assessed by each rater, with a single measurement being available for each rater and the raters being the only raters of interest'* [202]. For each automated framework examined, the ICC is computed based on the ground truth annotations and the predicted values of conflict intensity.

To obtain a 'human' baseline ICC result, i.e., a measure of 'level of consistency amongst 10 humans in assessing conflict intensity', we also compute the ICC amongst the 10 available annotations for each sequence. This facilitates a more fair evaluation of the various automated approaches examined in the experiments presented in Section 4.4.2. In particular, it enables us to compare the degree of conformity – in ICC terms – between the 'mean annotation' and the conflict intensity predictions yielded by the various frameworks to the degree of conformity amongst the measurements of conflict intensity obtained by 10 humans for the same data. The mean (standard deviation) of the 'inter-annotator' ICC is 0.495 (0.037) for the Set *two* and

0.414 (0.057) for the Set *three*, respectively.

4.4 Results

In this section, experimental results are reported and discussed separately for each of the two baseline experiments conducted on the CONFER Database for audio-visual continuous-time conflict intensity estimation.

4.4.1 Baseline Experiment I: Feature Comparison

In the first experiment of this study, we investigate the efficiency of the various audio and visual (shape- and appearance-based) descriptors as well as the (Video-Video and Audio-Visual) fusion of them described above, for the task of *continuous-time (frame-by-frame) estimation of continuous (real-valued) conflict intensity.* In total, 10 features (incl. fusion) are examined, namely *Audio*, *Pose*, *Expression (Expr.)*, *Points*, *SIFT*, *DCT*, *Expr.+SIFT* , *Expr.+DCT*, *Points+SIFT*, and *Fusion (AV)*. For the regression stage of this experiment, we use linear SVR which is one of the most commonly used regression frameworks in the literature for dimensional behavior and affect modeling [83]. We first examined the single-feature systems. Then, for Video-Video fusion, we chose to examine the combination of the whole shape feature vector (*Points*) with the best-performing appearance descriptor, i.e., *SIFT*, hence *Points+DCT* is not considered. Finally, for audio-visual fusion, we examined the combination of *Audio* features with the best-performing out of all visual features and fusion of them, i.e., *Expr.+SIFT*.

Conflict intensity estimation results, in terms of COR averaged over all 5 folds of the cross-validation experiment, are shown in the bar graph of Fig. 4.6 for the Sets *two* and *three* of the CONFER Database. Among the single-feature frameworks, the best performance of COR = 0.233 and COR = 0.302 for the Set *two* and *three* is achieved by *Audio* and *SIFT*, respectively. Note that *Audio* is the only feature that accounts for lower performance on the Set *three* than on the Set *two*, presumably due to the increased number of speakers in the former case incurring a larger number of speaker diarization errors. On the other hand, for all visual features there is a large discrepancy between the performances achieved on the Sets *two* and *three*, with the latter being higher in all cases. This finding makes sense upon observing that it is often the case with the recordings of the Set *three* that not all interactants are recorded in the same studio and thus some of them retain a (quasi-)frontal view during the conversation by looking straight at the camera rather than at their interlocutors. Under these conditions, the computer vision tasks of facial point tracking and image registration are rendered much easier and hence accurate, thus leading to more efficient and error-free visual feature extraction.

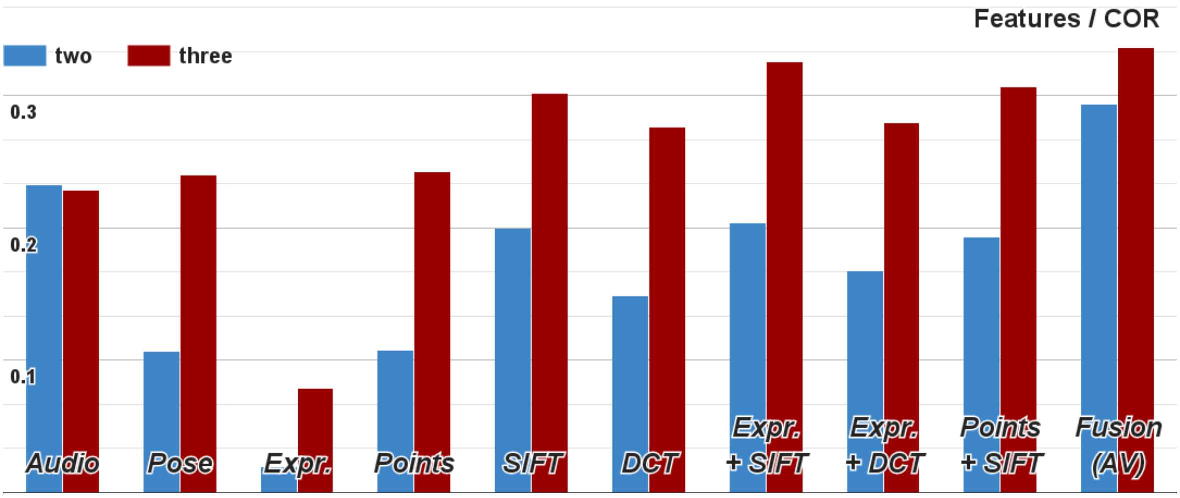


Figure 4.6: Baseline Experiment I: Conflict intensity estimation results, in terms of COR averaged over all 5 folds, as obtained by linear SVR trained with the various visual and audio features as well as fusion of them (Video-Video and Audio-Visual) examined herein, for the Sets *two* and *three* of the CONFER Database.

Among shape features, *Pose* features largely outperform *Expression* features, with the latter leading to a rather poor performance when considered alone. This conforms to recent evidence [25, 26] suggesting that head gestures (e.g., head nod, shake, roll, 'cut-off') are among the most common non-verbal cues through which (dis)agreement is manifested, hence the efficiency of head pose features in capturing the latter and conflict as well. Also, the poor performance yielded by *Expression* can be partially attributed to the high variation of expression-related facial deformations in the CONFER Database, which entails that a lot of the latter do not convey conflict information and thus are uninformative for the task at hand.

Appearance features perform more accurately than shape features. This is exactly as expected; while shape features are capable of capturing coarse deformations related to facial expression, appearance features are efficient in encapsulating finer movements and tale-telling transient features such as bulges, wrinkles and furrows [255, 163, 73]. Also, SIFT outperforms DCT. This is again not a surprising result given that SIFT features extracted from local patches around facial landmarks have been shown to be efficient for automatic face analysis "in-the-wild" [9]. Also, DCT being less efficient than SIFT in this experiment can be partially attributed to its Fourier-based transformation, which is applied locally, capturing energy characteristics in the visual scene which are unrelated to conflict (e.g., uninformative facial expressions, illumination changes caused by head movements). It is also worth mentioning that, the shape-based *Expression* descriptor, despite performing poorly when used in isolation, leads

Table 4.1: Baseline Experiment II: Conflict intensity estimation results, in terms of COR and ICC averaged over all 5 folds, as obtained by each feature-classifier combination examined herein for the Sets (a) *two*, and (b) *three* of the CONFER Database, respectively. The corresponding standard deviation values are reported inside parentheses. The best COR and ICC performances for the uni- and multi-modal frameworks (A: Audio, V: Visual, AV: Audio-Visual) are shown in boldface.

(a) Set *two*

Classifier / Feature	**Audio (A)**		**Expr.+SIFT (V)**		**Fusion (AV)**	
	COR	**ICC**	**COR**	**ICC**	**COR**	**ICC**
SVR	0.233 (0.064)	**0.774** (0.031)	0.204 (0.090)	0.174 (0.030)	**0.294** (0.065)	**0.781** (0.029)
RF	0.170 (0.054)	0.144 (0.020)	0.052 (0.024)	0.168 (0.042)	0.178 (0.053)	0.160 (0.020)
CCRF	**0.285** (0.177)	0.160 (0.355)	0.026 (0.067)	-0.001 (0.000)	0.221 (0.075)	0.163 (0.357)
LSTMs	0.232 (0.092)	0.178 (0.033)	0.126 (0.071)	0.183 (0.070)	0.251 (0.070)	0.195 (0.065)

(b) Set *three*

Classifier / Feature	**Audio (A)**		**Expr.+SIFT (V)**		**Fusion (AV)**	
	COR	**ICC**	**COR**	**ICC**	**COR**	**ICC**
SVR	0.229 (0.063)	**0.687** (0.036)	**0.326** (0.076)	0.357 (0.144)	**0.336** (0.033)	**0.296** (0.187)
RF	0.156 (0.061)	0.213 (0.050)	0.158 (0.108)	0.204 (0.092)	0.173 (0.092)	0.198 (0.089)
CCRF	0.213 (0.036)	0.045 (0.044)	0.153 (0.130)	0.014 (0.031)	0.211 (0.109)	0.014 (0.023)
LSTMs	0.259 (0.068)	0.221 (0.077)	0.185 (0.100)	0.186 (0.045)	0.148 (0.055)	0.195 (0.022)

to performance improvement when combined with either of the appearance descriptors. This behavior can be explained considering that *Expression* captures coarse non-rigid deformations from the whole face which are complementary to the local subtle movements encoded by the appearance descriptors extracted from the local patches.

Finally, audio-visual fusion outperforms all remaining frameworks, leading to COR = 0.294 and COR = 0.336 for the Set *two* and *three*, respectively. This result provides a strong indication that behavioral patterns associated with continuous-in-time manifestations of conflict under unconstrained recording conditions are more accurately recognized when cues from both the audio and video modality are considered, as is the case with other social behaviors such as (dis)agreement, mimicry, interest, and flirting [228, 161].

4.4.2 Baseline Experiment II: Classifier Comparison

In the second experiment of this study, we investigate the efficiency of the various classifiers described above in modeling and predicting conflict intensity in continuous time for each test sequence, approached again as a regression problem on a frame-by-frame basis. The features utilized to train the classifiers are those that performed best in the previous experiment, i.e., *Audio* and *Expr.+SIFT* for Audio and Video, respectively, and *Audio+Expr.+SIFT* for audio-visual (AV) fusion.

Conflict intensity estimation results, in terms of the COR and ICC metrics averaged over all 5 folds of the cross-validation experiment, are reported in Table 4.1a and Table 4.1b for the Set *two* and *three* of the CONFER Database, respectively. The best performances of COR = 0.294 and COR = 0.336 for the Set *two* and *three*, respectively, are those achieved by audio-visual fusion in the previous experiment. Interestingly, both the aforementioned best-performing frameworks employ SVR in the regression stage. However, it is worth noting that not for all classifiers does fusion result in improved performance (in terms of COR) over that furnished by the corresponding uni-modal systems. This can be partially attributed to different classifiers being to a different degree sensitive to (i) gross errors and outliers in the audio or/and the video stream which, in turn, result in erroneous estimates of the correlated components obtained by the classical CCA due to its reliance on least squares minimization, and (ii) errors induced by feature pro- and post-processing (e.g., normalization, AV synchronization). A partial remedy to the above-mentioned limitations could be sought in either applying more robust techniques for the extraction of individual and correlated components, such as the one proposed in [155], or 'delegating' both the tasks of modeling each stream separately and uncovering the correlations between them to the classifier by means of *model-level fusion* (see [255] for a survey of different types of fusion).

Regarding the uni-modal frameworks, the best performances of COR = 0.285 and COR = 0.326 are accounted for by the combination of *Audio* with CCRF and *Expr.+SIFT* with SVR for the Set *two* and *three*, respectively. The superiority of SVR among classifiers for the multi-modal frameworks and the high accuracy achieved by it also when trained with features from a single modality conforms to previous evidence indicating its robustness to overfitting and suitability for continuous prediction of behavior and affect dimensions [196, 145]. CCRF also yield accurate predictions in this experiment, presumably thanks to their ability to learn the conflict 'history' across successive observations of continuous conversational data given that they, like their discrete-output counterpart (CRF), relax the assumption of conditional independence of the features [112]. We argue that their performance for conflict intensity prediction could be improved by (i) examining different functions for the *vertex* and *edge* features (e.g., non-linear regressor for the *vertex* features), and (ii) investigating different normalization schemes, to which they have shown to be quite sensitive [240]. LSTMs trained with *Audio* features also achieve high COR values for both Sets, albeit on par with or not much higher than those achieved by SVR. This result might seem counter-intuitive at first sight, since LSTMs, similarly to CCRF, are capable of capturing long-range dependencies between successive observations and, as such, have been shown successful in continuous modeling of human behavior and affect [240, 145, 35]. However, we argue that the relatively low performance of LSTMs in this experiment is mainly due to them having been trained based on RMSE and that, by using an alternative implementation that allows COR-based objective function for LSTMs training, one will most probably achieve much higher performance. The same holds for the RF frameworks which have been also trained on the basis of mean-squared generalization error and thus are agnostic to contextual temporal information. The poor performance of RF for this experiment can be also attributed to the random feature selection process employed to determine the split at each node; this practice can result in sub-optimal partitioning of the feature space, especially in the case of insufficient training data [126]. To alleviate this limitation, one could resort to a semi-supervised approach to node splitting, such as the one proposed in [126], that is, to use also unlabeled data to guide the node splitting.

As for the results in terms of ICC, SVR combined with *AV Fusion* and *Audio* accounts for the best performances of ICC = 0.781 and ICC = 0.687 for the Set *two* and *three*, respectively. It is worth noting that the best ICC scores obtained by *Audio* are much higher than those obtained by the visual descriptor *Expr.+SIFT*. In other words, the predictions yielded by the former framework are much more ‘consistent’ in terms of ICC with the ground truth annotations than those yielded by the latter. This behavior can be partially attributed to the annotation process. In particular, it is highly likely that the annotators, who are all native speakers of Greek that is the language spoken in the CONFER Database, relied much more on the audio modality while annotating since in that alone they could easily identify informative cues associated with conflict escalation/resolution in terms of both the *physical* layer (e.g., interruptions, overlapping speech) and the *inferential layer* (e.g., sarcasm, rudeness, confrontation) of the conversation. The impact of this condition is larger in absolute terms for the ICC rather than the COR metric in the results reported in Table 4.1. This is presumably due to the random error associated with the ‘raters’ decreasing significantly for the audio-based system and thus leading to an increase in the ratio of variances to which ICC equals (see Section 4.3.5 and [202] for more details).

Furthermore, it is also worth noting that the aforementioned best performances in terms of ICC exceed the corresponding values of ICC = 0.495 and ICC = 0.414 measured amongst the 10 annotators for the Set *two* and *three*, respectively. This signifies that the corresponding frameworks, which were trained using the ‘mean annotator’ annotations, learned the trend of the ‘mean annotator’ better and were able to reproduce the trend accurately. This result is quite encouraging in that it reveals that even uni-modal systems based on a commonly used classifier can be more ‘consistent’ with the ‘mean human rating’ in assessing conflict intensity than several humans are with one another on the same dataset.

Overall, the relatively low results achieved in both experiments described above can be attributed to (i) the challenging nature of the CONFER Database, which consists of spontaneous conversational data where conflict naturally arises, (ii) the demanding subject- and session-independent experimental protocol adopted in this study, and (iii) the abundance of the data (106536 and 106404 frames in total for the Set *two* and *three*, resp.), which are all tested by means of cross-validation. However, these results indicate that there is much room for improvement for tools targeting the task at hand. We hope that these findings will encourage further research in the future in the development of audio-visual approaches to automatic analysis of conflict as well as similar behavioral and affective phenomena.

4.5 Conclusion

In this chapter, we presented the Conflict Escalation Resolution (CONFER) Database, a set of audio-visual recordings of naturalistic interactions from political debates suitable for the investigation of conflict behavior. The database contains 142 minutes of recordings in total and is the first of its kind to have been annotated in terms of continuous (real-valued) conflict intensity on a frame-by-frame basis. Audio-visual episodes and features as well as annotations are publicly available for non-commercial use at `http://ibug.doc.ic.ac.uk/resources/confer/`.
The CONFER Database contains naturalistic, competitive conversations from political debates where conflict naturally arises, and is the first database in the literature that is annotated in terms of real-valued conflict intensity on a frame-by-frame basis. As such, it is primarily intended for research targeting automatic conflict analysis and similar social attitudes such as (dis)agreement. However, it could also be a valuable source for studies of other social signals (e.g., turn-taking, back-channel communication, engagement, hot-spots) and social behaviors (e.g., interest, politeness, mimicry, social dominance, likeability). To this end, a suitable annotation effort should be put for each task. The provided audio-visual episodes of conflict have been filmed “in-the-wild”,that is, under unconstrained conditions, and thus involve a wide range of views, amenable lighting conditions, spontaneous and overlapping speech, and abrupt head and body movements or occlusions. Hence, the CONFER Database can also be used to facilitate research in automatic speech and speaker recognition, recognition of non-verbal behavioral cues (e.g., facial expressions, body postures, gestures, head nods, vocal outbursts and laughter) as well as related audio processing and computer vision tasks (e.g., speaker

diarization, face detection, facial point tracking, head pose estimation). All previous studies to date have approached conflict analysis within a classification framework generating discrete labels of conflict intensity for well-segmented episodes. The presented baseline experiments constitute the first audio-visual approach in the literature to *continuous-time* (frame-by-frame) estimation of *continuous-valued* conflict intensity. In our systematic study, we reported benchmark results of subject- and session-independent experiments by means of which the efficiency of commonly used audio and visual features and fusion of them as well as classifiers was examined for conflict intensity estimation in dyadic and multi-party conversations. Our results provide strong indications that there is much room for improvement in terms of both audio-visual representations as well as learning methodologies that can efficiently capture temporal dynamics of social phenomena manifested in spontaneous multi-party interactions. For the feature extraction stage, one could investigate descriptors stemming from the deep learning literature, such as convolutional neural networks [110, 204] and variational autoencoders [109], which can be trained either exclusively on our dataset or pre-trained on other datasets for similar tasks and, subsequently, undergo domain adaptation (see [49] for a survey). In this way, non-linear relationships in the training data distributions that be descriptive of the task in question could be discovered at a finer granularity depending on the architecture of the networks employed. On the other hand, for the temporal modeling task one could also examine different variants of recurrent neural networks (RNNs), e.g., substitute the LSTM units with Gated Recurrent Units (GRUs) [41]. Another interesting direction that could be followed is to use the recently proposed Variational RNNs (VRNNs) [42] which is a bayesian model combining representation learning via auto-encoders at each time step and recurrent connections with hidden memory units that can take an infinite amount of states. This model has been shown to provide good results on other sequential learning tasks such as speech and handwriting modeling [42], but its application on social signal processing applications involving high-dimensional noisy data and labels has not yet been pursued.

Overall, we can conclude that deep investigations of how best to reach satisfactory performance on automatic analysis of social attitudes like conflict and (dis)agreement are yet to be conducted. However, we believe that this chapter can serve as an introductory reading to researchers interested in the problem of automatic conflict intensity estimation based on nonverbal cues and their temporal dynamics. Most importantly, this benchmark paves the way for the investigation of social attitudes in continuous time and scale and provides an appropriate platform for the development of efficient temporal classifiers that can model social signals and behaviors at a finer granularity.

Chapter 5

Dynamic Behavior Analysis via Structured Rank Minimization

Contents

Human affect and behavior is inherently a dynamic phenomenon involving temporal evolution of patterns manifested through a multiplicity of non-verbal behavioral cues including facial expressions, body postures and gestures, and vocal outbursts, among others. While high performance is achieved by existing machine learning methodologies on datasets acquired under laboratory conditions, their performance drops significantly when they are assigned the task of modeling dynamic affect and behavior under unconstrained conditions, namely in the presence of grossly corrupted behavioral cues descriptors and possibly unreliable annotations. Aside from the susceptibility of existing models to such sources of gross noise, they do not model the temporal dynamics of affect and behavior in an explicit and interpretable way. In this chapter, we alleviate this problem and explicitly model the temporal dynamics by means of a linear dynamical system that generates continuous-time characterizations of affect or behavior as outputs when behavioral cues act as inputs. To this end, a novel robust structured rank minimization method and its scalable variant are proposed. The generalizability of the proposed framework is demonstrated by conducting experiments on three distinct dynamic behavior analysis tasks, namely conflict intensity prediction, prediction of valence and arousal, and multi-object/person tracking from detection.

5.1 Introduction

Analysis of human behavior concerns detection, tracking, recognition, and prediction of complex human behaviors including affect and social behaviors such as agreement and conflict escalation/resolution from audio-visual data captured in naturalistic, real-world conditions. Modeling human behavior for automatic analysis in such conditions is the prerequisite for next-generation human-centered computing and novel applications such as personalized natural interfaces (e.g., in autonomous cars), software tools for social skills enhancement including conflict management and negotiation, and assistive technologies (e.g., for independent living), to mention but a few.

Traditionally, research in behavior and affect analysis has focused on recognizing behavioral cues such as smiles, head nods, and laughter [53, 103, 127], pre-defined posed human actions (e.g., walking, running, and hand-clapping) [59, 151] or discrete, basic emotional states (e.g., happiness, sadness) [166, 43, 121] mainly from posed data acquired in laboratory settings. However, these models are deemed unrealistic as they are unable to capture the temporal evolution of non-basic, possibly atypical, behaviors and subtle affective states exhibited by humans in naturalistic settings. In order to accommodate such behaviors and subtle expressions, continuous-time and dimensional descriptions of human behavior and affect need to be employed. As explained in detail in Section 2.2, machine learning models commonly employed for automatic, continuous behavior and emotion analysis such as Hidden Markov Models (HMMs) [43], Dynamic Bayesian Networks (DBN) [171], Conditional Random Fields (CRFs) [141] and Long-Short Term Memory (LSTM) Neural Networks [145] and other regression-based approaches [147, 218, 102], despite their merits, they share a number of limitations; they rely on large sets of training data, involve learning of a large number of parameters, they do not model dynamics of human behavior and affect in an explicit way, and more importantly they are fragile in the presence of gross non-Gaussian noise and incomplete data, which is abundant in real-world (visual) data.

In the work presented in this chapter, we model and tackle the problem of *dynamic behavior analysis* in the presence of gross, but sparse, noise and incomplete visual data under a different perspective, making the following contributions.

1. The modeling assumption here is that for smoothly-varying dynamic behavior phenomena, such as conflict escalation and resolution, temporal evolution of human affect described in terms of valence and arousal, or motion of human crowds, among others, the observed data can be postulated to be trajectories (inputs and outputs) of a linear time-invariant (LTI) system. Recent advances in system theory [221, 70] indicate that such dynamics can be discovered by learning a low-complexity (i.e., low-order) LTI system based on its inputs and outputs via rank minimization of a Hankel matrix constructed from the observed data. Here, continuous-time annotations characterizing the temporal evolution of relevant behavior or affect are considered as system outputs, while features describing behavioral cues are deemed system inputs. In practice, visual data are often contaminated by gross, non-Gaussian noise mainly due to pixel corruptions, partial image texture occlusions or feature extraction failure (e.g., incorrect object localization, tracking errors), and human assessments of behavior or affect may be unreliable mainly due to annotator subjectivity or adversarial annotators. The existing structured rank minimization-based methods perform sub-optimally in the presence of gross corruptions. Therefore, to robustly learn a LTI system from grossly corrupted data, we formulate a novel *ℓ_q-norm regularized (Hankel) structured Schatten-p norm minimization* problem in Section 5.3. The Schatten p- and the sparsity promoting ℓ_q-norm act either as convex surrogates, when $p = q = 1$, or as non-convex approximations, when $p, q \in (0, 1)$, of the rank function and the ℓ_0-(quasi) norm, respectively.

2. To tackle the proposed optimization problem, an algorithm based on the Alternating-Directions Method of Multipliers (ADMM) [18] is developed in Section 5.4. Furthermore, in the same section a scalable version the algorithm is elaborated.

3. The proposed model is the heart of a general and novel framework for dynamic behavior modeling and analysis, which is detailed in Section 5.5. A common practice in behavioral and affective computing is to train machine learning algorithms by employing large sets of training data that comprehensively cover different subjects, contexts, interaction scenarios and recording conditions. The proposed approach allows us to depart from this practice. Specifically, we demonstrate for the first time that complex human behavior and affect, manifested by a single person or group of interactants, can be learned and predicted based on a small amount of person(s)-specific observations, amounting to a duration of just a few seconds.

4. The effectiveness and the generalizability of the proposed model is corroborated by means of experiments on synthetic and real-world data in Section 5.6. In particular, the

generalizability of the proposed framework is demonstrated by conducting experiments on 3 distinct dynamic behavior analysis tasks, namely (i) *conflict intensity prediction*, (ii) *prediction of valence and arousal*, and (iii) *tracklet matching*. The attained results outperform those achieved by other state-of-the-art methods on both synthetic and real-world data and, hence, evidence the robustness and effectiveness of the proposed approach.

5.2 Background and Related Work

In this section, notation conventions and mathematical formalism related to the Hankel matrix structure are first introduced. Next, in order to make the paper self-contained, we describe how learning of dynamical systems and, in particular, of a LTI system can be cast as a (Hankel)-structured rank minimization problem. Related works on structured rank minimization and their applications in visual information processing are also described.

5.2.1 Preliminaries

Notations. Matrices (vectors) are denoted by uppercase (lowercase) boldface letters, e.g., $\mathbf{X}, (\mathbf{x})$. $\mathbf{I}$ denotes the identity matrix of compatible dimensions. The ith element of vector $\mathbf{x}$ is denoted as x_i, the ith column of matrix $\mathbf{X}$ is denoted as $\mathbf{x_i}$, while the entry of $\mathbf{X}$ at position (i, j) is denoted by x_{ij}. For the set of real numbers, the symbol $\mathbb{R}$ is used. For two matrices $\mathbf{A}$ and $\mathbf{B}$ in $\mathbb{R}^{m\times n}$, $\mathbf{A} \circ \mathbf{B}$ denotes the Hadamard (entry-wise) product of $\mathbf{A}$ and $\mathbf{B}$, while $\langle \mathbf{A}, \mathbf{B}\rangle$ denotes the inner product $\mathrm{tr}(\mathbf{A}^T\mathbf{B})$, where $\mathrm{tr}(\cdot)$ is the trace of a square matrix. For a symmetric positive semi-definite matrix $\mathbf{A}$, we write $\mathbf{A} \succeq 0$. Regarding vector norms, $\|\mathbf{x}\|:= \sqrt{\sum_i x_i^2}$ denotes the Euclidean norm. The sign function is denoted by $\mathrm{sgn}(\cdot)$, while $|\cdot|$ denotes the absolute value operator. Regarding matrix norms, the ℓ_0-(quasi-) norm, which equals the number of non-zero entries, is denoted by $\|\cdot\|_0$. $\|\mathbf{X}\|_q:= \left(\sum_i \sum_j |X_{ij}|^q\right)^{1/q}$ is the matrix ℓ_q-norm, of which the Frobenius norm $\|\mathbf{X}\|_F:= \sqrt{\sum_i \sum_j X_{ij}^2} = \sqrt{\mathrm{tr}(\mathbf{X}^T\mathbf{X})}$ is a special case when $q = 2$. $\|\mathbf{X}\|$ denotes the spectral norm, which equals the largest singular value. If $\sigma_i(\mathbf{X})$ is the ith singular value of $\mathbf{X}$, $\|\mathbf{X}\|_{S_p}:= (\sum_i \sigma_i(\mathbf{X})^p)^{1/p}$ is the Schatten p-norm of $\mathbf{X}$, of which the nuclear norm $\|\mathbf{X}\|_*:= \sum_i \sigma_i(\mathbf{X})$ is a special case when $p = 1$. Linear maps are denoted by scripted letters. For a linear map $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, $\mathcal{A}^*$ denotes the adjoint map of $\mathcal{A}$, while $\sigma_{\max}(\mathcal{A})$ denotes the maximum singular value of $\mathcal{A}$. $\mathcal{I}$ denotes the identity map.

The Hankel matrix structure. Let $\mathbf{A} = [\mathbf{A_0}\ \mathbf{A_1}\ \ldots\ \mathbf{A_{j+k-2}}]$ be a $m \times n(j + k - 1)$ matrix, with each $\mathbf{A_t}$ being a $m \times n$ matrix for $t = 0, 1, \ldots , j + k - 2$. We define the Hankel linear map $\mathcal{H}(\mathbf{A}) := H_{m,n,j,k}(\mathbf{A})\mathbf{\Gamma}$, where

$$H_{m,n,j,k}(\mathbf{A}) = \begin{pmatrix} \mathbf{A_0} & \mathbf{A_1} & \cdots & \mathbf{A_{k-1}} \\ \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A_{j-1}} & \mathbf{A_j} & \cdots & \mathbf{A_{j+k-2}} \end{pmatrix} \in \mathbb{R}^{mj \times nk}, \tag{5.1}$$

and $\mathbf{\Gamma} \in \mathbb{R}^{nk \times q}$ with $\sigma_{\max}(\mathbf{\Gamma}) \leq 1$ [70]. Therefore, $H_{m,n,j,k}(\mathbf{A})$ is a block-Hankel matrix with $j \times k$ blocks, where each $\mathbf{A_i}$ is a matrix of dimension $m \times n$. Note that the Hankel structure enforces constant entries along the skew diagonals. We denote by $T = j + k - 1$ the total number of observations, while $M = mj$ and $N = nk$ denote the number of rows and columns of the Hankel matrix $H_{m,n,j,k}(\mathbf{A})$, respectively. For notational convenience, we write $H(\mathbf{A})$ to denote $H_{m,n,j,k}(\mathbf{A})$, when the dimensions m, n, j, k are clear from the context.

The adjoint map $\mathcal{H}^*$ is defined as $\mathcal{H}^*(\mathbf{\Lambda}) = H^*_{m,n,j,k}(\mathbf{\Lambda}\mathbf{\Gamma}^\mathbf{T})$, where for any matrix $\mathbf{B} \in \mathbb{R}^{mj \times nk}$

$$\begin{aligned} H^*_{m,n,j,k}(\mathbf{B}) &= H^*_{m,n,j,k} \begin{pmatrix} \mathbf{B_{00}} & \mathbf{B_{01}} & \cdots & \mathbf{B_{0,k-1}} \\ \mathbf{B_{10}} & \mathbf{B_{11}} & \cdots & \mathbf{B_{1,k-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B_{j-1,0}} & \mathbf{B_{j-1,1}} & \cdots & \mathbf{B_{j-1,k-1}} \end{pmatrix} \\ &= [\mathbf{B_{00}} \quad \mathbf{B_{01}} + \mathbf{B_{10}} \ldots \\ & \mathbf{B_{02}} + \mathbf{B_{11}} + \mathbf{B_{20}} \quad \cdots \quad \mathbf{B_{j-1,k-1}}] \in \mathbb{R}^{m \times n(j+k-1)}. \end{aligned} \tag{5.2}$$

It is proved in [70] that $\left\| H^*_{m,n,j,k}(\mathbf{B}) \right\|_F^2 \leq L \left\| \mathbf{B} \right\|_F^2$, where $L := \min\{j, k\}$. This finding, combined with $\sigma_{\max}(\mathbf{\Gamma}) \leq 1$, entails that the spectral norm of the adjoint map $\mathcal{H}^*$ is less than or equal to $\sqrt{L}$. Herein, the space of Hankel matrices is denoted by $\mathbb{S}_{\mathcal{H}}$.

5.2.2 LTI System Learning via Structured Rank Minimization

Dynamical systems, such as LTI systems, are able to compactly model the temporal evolution of time-varying data. While the dynamic model can be considered as known in some applications (e.g., Brownian dynamics in motion models), it is in general unknown and, hence, should be learned from the available data.

Consider a sequence of observed outputs $\mathbf{y_t} \in \mathbb{R}^m$ and inputs $\mathbf{u_t} \in \mathbb{R}^d$, respectively, for $t = 0, \ldots, T-1$. The goal is to find from the observed data, a state-space model, corresponding to a LTI system, given by

$$\begin{aligned} \mathbf{x_{t+1}} &= \mathbf{A}\mathbf{x_t} + \mathbf{B}\mathbf{u_t} \\ \mathbf{y_t} &= \mathbf{C}\mathbf{x_t} + \mathbf{D}\mathbf{u_t} \end{aligned} \tag{5.3}$$

such that the system is of low-order, i.e., it is associated with a low-dimensional state vector $\mathbf{x_t} \in \mathbb{R}^n$ at time t, where n is the *unknown* true system order. The order of the system (i.e., the dimension of the state vector) captures the memory of the system and it is a measure of its complexity. In (5.3), both the state and the measurement equations are linear and the parameters of the system, i.e., the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are constant over time but their dimensions are *unknown*. Therefore, to determine the model, we need to find the model order n, the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$, and the initial state $\mathbf{x_0}$. To this end, the model order should be estimated first. Next, the estimation of the system order using Hankel matrices is summarized.

Let us assume that the unknown state vectors has dimension $r > n$ and let $\mathbf{X} = [\mathbf{x_0} \quad \mathbf{x_1} \quad \dots \quad \mathbf{x_{T-1}}] \in \mathbb{R}^{r \times T}$, $\mathbf{Y} = [\mathbf{y_0} \quad \mathbf{y_1} \quad \dots \quad \mathbf{y_{T-1}}] \in \mathbb{R}^{m \times T}$, $\mathbf{U} = [\mathbf{u_0} \quad \mathbf{u_1} \quad \dots \quad \mathbf{u_{T-1}}] \in \mathbb{R}^{d \times T}$ be the matrices containing in their columns the unknown state vectors, the observed outputs, and the observed inputs of the system, respectively, for $t = 0, 1, \dots, T-1$. Let also $H_{m,1,r+1,T-r}(\mathbf{Y})$ and $H_{d,1,r+1,T-r}(\mathbf{U})$ be the Hankel matrices constructed from the observed system outputs and inputs, respectively, according to (5.1) and $\mathbf{U}^{\perp} \in \mathbb{R}^{(T-r) \times q}$ be the matrix whose columns form an orthogonal basis for the nullspace of $H_{d,1,r+1,T-r}(\mathbf{U})$. Then, the LTI in (5.3) can be expressed by employing the above mentioned Hankel matrices as follows.

$$H_{m,1,r+1,T-r}(\mathbf{Y}) = \mathbf{G}\mathbf{X} + \mathbf{L}H_{d,1,r+1,T-r}(\mathbf{U}), \tag{5.4}$$

where

$$\mathbf{G} = \begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA^r} \end{pmatrix}, \qquad \mathbf{L} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CB} & \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CAB} & \mathbf{CB} & \mathbf{D} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{CA^{r-1}B} & \mathbf{CA^{r-2}B} & \cdots & \cdots & \mathbf{D} \end{pmatrix} \tag{5.5}$$

By right-multiplying both sides of (5.4) with $\mathbf{U}^{\perp}$ and by setting $\mathcal{H}(\mathbf{Y}) = H(\mathbf{Y})\mathbf{U}^{\perp}$ we obtain

$$\mathcal{H}(\mathbf{Y}) = \mathbf{G}\mathbf{X}\mathbf{U}^{\perp}. \tag{5.6}$$

If the inputs are persistently exciting (i.e., $\mathbf{X}\mathbf{U}^{\perp}$ has full rank) and the outputs are exact, then by (5.6) it is clear that the system order, which is measured by the rank of $\mathbf{G}$ [221], is equal to $\operatorname{rank}(\mathcal{H}(\mathbf{Y}))$ [221] and from it a system realization (i.e., estimation of the unknown system parameters) is easily computed by solving a series of systems of linear equations following, for example, [221].

Table 5.1: List of structured rank minimization methods (including the proposed method) and the corresponding optimization problems. For all methods, the observed data matrix, its Hankel version, and the estimated (Hankel) structured low-rank approximate are denoted by $\mathbf{M} \in \mathbb{R}^{D \times T}$, $\mathbf{H} = \mathcal{H}(\mathbf{M}) \in \mathbb{R}^{M \times N}$ and $\hat{\mathbf{H}} = \mathcal{H}(\mathbf{L}) \in \mathbb{R}^{M \times N}$, respectively, unless otherwise stated.

	Method	Optimization Problem	Convex	Robust
Approximations of (5.7)	**Proposed**	$\min_{\mathbf{L},\mathbf{E}} \ \|\mathcal{H}(\mathbf{L})\|_{S_p}^p + \lambda \, \|\mathbf{W} \circ \mathbf{E}\|_q^q \quad \text{s.t.} \quad \mathbf{M} = \mathbf{L} + \mathbf{E}.$	depends on the choice of p and q	✔
	Hankel Rank Minimization (HRM) [70]	$\min_{\mathbf{L}} \frac{1}{2} \|\mathbf{M} - \mathcal{A}(\mathbf{L})\|_F^2 + \lambda \|\mathcal{H}(\mathbf{L})\|_*$, where $\mathcal{A}$ is a linear map.	✔	✗
	SVD-free [203]	$\min_{\mathbf{L},\mathbf{Q},\mathbf{R}} \frac{1}{2}(\mathbf{M} - \mathbf{L})^T \mathbf{W}(\mathbf{M} - \mathbf{L}) + \frac{1}{2}(\|\mathbf{Q}\|_F^2 + \|\mathbf{R}\|_F^2) \quad \text{s.t.} \quad \mathcal{H}(\mathbf{L}) = \mathbf{Q}\mathbf{R}^T.$	✗	✗
	[252]	$\min_{\mathbf{Q},\mathbf{R}} \frac{1}{2}(\|\mathcal{A}(\mathbf{C}\mathbf{g}) - \mathbf{J}\|_F^2 + \frac{\lambda}{2}(\|\mathbf{J}\mathbf{g}\|_F^2 + \frac{\mu}{2}(\|\mathbf{Q}\|_F^2 + \|\mathbf{R}\|_F^2),$ where $\mathbf{g} = \mathbf{vec}(\mathbf{Q}\mathbf{R}^T)$ and $\mathcal{A}$ is a linear map. Return $\hat{\mathbf{H}} = \mathbf{Q}\mathbf{R}^T$.	✗	✗
	Structured Robust PCA (SRPCA) [11]	$\min_{\hat{\mathbf{H}},\mathbf{E}} \sum_i w_i \sigma_i(\hat{\mathbf{H}}) + \|\mathbf{W_e} \circ \mathbf{E}\|_1 + \frac{1}{2}\|\mathbf{W_F} \circ \mathbf{E}\|_F^2$ s.t. $\mathbf{H} = \hat{\mathbf{H}} + \mathbf{E} \ ; \ \hat{\mathbf{H}}, \mathbf{E} \in \mathbb{S}_{\mathcal{H}}.$	✔	✔
Related Methods	Iterative Hankel Total Least Squares (IHTLS) [54]	Given $\mathbf{H} = [\mathbf{F} \,\vert\, \mathbf{g}] \in \mathbb{S}_{\mathcal{H}}$, estimate $\hat{\mathbf{H}} = [\mathbf{F} + \mathbf{E} \,\vert\, \mathbf{g} + \mathbf{k}]$ by solving $\min_{\mathbf{x},\mathbf{E},\mathbf{k}} \|\mathbf{W} \circ [\mathbf{E} \,\vert\, \mathbf{k}]\|_F^2$ s.t. $(\mathbf{F} + \mathbf{E})\mathbf{x} = \mathbf{g} + \mathbf{k} \ ; \ [\mathbf{F} \,\vert\, \mathbf{g}], [\mathbf{E} \,\vert\, \mathbf{k}] \in \mathbb{S}_{\mathcal{H}}.$	✗	✗
	Structured Low-Rank Approximation (SLRA) [134]	$\min_{\mathbf{G}} F(\mathbf{G}) \quad \text{s.t.} \quad \mathbf{G} \in \mathbb{R}^{(M-K) \times M}$ has full row rank, where $F(\mathbf{G}) := \min_{\mathbf{L}} \|\mathbf{W} \circ (\mathbf{M} - \mathbf{L})\|_F^2 \quad \text{s.t.} \quad \mathbf{G}\mathcal{H}(\mathbf{M}) = \mathbf{0}.$	✗	✗

However, real-world data are not exact and thus $\mathcal{H}(\mathbf{Y})$ is full-rank. Therefore, to find the minimum order realization of the system, we seek a matrix $\hat{\mathbf{Y}}$ which is as close as possible, in the least square sense, to the observed data and the rank of $\mathcal{H}(\hat{\mathbf{Y}})$ is minimal. Formally, we seek to solve the following Hankel structured rank minimization problem

$$\min_{\hat{\mathbf{Y}}} \ \operatorname{rank}(\mathcal{H}(\hat{\mathbf{Y}})) + \frac{\lambda}{2}\|\hat{\mathbf{Y}} - \mathbf{Y}\|_F^2, \tag{5.7}$$

where $\lambda > 0$. Assuming that $\hat{\mathbf{Y}}$ is a solution of (5.7), then $\operatorname{rank}(\mathcal{H}(\hat{\mathbf{Y}}))$ acts as the estimated system order[1] and $\hat{\mathbf{Y}}$ is used next to estimate the system parameters $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \hat{\mathbf{D}}$ and the initial state vector $\hat{\mathbf{x}}_\mathbf{0}$ by solving a series of systems of linear equations [221].

5.2.3 Hankel Rank Minimization Models and Applications

Problem (5.7) is combinatorial due to the discrete nature of the rank function and thus difficult to be solved [69]. To tackle this problem, several approximations have been proposed. In particular, by employing the nuclear norm, which is the convex surrogate of the rank function [69], a convex approximation of (5.7) has been proposed in [70]. By adopting the variational norm of the nuclear norm (i.e., $\|\hat{\mathbf{Y}}\|_* = \min_{\hat{\mathbf{Y}}=\mathbf{U}\mathbf{V}} \|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2$), non-linear approximations to (5.7) have been developed [203, 252]. Furthermore, to estimate the rank of an incomplete Hankel matrix (i.e., in the presence of missing data), the models in [134, 54, 11] have been proposed.

[1] Note that for all experiments presented in this paper, the system order is defined as the rank of the estimated low-rank Hankel matrix, which is calculated as the number of singular values that are larger than 0.5% of the spectral norm, following [70].

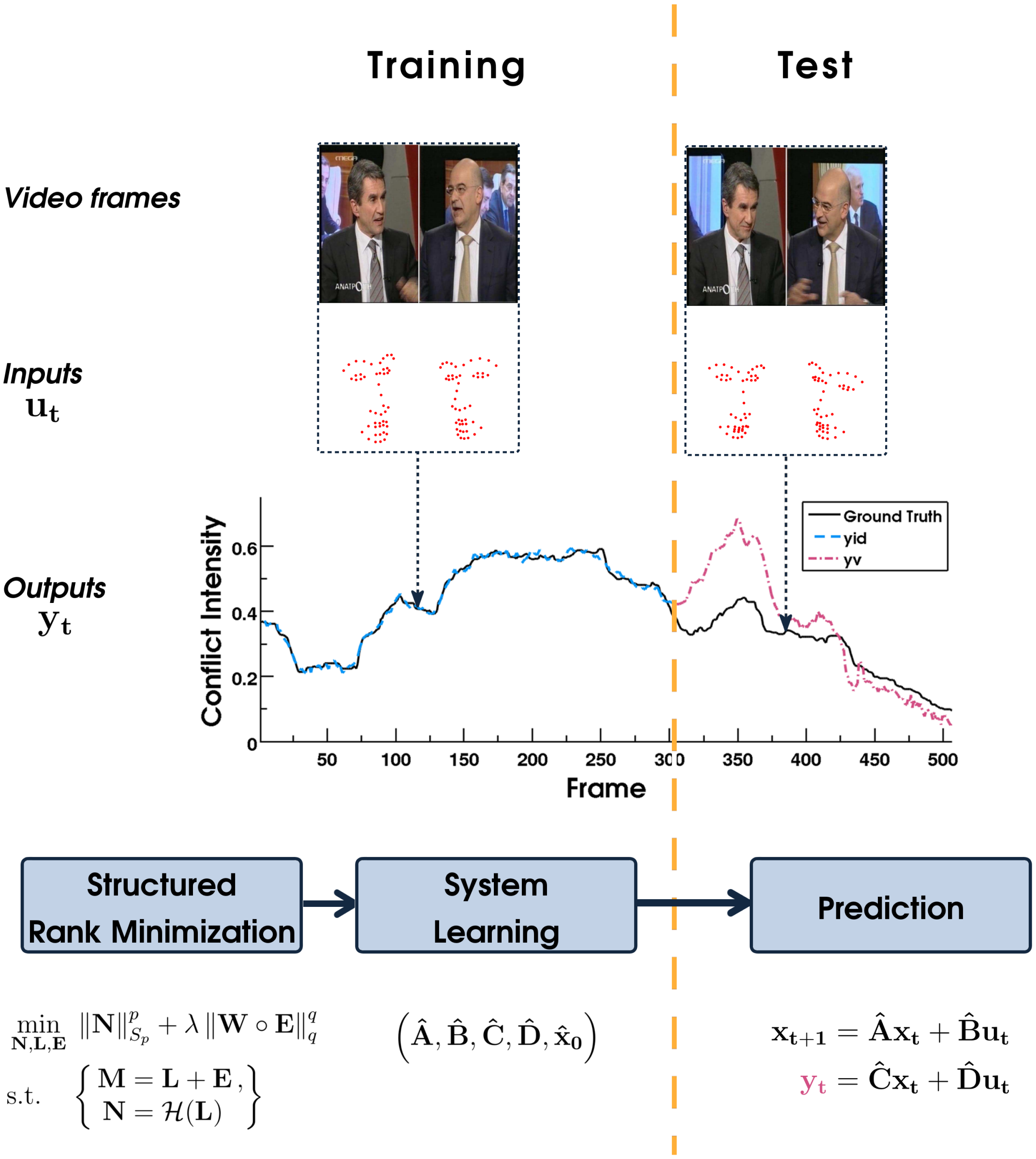

Training
Test
Video frames
Inputs
$\mathbf{u_t}$
Outputs
$\mathbf{y_t}$
Conflict Intensity
0
0.2
0.4
0.6
50
100
150
200
250
300
350
400
450
500
Frame
Ground Truth
yid
yv
Structured Rank Minimization
System Learning
Prediction
$\min_{\mathbf{N},\mathbf{L},\mathbf{E}} \|\mathbf{N}\|_{S_p}^p + \lambda \|\mathbf{W} \circ \mathbf{E}\|_q^q$
s.t. $\begin{Bmatrix} \mathbf{M} = \mathbf{L} + \mathbf{E}, \\ \mathbf{N} = \mathcal{H}(\mathbf{L}) \end{Bmatrix}$
$\left(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \hat{\mathbf{D}}, \hat{\mathbf{x}}_0\right)$
$\mathbf{x_{t+1}} = \hat{\mathbf{A}}\mathbf{x_t} + \hat{\mathbf{B}}\mathbf{u_t}$
$\mathbf{y_t} = \hat{\mathbf{C}}\mathbf{x_t} + \hat{\mathbf{D}}\mathbf{u_t}$

Figure 5.1: Illustration of the proposed dynamic behavior analysis framework, as applied on the task of conflict intensity prediction for a sequence from CONFER dataset. A portion of the sequence frames is used for LTI system learning through the proposed structured rank minimization method (training), while the remaining frames are used for prediction (test).

Representative structured rank minimization models along with the optimization problems that they solve are listed in Table 5.1. The aforementioned models have been mainly applied in the fields of system analysis and control theory for *system identification and realization*, in finance for *time-series analysis and forecasting*, and more recently for computer vision problems (see Section 2.2.2 for details). However, none of these methods has been exploited to learn behavior dynamics based on continuous annotations of behavior or affect and visual features. This will be investigated shortly in Section 5.6.

Remark. Despite their merits, the aforementioned models exhibit the following limitations. By adopting the least squares error, the majority of the models in Table 5.1 assume Gaussian distributions with small variance [90]. Such an assumption rarely holds in real-world data that are often corrupted by sparse, non-Gaussian noise (cf. Section 1). This drawback is partially alleviated in SRPCA [11], where a sparsity promoting norm is incorporated into the nuclear norm minimization problem in order to account for sparse noise of large magnitude. Furthermore, the convex relaxation of the rank function with the nuclear norm in [70, 11] may introduce a relaxation gap. Therefore, due to the above reasons, the estimated rank of the Hankel matrix obtained by the models in [70, 11] may be arbitrarily away from the true one [50]. On the other hand, since the models in [203, 252, 134] rely on factorizations of the Hankel matrix, they implicitly assume that the rank of the Hankel matrix is known in advance; obviously this is not the case in practice. To alleviate the aforementioned limitations and robustly estimate the rank of the Hankel matrix in the presence of gross noise and missing data, a novel structured rank minimization model is detailed next.

5.3 Problem Formulation

Let $\mathbf{M} = [\mathbf{m_0}\ \mathbf{m_1}\ \ldots\ \mathbf{m_{T-1}}] \in \mathbb{R}^{D\times T}$ be a matrix containing in its columns contaminated by gross but sparse noise, time varying data. The goal is to robustly learn the dynamics underlying the data, in the presence of sparse, non-Gaussian noise and missing data.

To this end, we seek to decompose $\mathbf{M}$ as a superposition of two matrices: $\mathbf{M} = \mathbf{L} + \mathbf{E}$, where $\mathbf{L} \in \mathbb{R}^{D\times T}$ and $\mathbf{E} \in \mathbb{R}^{D\times T}$, such that the Hankel matrix of $\mathbf{L}$ (i.e., $\mathcal{H}(\mathbf{L}) \in \mathbb{R}^{M\times N}$) be of minimum rank and $\mathbf{E}$ be sparse. The minimum rank of $\mathcal{H}(\mathbf{L})$ correspond to the minimum-order LTI system that describes the data, while by imposing $\mathbf{E}$ to be sparse, we account for sparse, non-Gaussian noise.

A natural estimator accounting for the low-rank of the Hankel matrix $\mathcal{H}(\mathbf{L})$ and the sparsity of $\mathbf{E}$ is to minimize the rank of $\mathcal{H}(\mathbf{L})$ and the number of non-zero entries of $\mathbf{E}$, measured by the ℓ_0 (quasi)-norm. This is equivalent to solving the following non-convex optimization problem.

$$\min_{\mathbf{L}}\ \operatorname{rank}(\mathcal{H}(\mathbf{L})) + \lambda\|\mathbf{M} - \mathbf{L}\|_0\,, \tag{5.8}$$

where λ is a positive parameter. Clearly, (5.8) is a robust version of the Hankel structured rank minimization problem (5.7).

Problem (5.8) is intractable, as both rank and ℓ_0-norm minimization are NP-hard [222, 143]. In order to tackle this NP-hard problem, both convex and non-convex relaxations of the rank function and the ℓ_0-norm are considered. To this end, we choose to approximate the rank function and the ℓ_0-norm by the Schatten p- and the ℓ_q-norm, respectively, and solve

$$\min_{\mathbf{L}} \ \|\mathcal{H}(\mathbf{L})\|_{S_p}^p + \lambda \, \|\mathbf{M} - \mathbf{L}\|_q^q \,, \tag{5.9}$$

which is a convex optimization problem for $p = q = 1$ (i.e., the Schatten 1-norm is by definition the nuclear norm) and non-convex for $0 < p, q < 1$.

Convex approximations of the rank function and the ℓ_0-(quasi)-norm by means of the nuclear norm (i.e., Schatten 1-norm) [69] and the ℓ_1-norm [60] have been widely applied in several rank and sparsity minimization problems (e.g., [32]). The main advantage of this approach is that the global optimum of the convex problems can be found relatively easily by using off-the-shelf optimization methods such as the ADMM. However, the convexification of rank minimization problems may suffer from the following two drawbacks. First, the recoverability of the low-rank solutions via nuclear norm minimization is only guaranteed under *incoherence assumptions* (e.g., [32]). Such assumptions regarding incoherence may not be guaranteed in practical scenarios [50]. For example in the proposed model, the resulting global optimal solution of the convex instance of (5.9) ($p, q \geq 1$) may be arbitrarily away from the actual solution of (5.8). Second, it is known that the ℓ_1-norm is a biased estimator (e.g., [256]). Since the nuclear norm (or equivalently the Schatten-1 norm) is essentially the application of theℓ_1 norm on the singular values, it may only find a biased solution. To alleviate the aforementioned issues of the convex instance of (5.9), we further consider the non-convex approximation of (5.8) by employing the Schatten-p norm and ℓ_q-norm with $p, q \in (0, 1)$. Such non-convex functions have been shown to provide better estimation accuracy and variable selection consistency [236] in related approximations of ℓ_0-norm regularized rank minimization problems [149, 150, 168].

To disentangle the Schatten p- and ℓ_q-norm minimization sub-problems in (5.9) from the matrix structure and data-fitting requirements, respectively, (5.9) is equivalently written as

$$\min_{\mathbf{N},\mathbf{L},\mathbf{E}} \ \|\mathbf{N}\|_{S_p}^p + \lambda \, \|\mathbf{E}\|_q^q \quad \text{s.t.} \quad \begin{Bmatrix} \mathbf{M} = \mathbf{L} + \mathbf{E} \,, \\ \mathbf{N} = \mathcal{H}(\mathbf{L}) \,. \end{Bmatrix} \tag{5.10}$$

To account also for (partially) missing observations in $\mathbf{M}$, we introduce the matrix $\mathbf{W} \in \mathbb{R}^{D \times T}$ which is given by

$$w_{ij} = \begin{cases} 1 \,, & \text{if } (i,j) \in \Omega \,, \\ 0 \,, & \text{otherwise} \,, \end{cases} \tag{5.11}$$

where $\Omega \subset [1, D] \times [1, T]$ is the set containing the indices of the observed (available) entries in $\mathbf{M}$. By incorporating $\mathbf{W}$ inside the ℓ_q-norm term in (5.10) as a multiplicative weight matrix for $\mathbf{E}$, we arrive at the following problem.

$$\min_{\mathbf{N},\mathbf{L},\mathbf{E}} \ \|\mathbf{N}\|_{S_p}^p + \lambda \|\mathbf{W} \circ \mathbf{E}\|_q^q \quad \text{s.t.} \quad \begin{Bmatrix} \mathbf{M} = \mathbf{L} + \mathbf{E}\,, \\ \mathbf{N} = \mathcal{H}(\mathbf{L})\,. \end{Bmatrix} \tag{5.12}$$

Remark. Note that the choice of the Hankel map $\mathcal{H}(\cdot)$ depends on the application (see Section 5.2.2 and 5.5). In any case, the Hankel matrix $H_{D,1,j,k}(\mathbf{L}) \in \mathbb{R}^{(M=Dj)\times(N=k)}$ is computed according to (5.1); the number of blocks along the row and column dimension j and k, respectively, are set to $j = r + 1$ and $T - r$, where T is the number of observations and $r > n$, with n denoting the system order.

5.4 Algorithmic Frameworks

In this section, the proposed Alternating-Directions Method of Multipliers (ADMM)-based [18] solver is described along its scalable version.

5.4.1 Alternating-Direction Method-Based Algorithm

The ADMM is employed to solve (5.12). To this end, the augmented Lagrangian function for (5.12) is defined as follows.

$$\begin{aligned} \mathcal{L}(\mathbb{V}, \mathbb{Y}, \mu) &= \|\mathbf{N}\|_{S_p}^p + \lambda \|\mathbf{W} \circ \mathbf{E}\|_q^q \\ &+ \langle \mathbf{M} - \mathbf{L} - \mathbf{E}, \mathbf{\Lambda_1} \rangle + \langle \mathbf{N} - \mathcal{H}(\mathbf{L}), \mathbf{\Lambda_2} \rangle \\ &+ \frac{\mu}{2}\Big(\|\mathbf{M} - \mathbf{L} - \mathbf{E}\|_F^2 + \|\mathbf{N} - \mathcal{H}(\mathbf{L})\|_F^2 \Big)\,, \end{aligned} \tag{5.13}$$

where μ is a positive parameter and $\mathbb{V} := \{\mathbf{N} \in \mathbb{R}^{M\times N}, \mathbf{L} \in \mathbb{R}^{D\times T}, \mathbf{E} \in \mathbb{R}^{D\times T}\}$, $\mathbb{Y} := \{\mathbf{\Lambda_1} \in \mathbb{R}^{D\times T}, \mathbf{\Lambda_2} \in \mathbb{R}^{M\times N}\}$ are the sets containing all the unknown variables and the Lagrange multipliers for the equality constraints in (5.12), respectively. Specifically, at each iteration of the proposed ADMM-based solver, (5.13) is minimized with respect to each variable in $\mathbb{V}$ in an alternating fashion and, subsequently, the Lagrange multipliers in $\mathbb{Y}$ and the parameter μ are updated. The iteration index is denoted herein by i. The notation $\mathbb{L}(\mathbf{N}, \mathbb{Y}[i], \mu[i])$ is used to denote the solution stage in which all other variables but $\mathbf{N}$ are kept fixed, and similarly for the other unknown variables.

The solutions of minimization of (5.13) with respect to $\mathbf{E}$ and $\mathbf{N}$ are based on the operators and Lemmas that are introduced next. Minimizing (5.13) with respect to $\mathbf{L}$ does not admit a closed form solution due to the presence of the quadratic terms. Similarly to [70], to 'cancel out' these terms we add a proximal term to the respective partial augmented Lagrangian. The additive term is based on the (semi-) norm $\|\cdot\|_{\mathcal{Q}_0}$ induced by the (semi-) inner product $\mathbf{P}^T\mathcal{Q}_0\mathbf{P}$, with $\mathcal{Q}_0$ being the positive (semi-) definite matrix given by

$$\mathcal{Q}_0 = L\mathcal{I} - \mathcal{H}^*\mathcal{H} \succeq 0, \tag{5.14}$$

where $L := \min\{j, k\}$. As shown in Section 5.2.1, $\sqrt{L}$ is the upper bound of the spectral norm of the Hankel adjoint map $\mathcal{H}^*$.

Thus, given the variables $\mathbb{V}[i]$, the Lagrange multipliers $\mathbb{Y}[i]$ and the parameter $\mu[i]$ at iteration i, the updates of the proposed solver, summarized in Algorithm (4), are as follows.

Update the primal variables.

$$\begin{aligned}\mathbf{E}[i+1] &= \arg\min_{\mathbf{E}} \mathcal{L}(\mathbf{E}, \mathbb{Y}[i], \mu[i]) \\ &= \arg\min_{\mathbf{E}} \lambda\mu[i]^{-1} \|\mathbf{W} \circ \mathbf{E}\|_q^q \\ &+ \frac{1}{2}\left\|\mathbf{E} - \left(\mathbf{M} - \mathbf{L} + \mu[i]^{-1}\mathbf{\Lambda_1}[i]\right)\right\|_F^2\end{aligned} \tag{5.15}$$

$$\begin{aligned}\mathbf{N}[i+1] &= \arg\min_{\mathbf{N}} \mathcal{L}(\mathbf{N}, \mathbb{Y}[i], \mu[i]) \\ &= \arg\min_{\mathbf{N}} \mu[i]^{-1} \|\mathbf{N}\|_{S_p}^p \\ &+ \frac{1}{2}\left\|\mathbf{N} - \left(\mathcal{H}(\mathbf{L}) - \mu[i]^{-1}\mathbf{\Lambda_2}[i]\right)\right\|_F^2\end{aligned} \tag{5.16}$$

$$\mathbf{L}[i+1] = \arg\min_{\mathbf{L}} \mathcal{L}(\mathbf{L}, \mathbb{Y}[i], \mu[i]) + \frac{\mu[i]}{2}\|\mathbf{L} - \mathbf{L}[i]\|_{\mathcal{Q}_0}^2 \tag{5.17}$$

Update the Lagrange multipliers.

$$\mathbf{\Lambda_1}[i+1] = \mathbf{\Lambda_1}[i] + \mu[i]\left(\mathbf{M} - \mathbf{L} - \mathbf{E}\right) \tag{5.18}$$

$$\mathbf{\Lambda_2}[i+1] = \mathbf{\Lambda_2}[i] + \mu[i]\left(\mathbf{N} - \mathcal{H}(\mathbf{L})\right) \tag{5.19}$$

Equation (5.15), which offers the update for $\mathbf{E}$, is solved based on the *generalized soft thresholding operator* proposed in [150] and briefly described next. Consider the following problem.

$$\arg\min_{\mathbf{B}} \alpha\|\mathbf{B}\|_q^q + \frac{1}{2}\|\mathbf{B} - \mathbf{Z}\|_F^2, \tag{5.20}$$

with $\mathbf{B} \in \mathbb{R}^{m\times n}$ and α a positive parameter. Problem (5.20) is separable with respect to the elements of $\mathbf{B}$ and is thereby decomposed into $m \times n$ sub-problems of the form

$$\min_{b_{ij}} \alpha|b_{ij}|^q + \frac{1}{2}(b_{ij} - z_{ij})^2 . \tag{5.21}$$

Let us now define $h(b_{ij}) = \alpha|b_{ij}|^q + \frac{1}{2}(b_{ij} - z_{ij})^2$, $c_1 = (\alpha q(1-q))^{\frac{1}{2-q}}$ and $c_2 = c_1 + \alpha q|c_1|^{q-1}$. Equation (5.21) admits an analytical solution for $q \in (0, 1]$ given by

$$b^*_{ij} = \begin{cases} 0 & \text{if } |b_{ij}| \leq c_2 \\ \arg\min_{b_{ij}\in\{0,\rho_1\}} h(b_{ij}) & \text{if } b_{ij} > c_2 \\ \arg\min_{b_{ij}\in\{0,\rho_2\}} h(b_{ij}) & \text{if } b_{ij} < -c_2 \, , \end{cases} \tag{5.22}$$

where ρ_1 and ρ_2 are the roots of $h'(b_{ij}) = \alpha q|b_{ij}|^{q-1}\text{sgn}(b_{ij}) + b_{ij} - z_{ij} = 0$ in $[c_1, z_{ij}]$ and $[z_{ij}, -c_1]$, respectively. The roots can easily be found by applying the iterative Newton-Raphson root-finding method initialized at z_{ij}. Similarly to [168], we henceforth call the element-wise solver (5.22) *generalized q-shrinkage operator* and denote it by $\mathcal{S}^q_\alpha\{\cdot\}$. Note that when $q = 1$ the aforementioned operator reduces to the element-wise application of the well-known *shrinkage operator* [32], defined by

$$\mathcal{S}_\alpha\{x\} := \text{sgn}(x)\max\{|x| - \alpha, 0\}. \tag{5.23}$$

We shall denote by $\mathcal{S}^q_{(\alpha,\mathbf{W})}\{\cdot\}$ the operator for which $\bar{\alpha} = \alpha w_{ij}$, with $\mathbf{W} \in \mathbb{R}^{m\times n}$ known, is used instead of α for the solution of each respective b_{ij} in (5.22).

The solution of (5.16), that is, the minimization of (5.13) with respect to $\mathbf{N}$, is based on the following Lemma.

Lemma 5.4.1 *[150] The solution of the optimization problem*

$$\arg\min_{\mathbf{B}} a \, \|\mathbf{B}\|^p_{S_p} + \frac{1}{2} \|\mathbf{B} - \mathbf{Z}\|^2_F \, , \tag{5.24}$$

with $p \in (0, 1]$, is given by $\mathbf{B} = \mathbf{U_S}\mathcal{S}^p_\alpha\{\mathbf{\Sigma}\}\mathbf{V_S}^T$, where $\mathbf{U_S}\mathbf{\Sigma}\mathbf{V_S}^T = \mathbf{Z}$ is the SVD of $\mathbf{Z}$.

We shall denote by $\mathcal{D}^p_\alpha\{\cdot\}$ the operator – henceforth called *generalized singular value p-shrinkage operator* – that solves (5.24).

Clearly, problem (5.17) admits a closed-form solution.

Algorithm 4 ADMM solver for (5.12).

```
Input: Data: $\mathbf{M} \in \mathbb{R}^{D\times T}$. Weights: $\mathbf{W} \in \mathbb{R}^{D\times T}$. Parameters: $\{p, q, \lambda\}$. Definitions: $\mathcal{H}(\cdot)$.
1: Set $r = \frac{T+2}{d+m+1}$, $j = r+1$, $k = T-j+1$, $M = Dj$, $N = k$, $L = \min\{j,k\}$, $\rho = 1.05$, $\mu_{\max} = 10^{10}$, $\epsilon_1 > 0, \epsilon_2 > 0$.
2: Initialize: Set $\mathbf{L}[0] = 1.1\mathbf{M}$ and $\boldsymbol{\Lambda}_\mathbf{1}[0], \boldsymbol{\Lambda}_\mathbf{2}[0]$ to zero matrices. Set $\mu[0] = L(2\lambda\|\mathbf{M}\|)^{-1}$.
3: while not converged do
4:     $\mathbf{E}[i+1] \leftarrow \mathcal{S}^q_{(\lambda\mu[i]^{-1},\mathbf{W})}\left\{\mathbf{M} - \mathbf{L}[i] + \mu[i]^{-1}\boldsymbol{\Lambda}_\mathbf{1}[i]\right\}$.
5:     $\mathbf{N}[i+1] \leftarrow \mathcal{D}^p_{(\mu[i]^{-1})}\left\{\left(\mathcal{H}(\mathbf{L}[i]) - \mu[i]^{-1}\boldsymbol{\Lambda}_\mathbf{2}[i]\right)\right\}$.
6:     $\mathbf{L}[i+1] \leftarrow \frac{1}{L+1}\left(\mathcal{H}^*\left(\mathbf{N}[i+1]+\mu[i]^{-1}\boldsymbol{\Lambda}_\mathbf{2}[i]-\mathcal{H}(\mathbf{L}[i])\right)+\mu[i]^{-1}\boldsymbol{\Lambda}_\mathbf{1}[i]+\mathbf{M}-\mathbf{E}[i+1]+L\mathbf{L}[i]\right)$.
7:     Update the Lagrange multipliers by (5.18), (5.19).
8:     Update $\mu$: $\mu[i+1] = \min(\rho\mu[i], \mu_{\max})$.
9: end while
Output: $\mathbb{V} = \{\mathbf{N} \in \mathbb{R}^{M\times N}, \mathbf{L} \in \mathbb{R}^{D\times T}, \mathbf{E} \in \mathbb{R}^{D\times T}\}$.
```

The proposed ADMM-based solver is summarized in Algorithm 4. The latter is terminated when the following conditions are met

$$\left\{\begin{array}{l} \max\left\{\dfrac{\|\mathbf{M}-\mathbf{L}[i+1]-\mathbf{E}[i+1]\|_F}{\|\mathbf{M}\|_F},\right. \\ \left.\dfrac{\|\mathbf{N}[i+1]-\mathcal{H}(\mathbf{L}[i+1])\|_F}{\|\mathbf{M}\|_F}\right\} < \epsilon_1, \\ \max\left\{\dfrac{\|\mathbf{N}[i+1]-\mathbf{N}[i]\|_F}{\|\mathbf{M}\|_F}, \dfrac{\|\mathbf{L}[i+1]-\mathbf{L}[i]\|_F}{\|\mathbf{M}\|_F},\right. \\ \left.\dfrac{\|\mathbf{E}[i+1]-\mathbf{E}[i]\|_F}{\|\mathbf{M}\|_F}\right\} < \epsilon_2\,, \end{array}\right\} \tag{5.25}$$

where ϵ_1 and ϵ_2 are small positive parameters, or a maximum of 1000 iterations are reached.

Computational Complexity and Convergence. The cost of each iteration in Algorithm 4 is dominated by the calculation of the *generalized singular value p-shrinkage operator* in Step 5, which involves a complexity equal to that of SVD, i.e., $\mathcal{O}\left(\max\{M^2N, MN^2\}\right)$. The *generalized q-shrinkage operator*, utilized in Step 4, entails linear complexity $\mathcal{O}(DT)$.

Regarding the convergence of Algorithm 4, there is no established convergence proof of the ADMM for problems in the form of (5.12). Indeed, the ADMM is only known to converge for convex separable problems with up to two blocks of variables (e.g., [18], [32]). However, this is not the case even in the convex instance of (5.12) (i.e., when $p = q = 1$), since the optimization problem involves more than two blocks of variables. For the multi-block separable convex problems, with three or more blocks of variables, it is known that the original ADMM is not necessarily convergent [37]. On the other hand, theoretical convergence analysis of the ADMM for non-convex problems is rather limited, making either assumptions on the iterates of the algorithm [244, 131] or dealing with special non-convex models [119, 231, 232], none of which is applicable for the proposed optimization problem (5.12). However, it is worth noting that the ADMM exhibits good numerical performance in non-convex problems such as non-negative matrix factorization [208], tensor decomposition [120], matrix separation [201, 168], matrix completion [244], motion segmentation [94], to mention but a few.

To the best of our knowledge, the only work which focuses on the convergence analysis of the ADMM when applied for the optimization of piecewise linear functions such as the Schatten p-norm and the ℓ_q–norm (when $0 < p, q \leq 1$) is the recent preprint of [235]. However, since a systematic convergence analysis is out of the scope of this paper, we plan to adapt the analysis in [235] in order to analyze the convergence of the proposed algorithm in the future.

Even though we cannot theoretically guarantee the convergence of the proposed solver, the experimental results on synthetic data in Section 5.6.1 show that its numerical performance is good in practice. Specifically, the empirical convergence of the proposed solver is evidenced, where both the primal residual and the primal objective are non-increasing after the very few iterations (see Fig. 5.4). Similar convergence behavior characterizes also the experiments on real-world data presented in Section 5.6, where we have observed that even the non-convex variant with $p = q = 0.1$ of the proposed method (5.12) needs no more than 180 iterations to converge in most cases.

5.4.2 Scalable Version of the Algorithm

To improve the scalability and reduce the computational complexity of the ADMM-based Algorithm 4, we develop here a scalable version. Depending on the application, and more specifically, the number of inputs and/or outputs employed and the number of observations, the dimension of the Hankel matrix $\mathcal{H}(\mathbf{L}) \in \mathbb{R}^{M \times N}$ can rise largely, which makes the calculation of SVD prohibitive. To alleviate the aforementioned computational complexity issue, we further impose that $\mathcal{H}(\mathbf{L}) \in \mathbb{R}^{M \times N}$ is factorized into an orthonormal matrix and a low-rank matrix as $\mathcal{H}(\mathbf{L}) = \mathbf{QR}$, with $\mathbf{Q} \in \mathbb{R}^{M \times K}$, $\mathbf{R} \in \mathbb{R}^{K \times N}$ and $K \ll M, N$. In this factorization, $\mathbf{Q} \in \mathbb{R}^{M \times K}$ is a column-orthogonal matrix satisfying $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ and $\mathbf{R} \in \mathbb{R}^{K \times N}$ is a low-rank matrix representing the embedding of $\mathcal{H}(\mathbf{L})$ onto the K-dimensional subspace spanned by the columns of $\mathbf{Q}$.

Due to the unitary invariance of the Schatten p-norm, the following equality holds

$\|\mathbf{QR}\|_{S_p} = \|\mathbf{R}\|_{S_p}$. Thus, by incorporating the factorization $\mathcal{H}(\mathbf{L}) = \mathbf{QR}$ and adding the orthonormality constraint for $\mathbf{Q}$, (5.12) is written as

$$\min_{\mathbf{R},\mathbf{L},\mathbf{E},\mathbf{Q}} \quad \|\mathbf{R}\|_{S_p}^p + \lambda \|\mathbf{W} \circ \mathbf{E}\|_q^q \\ \text{s.t.} \quad \left\{ \begin{aligned} \mathbf{M} &= \mathbf{L} + \mathbf{E}\,, \\ \mathbf{QR} &= \mathcal{H}(\mathbf{L})\,, \\ \mathbf{Q}^\top \mathbf{Q} &= \mathbf{I}. \end{aligned} \right\} \tag{5.26}$$

Since $MK + KN \ll MN$, the number of variables has been significantly reduced. Clearly, this modification reduces the overall complexity of the method, since the SVD is now applied on $M \times K$ and $K \times N$ matrices as opposed to a $M \times N$ matrix.

The ADMM is employed to solve (5.26). With $\mathbb{V} := \{\mathbf{R} \in \mathbb{R}^{K\times N}, \mathbf{L} \in \mathbb{R}^{D\times T}, \mathbf{E} \in \mathbb{R}^{D\times T}, \mathbf{Q} \in \mathbb{R}^{M\times K}\}$ and $\mathbb{Y} := \{\mathbf{\Lambda_1} \in \mathbb{R}^{D\times T}, \mathbf{\Lambda_2} \in \mathbb{R}^{M\times N}\}$ defined as the sets containing all the unknown variables and the Lagrange multipliers for the first two equality constraints in (5.26), respectively, the (partial) augmented Lagrangian function is defined as

$$\begin{aligned} \mathcal{L}^{\text{sc}}(\mathbb{V}, \mathbb{Y}, \mu) &= \|\mathbf{R}\|_{S_p}^p + \lambda \|\mathbf{W} \circ \mathbf{E}\|_q^q \\ &+ \langle \mathbf{M} - \mathbf{L} - \mathbf{E}, \mathbf{\Lambda_1} \rangle + \langle \mathbf{QR} - \mathcal{H}(\mathbf{L}), \mathbf{\Lambda_2} \rangle \\ &+ \frac{\mu}{2}\Big(\|\mathbf{M} - \mathbf{L} - \mathbf{E}\|_F^2 + \|\mathbf{QR} - \mathcal{H}(\mathbf{L})\|_F^2 \Big)\,, \end{aligned} \tag{5.27}$$

where μ is a positive parameter. Therefore, at each iteration of the ADMM-based solver for (5.26), we solve

$$\min_{\mathbb{V}} \mathcal{L}^{\text{sc}}(\mathbb{V}, \mathbb{Y}, \mu) \quad \text{s.t.} \quad \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}\,, \tag{5.28}$$

with respect to each variable in $\mathbb{V}$ in an alternating fashion and, subsequently, the Lagrange multipliers in $\mathbb{Y}$ and the parameter μ are updated.

The proposed solver for (5.26) is summarized in Algorithm 5. The updates for $\mathbf{R}, \mathbf{L}, \mathbf{E}$ are similar to those employed to solve (5.12). The solution of (5.28) with respect to $\mathbf{Q}$ is based on the *Procrustes operator*, which is defined as $\mathcal{P}[\mathbf{L}] = \mathbf{AB}^T$ for a matrix $\mathbf{L}$ with SVD $\mathbf{L} = \mathbf{A\Sigma B}^T$ and solves the problem in the following Lemma.

Lemma 5.4.2 *[261] The constrained minimization problem:*

$$\arg\min_{\mathbf{B}} \|\mathbf{A} - \mathbf{B}\|_F^2 \quad \textit{s.t.} \quad \mathbf{B}^T\mathbf{B} = \mathbf{I} \tag{5.29}$$

has a closed-form solution given by $\mathbf{P} = \mathcal{P}[\mathbf{A}]$.

Algorithm 5 ADMM solver for (5.26) (scalable version).

```
Input: Data: $\mathbf{M} \in \mathbb{R}^{D\times T}$. Weights: $\mathbf{W} \in \mathbb{R}^{D\times T}$. Parameters: $\{p, q, \lambda\}$, number of components $K$. Definitions: $\mathcal{H}(\cdot)$.
1: Set $r = \dfrac{T+2}{d+m+1}$, $j = r+1$, $k = T-j+1$, $M = Dj$, $N = k$, $L = \min\{j,k\}$, $\rho = 1.05$, $\mu_{\max} = 10^{10}$, $\epsilon_1 > 0, \epsilon_2 > 0$.
2: Initialize: Set $\mathbf{Q}[0], \mathbf{\Lambda_1}[0], \mathbf{\Lambda_2}[0]$ to zero matrices and $\mathbf{L}[0] = 1.1\mathbf{M}$. Set $\mu[0] = L(2\lambda\|\mathbf{M}\|)^{-1}$.
3: while not converged do
4:     $\mathbf{E}[i+1] \leftarrow \mathcal{S}^q_{(\lambda\mu[i]^{-1},\mathbf{W})}\left\{\mathbf{M} - \mathbf{L}[i] + \mu[i]^{-1}\mathbf{\Lambda_1}[i]\right\}$.
5:     $\mathbf{R}[i+1] \leftarrow \mathcal{D}^p_{(\mu[i]^{-1})}\left\{\mathbf{Q}^T[i]\left(\mathcal{H}(\mathbf{L}[i]) - \mu[i]^{-1}\mathbf{\Lambda_2}[i]\right)\right\}$.
6:     $\mathbf{Q}[i+1] \leftarrow \mathcal{P}\left\{\left(\mathcal{H}(\mathbf{L}[i]) - \mu[i]^{-1}\mathbf{\Lambda_2}[i]\right)\mathbf{R}^T[i+1]\right\}$.
7:     $\mathbf{L}[i+1] \leftarrow \frac{1}{L+1}\left(\mathcal{H}^*\left(\mathbf{Q}[i+1]\mathbf{R}[i+1] + \mu[i]^{-1}\mathbf{\Lambda_2}[i] - \mathcal{H}(\mathbf{L}[i])\right) + \mu[i]^{-1}\mathbf{\Lambda_1}[i] + \mathbf{M} - \mathbf{E}[i+1] + L\mathbf{L}[i]\right)$.
8:     Update the Lagrange multipliers by (5.18), (5.19).
9:     Update $\mu$: $\mu[i+1] = \min(\rho\mu[i], \mu_{\max})$.
10: end while
Output: $\mathbb{V} = \{\mathbf{R} \in \mathbb{R}^{K\times N}, \mathbf{L} \in \mathbb{R}^{D\times T}, \mathbf{E} \in \mathbb{R}^{D\times T}, \mathbf{Q} \in \mathbb{R}^{M\times K}\}$.
```

Computational Complexity and Convergence. The cost of each iteration in Algorithm 5 is dominated by the calculation of the *generalized singular value p-shrinkage operator* and the *Procrustes operator* in Step 5 and 6, respectively, which both rely on SVD, thus involving respective complexities of $\mathcal{O}\left(\max\{K^2N, KN^2\}\right)$ and $\mathcal{O}\left(\max\{M^2K, MK^2\}\right)$. It is worth stressing again that choosing $K \ll M, N$, which implies $MK + KN \ll MN$, leads to a significantly reduced overall complexity for Algorithm 5 compared to that of Algorithm 4, which is instead dominated by a SVD on a $M \times N$ matrix, hence $\mathcal{O}\left(\max\{M^2N, MN^2\}\right)$. Again, the *generalized q-shrinkage operator*, utilized in Step 4, entails linear complexity $\mathcal{O}(DT)$.

Regarding the convergence of Algorithm 5 which solves the scalable version of the proposed model (5.26), there is no yet established convergence proof of the ADMM for problems in the form of (5.26). The discussion provided above on the convergence of Algorithm 4 applies to a large extent for Algorithm 5 as well. As a matter of fact, theoretical analysis for the convergence of Algorithm 5 becomes more challenging, compared to the case of Algorithm 4, considering that the factorization $\mathbf{QR} = \mathcal{H}(\mathbf{L})$ and the non-linear orthonormality constraint $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ are introduced in the scalable version of the proposed model (5.26). It is also worth noting that problem (5.26) is always non-convex due to these two equality constraints, and thus the solutions yielded by the optimization problems (5.12) and (5.26) cannot be related.

However, it has been shown in [124] that the ADMM converges to a local minimum for a problem similar to problem (5.26) with convex objective function, i.e., $p, q \geq 1$. To the best of our knowledge, for the case $0 < p, q < 1$, i.e., when the Schatten p-norm and the ℓ_q-norm act as non-convex approximations of the rank function and the ℓ_0-(quasi) norm, respectively, there has been no theoretical evidence for the convergence of the ADMM for the problem (5.26) and further investigation is needed.

Nevertheless, the ADMM has been shown to achieve good numerical performance in non-convex subspace learning problems employing a similar matrix factorization approach with one of the factors being orthonormal [187, 168]. Also, experimental results on synthetic data evidence the empirical convergence of Algorithm 5, which has been found to be similar to that shown for Algorithm 4 ($p = q = 0.5$) in Fig. 5.4. Good numerical performance is also achieved by the scalable solver in the experiments presented in Section 5.6.

5.5 Dynamic Behavior Analysis Frameworks based on Hankel Structured Rank Minimization

In this section, we develop two frameworks for dynamic behavior analysis.

5.5.1 Dynamic Behavior Prediction

Consider the case where continuous-time, real-valued annotations characterizing dynamic behavior or affect (e.g., conflict, valence, arousal), manifested in a video sequence of T frames, are available for a number of consecutive frames $t = 0, 1, \dots, T_{train} - 1$ (training set). The goal herein is to first learn a low-order LTI system that generates the annotations as outputs $\mathbf{Y} = [\mathbf{y_0}, \mathbf{y_1}, \dots, \mathbf{y_{T_{train}-1}}] \in \mathbb{R}^{m \times T_{train}}$ when visual features act as inputs $\mathbf{U} = [\mathbf{u_0}, \mathbf{u_1}, \dots, \mathbf{u_{T_{train}-1}}] \in \mathbb{R}^{d \times T_{train}}$, and next use it to predict behavior measurements $\hat{\mathbf{y}}_\mathbf{t}$ for the remaining frames of the sequence $t = T_{train}, \dots, T - 1$ (test set), based on the respective features $\mathbf{u_t}$. To this end, the following framework is proposed.

First, the proposed structured minimization problem (5.10) is solved, with $\mathbf{M} = \mathbf{Y}$ and the Hankel map $\mathcal{H}(\cdot)$ defined as in Section 5.2.2, to estimate the system order. Second, the low-rank solution $\mathcal{H}(\mathbf{L})$ is used to estimate the system matrices $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \hat{\mathbf{D}}$ and the initial state vector $\hat{\mathbf{x}}_\mathbf{0}$ by solving a system of linear equations, following, for example, [221]. Finally, test set predictions $\hat{\mathbf{y}}$ ($t = T_{train}, \dots, T - 1$) for dynamic behavior are obtained by applying the equations of the learned state-space model (5.3) for $t = 0, 1, \dots, T_{train} - 1$, with the visual features used as inputs $\mathbf{u_t}$.

Applications. The aforementioned framework can be used for continuous prediction of any number or type of real-valued behavioral attributes manifested in a video sequence, by employing a portion of consecutive frames (even a few seconds) to learn a LTI system as described above (see Section 5.6).

5.5.2 Dynamic Behavior Prediction with Partially Missing Outputs

Consider now the scenario in which the goal is to predict *missing* (or unreliable) and not necessarily consecutive real-valued measurements of dynamic behavior or affect, viewed as missing outputs $\bar{\mathbf{y}}_\mathbf{t}$ of a low-order LTI system, directly by employing the observed visual features as inputs $\mathbf{u}_\mathbf{t}$ and the available annotations as outputs $\mathbf{y}_\mathbf{t}$, without explicitly learning the system. Herein, we approach this task as a (Hankel) structured low-rank *matrix completion* problem and address it by means of the following predictive framework that is based on the proposed model (5.12).

Let $\mathbf{Y} = [\mathbf{y_0}, \mathbf{y_1}, \dots, \mathbf{y_{T-1}}] \in \mathbb{R}^{m \times T}$ and $\mathbf{U} = [\mathbf{u_0}, \mathbf{u_1}, \dots, \mathbf{y_{T-1}}] \in \mathbb{R}^{m \times T}$ be the matrices containing all T observations (available and missing) of inputs and outputs, respectively, and let $\mathbf{M} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{U} \end{bmatrix} \in \mathbb{R}^{D \times T}$ and $\mathcal{H}(\mathbf{M}) = H_{D,1,r+1,T-r}\left(\begin{bmatrix} \mathbf{Y} \\ \mathbf{U} \end{bmatrix}\right)$, with $D = m + d$. Let also $\Omega \subset [1, D] \times [1, T]$ be the set containing the indices of the observed (available) entries in $\mathbf{M}$. When outputs are noisy, the following property holds only approximately [221], under the assumption of persistently exciting inputs.

$$\operatorname{rank}\left(\mathcal{H}(\mathbf{M})\right) = n + \operatorname{rank}\left(H(\mathbf{U})\right). \tag{5.30}$$

Thus, a low-rank approximation of $\mathcal{H}(\mathbf{M})$ should be obtained to estimate the true order of the system n.

To this end, the proposed model (5.12) is solved, with $\mathbf{M}$ defined as above and $\mathbf{W}$ computed according to (5.11). Note that this process simultaneously 'completes' the missing observations of $\mathbf{M}$, by forcing the approximation of $\mathcal{H}(\mathbf{M})$ to be low-rank, or in other words, the 'completed' trajectory $\mathbf{L}$ to follow the same linear dynamics underlying the observed trajectory $\mathbf{M}$. Finally, the missing outputs are recovered from the respective entries of the low-rank approximation $\mathcal{H}(\mathbf{L})$. Notably, this framework has the advantage that the missing observations are obtained directly by solving (5.12), thus avoiding the computational load associated with learning a minimum order realization of the system.

Applications. The aforementioned framework can achieve prediction of missing (past or future) observations pertaining to dynamic human behavior or affect, with the latter used as outputs of a low-order LTI system. For instance, a computer vision problem that can be addressed by means of the proposed framework is the problem of *tracklet matching* [56, 57, 54], which consists of stitching trajectories of detections belonging to the same target. For this task, one needs to assess whether the joint trajectory of detections $\mathbf{M} = \left[\mathbf{Y}_{\mathbf{start}} \bar{\mathbf{Y}}_{\mathbf{inter}} \mathbf{Y}_{\mathbf{end}}\right]$, where $\mathbf{Y}_{\mathbf{start}}$ and $\mathbf{Y}_{\mathbf{end}}$ are the observed trajectories and $\bar{\mathbf{Y}}_{\mathbf{inter}}$ is a zero-valued matrix corresponding to the 'missing' intermediate trajectory, is the output of the same autonomous (output-only) LTI system that generated $\mathbf{Y}_{\mathbf{start}}$ and $\mathbf{Y}_{\mathbf{end}}$. This is achieved by solving (5.12) for $\bar{\mathbf{Y}}_{\mathbf{inter}}$, with $\mathbf{M}$ defined as above, and subsequently comparing $\operatorname{rank}(\mathcal{H}(\mathbf{L}))$ with $\operatorname{rank}(\mathcal{H}(\mathbf{Y}_{\mathbf{start}}))$ and $\operatorname{rank}(\mathcal{H}(\mathbf{Y}_{\mathbf{end}}))$ (see Section 5.6.4).

5.6 Experiments

The efficiency of the proposed structured rank minimization methods is evaluated on synthetic data corrupted by sparse, non-Gaussian noise (Section 5.6.1), as well as on real-world data with applications to: i) *conflict intensity prediction* (Section 5.6.2), ii) *valence-arousal prediction* (Section 5.6.3), and iii) *tracklet matching* (Section 5.6.4). For the case of dynamic behavior analysis experiments on real-world data, for the first two tasks, the framework described in Section 5.5.1 is employed, while for the last we utilize the framework described in Section 5.5.2.

Aside from the proposed methods, five structured minimization methods are also examined, namely HRM[2] [70], SVD-free [203], SRPCA [11], IHTLS [54], and SLRA [134] (see further details on these methods in Table 5.1). For all experiments presented in our paper, a grid search is employed to tune the parameter λ of the proposed methods or any other parameters of the compared methods that need tuning. Tuning is performed by following an *out-of-sample evaluation*, that is, the last portion of the training frames is withheld for validation and the best-performing model is used for testing. Specifically, the last $2r$ training observations, with r defined in Section 5.3, are kept out for validation in all our experiments.

5.6.1 Experiment on Synthetic Data

In the experiments presented in this section, the efficiency of the proposed method (5.12) is evaluated on synthetic data corrupted with sparse, non-Gaussian noise. In order to generate Hankel matrices of given rank n, we follow the methodology proposed in [169], that is, T outputs $y(t)$ of an autonomous stable LTI system of order n are generated by applying the

[2] The Dual AGP algorithm in [70] is used.

following formula

$$y(t) = \sum_{k=1}^{n} z_k^t\,, \quad t = 1, 2, \ldots, T\,, \tag{5.31}$$

where z_k appear in pairs of conjugate numbers so that the observations $y(t)$ are real numbers. It follows naturally that a $M \times N$ Hankel matrix $\mathbf{Y} = \mathcal{H}(\mathbf{y}) = H_{1,1,M,N}(\mathbf{y})$ with $\mathbf{y}$ derived according to (5.31) has rank equal to n [169]. Subsequently, sparse, non-Gaussian noise $\boldsymbol{\eta} \in \mathbb{R}^{1\times T}$ is added to the original signal $\mathbf{y}$, with the non-zero entries following the Bernoulli model with probability $\rho = 0.2$, as in [32]. The final corrupted signal is formed as $\tilde{\mathbf{y}} = \mathbf{y} + \boldsymbol{\eta}$, with the corresponding noisy Hankel matrix $\tilde{\mathbf{Y}} = \mathcal{H}(\tilde{\mathbf{y}})$ being full-rank.

In what follows, the efficiency of various structured rank minimization methods in reconstructing the noiseless system outputs $y(t)$, $t = 1, 2, \ldots, T$, by finding a low-rank approximation $\hat{\mathbf{Y}} = \mathcal{H}(\hat{\mathbf{y}})$ given the noisy Hankel matrix $\tilde{\mathbf{Y}}$, is experimentally assessed in various scenarios. The reconstruction error, for both the noiseless observations $\mathbf{y}$ and the noise $\boldsymbol{\eta}$, is measured in terms of relative reconstruction error as follows.

$$\operatorname{err}(\mathbf{s}, \hat{\mathbf{s}}) = \frac{\|\mathbf{s} - \hat{\mathbf{s}}\|}{\|\mathbf{s}\|}\,, \tag{5.32}$$

with $\mathbf{s}$ denoting the original signal and $\hat{\mathbf{s}}$ denoting the estimated signal by the algorithm.

Experiment with varying system orders. Herein, experiments are conducted for various orders of the LTI system generating the 'clean' data, as described above. Specifically, the system order n is varied in $\{6, 12, 18\}$. For each value of n the experiment is repeated 10 times, that is, for 10 different output trajectories $\mathbf{y} \in \mathbb{R}^{1\times T}$ computed by randomly selecting the complex coefficients in (5.31). For the proposed model, Algorithm 1 is used and the following combinations are examined for the p and q values corresponding to the Schatten p- and ℓ_q-norm, respectively: $(p, q) \in \{(1, 1), (0.9, 0.9), (0.5, 0.5), (0.1, 0.1)\}$. The methods HRM, SVD-free, SRPCA, IHTLS, and SLRA (listed in Table 1) are also evaluated for comparison. For each method, results are reported in terms of minimum reconstruction error $\operatorname{err}(\mathbf{y}, \hat{\mathbf{y}})$ computed according to (5.32). Performance is also evaluated in terms of reconstruction error for the noise signal $\operatorname{err}(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})$ and the Pearson Correlation Coefficient (COR) measured between the noiseless observations $\mathbf{y}$ and the reconstructed outputs $\hat{\mathbf{y}}$.

Tables 5.2a, 5.2b and 5.2c contain the results obtained by the various methods for system order $n = 6$, $n = 12$ and $n = 18$, respectively. Specifically, mean and standard deviation values of the reconstruction errors $\operatorname{err}(\mathbf{y}, \hat{\mathbf{y}})$ and $\operatorname{err}(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})$ and the COR values computed over the 10 trials of each experiment are reported. The mean values of the estimated system order (rank of $\hat{\mathbf{Y}} = \mathcal{H}(\hat{\mathbf{y}})$), number of iterations and execution time are also reported. Firstly, we observe

Table 5.2: Recovery results obtained by the proposed method and the compared methods corresponding to system order a) $n = 6$, b) $n = 12$ and c) $n = 18$. Results are reported in terms of mean values over 10 repetitions of the experiment, while standard deviation values are reported inside parentheses.

(a) System order $n = 6$

Method	err($\mathbf{y}, \hat{\mathbf{y}}$)	err($\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}$)	**COR**	**Order**	**Iter**	**Time**
HRM	0.630 (0.161)	0.259 (0.119)	0.773 (0.145)	8 (2.3)	49 (32)	0.008 (0.005)
SVD-free	0.894 (0.181)	0.365 (0.167)	0.809 (0.169)	1 (0.4)	905 (301)	0.448 (0.162)
SRPCA	0.922 (0.142)	0.372 (0.137)	0.677 (0.492)	7 (2.1)	101 (16)	0.030 (0.004)
IHTLS	0.629 (0.301)	0.267 (0.177)	0.810 (0.203)	2 (0.5)	41 (42)	0.011 (0.011)
SLRA	0.612 (0.292)	1.094 (0.085)	0.816 (0.190)	1 (0.5)	33 (23)	0.002 (0.002)
ours ($p = 1,\ q = 1$)	0.395 (0.218)	0.173 (0.137)	0.900 (0.093)	6 (2.2)	90 (10)	0.016 (0.002)
ours ($p = 0.9,\ q = 0.9$)	0.313 (0.232)	0.141 (0.136)	0.926 (0.079)	5 (3.2)	130 (17)	0.026 (0.003)
ours ($p = 0.5,\ q = 0.5$)	0.299 (0.220)	0.129 (0.141)	0.933 (0.066)	6 (2.7)	215 (90)	0.047 (0.014)
ours ($p = 0.1,\ q = 0.1$)	**0.233 (0.218)**	**0.107 (0.132)**	**0.952 (0.061)**	5 (1.8)	217 (19)	0.043 (0.004)

(b) System order $n = 12$

Method	err($\mathbf{y}, \hat{\mathbf{y}}$)	err($\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}$)	**COR**	**Order**	**Iter**	**Time**
HRM	0.692 (0.234)	0.205 (0.097)	0.637 (0.352)	10 (7.5)	57 (31)	0.022 (0.012)
SVD-free	0.942 (0.104)	0.273 (0.077)	0.634 (0.343)	2 (0.7)	703 (478)	0.544 (0.378)
SRPCA	0.655 (0.211)	0.181 (0.051)	0.848 (0.167)	6 (2.6)	102 (7)	0.064 (0.004)
IHTLS	0.719 (0.299)	0.217 (0.120)	0.616 (0.35)9	1 (0.5)	50 (43)	0.042 (0.030)
SLRA	0.832 (0.355)	1.071 (0.060)	0.416 (0.500)	1 (0.4)	58 (40)	0.006 (0.005)
ours ($p = 1,\ q = 1$)	0.414 (0.333)	0.120 (0.096)	0.813 (0.278)	6 (3.1)	107 (4)	0.042 (0.002)
ours ($p = 0.9,\ q = 0.9$)	0.365 (0.338)	0.103 (0.097)	0.856 (0.213)	6 (1.8)	148 (8)	0.063 (0.004)
ours ($p = 0.5,\ q = 0.5$)	**0.333 (0.363)**	**0.094 (0.105)**	**0.863 (0.199)**	5 (2.2)	210 (24)	0.089 (0.011)
ours ($p = 0.1,\ q = 0.1$)	0.341 (0.298)	0.111 (0.094)	0.859 (0.250)	13 (3.0)	181 (91)	0.088 (0.047)

(c) System order $n = 18$

Method	err($\mathbf{y}, \hat{\mathbf{y}}$)	err($\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}$)	**COR**	**Order**	**Iter**	**Time**
HRM	0.780 (0.238)	0.216 (0.108)	0.483 (0.364)	8 (8.9)	87 (39)	0.063 (0.031)
SVD-free	0.889 (0.203)	0.242 (0.107)	0.567 (0.301)	1 (0.5)	619 (493)	0.789 (0.648)
SRPCA	0.626 (0.238)	0.160 (0.065)	0.752 (0.247)	8 (3.7)	107 (10)	0.127 (0.023)
IHTLS	0.945 (0.309)	0.247 (0.093)	0.479 (0.390)	2 (1.6)	41 (36)	0.082 (0.056)
SLRA	0.958 (0.263)	1.082 (0.057)	0.471 (0.354)	2 (3.1)	65 (39)	0.012 (0.009)
ours ($p = 1,\ q = 1$)	0.572 (0.312)	0.151 (0.088)	0.723 (0.269)	6 (4.7)	108 (10)	0.076 (0.009)
ours ($p = 0.9,\ q = 0.9$)	0.552 (0.322)	0.144 (0.087)	0.736 (0.273)	6 (3.0)	154 (8)	0.133 (0.028)
ours ($p = 0.5,\ q = 0.5$)	0.534 (0.327)	0.141 (0.088)	0.739 (0.239)	6 (3.0)	154 (8)	0.133 (0.028)
ours ($p = 0.1,\ q = 0.1$)	**0.524 (0.346)**	**0.135 (0.091)**	**0.744 (0.241)**	6 (4.1)	223 (9)	0.171 (0.021)

that the non-convex instances of the proposed method, i.e., when $p, q < 1$, consistently account for the most accurate reconstruction of both the clean signal, in terms of both reconstruction error and correlation, as well as the recovery of the sparse noise.

In most cases, the performance is improved when smaller values for p and q are chosen for the proposed model. Secondly, all the compared methods (HRM, SVD-free, SRPCA, IHTLS and SLRA) achieve much lower performance in terms of all the three metrics employed. Furthermore, it is worth noting that, in the scenarios corresponding to orders $n = 12$ and $n = 18$, SRPCA recovers the noise more accurately than the HRM, SVD-free, IHTLS and SLRA. This is expected since the former is the only method amongst the compared ones that is robust to sparse, non-Gaussian noise. It is also worth mentioning that the system order pertaining to the recovered observations varies significantly amongst different methods. Amongst the different instances of the proposed method, this variation is much smaller, with the only exception being the result obtained by our method with $(p, q) = (0.1, 0.1)$ for the case $n = 12$. Regarding the number of iterations, which varies largely across methods, we observe that the non-convex instances of the proposed method require a larger amount of iterations to converge, as compared to the convex instance ($p = q = 1$). However, even in the scenario of order $n = 18$, the best-performing instance of the proposed method ($p = q = 0.1$) needs 223 iterations in average to converge. Finally, the execution times corresponding to the best-performing, non-convex instances of the proposed method in all three experiments are comparable to those accounted for by even convex compared methods, such as SRPCA.

In Fig. 5.2, characteristic signal reconstruction results, as produced by all the different variants of the proposed method as well as the compared methods, are illustrated for a trajectory of synthetic outputs corresponding to a system of order $n = 6$. The noisy observations $\tilde{\mathbf{y}}$, which are the given data to the structured minimization methods, are depicted in each graph along with the original 'clean' observations $\mathbf{y}$ and their reconstruction $\hat{\mathbf{y}}$ obtained by the corresponding method. The error $\mathrm{err}(\mathbf{y}, \hat{\mathbf{y}})$ in recovering the noiseless observations is reported for each method in the corresponding sub-caption. In Fig. 5.3, the corresponding noise reconstruction results obtained by each method are shown for the same example as that of Fig. 5.2.

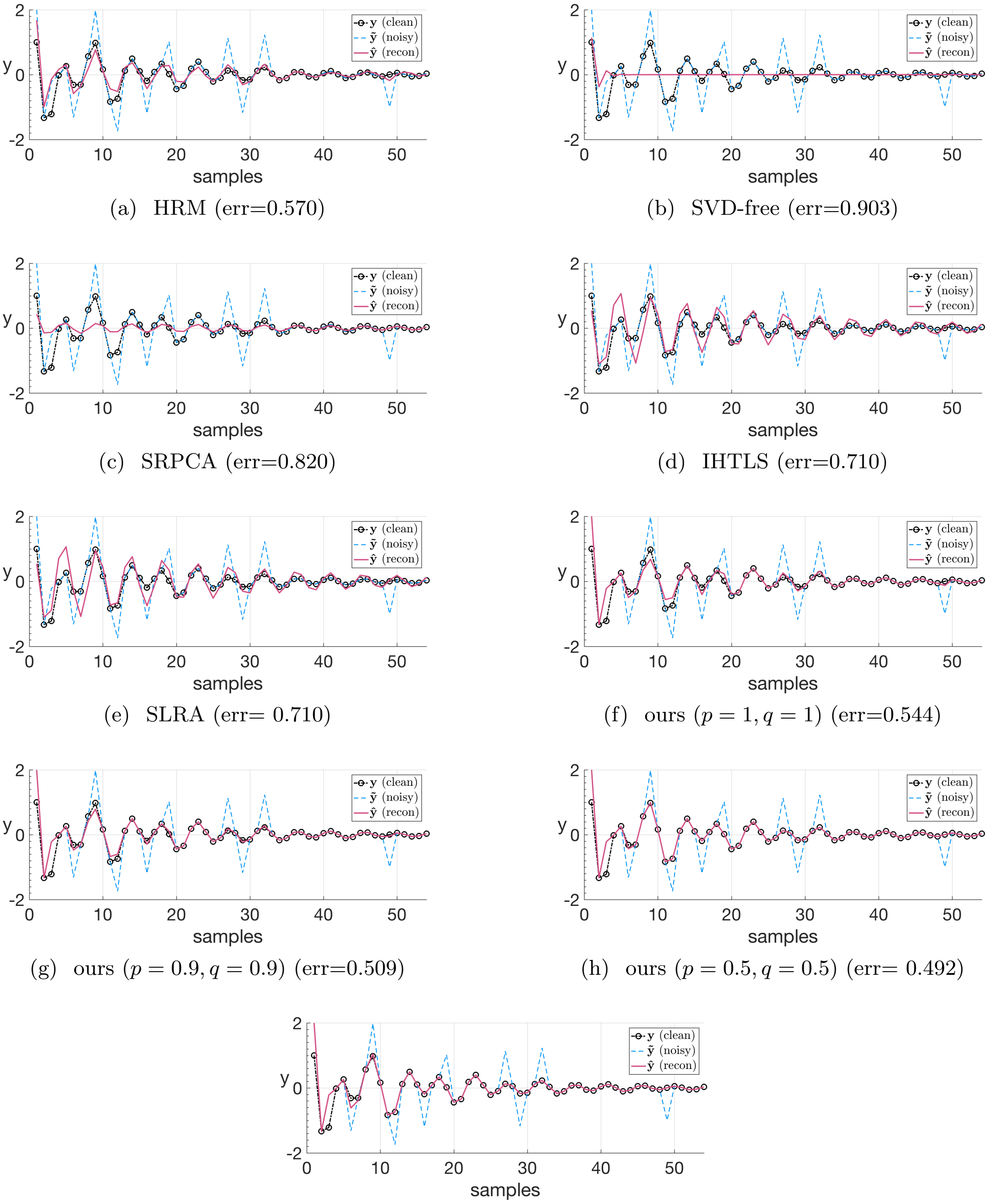
By inspecting Fig. 5.2, we notice that the lowest reconstruction errors amongst all methods are obtained by the non-convex instances of the proposed method, which perform similarly to one another and better than the convex instance ($p = q = 1$). The former manage to perfectly fit the noiseless observations, with the only significant deviations taking place for the data points corresponding to $t \in \{1, 3, 6\}$. Instead, the reconstructed signals produced by the compared methods show considerable divergence from the original noiseless observations for most of the samples. Poor performance is exhibited even for SRPCA, despite its design offering robustness to sparse non-Gaussian noise.

The susceptibility of the compared methods to sparse, gross corruptions becomes more evident by observing the corresponding noise reconstruction results in Fig. 5.3. It is immediately apparent that all compared methods fail to recover the noise, as the recovered noise is far from sparse and largely deviates from the original noise signal. Even the noise signal recovered by SRPCA, which is the only out of the compared methods that is robust to gross noise, is only relatively sparse for the second half of the observations with the recovery of the noise in the first half being rather poor. Instead, the non-convex instances of the proposed method, especially those corresponding to $p = q = 0.5$ and $p = q = 0.1$, succeed in recovering a noise signal that is mostly sparse and also fits accurately the magnitude of the original corruptions. As a matter of fact, by observing Fig. 5.3h and Fig. 5.3i, one can see that all noise entries except for those lying at the entries $t \in \{1, 3, 6\}$ are perfectly recovered, with the rest of the recovered noise signal elements correctly estimated to be zero. On the other hand, the reconstructed noise signal obtained by the convex variant of the proposed method ($p = q = 1$), despite being mostly sparse, differs significantly from the original noise signal at the entries where the noise occurs.

Empirical convergence analysis. In this experiment, the convergence of the proposed method is assessed by employing various types of initialization. To this end, we employ synthetic data corrupted with sparse, non-Gaussian noise, generated similarly to the previous experiment. We clarify here that the only variable that needs to be initialized in Algorithm 1, except for the Lagrange multipliers, is the matrix $\mathbf{L}$. All other variables are calculated in the 1st iteration of the ADMM loop according to the respective updates.

The proposed solver is executed using the following three types of initialization, namely, 'original signal': $\mathbf{L}[\mathbf{0}] = 1.1\tilde{\mathbf{y}}$, 'zeros': $\mathbf{L}[\mathbf{0}] = \mathbf{0}$, 'gaussian': $\mathbf{L}[\mathbf{0}][t] \sim \mathcal{N}(0, 1)$, $t = 1, 2, \ldots, T$, where $\tilde{\mathbf{y}}$ denote the noisy system outputs constructed as in the previous experiments and $\mathcal{N}(0, 1)$ denotes the normal distribution. For each type of initialization, the values of the *primal objective* ($\|\mathbf{N}\|_{S_p}^{p} + \lambda \|\mathbf{E}\|_{q}^{q}$) and the *primal residual* ($\|\mathbf{M} - \mathbf{L} - \mathbf{E}\|_F$) are plotted as a function of the iteration index in Fig. 5.4. Here $\mathbf{M} = \tilde{\mathbf{y}}$ denotes the given noisy data and $\mathbf{L} = \hat{\mathbf{y}}$ the reconstruction. These plots enable us to demonstrate the convergence of the proposed solver. Note that for the last initialization scenario, the experiment is repeated 10 times. and the average convergence curve is plotted.

By inspecting both graphs, it is evident that all three initializations lead to similar convergence behavior in the sense that both the primal objective and the primal residual are non-increasing after the first few iterations. However, by initializing the algorithm using the scaled version of the original signal ($\mathbf{L}[\mathbf{0}] = 1.1\tilde{\mathbf{y}}$) the primal objective attains smaller values than the other two types of initialization. This justifies our choice of initialization as $\mathbf{L}[\mathbf{0}] = 1.1\tilde{\mathbf{y}}$ in the proposed



(a) HRM (err=0.570)

(b) SVD-free (err=0.903)

(c) SRPCA (err=0.820)

(d) IHTLS (err=0.710)

(e) SLRA (err= 0.710)

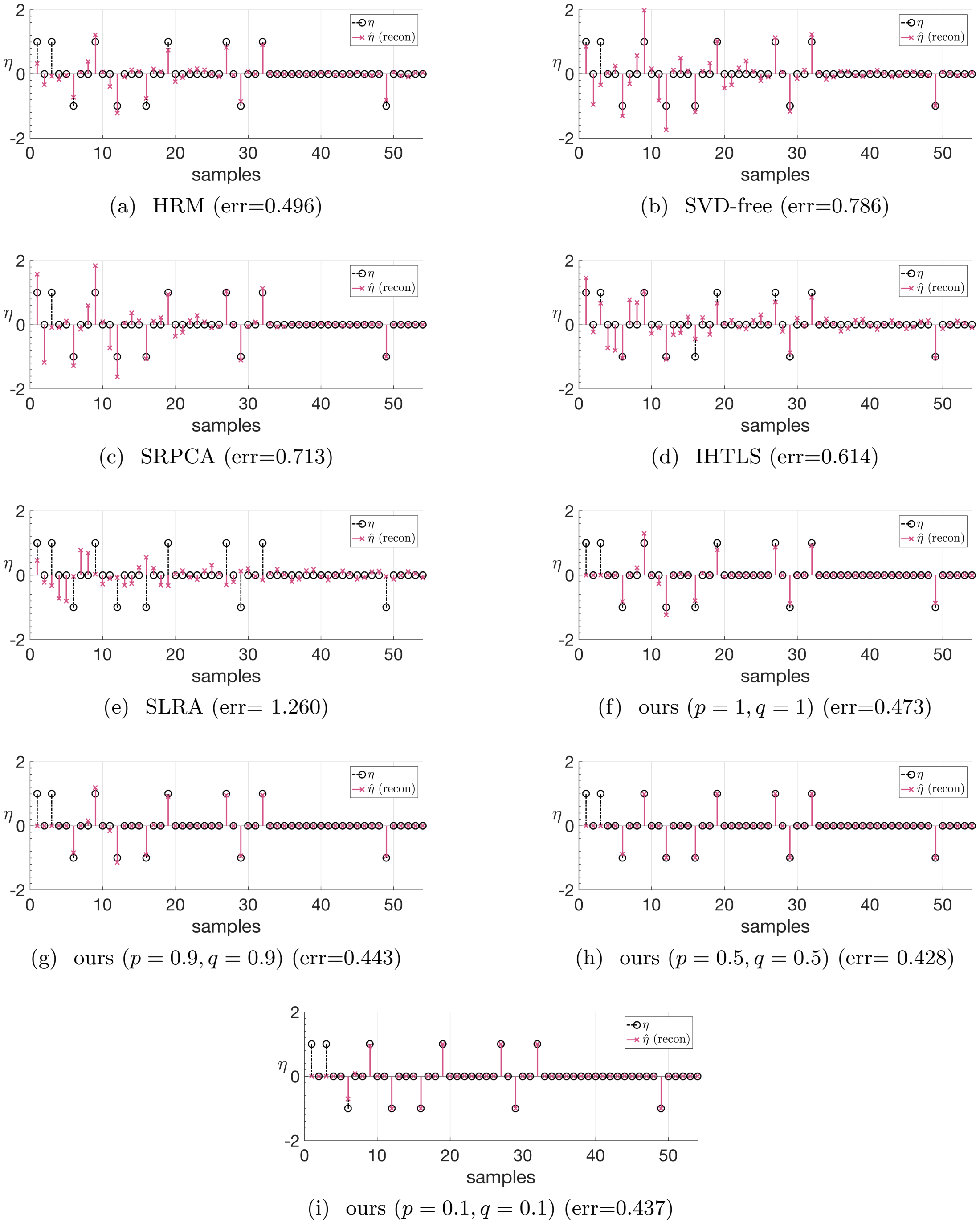
(f) ours ($p = 1, q = 1$) (err=0.544)

(g) ours ($p = 0.9, q = 0.9$) (err=0.509)

(h) ours ($p = 0.5, q = 0.5$) (err= 0.492)

(i) ours ($p = 0.1, q = 0.1$) (err=0.502)

Figure 5.2: (Better viewed in color). Signal reconstruction results as obtained by the different variants of the proposed method as well as the compared methods for an instance of the experiment on synthetic data corresponding to system order $n = 6$. The trajectories of noiseless and noisy observations are plotted along with the reconstruction of the former by each method on each respective graph.



(a) HRM (err=0.496)

(b) SVD-free (err=0.786)

(c) SRPCA (err=0.713)

(d) IHTLS (err=0.614)

(e) SLRA (err= 1.260)

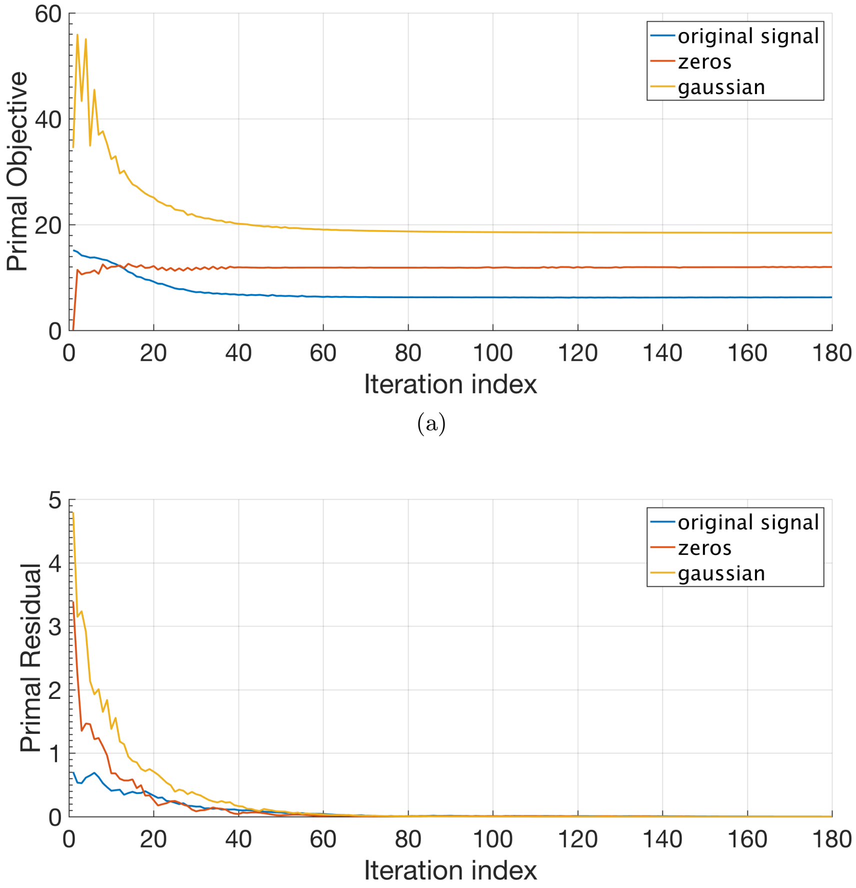
(f) ours $(p = 1, q = 1)$ (err=0.473)

(g) ours $(p = 0.9, q = 0.9)$ (err=0.443)

(h) ours $(p = 0.5, q = 0.5)$ (err= 0.428)

(i) ours $(p = 0.1, q = 0.1)$ (err=0.437)

Figure 5.3: (Better viewed in color). Noise reconstruction results corresponding to the same experiment as that of Fig. 5.2, as obtained by the different variants of the proposed method as well as the compared methods. The trajectories of original noise signal and its reconstruction are plotted on each respective graph.



(a)

(b)

Figure 5.4: (Better viewed in color). Empirical convergence analysis results for 3 different initializations of the proposed solver (Algorithm 1 with $(p, q) = (0.5, 0.5)$) illustrated for the reconstruction of synthetic data corresponding to system order $n = 6$. The graphs illustrated are plots of the value of (a) the *Primal Objective* $\|\mathbf{N}\|_{S_p}^p + \lambda \|\mathbf{E}\|_q^q$, and (b) the *Primal Residual* $\|\mathbf{M} - \mathbf{L} - \mathbf{E}\|_F$ of the proposed method (5.12), with the iteration index. Note that $\mathbf{M} = \tilde{\mathbf{y}}$ denotes the given noisy data and $\mathbf{L} = \hat{\mathbf{y}}$ the reconstruction in this experiment. The different initializations of the matrix $\mathbf{L}$ in Algorithm 1 correspond to the following scenarios: {'multiple': $\mathbf{L}[\mathbf{0}] = 1.1\tilde{\mathbf{y}}$, 'zeros': $\mathbf{L}[\mathbf{0}] = \mathbf{0}$, 'gaussian': $\mathbf{L}[\mathbf{0}][t] \sim \mathcal{N}(0, 1)$, $t = 1, 2, \ldots, T$ (mean value over 10 repetitions)}, where T denotes the number of observations.

algorithms.

5.6.2 Conflict Intensity Prediction

In this section, we address the problem of continuous *conflict* intensity prediction based on the visual modality only. To the best of our knowledge, the presented experiments constitute the first work that i) addresses *continuous* conflict intensity prediction through a dynamic modeling framework (as opposed to frame-by-frame classification or regression), and ii) uses *visual features only*.

MEGA
ΑΝΑΤΡΟΠΗ
ΓΙΑΤΙ ΕΓΙΝΑΝ ΟΙ ΔΙΑΓΡΑΦΕΣ ;
MEGA
Θ. ΠΑΓΚΑΛΟΣ: ΟΤΑΝ ΜΙΛΑΩ ΓΙΑ ΣΥΝΕΝΟΧΗ
ΕΝΝΟΩ ΚΑΙ ΜΕΤΑΞΥ ΤΟΥ ΠΟΛΙΤΙΚΟΥ ΠΡΟΣΩΠΙΚΟΥ
ΑΝΑΤΡΟΠΗ
MEGA
ΑΝΑΤΡΟΠΗ

Figure 5.5: Three sample snapshots from the CONFER dataset, corresponding to dyadic conversations of two guests in conflict.

Data. Video excerpts from live political debates from the CONFER Database, presented in Chapter 4, are utilized. Only episodes from the Set *two* of the database involving exactly two interlocutors are considered herein. The temporal resolution of the video stream is 25 frames per second. For each sequence, the corresponding CCA-derived annotation normalized to $[0, 1]$, is used as ground truth for conflict intensity. Three sample snapshots from the CONFER Database are depicted in Fig. 5.5.

Features & Experimental Protocol. For visual feature extraction, we use the Gauss-Newton Deformable Part Model in [215] for facial landmark detection, which when combined with a person-specific face detector produces very accurate results [39], to detect 49 fiducial facial points in each frame of an input video for each of the two interactants. The points are subsequently globally registered, using a 2-D non-reflective similarity transformation with respect to 4 reference points (centers of the eyes, center of the nose and top of the nose), to remove the effects of head translation, scale and in-plane rotation. This way, *yaw* and *pitch* pose angles, which are expected to be informative in terms of conflict, are retained in the shape configuration. Finally, Principal Component Analysis (PCA) is used at each frame to reduce dimensionality for the points of each speaker to 7, based on the components collectively accounting for 98% of the total variance.

The dynamic behavior prediction framework described in Section 5.5.1 is applied separately for each sequence used in the experiments of this section. During training, the stacked feature vectors corresponding to the two interlocutors are used as inputs $\mathbf{u_t}$ at each time frame t of the training set ($t \in [0, T_{train} - 1]$), while the ground truth is used as output $\mathbf{y_t}$ of a LTI system. The goal is to predict the output $\hat{\mathbf{y}}_\mathbf{t}$ (conflict intensity) for each frame of the sequence ($t \in [0, T - 1]$), based on the learned system parameters and the respective inputs (features).

For our experiments, 43 non-overlapping segments have been extracted from the 73 available episodes of the Set *two*, based on the following condition: they are at least 400 frames long, so that the predictive capability of the proposed framework can be evaluated on long temporal segments portraying frequent conflict intensity fluctuations and conflict escalation/resolution. The resulting subset of clips has a mean and standard deviation of duration of 804 frames and 561 frames, respectively, and corresponds to 22 subjects. For each of the 43 video sequences, the first $P = 60\%$ of the frames are used for training, while the remaining frames are used for testing. This choice establishes a *subjects-dependent* experimental setting. It is worth mentioning that the experimental setting is challenging given that the proposed framework learns temporal behavioral patterns related to conflict escalation/resolution, which vary largely among different persons and contexts, from a single dyadic interaction with average duration of about 19 seconds. This is in contrast to relying on a large set of training instances containing multiple interactants exhibiting conflicting behavior in various contexts.

For the proposed model (5.12), the following combinations are examined for the p and q values corresponding to the Schatten p- and ℓ_q-norm, respectively: $(p, q) \in \{(1, 2), (1, 1), (0.9, 0.9), (0.5, 0.5), (0.1, 0.1)\}$. The scalable Algorithm 5 is also used for this experiment, with the dimension of the column space of $\mathbf{Q}$ in (5.26) set to $K = 10$. The convergence parameters ϵ_1 and ϵ_2 are set to 10^{-4} and 10^{-7}, respectively. For each sequence, 150 values, logarithmically spaced in the interval $[10^{-3}, 1]$ are examined for the tuning of parameter λ in Algorithms 4 and 5. Similarly, a suitable grid search is conducted to tune the parameters of the compared methods. For details on methods to which we compare, see Table 5.1.

For evaluation, the Pearson Correlation Coefficient (COR) is used, measured between the ground truth $\mathbf{y_t}$ (mean over the 10 annotations) and the predicted output $\hat{\mathbf{y}}_\mathbf{t}$ on the test set frames ($t \in [T_{train}, T - 1]$) of each sequence. Motivated by recent works on predictive analysis of human behavior [136, 101], we choose to also report the Intra-Class Correlation Coefficient (ICC) [202]. Similarly to the experiments presented in Chapter 4, we employ the coefficient ICC(3,1) that corresponds to the case "Each target is rated by each of the same k judges, who are the only judges of interest" [202]. For each sequence and method, the ICC(3,1) (henceforth denoted by ICC) is calculated by considering the 'method' and the 'mean annotator' as the only 'judges' of interest and the conflict intensity values for the test set frames as 'targets' in the definition above. To obtain a 'human' baseline ICC result, i.e., a measure of 'level of consistency amongst 10 humans in measuring conflict intensity', we also compute the ICC amongst the 10 annotations for the test frames of each sequence. The average value of the inter-annotator ICC, denoted by ICC$_h$, over all 43 sequences, was found ICC$_h$=0.740. Finally, note that each method is separately optimized in terms of each metric.

Results & Discussion. Results in terms of mean value of COR and ICC over all 43 sequences are reported in Table 5.3 for all methods examined. For details on methods to which

Table 5.3: Conflict intensity prediction results in terms of COR and ICC, averaged over all 43 sequences used from the CONFER dataset. Averaged values for the resulting system order and execution time (Time: secs per frame × 100) are also shown for each (COR-optimized) structured rank minimization method. For details on methods to which we compare, see Table 5.1.

Method	**Order**	**Time**	**COR**	**ICC**
HRM	12	0.08	0.630	0.748
SVD-free	3	0.02	0.005	0.492
SRPCA	14	1.12	0.491	0.721
IHTLS	6	7.77	0.724	0.775
SLRA	7	1.34	0.637	0.708
ours ($p = 1$, $q = 2$)	4	0.22	0.565	0.762
ours ($p = 1$, $q = 1$)	5	0.26	0.771	0.817
ours ($p = 0.9$, $q = 0.9$)	6	0.35	0.800	0.824
ours ($p = 0.5$, $q = 0.5$)	7	0.59	0.805	0.811
ours ($p = 0.1$, $q = 0.1$)	9	0.70	0.801	0.822
ours^{sc} ($p = 1$, $q = 2$)	4	0.19	0.671	0.772
ours^{sc} ($p = 1$, $q = 1$)	5	0.26	0.789	0.813
ours^{sc} ($p = 0.9$, $q = 0.9$)	6	0.34	0.788	0.827
ours^{sc} ($p = 0.5$, $q = 0.5$)	5	0.68	0.781	0.815
ours^{sc} ($p = 0.1$, $q = 0.1$)	5	0.83	**0.806**	**0.833**

we compare, see Table 5.1. The values of the resulting LTI system order and execution time (Time: secs per frame × 100) for the respective best-performing structured rank minimization solution are also reported, again averaged over all sequences[3]. As can be seen, the proposed methods outperform all methods that are compared to, in terms of both COR and ICC. The second-best-performing method in terms of both metrics is IHTLS, with all remaining methods yielding lower scores. Results obtained by the scalable Algorithm 5 (denoted by ours^{sc}) are on par with those yielded by Algorithm 4. As a matter of fact, the best overall performance in terms of both metrics is achieved by the scalable algorithm with $p = q = 0.1$. Furthermore, the non-convex instances of the proposed methods (5.12) and (5.26) $(p, q < 1)$ yield superior performance, as compared to that obtained by the convex model instances ($p, q = 1$ and $p = 1, q = 2$). These results indicate that the dynamic model learned with the non-convex instances explain better the observed data thus providing a better estimate for the system order than that learned with the convex instances. This may be attributed to the relaxation gap entailed by replacing the rank and ℓ_0-norm with the Schatten p- and ℓ_q-norm, respectively, is tighter than that entailed by convex approximations. Also, it is interesting to observe that the choice $q = 2$, which corresponds to a Frobenius-norm based fitting measure, consistently results in the lowest performance amongst the values examined for the ℓ_q-norm. Presumably, this is due to the susceptibility of the corresponding fitting measures to gross, sparse noise [90].

Regarding run time efficiency, it is worth noting that the execution time accounted for by the best-performing variant of the proposed methods (ours^{sc} with $p = q = 0.1$) is close to a degree of magnitude smaller than that of the best-performing out of the compared methods (IHTLS). As expected, execution time increases as p and q values move closer to zero. Moreover, the high COR and ICC scores achieved by the proposed methods are accompanied by low values for the resulting system orders (e.g., $n \in [4, 6]$ for ours^{sc}). This property is crucial for both the generalizability and execution time efficiency of the overall predictive framework.

Notably, IHTLS, HRM and the proposed methods lead to an average ICC which is higher than the mean inter-annotator ICC$_h$ of 0.740. This means that these methods, which were trained using the 'mean annotator' annotations, have learned the trend of the 'mean annotator' exceptionally well and were able to reproduce the trend accurately. This clearly demonstrates the suitability of these methods for modeling the human behavior analysis task at hand (i.e., conflict intensity prediction).

Effect of the training set size on prediction accuracy. The results reported in Table 5.3 correspond to using the first $P = 60\%$ of each sequence's frames for training (structured rank

[3] The Order and Time values reported correspond to the COR-optimized methods.

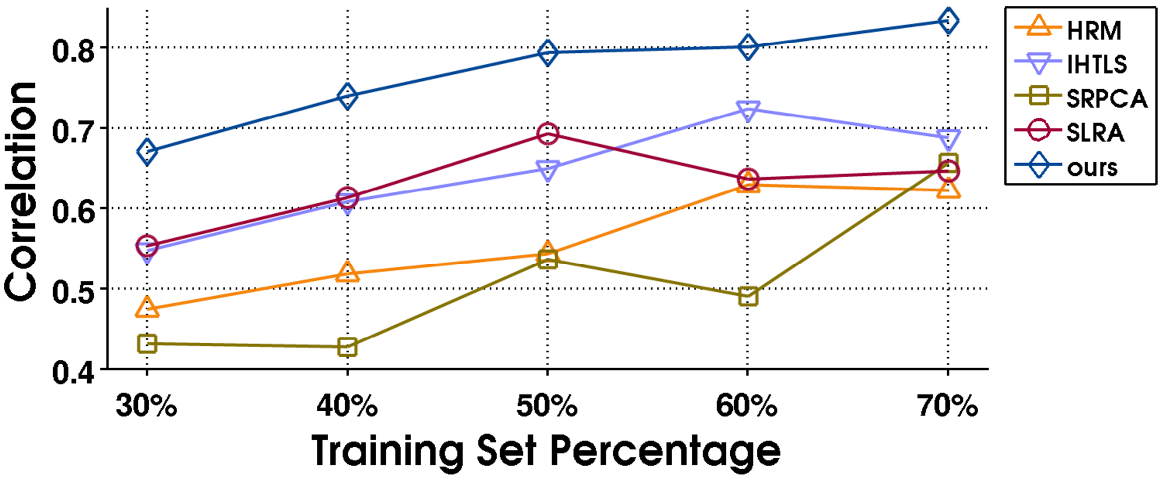
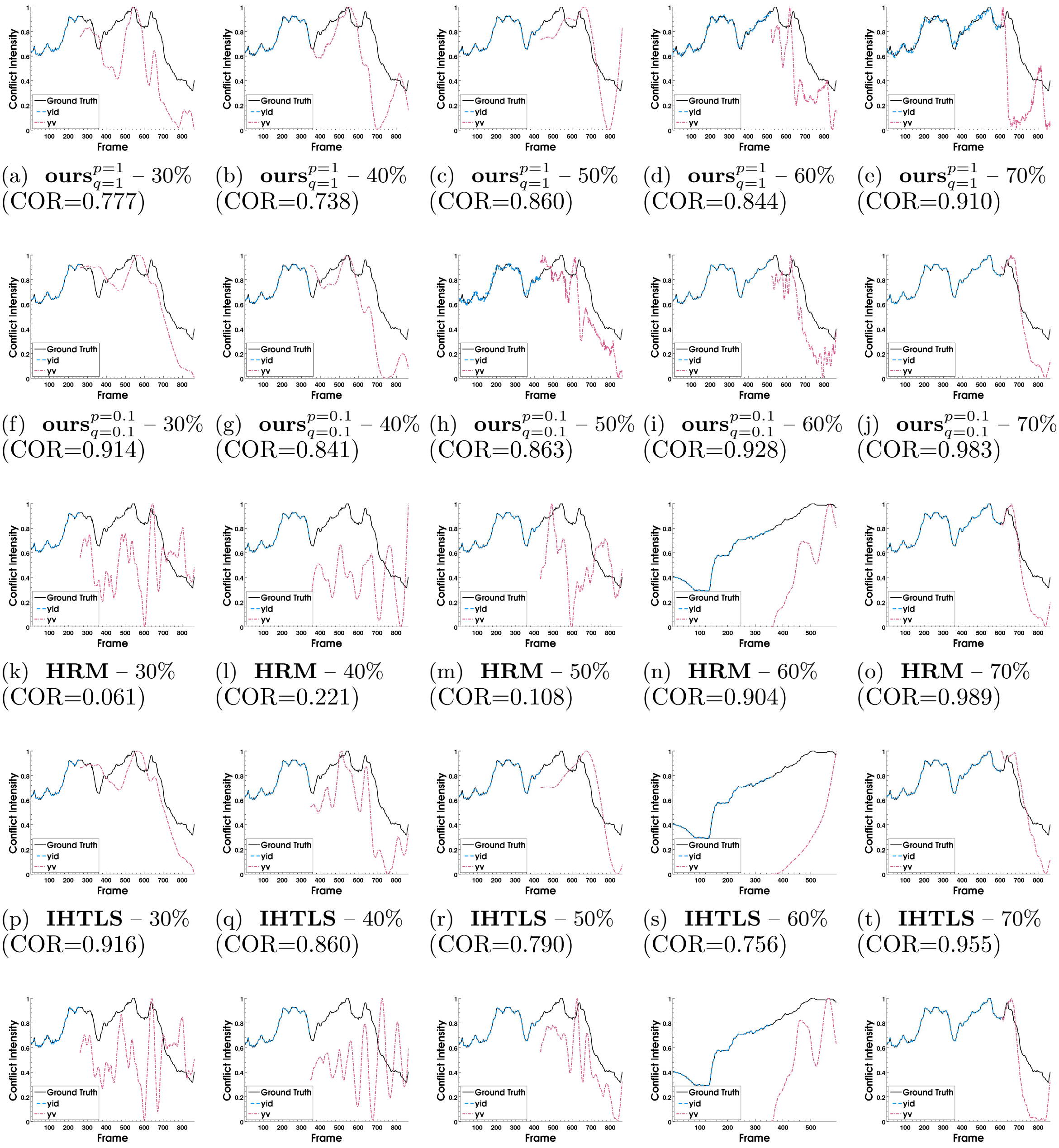


Figure 5.6: Average correlation (COR) values plotted as a function of the training set percentage, for the conflict intensity prediction experiment on the CONFER dataset with varying training size. For details on methods to which we compare, see Table 5.1. Results for the proposed method (5.12) were obtained by using Algorithm 4 with $p = q = 0.1$.

minimization and LTI system learning) and the remaining frames for predicting the respective conflict intensity values. To investigate how the choice of the portion of frames used for training affects the predictive capability of the structured rank minimization-based framework, we vary the training set percentage P in $\{30\%, 40\%, 50\%, 60\%, 70\%\}$ of the sequence length. The test set percentages vary also according to 100-P. The resulting training (test) set sizes, averaged over all 43 sequences, are 240, 322, 402, 483, 563 (559, 482, 401, 321, 241) frames, respectively. For this experiment, the proposed method with $p = q = 0.1$ is examined along with the same five compared methods, while performance is evaluated in terms of the COR metric only. For details on methods to which we compare, see Table 5.1.

A graph that shows the COR values (averaged over all sequences) obtained for each percentage P by the various methods[4] is illustrated in Fig. 5.6. The proposed method consistently outperforms the compared methods in all five scenarios. The second-best-performing method is SLRA and IHTLS for P in $\{30\%, 40\%, 50\%\}$ and P in $\{60\%, 70\%\}$, respectively. The superiority of the proposed method over the compared methods for this experiment is more evident in the cases where 30% or 40% of the frames are used for training; the discrepancy in performance achieved by the proposed method and SLRA reaches 0.117 and 0.126 in absolute COR terms, respectively. Overall, in most of the cases, a higher COR value is achieved by all methods when more data are used for training. For our method, the obtained COR values increase

[4] COR values obtained by the SVD-free method are omitted from this discussion, as they were much lower compared to the other methods.



(a) $\textbf{ours}^{p=1}_{q=1}$ – 30% (COR=0.777)
(b) $\textbf{ours}^{p=1}_{q=1}$ – 40% (COR=0.738)
(c) $\textbf{ours}^{p=1}_{q=1}$ – 50% (COR=0.860)
(d) $\textbf{ours}^{p=1}_{q=1}$ – 60% (COR=0.844)
(e) $\textbf{ours}^{p=1}_{q=1}$ – 70% (COR=0.910)

(f) $\textbf{ours}^{p=0.1}_{q=0.1}$ – 30% (COR=0.914)
(g) $\textbf{ours}^{p=0.1}_{q=0.1}$ – 40% (COR=0.841)
(h) $\textbf{ours}^{p=0.1}_{q=0.1}$ – 50% (COR=0.863)
(i) $\textbf{ours}^{p=0.1}_{q=0.1}$ – 60% (COR=0.928)
(j) $\textbf{ours}^{p=0.1}_{q=0.1}$ – 70% (COR=0.983)

(k) **HRM** – 30% (COR=0.061)
(l) **HRM** – 40% (COR=0.221)
(m) **HRM** – 50% (COR=0.108)
(n) **HRM** – 60% (COR=0.904)
(o) **HRM** – 70% (COR=0.989)

(p) **IHTLS** – 30% (COR=0.916)
(q) **IHTLS** – 40% (COR=0.860)
(r) **IHTLS** – 50% (COR=0.790)
(s) **IHTLS** – 60% (COR=0.756)
(t) **IHTLS** – 70% (COR=0.955)

(u) **SRPCA** – 30% (COR=-0.020)
(v) **SRPCA** – 40% (COR=-0.105)
(w) **SRPCA** – 50% (COR=0.880)
(x) **SRPCA** – 60% (COR=0.839)
(y) **SRPCA** – 70% (COR=0.966)

Figure 5.7: (Better viewed in color). Conflict intensity prediction results for a single sequence of the CONFER dataset, as produced by the proposed method ($(p,q) \in \{(1,1),(0.1,0.1)\}$), HRM, IHTLS and SRPCA for different portions of frames used for training (reported as percentages in the sub-captions along with the respective COR). For details on methods to which we compare, see Table 5.1. In each graph, the curve designated by 'yid' ('yv') corresponds to the training (test) predictions, while the third, solid-line curve corresponds to the ground truth annotations (mean over 10 ratings). The test set predictions have been normalized to the range [0,1] for better visualization.

strictly monotonically with P, reaching COR $= 0.834$ at $P = 70\%$.

In Fig.5.7, conflict intensity predictions, as obtained by the proposed method ($(p, q) \in \{(1, 1), (0.1, 0.1)\}$), HRM, IHTLS and SRPCA for a sequence of the CONFER dataset, are illustrated along with the ground truth annotations as line plots for the various training set percentages examined. The COR values obtained are also shown in the respective sub-captions. As can be seen, the test sequence in question establishes a challenging scenario, since it involves instances of both conflict escalation and resolution, either short- or long-term. One can easily notice that for all scenarios the trends of conflict intensity along the test frames are accurately predicted by the non-convex instance of the proposed method ($p = q = 0.1$), while the convex model instance ($p = q = 1$) yields smaller COR values in all five cases examined. The former achieves a COR value as high as 0.914 (Fig. 5.7f) for a total of 604 test frames when trained on just the first 30% of the sequence (260 frames). In the same scenario, IHTLS performs similarly, while other methods such as HRM and SRPCA yield COR values that lie just above or below zero, respectively. The various compared methods exhibit different patterns in performance as the amount of video frames used for training increases. For instance, IHTLS outperforms the other methods when less training data are used (30% and 40%) while SRPCA and HRM show a dramatic increase in performance at the point where 50% and 60% of the video frames are employed for training, respectively. The effectiveness of IHTLS in the scenarios involving less training data for the sequence in question is as expected. IHTLS is more likely to find a local approximation for the 'low-complexity' temporal dynamics of the first portion of the sequence that be low-rank and hence a simpler, more generalizable system than the convex, nuclear-norm based methods SRPCA and HRM, since the former searches for the desired rank iteratively starting from rank 1 [54]. Finally, as expected, the highest COR values obtained overall correspond to the highest training percentage of 70% and are similar across all methods.

5.6.3 Valence and Arousal Prediction

In this section, the efficiency of the proposed dynamic behavior analysis framework is validated on the problem of *continuous prediction of valence and arousal* based on *visual features only*. As explained in Section 2.2, *Valence* (how positive or negative the affect is) and *Arousal* (how excited or apathetic the affect is) are the latent dimensions that are most widely used to measure emotional experience, since they are considered to encapsulate most of the affect variance [113]. In this chapter, we address continuous prediction of valence and arousal using visual features only. Motivated by evidence suggesting that valence and arousal exhibit high correlation [164], we treat them in a joint framework, that is, as outputs generated by the same LTI system.

Figure 5.8: Example images from the SEMAINE database portraying three subjects from Session 46 (left), 82 (middle), and 94 (right).

Data. The SEMAINE database [138], which contains audio-visual recordings of emotionally colored conversations between a human and an operator, is employed. The operator plays the role of an avatar and, depending on the choice of the latter, acts assuming one of 4 distinct personalities (happy, gloomy, angry or pragmatic). Since the goal of the operator is to elicit emotional reactions by the user, naturalistic dyadic conversations are developed, which are suitable for spontaneous affect analysis. Each video has been recorded at 50 frames per second, and has been annotated frame by frame by six raters in terms of real-valued valence and arousal ranging from -1 to 1. A subset of SEMAINE, containing 40 sequences that are at least 3000 frames ($\sim$ 1min) long from a total of 10 subjects, is used. For each sequence, the mean values of valence and arousal annotations over the six ratings are utilized as ground truth. Three sample video frames corresponding to three different users from the SEMAINE database are depicted in Fig. 5.8.

Features & Experimental Protocol. The Active Appearance Model-based tracker [153], which performs simultaneous tracking of 3D head pose, lips, eyebrows, eyelids and irises in videos, is employed to extract facial features. For each frame, 113 2D characteristic facial landmarks are obtained. To ensure that only expression-related information is retained in the feature representation, we use the tracker's estimates of 3D head pose values to remove pose angles. Scale and translation effects are subsequently removed from the 226 coordinates of the pose-normalized points, according to the procedure described for the experiment in Section 5.6.2. Finally, dimensionality reduction is performed by means of PCA. Again, 98% of the total energy is retained resulting to a 12-dimensional feature vector.

For each of the 40 sequences, the framework described in Section 5.5.1 is employed for continuous valence and arousal prediction. Only the first 3000 frames are considered for each sequence. The experimental protocol is similar to that established for the conflict intensity prediction experiment. The first 2000 frames of each sequence are used for training, while the remaining 1000 frames ($\sim$ 20 secs) are used for V-A prediction. For this experiment, the visual feature vectors are used as inputs and the V-A values are used as outputs. Predictive performance for both valence and arousal is assessed again by means of both COR and ICC. To facilitate the evaluation and discussion with respect to each of the affect dimensions, we choose to optimize each method separately for each dimension and performance metric. For details on methods to which we compare, see Table 5.1. For the proposed method, only Algorithm 4 is examined in this experiment. The mean value over all 40 sequences of the inter-annotator ICCh, calculated amongst the six available ratings, was found to be ICC $_h^{\mathrm{V}}$=0.778 for valence and ICC$_h^{\mathrm{A}}$=0.893 for arousal, respectively. The higher inter-annotator reliability for arousal is expected in the case of the SEMAINE data due to the three interlinked facts: (i) the majority of SEMAINE annotated data relate to high aroused emotions, (ii) the annotators were presented with audio-visual recordings to be annotated, and (iii) the arousal is better recognized when audio modality is available [195, 15].

Results & Discussion. Valence and arousal prediction results, in terms of mean value of COR and ICC over all 40 SEMAINE sequences, are reported in Table 5.4 for all methods examined. For details on methods to which we compare, see Table 5.1. Mean values for the resulting system order and execution time (Time: secs per frame × 100) are also reported[5]. As can be seen, the best performance, in terms of both metrics, is obtained by the proposed method, for both valence and arousal prediction. The second-best-performing method in terms of COR (ICC) is HRM (SLRA) for both affect dimensions. Overall, valence and arousal are predicted with similar accuracies by almost all the methods. Again, the non-convex instances of the proposed method ($p, q < 1$) account for significant performance boost over convex model instances ($p, q = 1$ and $p = 1, q = 2$), yet accompanied by an increase in model complexity and execution time. Still, in most of the cases the proposed method results in systems of lower-complexity, as compared to those accounted for by the remaining methods. Regarding execution time, the various methods achieve comparable performances, with the exception of IHTLS that is much slower for this experiment, probably due to the increased dimensions of the data Hankel matrices.

Finally, it is worth noting that the inter-annotator ICC$_h^{\mathrm{V}}$ for valence is exceeded by HRM, SLRA and our method, whereas no method furnishes an ICC value greater than ICC$_h^{\mathrm{A}}$ for arousal. This result is exactly as expected. Namely, as explained above, in the case of the utilized SEMAINE data, human annotators were presented with audio-visual (rather than visual-only) recordings when they were conducting the annotation. The presence of audio data does not affect the human performance in recognition of valance, but it does affect the recognition of arousal – arousal is better recognized when audio cues are available to humans to

[5] The Order and Time values reported correspond to the COR-optimized methods.

Table 5.4: Valence (Val.) and Arousal (Ar.) prediction results in terms of COR and ICC, averaged over all 40 sequences used from the SEMAINE dataset. Averaged values for the resulting system order (Val. and Ar.), and execution time (Time: secs per frame × 100) (Val.) are also shown for each (COR-optimized) structured rank minimization method. For details on methods to which we compare, see Table 5.1.

Method	**Order**		**Time**	**COR**		**ICC**	
	Val.	**Ar.**	**Val.**	**Val.**	**Ar.**	**Val.**	**Ar.**
HRM	19	17	1.49	0.812	0.794	0.805	0.801
SVD-free	2	3	0.46	-0.024	0.001	0.504	0.412
SRPCA	16	21	5.95	0.771	0.743	0.774	0.765
IHTLS	10	9	121.14	0.727	0.739	0.739	0.734
SLRA	14	15	4.46	0.737	0.728	0.830	0.823
ours ($p = 1$, $q = 2$)	5	6	3.80	0.834	0.818	0.823	0.819
ours ($p = 1$, $q = 1$)	8	7	4.56	0.844	0.838	0.835	**0.835**
ours ($p = 0.9$, $q = 0.9$)	8	8	6.32	0.851	0.842	0.828	0.824
ours ($p = 0.5$, $q = 0.5$)	9	9	9.43	0.857	**0.871**	0.821	0.830
ours ($p = 0.1$, $q = 0.1$)	13	13	12.27	**0.866**	0.869	**0.837**	0.824

rely on [15]. Hence, while automated methods like HRM and our methods are highly suitable for modeling human behavior analysis tasks at hand (i.e., valence intensity prediction), they could not learn the trends of the 'mean annotator' well enough for the case of arousal intensity prediction, because these were relying on audio data unavailable to the tested automated methods.

5.6.4 Tracklet matching

In this section, the efficiency of the proposed method is evaluated on the task of multi-object/person tracking from detection, alternatively called *tracklet matching*. The goal is to identify targets in the visual stream across occlusions from a set of given detections. This application serves to demonstrate that the proposed framework (i) is capable of robustly distinguishing motion dynamics corresponding to different targets in visually cluttered scenarios, aside from modeling appearance and shape dynamics in videos of human faces, and (ii) can operate as an unsupervised learning algorithm based on the assumption of an output-only linear dynamical system generating the sequential observations (frame-by-frame object/person detections in our case).

Data. Experiments are conducted on the recently published Similar MultiObject Tracking (SMOT) dataset [54], which consists of 8 videos[6]showing multiple targets with identical or very similar appearance. For each video, the provided hand-labeled detections for the targets appearing in each frame are employed. Overall, the task is challenging due to the presence of multiple targets, long trajectories, object occlusions and crossings, missing data and camera motion.

Features & Experimental Protocol. We follow the tracklet matching framework proposed in [54], which is based on a Generalized Linear Assignment (GLA) Problem. Thus, given N tracklets (trajectories of system outputs) $\{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(N)}\}$, GLA solves

$$\begin{aligned} &\max_{\mathbf{K}} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} k_{ij} \\ &\text{s.t.} \sum_{i=1}^{N} k_{ij} \leq 1\,; \sum_{j=1}^{N} k_{ij} \leq 1\,; \, k_{ij} \in \{0,1\}\,, \end{aligned} \tag{5.33}$$

where $\mathbf{K}$ is an adjacency matrix, with $k_{ij} = 1$ denoting that $\mathbf{Y}^{(i)}$ is the predecessor of $\mathbf{Y}^{(j)}$, and $\mathbf{P}$ is a similarity matrix given by

$$p_{ij} = \begin{cases} -\infty & \text{if } \mathbf{Y}^{(i)} \text{ and } \mathbf{Y}^{(j)} \text{ conflict} \\ \dfrac{\operatorname{rank}(\mathcal{H}(\mathbf{Y}^{(i)})) + \operatorname{rank}(\mathcal{H}(\mathbf{Y}^{(j)}))}{\min_{\bar{\mathbf{Y}}_i^j} \operatorname{rank}(\mathcal{H}(\mathbf{Y}^{(ij)}))} - 1 & \text{otherwise,} \end{cases} \tag{5.34}$$

with $\mathbf{Y}^{(ij)} = [\mathbf{Y}^{(i)} \; \bar{\mathbf{Y}}_i^j \; \mathbf{Y}^{(j)}]$ being the joint tracklet of detections, padded with zeros at the entries of the tracklet $\bar{\mathbf{Y}}_i^j$ of missing data. Hence, the critical point of the aforementioned algorithm is the solution of the low-rank Hankel *matrix completion* problem $\min_{\bar{\mathbf{Y}}_i^j} \operatorname{rank}(\mathcal{H}(\mathbf{Y}^{(ij)}))$ in (5.34). This is solved according to the framework described in Section 5.5.2, in which the underlying LTI system is assumed to be autonomous and the data Hankel matrices are composed of the respective outputs (2D tracking point coordinates).

Two experimental scenarios are considered, similarly to [54]. In the first experiment, *false positives* are increased by injecting uniformly distributed false detections with percentage varying as $[0\%, 10\%, \ldots, 50\%]$. In the second scenario, *false negatives* are increased by removing, again uniformly, true detections with percentage varying as $[0\%, 6\%, \ldots, 30\%]$. For each scenario,

[6] 1. *slalom* (three skiers), 2. *juggling* (3-ball juggling scene), 3. *acrobats*, 4. *seagulls*, 5. *TUD-Campus* (pedestrians), 6. *TUD-Crossing* (pedestrians), 7. *crowd* (from the crowd UCF dataset), 8. *balls* (bouncing identical ping pong balls).

Table 5.5: Tracklet matching results, in terms of MOTA (Eq. (5.35)), on the SMOT dataset for each experimental scenario. For each noise type, the results are averaged over 6 noise levels, with each of the latter examined 10 times. Average execution time (Time: secs per frame) accounted for by each structured rank minimization method is also shown. For details on methods to which we compare, see Table 5.1.

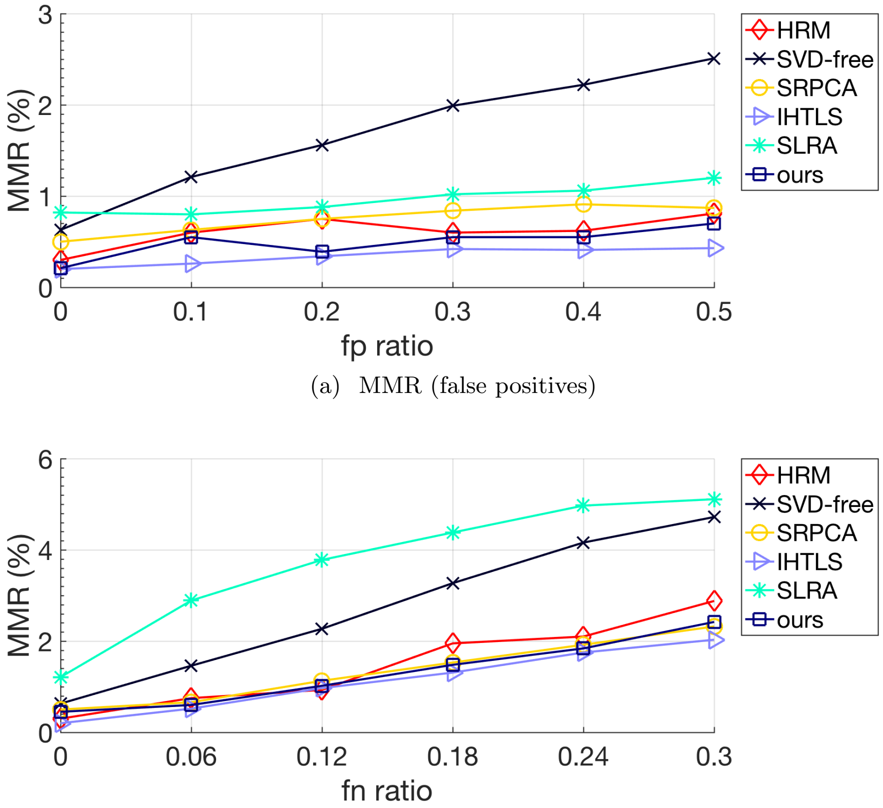
Method	**False Positives**		**False Negatives**	
	Time	**MOTA**	**Time**	**MOTA**
HRM	0.202	0.9749	0.419	0.8687
SVD-free	0.033	0.9602	0.023	0.8422
SRPCA	0.104	0.9734	0.200	**0.8812**
IHTLS	0.174	**0.9799**	0.334	0.8712
SLRA	0.051	0.9646	0.230	0.7731
ours ($p = 1$, $q = 2$)	0.113	0.9733	0.249	0.8591
ours ($p = 0.5$, $q = 2$)	0.169	0.9745	0.277	0.8826
ours ($p = 0.1$, $q = 2$)	0.211	**0.9779**	0.311	**0.8880**

the experiment is repeated 10 times for each noise level, and the average performance over the 60 runs is reported. The same five methods used for comparison in the previous experiments are examined. For details on methods to which we compare, see Table 5.1. For the proposed method, Algorithm 4 is used, with the weight matrix $\mathbf{W}$ in (5.12) formed by setting its entries corresponding to the 'missing' tracklet $\dot{\mathbf{Y}}_i^j$ to zeros and all remaining entries to ones. Various values are examined for the parameters, that is, $(p, q) \in \{(1, 2), (0.5, 2), (0.1, 2)\}$ and $\lambda \in \{10^{-6}, 5 \cdot 10^{-6}, 10^{-5}, \ldots, 10^3\}$, for each video and noise level. The convergence parameters ϵ_1 and ϵ_2 in Algorithm 4 are set to 10^{-7}. For all methods examined, a Frobenius-norm based fitting measure is adopted ($q = 2$ for the proposed method). This experimental choice was motivated by preliminary experiments, in which it was observed that the use of sparsity promoting norms for approximation error resulted in trivial solutions when a large amount of missing data was involved.

For evaluation, the MOTA measure [17] is used, which is given by

$$\text{MOTA} = 1 - \frac{\sum_t (fn_t + fp_t + mm_t)}{\sum_t g_t}, \tag{5.35}$$

where fn_t, fp_t, mm_t and g_t denote the false positives, false negatives, mismatches and ground truth detections for frame t, respectively.

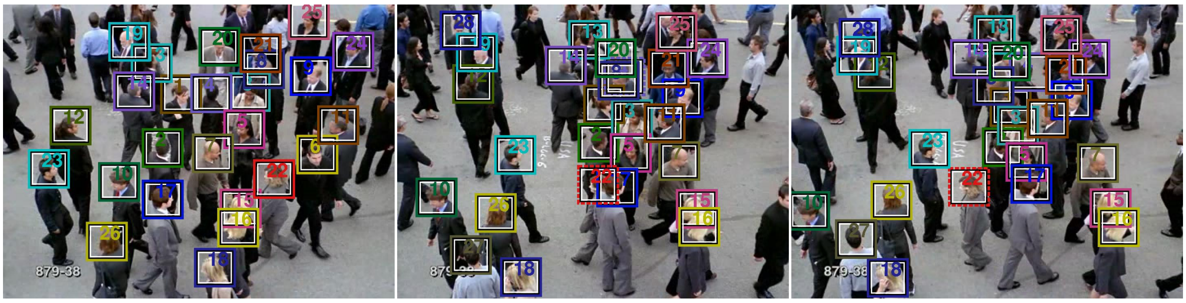


(a) MMR (false positives)

(b) MMR (false negatives)

Figure 5.9: Tracklet matching results, as obtained by the proposed method ($p = 0.1, q = 2$) and the various compared methods, in terms of *MissMatch Ratio* MMR $= \frac{\sum_t (mm_t)}{\sum_t g_t}$ plotted as a function of noise level for the (a) false positives and (b) false negatives scenario, respectively.

Results & Discussion. Tracklet matching results in terms of the MOTA measure – averaged over all 8 videos, noise levels and experiment runs – are reported for each scenario in Table 5.5. For details on methods to which we compare, see Table 5.1. Run time performance (Time: secs per frame) of each respective algorithm, averaged similarly, is also reported. Overall, performance varies less amongst different methods for the false positives case, as compared to the false negatives case. This can be partially ascribed to the former case corresponding to a less demanding task of tracklet matching, since it involves a smaller amount of missing data. The proposed method performs similarly to IHTLS in terms of MOTA for both experimental scenarios, with the difference in performance for all 8 videos calculated as not statistically significant according to a paired $t$test. All remaining methods achieve lower scores. The computational efficiency of the proposed method ($p = 0.1, q = 2$) is comparable to that accounted for by the best-performing amongst the compared methods, for both scenarios. Similarly to the previous experiments, the convex instance of our method ($p = 1, q = 2$)



(a) frame 87 (b) frame 127 (c) frame 140

Figure 5.10: (Better viewed in color). Tracklet matching results, as produced by the proposed method (Algorithm (4) with $p = 0.1$, $q = 2$), illustrated on three frames of the *crowd* sequence from the SMOT dataset. The estimated trajectory index corresponding to each detection is shown inside a bounding box. Solid line boxes indicate given detections, while dashed line boxes indicate detections estimated by our method.

corresponds to a smaller execution time than that of the non-convex instances, albeit to a poorer performance.

Results in terms of *MissMatch Ratio* MMR $= \frac{\sum_t (mm_t)}{\sum_t g_t}$ plotted as a function of noise level, as obtained by the proposed method ($p = 0.1, q = 2$) and the various compared methods, are shown separately for the false positives and false negatives scenario in Fig. 5.9. By comparatively inspecting the two graphs, it is evident that more mismatches consistently occur in the false negatives scenario for all methods, which is exactly as expected. Also, MMR values vary slightly across noise levels in the false positives scenario for most methods, while in the most demanding false negatives scenario mismatches increase at a higher rate with the noise level. The best-performing methods for both cases are IHTLS and the proposed method, with the difference in MMR values being statistically insignificant according to a paired $t$test for all noise levels in both cases. On the other hand, the poorest performance for both cases is accounted for by the SVD-free and SLRA methods.

Tracklet matching results accounted for by the proposed method ($p = 0.1$, $q = 2$), shown as bounding boxes containing the estimated trajectory indices for the corresponding detections, are depicted on three characteristic frames of the *crowd* sequence from the SMOT dataset. The bounding boxes drawn with dashed lines correspond to detections estimated by the proposed method. One can observe that tracklets have been merged accurately in this challenging scenario that involves a heavily occluded surveillance scene. It is also worth noting that trajectory 22 (shown in red box) has been accurately 'completed' for frames 127 and 140 (Fig. 5.10b and 5.10c, resp.), despite the intense occlusion occurring at frame 127.

5.7 Conclusions

A framework for dynamic behavior analysis in real-world conditions was presented in this chapter. Specifically, the presented framework essentially employs a novel structured rank minimization method to learn a low-complexity system from time-varying data, in the presence of gross sparse noise and possibly missing data. By resorting to the ADMM, an efficient algorithm for the proposed structured rank minimization model along with a scalable version has been developed. Regarding applications, focus was placed on vision-based conflict intensity prediction, valence and arousal prediction, and tracklet matching. Extensive experiments on real-world data drawn from these application domains demonstrate the robustness and the effectiveness of the proposed framework.

Overall, the predictive framework proposed herein is the first machine learning approach to dynamic analysis of dimensional affect and behavior in which annotations and features act as outputs and inputs, respectively, of a low-order linear dynamical system that explicitly models the latent temporal structure. A robust sequential learning framework was proposed that explicitly recovers the temporal dynamics from (possibly) grossly corrupted and missing observations. The optimization problem that is the core of the presented framework (i) can take both convex and non-convex instances, thus allowing flexibility in the trade-off between accuracy and computational efficiency, and (ii) can robustly estimate the memory of an underlying low-order auto-regressive process for the data. The latter is in turn used to learn an explicit systemic representation of the displayed dynamics, which allows for representation, categorization and comparison of affective and behavioral displays on the basis of their dynamics. Most importantly, we demonstrate for the first time that naturalistic human behavior and affect, manifested by a single person or group of interactants, can be learned and predicted based on a small amount of person(s)-specific observations, amounting to a duration of just a few seconds.

The proposed learning framework has been primarily designed to address smoothly-varying dynamic phenomena. However, one could use our modeling paradigm to design extensions that can achieve more comprehensive and localized both in time and frequency spectral learning of the displayed dynamics by means of e.g., multi-linear or tensor decomposition. In this way, one could identify multiple latent auto-regressive components manifested either in parallel or consecutively in the sequential observations, thus tackling behaviors that involve distinct, consecutive micro-behaviors or that necessitate more than one co-occurring frequency components for accurate modeling of their dynamics. Also, one could employ dynamical system learning in a sliding window approach on sequential data to perform anomaly detection. For instance, meaningful events could be sought in moments when the corresponding system order rises or falls abruptly with respect to the neighboring windows. Another interesting direction to explore would be to use more sophisticated features for the feature representation stage compared to the PCA coefficients of facial tracking points employed herein. For instance, one could employ variational autoencoders [109] or very deep CNN features (e.g., [204]) for the visual representation learning on each time step and apply our dynamical learning framework on these. Finally, one could employ the linear dynamical system representations learned for different behaviors to perform behavior similarity estimation. In other words, one could first identify a handful of sequences-templates to serve as typical examples of the behaviors of interest in terms of their dynamics and classify never-before-seen sequences based on the similarity its systemic representation bears to that of each one of the training behavioral templates.

Chapter 6

Conclusion and Future Work

Contents

In this chapter, we provide a summary of the work presented in this thesis highlighting the most significant research outcomes stemming from our study. We also identify areas of our research that can be improved upon, thus offering valuable insights for future directions based on our findings.

6.1 Thesis Summary

We have presented here our work on robust machine learning methods for human face, affect and behavior analysis. Our contributions have been with respect to both static and dynamic modeling of facial, affective and behavioral attributes in data captured under real-world, unconstrained conditions. For the latter problem, we have also contributed in terms of providing a new dataset suitable for analyzing subtle, spontaneous expressions of a social attitude in continuous time.

In Chapter 2 we reviewed existing machine learning approaches to static face analysis and dynamic behavior and affect analysis as well as databases of dyadic or multi-party social interactions. A method for recovering mutually incoherent and structured components in still face images, relying on discriminant information as well as structure-inducing norms on the facial aspects, was presented in Chapter 3. A dictionary-based framework that uses the extracted components corresponding to facial attributes such as facial identity, facial expression and AU activation, to jointly address interrelated multi-label classification tasks for static face analysis, was also presented. By conducting experiments on four datasets, we discovered that the proposed learning algorithm is i) robust in recovering low-dimensional components associated with facial attributes in images corrupted by gross noise (e.g., non-uniform illumination, contiguous occlusion) and ii) efficient in recognizing the attributes in a variety of settings, namely joint face and expression recognition, face recognition under varying percentages of training data corruption, subject-independent expression recognition, and action unit detection. Overall, the proposed *Discriminant Incoherent Component Analysis (DICA)* constitutes a robust framework that can generalize to classification of any number or type of labeled affective or behavioral attributes that manifest themselves in the visual stream through specific structures associated with mutually incoherent modes of variation.

Having identified a gap in the literature which is the lack of databases annotated in terms of dimensional descriptions of social behavior in continuous time, we released the *Conflict Escalation Resolution (CONFER) Database* to facilitate social, cognitive and computer science studies on a preeminent social attitude, namely interpersonal *conflict* arising in naturalistic dyadic or multi-party conversations. The CONFER Database, which we presented in Chapter 4, is the first audio-visual database to have been annotated on a frame-by-frame basis in terms of dimensional rather than categorical characterizations of a social attitude. To establish a research platform for continuous-time and dimensional social behavior recognition, we went one step further by using the CONFER Database to conduct the first systematic experimental study on continuous-time conflict intensity estimation. In our experiments, the effectiveness of various audio and visual features and fusion of them as well as classifiers was evaluated for the problem at hand. Our findings validated previous evidence suggesting the importance of the temporal aspect in recognizing spontaneous human affect and behavior and brought into view the limitations of existing machine learning classifiers in capturing temporal dependencies when assigned the task of modeling affect and social behavior at a finer granularity.

Motivated by the desire to describe the inherent dynamic structure of human affect and behavior manifested in real-world scenarios, we steered the bulk of our research efforts into devising a model that can explicitly model the temporal dynamics of affective and behavioral displays. We approached this problem on the basis of the natural assumption that continuous-time annotations characterizing the temporal evolution of smoothly-varying affect or behavior phenomena can be viewed as outputs of a low-complexity *linear dynamical system* when behavioral cues (features) act as system inputs. The dynamic behavior analysis framework presented in Chapter 5 robustly learns the system describing this latent auto-regressive process.

This is achieved by a novel structured rank minimization method for linearly (Hankel)-structured data matrices which, unlike existing methods, can accurately estimate the most crucial hidden variable, that is, the memory of this system, in the presence of grossly corrupted features and annotations and (possibly) partially missing data. Having learned this dynamical system from the training observations, unknown future values of dimensional affect or behavior (system outputs) manifested in a video sequence can be predicted by applying the system equations for the respective features (system inputs).

Aiming to evaluate the generalizability and effectiveness of the predictive framework proposed in Chapter 5 in challenging scenarios, we conducted extensive experiments on three distinct dynamic behavior analysis tasks, namely (i) *conflict intensity prediction*, (ii) *prediction of valence and arousal*, and (iii) multi-(object/person) tracking from detection. In the first two tasks, our method was evaluated as a supervised learning algorithm in recovering the functional mapping of behavioral cues to real-valued annotations of affect and behavior, while in the last task it was assigned the role of an unsupervised learning model in distinguishing dynamics corresponding to motion trajectories of different objects/persons in visually cluttered scenarios. All three tasks were posited as frame-by-frame regression problems, consisting of predicting future values or 'completing' missing intermediate values characterizing human affect, behavior or motion from a small amount of observations captured under 'in-the-wild' conditions. Our experimental findings serve as a testament to a compelling research achievement in an era where 'big data' is regarded as a prerequisite for the efficiency of a machine learning model. Specifically, we demonstrated for the first time that complex human behavior and affect, manifested by a single person or group of interactants, can be learned and predicted based on small training sets of person(s)-specific observations, amounting to a duration of just a few seconds.

6.2 Future Work

There are numerous possible extensions of this work with respect to both the conceptual platform on the basis of which analysis of high-level affective displays and social behaviors should be posited as well as the assets with which machine learning paragons should be endowed to approach these tasks in a principled and efficient manner. In what follows, we list directions that merit most attention for future investigation.

First of all, we regard as indispensable the need for the release of bigger and better datasets comprising data that are (i) captured in real-world, 'in-the-wild' rather than laboratory conditions based on handy rather than intrusive acquisition devices (e.g., mobile devices) and (ii) annotated rigorously by multiple experts in terms of spontaneous rather than posed expression of affective and social internal states. In view of the multi-modal nature of human affect and behavior signals, the datasets should include sensory information from multiple modalities so as to facilitate the deployment of multi-modal interfaces. On the other hand, human assessments of affect and behavior should also encapsulate contextual and temporal information which is regarded as crucial in the human and machine perception of the relevant subtle, highly-ambiguous cues. Motivated by our findings on automatic analysis of a social attitude, namely conflict, we argue that the social signal processing community should imitate the affective computing community in investing more research efforts in establishing dimensional rather than categorical descriptions of social signals and behaviors to goad the development of temporal modeling paradigms for the relevant problems. To this end, effective annotation tools should be developed to ensure that continuous, real-time annotations devoid of meaningless artifacts be generated with the minimum amount of effort from the human raters. Finally, more sophisticated annotation fusion techniques should be devised to reduce the noise effects naturally incurred by the process of combining multiple human characterizations of elusive human affect and behavior phenomena into a single ground truth annotation.

Modeling naturalistic human affect and behavior based on real-world data and erratic human annotations unavoidably comes with the expense of having to deal with various types of noise that can be unbounded in magnitude and having a random support in the measurement domain, thus rendering the assumption of a Gaussian model unrealistic. The presence of such outlying measurements can lead to solutions that are arbitrarily skewed from the desired solution and thus be detrimental to the performance achieved by classical machine learning approaches based on least squares estimation techniques. Hence, it becomes evident that applications such as face recognition facial expression recognition under contiguous occlusion and non-uniform illumination or recognition of social signals and behaviors in multi-party conversations involving extreme head pose angles and abrupt head, hand and body movements, necessitate machine learning models that be robust to such gross but sparse noise. In this light, robustifying existing approaches or employing robust statistics and optimization techniques to design new robust models should be considered as one of the top priorities of research in machine analysis of naturalistic human non-verbal behavior.

Regarding static face analysis, which is one of the two main application domains with which this thesis has dealt with, we strongly believe that multiple future research avenues can have as a starting point the Discriminant Incoherent Component Analysis (DICA) presented in

Chapter 3. In particular, we have seen that discriminant dictionary learning and sparsity-based recognition can serve as the main premises for the construction of an efficient, in terms of both accuracy and computational load, unified learning framework that can jointly address intertwined classification tasks of labeled facial attributes such as facial identity, facial expression and activation of AUs. Future research could investigate the appropriateness of alternative structures for the extraction of class-specific components related to other type of facial attributes such as pose and illumination in a supervised manner as well as extend the DICA to the temporal dimension, building on tensor rather than matrix decomposition. Overall, ideas stemming from multi-task learning techniques could be applied to couple the DICA with deeper, hierarchical architectures for the concurrent extraction of multiple, spatial and temporal, components in video sequences capturing different aspects of spatio-temporal (not just facial) human behavior.

Regarding dynamic affect and behavior analysis, we maintain that designing classifiers that can explicitly model the temporal dynamics of non-verbal cues through which spontaneous emotion and social behavior, characterized by means of dimensional descriptions, is manifested in longer temporal intervals should be a top priority in the fields of affective computing and social signal processing. More research efforts should be invested on devising classifiers that can capture the hidden temporal structures, the synergy of multi-modal cues, and the contextual information collectively signifying expressions of affective and social behavior internal states displayed in data acquired under real-world conditions. The dynamic behavior analysis framework presented in Chapter 5 can serve as a leading exemplar for the development of automated frameworks capable of explicitly modeling temporal dynamics so as to address dimensional affect and behavior analysis in continuous-time from small training sets, viewed as a frame-by-frame regression problem rather than a sequence classification problem.

Various extensions to our dynamic affect and behavior analysis framework could be explored in order to endow it with the aforementioned desirable properties. A natural extension would be to equip it with the ability to recover a latent auto-regressive process describing highly correlated temporal patterns of affect and behavior from various modalities, thus allowing it to perform efficient model-level feature fusion for multi-modal settings. Furthermore, ideas from factor analysis, tensor/multi-linear decomposition and source separation techniques could be exploited in order to enable our method to decouple components describing affect- and behavior-related temporal dynamics from components encoding contextual information related to e.g., the identity or culture of person exhibiting an observed behavior or the stimulus that caused it, in naturalistic, multi-party conversations. This property, aside from enhancing the accuracy of the our framework in highly context-dependent scenarios, could facilitate research in social and cognitive studies targeting social role recognition and causality detection in social interactions such as those arising in group meetings. On the other hand, we have seen that our method can accurately learn a dynamical system describing the continuous-time annotations characterizing person(s)-specific behavior or affect as outputs when behavioral cues (features) act as system inputs from a single image sequence. Extending the proposed structured rank minimization method, which is the core of this learning framework, so that it can learn such a system from multiple sequences of different length portraying the same or even different subjects, would have numerous benefits. First of all, this extension would make it possible for our method to learn templates of prototypic subject(s)-dependent or subject(s)-independent observed behaviors, which could be used for various applications such as behavioral biometrics, personality recognition or group behavior and crowd analysis, respectively, to mention but a few. Moreover, learning grammars of dynamic affect and behavior from multiple image sequences and describing them explicitly in terms of the parameters of linear dynamical systems, could open numerous possibilities for measuring behavior similarity. In other words, by using metric learning techniques one could directly compare the learned systems to identify the match/mismatch or the degree of similarity between two behavior prototypes or instances. This way, one could use this approach to distinguish triple jump from long jump or ball dribbling from ball shooting for sports video analytics, to identify early signs of depression in an observed individual or to measure similarity in the responses of different individuals to machine-mediated communication in the work environment. Finally, another fascinating research avenue would be to use ideas inspired by the recent success of generative adversarial networks to combine our system learning method with a generative neural network-based model where the space of system parameters would act as the feature domain for both classifiers. A clear advantage of this combined model would be its straightforward capability of learning systems from multiple sequences as well as of generating new, synthetic instances of the same dynamical system. It would be interesting to explore the discriminative power and data generating ability of such unified adversarial learning framework as compared to generative- or discriminative-only learning approaches, such as those employed in this thesis.

6.3 Epilogue

We hope that we have revealed new research avenues, provided valuable insights and inspired eagerness for machine analysis of spontaneous affect and behavior. Our research journey convinced us that endowing machines with emotional and social competence necessitates robust, scalable and interpretable models that can efficiently capture latent contextual and temporal regularities in the observed affective and behavioral displays. We hope that the machine learning techniques presented in this thesis have enlightened crucial aspects of this modeling paradigm and will constitute the alpha for its future advancement.

Appendix A

Appendices

A.1 Solution of Problem (3.8)

Let us consider the problem (3.8). In this step of ADMM, we are minimizing w.r.t. $\mathbf{V}^{(i)}$ at iteration t, with $\{\mathbf{U}^{(i)}\}_{i=1}^{n_c}$, $\{\mathbf{V}^{(j)}[t]\}_{j\neq i}$, and $\mathbf{O}$ kept fixed. Let us re-write the problem for clarity of presentation:

$$\begin{aligned}
\mathbf{V}^{(i)}[t+1] &= \underset{\mathbf{V}^{(i)}}{\arg\min}\, \mathcal{L}(\mathbf{V}^{(i)}, \mathbf{Y}[t], \mu[t]) \\
&= \underset{\mathbf{V}^{(i)}}{\arg\min}\, \lambda^{(i)}\|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta \sum_{i\neq j} \|\mathbf{V}^{(i)}\mathbf{V}^{(j)^T}\|_F^2 \\
&+ \frac{\mu[t]}{2}\|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1}\mathbf{Y}[t]\|_F^2 \\
&= \underset{\mathbf{V}^{(i)}}{\arg\min}\, \lambda^{(i)}\|\mathbf{V}^{(i)}\|_{(\cdot)} + f(\mathbf{V}^{(i)})
\end{aligned} \tag{A.1}$$

The minimizer (A.1) consists of a non-smooth term, induced by a norm function $\|\cdot\|_{(\cdot)}$, and a smooth, twice differentiable term described by the function f. It can easily be proved that the gradient ∇f is Lipschitz-continuous.

By linearizing f in the vicinity of the current point $\mathbf{V}^{(i)}[t]$, and by exploiting the Lipschitzcontinuity of ∇f, we obtain the following equivalent problem

$$\begin{aligned} &\min_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + f(\mathbf{V}^{(i)}[t]) \\ &+ \operatorname{tr}\left(\nabla f(\mathbf{V}^{(i)}[t])^T (\mathbf{V}^{(i)} - \mathbf{V}^{(i)}[t])\right) \\ &+ \frac{L}{2}\|\mathbf{V}^{(i)} - \mathbf{V}^{(i)}[t]\|_F^2 \end{aligned} \tag{A.2}$$

where $L > 0$ is an upper bound on the Lipschitz constant of ∇f. Problem (A.2) is re-written as

$$\min_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \frac{1}{2}\|\mathbf{V}^{(i)} - (\mathbf{V}^{(i)}[t] - \frac{1}{L}\nabla f(\mathbf{V}^{(i)}[t])\|_F^2 \tag{A.3}$$

Having expressed the minimizer in this form, we now directly apply the SVT (shrinkage) operator, in case the nuclear- (ℓ_1-) norm is chosen for the first term of (A.3). For the nuclear norm, the solution is given by

$$\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{S}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - \frac{1}{L}\nabla f(\mathbf{V}^{(i)}[t])\right], \tag{A.4}$$

whereas for the ℓ_1-norm the solution is given by

$$\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{D}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - \frac{1}{L}\nabla f(\mathbf{V}^{(i)}[t])\right] \tag{A.5}$$

The gradient $\nabla f(\mathbf{V}^{(i)}[t])$ is computed as

$$\begin{aligned} \nabla f(\mathbf{V}^{(i)}[t]) = \left(-\mu[t]\mathbf{U}^{(i)}[t]^T\right)\Bigg(\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)}[t]\mathbf{V}^{(i)}[t]\mathbf{X}_{\mathcal{S}^{(i)}} \\ - \mathbf{O}[t] + \mu[t]^{-1}\mathbf{Y}[t]\Bigg)\mathbf{X}_{\mathcal{S}^{(i)}}^T + 2\eta\sum_{j\neq i}\mathbf{V}^{(j)}[t]^T\mathbf{V}^{(j)}[t], \end{aligned} \tag{A.6}$$

whereas an upper bound on the Lipschitz constant of ∇f is given by

$$L = 1.02\lambda_{\max}\left[\mu[t]\mathbf{X}_{\mathcal{S}^{(i)}}\mathbf{X}_{\mathcal{S}^{(i)}}^T + 2\eta\sum_{j\neq i}\mathbf{V}^{(j)}[t]^T\mathbf{V}^{(j)}[t]\right] \tag{A.7}$$

The respective closed-form solutions are obtained by substituting (A.6) and (A.7) into (A.4) or (A.5).

Bibliography

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. The K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 21, 59

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *Computer vision-ECCV 2004*, pages 469–481. Springer, 2004. 79

[3] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006. 67

[4] Joan Alabort-i Medina, Epameinondas Antonakos, James Booth, Patrick Snape, and Stefanos Zafeiriou. Menpo: A Comprehensive Platform for Parametric Image Alignment and Visual Deformable Models. In *Proceedings of the ACM International Conference on Multimedia (ACMMM), Orlando, Florida, USA*, pages 679–682, 2014. 77

[5] Jens Allwood. Cooperation, competition, conflict and communication. *Gothenburg Papers in Theoretical Linguistics*, 94:1–14, 2007. 9, 70

[6] Zara Ambadar, Jonathan W Schooler, and Jeffrey F Cohn. Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions. *Psychological science*, 16(5):403–410, 2005. 25

[7] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis., 1992. 2

[8] Theodore Wilbur Anderson. An introduction to multivariate statistical analysis. Technical report, Wiley New York, 1962. 76

[9] E. Antonakos, J. Alabort-i medina, and S. Zafeiriou. Active Pictorial Structures. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5435–5444, Boston, MA, USA, 2015. 79, 85

[10] Mustafa Ayazoglu, Binlong Li, Caglayan Dicle, Mario Sznaier, Octavia Camps, et al. Dynamic subspace-based coordinated multicamera tracking. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2462–2469, 2011. 30

[11] Mustafa Ayazoglu, Mario Sznaier, and Octavia Camps. Fast algorithms for structured robust principal component analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1704–1711. IEEE, 2012. 30, 97, 99, 110

[12] Mustafa Ayazoglu, Burak Yilmaz, Mario Sznaier, and Octavia Camps. Finding causal interactions in video sequences. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3575–3582, 2013. 30

[13] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002. 16

[14] Tadas Baltrusaitis, Ntombikayise Banda, and Peter Robinson. Dimensional affect recognition using continuous conditional random fields. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013. 26, 81, 82

[15] Tanja Bänziger and Klaus R Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook and manual*, pages 271–294, 2010. 126, 127

[16] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 15, 16

[17] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *Journal on Image and Video Processing*, 2008:1, 2008. 129

[18] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014. 47, 49, 50, 93, 101, 104

[19] Surya Bhattacharya, Mahdi M Kalayeh, Rahul Sukthankar, and Mubarak Shah. Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2243–2250, 2014. 30

[20] Joan-Isaac Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2013. 3

[21] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual Detection of Behavioural Mimicry. In *Affective Computing and Intelligent Interaction (ACII 2013)*, pages 123–128, Geneva, Switzerland, September 2013. 77

[22] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61, 2015. 3, 37

[23] Matthew Black, Athanasios Katsamanis, Chi-Chun Lee, Adam C Lammert, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. Automatic classification of married couples' behavior using audio features. In *INTERSPEECH*, pages 2030–2033, 2010. 27

[24] Konstantinos Bousmalis. *Infinite Hidden Conditional Random Fields for the Recognition of Human Behaviour*. PhD thesis, Imperial College London, 2014. 8, 24, 27

[25] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–9, 2009. 3, 26, 34, 70, 77, 85

[26] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, 31(2):203–221, 2013. 27, 28, 34, 36, 85

[27] Konstantinos Bousmalis, Louis Philippe Morency, and Maja Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, pages 746–752, 2011. 3, 9, 26, 27, 28, 36, 37, 39, 70

[28] Konstantinos Bousmalis, Stefanos Zafeiriou, Louis-Philippe Morency, and Maja Pantic. Infinite hidden conditional random fields for human behavior analysis. *IEEE transactions on neural networks and learning systems*, 24(1):170–177, 2013. 3, 9, 28

[29] Thierry Bouwmans and El Hadi Zahzah. Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014. 7

[30] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 81, 82

[31] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. 48

[32] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011. 7, 17, 20, 21, 22, 48, 50, 54, 55, 59, 100, 103, 104, 111

[33] Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007. 35, 36, 37

[34] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 81

[35] Linlin Chao, Jianhua Tao, Minghao Yang, and Ya Li. Multi task sequence learning for depression scale prediction from video. In *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 526–531, 2015. 82, 88

[36] Michael Chau and Margrit Betke. Real time eye tracking and blink detection with usb cameras. *Boston University Computer Science*, 2215(2005-2012):1–10, 2005. 3, 26, 70

[37] Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016. 105

[38] Hui Chen, Jiangdong Li, Fengjun Zhang, Yang Li, and Hongan Wang. 3D model-based continuous emotion recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1836–1845, June 2015. 82

[39] G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline Deformable Face Tracking in Arbitrary Videos. In *Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15)*, Santiago, Chile, December 2015. 77, 118

[40] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Selective transfer machine for personalized facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3515–3522. IEEE, 2013. 65, 66

[41] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 90

[42] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015. 31, 90

[43] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1):160–187, 2003. 9, 23, 32, 92

[44] Virginia W Cooper. Participant and observer attribution of affect in interpersonal conflict: an examination of noncontent verbal behavior. *Journal of Nonverbal Behavior*, 10(2):134–144, 1986. 70, 75, 77

[45] Tim Cootes, E Baldock, and J Graham. An introduction to active shape models. *Image processing and analysis*, pages 223–248, 2000. 78

[46] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 26

[47] R Cowie, H Gunes, G McKeown, L Vaclau-Schneider, J Armstrong, and E Douglas-Cowie. The emotional and communicative significance of head nods and shakes in a naturalistic database. In *LREC Int. Workshop on Emotion*, pages 42–46, 2010. 25

[48] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000. 8

[49] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. 90

[50] Yuchao Dai and Hongdong Li. Rank minimization or nuclear-norm minimization: Are we solving the right problem? In *International Conference on Digital lmage Computing: Techniques and Applications (DlCTA)*, pages 1–8. IEEE, 2014. 99, 100

[51] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005. 67

[52] Fernando De La Torre and Michael J Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1-3):117–142, 2003. 4, 17

[53] Oscar Déniz, M Castrillon, J Lorenzo, L Anton, and Gloria Bueno. Smile detection for user interfaces. In *Advances in Visual Computing*, pages 602–611. Springer, 2008. 3, 23, 26, 70, 92

[54] Caglayan Dicle, Octavia Camps, Mario Sznaier, et al. The way they move: tracking multiple targets with similar appearance. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2304–2311, 2013. 30, 97, 110, 124, 128

[55] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *International Conference on Machine Learning*, pages 281–288. ACM, 2006. 17, 22

[56] Tao Ding, Mario Sznaier, and Octavia Camps. A rank minimization approach to fast dynamic event detection and track matching in video sequences. In *IEEE Conference on Decision and Control (CDC)*, pages 4122–4127, 2007. 30, 110

[57] Tao Ding, Mario Sznaier, and Octavia Camps. Receding horizon rank minimization based estimation with applications to visual tracking. In *IEEE Conference on Decision and Control (CDC)*, pages 3446–3451, 2008. 30, 110

[58] Tao Ding, Mario Sznaier, Octavia Camps, et al. A rank minimization approach to video inpainting. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 30

[59] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005. 23, 92

[60] David L Donoho. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006. 7, 17, 18, 45, 100

[61] P Ekman, WV Friesen, and JC Hager. Facial action coding system. *Salt Lake City: Research Nexus eBook*, 2002. 7, 14

[62] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003. 26

[63] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005. 3, 9, 28

[64] Ehsan Elhamifar and René Vidal. Robust classification using structured sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1873–1879, 2011. 20, 22

[65] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2001. 65, 66

[66] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the ACM international conference on Multimedia (ACMMM)*, pages 1459–1462, 2010. 76

[67] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 54, 65

[68] Maryam Fazel. *Matrix rank minimization with applications.* PhD thesis, PhD thesis, Stanford University, 2002. 6, 45

[69] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001. 6, 17, 30, 97, 100

[70] Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013. 29, 30, 93, 95, 97, 99, 102, 110

[71] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 669. Association for Computational Linguistics, 2004. 70, 75

[72] C. Georgakis, S. Petridis, and M. Pantic. Discriminating Native from Non-Native Speech Using Fusion of Visual Cues. In *ACM International Conference on on Multimedia (ACMMM)*, pages 1177–1180, Orlando, Florida, USA, November 2014. 77, 82

[73] C. Georgakis, S. Petridis, and M. Pantic. Discrimination Between Native and Non-Native Speech Using Visual Features Only. *IEEE Transactions on Cybernetics (TCYB)*, 46(12):2758–2771, December 2016. 77, 85

[74] Christos Georgakis, Yannis Panagakis, and Maja Pantic. Discriminant Incoherent Component Analysis. *IEEE Transactions on Image Processing*, 25(5):2021–2034, 2016. 74, 79

[75] Christos Georgakis, Stavros Petridis, and Maja Pantic. Visual-only discrimination between native and non-native speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4828–4832, 2014. 77

[76] Panayiotis G Georgiou, Matthew P Black, Adam C Lammert, Brian R Baucom, and Shrikanth S Narayanan. "That's Aggravating, Very Aggravating": Is It Possible to Classify Behaviors in Couple Interactions Using Automatically Derived Lexical Features? In *Affective Computing and Intelligent Interaction*, pages 87–96. Springer, 2011. 36, 37

[77] Sayan Ghosh, Eugene Laksana, Stefan Scherer, and Louis-Philippe Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. In *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 609–615, 2015. 23

[78] Alex Graves. Rnnlib: A recurrent neural network library for sequence learning problems, 2013. 81

[79] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. 82

[80] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 51, 54, 62

[81] H. Gunes, M. A. Nicolaou, and M. Pantic. *Continuous Analysis of Affect from Voice and Face*, pages 255–292. Springer-Verlag, 2011. 24, 25

[82] Hatice Gunes and Maja Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Intelligent virtual agents*, pages 371–377. Springer, 2010. 26

[83] Hatice Gunes and Björn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013. 2, 3, 8, 24, 25, 26, 70, 77, 82, 83, 84

[84] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 827–834. IEEE, 2011. 14

[85] M Hassaballah and Saleh Aly. Face recognition: challenges, achievements and future directions. *IET Computer Vision*, 9(4):614–626, 2015. 16

[86] Ran He, Bao-Gang Hu, Wei-Shi Zheng, and Xiang-Wei Kong. Robust principal component analysis based on maximum correntropy criterion. *IEEE Transactions on Image Processing*, 20(6):1485–1494, 2011. 17, 22

[87] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1208–1213. IEEE, 2005. 16

[88] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005. 16

[89] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 41(6):765–781, 2011. 79

[90] Peter J Huber. *Robust statistics*. Springer, 2011. 4, 6, 17, 99, 121

[91] Hayley Hung and Gokul Chittaranjan. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *ACM International Conference on Multimedia*, pages 879–882, 2010. 36, 37

[92] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The ICSI meeting corpus. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–364, 2003. 36, 37

[93] DB Jayagopi, H Hung, C Yeo, and D Gatica-Perez. Modeling Dominance in Group Conversations from Nonverbal Activity Cues. *IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on Multimodal Processing for Speech-based Interactions*, 2009. 3, 27

[94] Pan Ji, Hongdong Li, Mathieu Salzmann, and Yuchao Dai. Robust motion segmentation with unknown correspondences. In *European Conference on Computer Vision*, pages 204–219. Springer, 2014. 105

[95] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic. A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modelling. *IEEE Transactions on Cybernetics*, 44(2):161–174, 2014. 16

[96] Bin Jiang and Ke-bin Jia. Research of robust facial expression recognition under facial occlusion condition. In *International Conference on Active Media Technology*, pages 92–100. Springer, 2011. 17

[97] Xudong Jiang and Jian Lai. Sparse and Dense Hybrid Representation via Dictionary Decomposition for Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):1067–1079, 2015. 20, 22

[98] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002. 5, 15, 78

[99] Charles M Judd. Cognitive effects of attitude conflict resolution. *Journal of Conflict Resolution*, 22(3):483–498, 1978. 9, 70, 75

[100] Patrik N Juslin, Klaus R Scherer, J Harrigan, R Rosenthal, and K Scherer. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, pages 65–135, 2005. 2

[101] S. Kaltwang, S. Todorovic, and M. Pantic. Doubly Sparse Relevance Vector Machine for Continuous Facial Behavior Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015. (to appear). 25, 26, 27, 77, 79, 83, 119

[102] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377. Springer, 2012. 9, 27, 92

[103] Shinjiro Kawato and Jun Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 40–45, 2000. 3, 23, 26, 70, 92

[104] Dacher Keltner and P Ekman. Expression of emotion. *Handbook of Affective Sciences. Oxford University Press, New York*, pages 411–414, 2003. 2

[105] Minyoung Kim and Vladimir Pavlovic. Discriminative learning for dynamic state prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1847–1861, 2009. 28

[106] Samuel Kim, Fabio Valente, and Alessandro Vinciarelli. Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, pages 5089–5092, 2012. 3, 26, 27, 35, 70, 71, 72, 75, 76, 77

[107] Samuel Kim, Sree Harsha Yella, and Fabio Valente. Automatic detection of conflict escalation in spoken conversations. In *INTERSPEECH*, pages 1167–1170, 2012. 3, 26, 27, 70, 71, 76, 77

[108] Sungho Kim, Filipe Valente, Maurizio Filippone, and Alessandro Vinciarelli. Predicting Continuous Conflict Perception with Bayesian Gaussian Processes. *IEEE Transactions on Affective Computing*, 5(2):187–200, 2014. 3, 27, 28, 36, 70, 71, 76, 77

[109] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 32, 90, 133

[110] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 23, 90

[111] Nojun Kwak. Principal component analysis based on l1-norm maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(9):1672–1680, 2008. 17, 22

[112] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 1, pages 282–289, 2001. 82, 87

[113] Richard D Lane and Lynn Nadel. *Cognitive neuroscience of emotion*. Oxford University Press, USA, 2002. 8, 24, 124

[114] Chan-Su Lee and Rama Chellappa. Sparse localized facial motion dictionary learning for facial expression recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3548–3552, 2014. 19

[115] Annick M Leroy and Peter J Rousseeuw. Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*, 1987. 4

[116] John M Levine and Richard L Moreland. *Small groups: key readings.* Psychology Press, 2008. 9, 70

[117] Binlong Li, Mustafa Ayazoglu, Teresa Mao, Octavia Camps, Mario Sznaier, et al. Activity recognition using dynamic subspace angles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3200, 2011. 30

[118] Binlong Li, Octavia Camps, Mario Sznaier, et al. Cross-view activity recognition using hankelets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1362–1369, 2012. 30

[119] Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015. 105

[120] Athanasios P Liavas and Nicholas D Sidiropoulos. Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(20):5450–5463, 2015. 105

[121] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006. 23, 92

[122] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013. 6

[123] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670, 2010. 6

[124] Guangcan Liu and Shuicheng Yan. Active subspace: Toward scalable low-rank learning. *Neural computation*, 24(12):3371–3394, 2012. 6, 49, 108

[125] Jun Liu and Jieping Ye. Efficient Euclidean projections in linear time. In *ACM International Conference on Machine Learning*, pages 657–664, 2009. 55

[126] Xiao Liu, Mingli Song, Dacheng Tao, Zicheng Liu, Luming Zhang, Chun Chen, and Jiajun Bu. Semi-supervised node splitting for random forest construction. In *IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 492–499, 2013. 88

[127] Andrea Lockerd and Florian Mueller Mueller. LAFCam: Leveraging affective feedback camcorder. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pages 574–575. ACM, 2002. 3, 23, 26, 70, 92

[128] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 79

[129] Juwei Lu, Konstantinos N Plataniotis, Anastasios N Venetsanopoulos, and Stan Z Li. Ensemble-based discriminant learning with boosting for face recognition. *IEEE transactions on neural networks*, 17(1):166–178, 2006. 16

[130] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101, 2010. 51, 54, 57

[131] Sindri Magnusson, Pradeep Chathuranga Weeraddana, Michael Rabbat, and Carlo Fischione. On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. *IEEE Transactions on Control of Network Systems* , 2015. 105

[132] Mohammad H Mahoor, Mu Zhou, Kevin L Veon, S Mohammad Mavadati, and Jeffrey F Cohn. Facial action unit recognition with sparse representation. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)* , pages 336–342, 2011. 16, 19, 65

[133] Xia Mao, YuLi Xue, Zheng Li, Kang Huang, and ShanWei Lv. Robust facial expression recognition based on rpca and adaboost. In *2009 10th Workshop on Image Analysis for Multimedia Interactive Services*, pages 113–116. IEEE, 2009. 17

[134] Ivan Markovsky. Recent progress on variable projection methods for structured low-rank approximation. *Signal Processing*, 96:406–419, 2014. 30, 97, 99, 110

[135] Aleix M Martinez. The AR face database. *CVC Technical Report*, 24, 1998. 51, 54, 60

[136] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 119

[137] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. The AMI meeting corpus. In *International Conference on Methods and Techniques in Behavioral Research*, volume 88, 2005. 35, 36, 37

[138] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012. 26, 125

[139] Kerui Min, Zhengdong Zhang, John Wright, and Yi Ma. Decomposing background topics from keywords by principal component pursuit. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 269–278. ACM, 2010. 7

[140] Mohammad Reza Mohammadi, Emad Fatemizadeh, and Mohammad H Mahoor. Intensity estimation of spontaneous facial action units based on their sparsity properties. *IEEE transactions on cybernetics*, 46(3):817–826, 2016. 20

[141] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84, 2010. 26, 92

[142] Imran Naseem, Roberto Togneri, and Mohammed Bennamoun. Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2106–2112, 2010. 18, 54, 57, 60, 62, 64, 66

[143] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995. 100

[144] Yuzhao Ni, Ju Sun, Xiaotong Yuan, Shuicheng Yan, and Loong-Fah Cheong. Robust low-rank subspace segmentation with semidefinite guarantees. In *IEEE International Conference on Data Mining Workshops*, pages 1179–1188, 2010. 6

[145] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Transactions on Affective Computing*, pages 92–105, 2011. 9, 26, 76, 77, 82, 83, 87, 88, 92

[146] M. A. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic Probabilistic CCA for Analysis of Affective Behaviour and Fusion of Continuous Annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1299–1311, 2014. 3, 4, 8, 24, 33, 76

[147] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30(3):186–196, 2012. 9, 25, 26, 92

[148] Mihalis A Nicolaou, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Robust Canonical Correlation Analysis: Audio-visual fusion for learning continuous interest. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1522–1526, 2014. 27, 77

[149] Feiping Nie, Heng Huang, and Chris Ding. Low-Rank Matrix Recovery via Efficient Schatten p-Norm Minimization. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. 100

[150] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Joint Schatten p-norm and \ ell _p-norm robust matrix completion for missing value recovery. *Knowledge and Information Systems*, 42(3):525–544, 2013. 100, 102, 103

[151] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision (IJCV)*, 79(3):299–318, 2008. 23, 92

[152] X Niyogi. Locality preserving projections. In *Neural information processing systems*, volume 16, page 153. MIT, 2004. 15

[153] J. Orozco, O. Rudovic, J. Gonzàlez, and M. Pantic. Hierarchical On-line Appearance-Based Tracking for 3D Head Pose, Eyebrows, Lips, Eyelids and Irises. *Image and Vision Computing*, 31(4):322–340, February 2013. 125

[154] Weihua Ou, Xinge You, Dacheng Tao, Pengyue Zhang, Yuanyan Tang, and Ziqi Zhu. Robust face recognition via occlusion dictionary learning. *Pattern Recognition*, 47(4):1559–1572, 2014. 20, 22

[155] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic. Robust Correlated and Individual Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Special Issue in Multimodal Pose Estimation and Behaviour Analysis, (accepted)*, 2016. 6, 27, 42, 44, 71, 72, 77, 87

[156] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce. Music genre classification via joint sparse low-rank representation of audio features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1905–1917, 2014. 42

[157] Yannis Panagakis, Mihalis Nicolaou, Stefanos Zafeiriou, and Maja Pantic. Robust canonical time warping for the alignment of grossly corrupted sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 540–547, 2013. 42

[158] Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Audiovisual Conflict Detection in Political Debates. In *Computer Vision-ECCV 2014 Workshops*, pages 306–314. Springer, 2014. 72

[159] M. Pantic. *Affective Computing (revisited)*, volume 1, pages 15–21. Information Science Reference, 2009. 2

[160] M. Pantic, R. Cowie, F. D'ericco, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroder, and A. Vinciarelli. *Social Signal Processing: The Research Agenda*, pages 511–538. Springer, 2011. 1, 2, 26, 34, 36, 37, 70

[161] M. Pantic and A. Vinciarelli. *Social Signal Processing*, pages 84–93. Springer, 2014. 1, 2, 3, 4, 8, 24, 26, 27, 34, 70, 86

[162] Maja Pantic. Facial expression recognition. In *Encyclopedia of biometrics*, pages 400–406. Springer, 2009. 14, 42

[163] Maja Pantic. Automatic analysis of facial expressions. *Encyclopedia of Biometrics*, pages 128–134, 2015. 2, 14, 26, 42, 85

[164] Maja Pantic and Marian Stewart Bartlett. *Machine analysis of facial expressions*. I-Tech Education and Publishing, 2007. 2, 15, 25, 26, 124

[165] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S Huanag. Human-Centred Intelligent Human? Computer Interaction (HCI^2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008. 1, 3, 26, 70

[166] Maja Pantic and Leon JM Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(12):1424–1445, 2000. 2, 23, 92

[167] Maja Pantic and Leon JM Rothkrantz. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424–1445, 2000. 14

[168] G. Papamakarios, Y. Panagakis, and S. Zafeiriou. Generalised Scalable Robust Principal Component Analysis. In *British Machine Vision Conference (BMVC 2014)* , 9 2014. 6, 42, 74, 100, 103, 105, 108

[169] Haesun Park, Lei Zhang, and J Ben Rosen. Low rank approximation of a hankel matrix by structured total least norm. *BIT Numerical Mathematics*, 39(4):757–779, 1999. 110, 111

[170] Vishal M Patel and Ramalingam Chellappa. Sparse representations, compressive sensing and dictionaries for pattern recognition. In *First Asian Conference on Pattern Recognition (ACPR)*, pages 325–329. IEEE, 2011. 7

[171] Vladimir Pavlović, James M Rehg, Tat-Jen Cham, and Kevin P Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 94–101, 1999. 92

[172] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 81

[173] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012. 49

[174] Alex Pentland. Social signal processing [exploratory dsp]. *IEEE Signal Processing Magazine*, 24(4):108–111, 2007. 2, 3, 26

[175] Anna Pesarin, Marco Cristani, Vittorio Murino, and Alessandro Vinciarelli. Conversation analysis at work: detection of conflict in competitive discussions through semi-automatic turn-organization analysis. *Cognitive processing*, 13(2):533–540, 2012. 36

[176] Stavros Petridis and Maja Pantic. Audiovisual discrimination between speech and laughter: Why and when visual information might help. *IEEE Transactions on Multimedia*, 13(2):216–234, 2011. 78

[177] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *ACM International Conference on Multimodal Interfaces*, pages 53–60, 2008. 3, 27

[178] Fabio Pianesi, Massimo Zancanaro, Bruno Lepri, and Alessandro Cappelletti. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3-4):409–429, 2007. 36, 37

[179] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997. 1, 2

[180] Gerasimos Potamianos, Ashish Verma, Chalapathy Neti, Giridharan Iyengar, and Sankar Basu. A cascade image transform for speaker independent automatic speechreading. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1097–1100, 2000. 79

[181] Liliana Lo Presti and Marco La Cascia. Ensemble of Hankel Matrices for Face Emotion Recognition. In *Image Analysis and Processing (ICIAP)*, pages 586–597. Springer, 2015. 30

[182] Liliana Lo Presti, Marco La Cascia, Stan Sclaroff, and Octavia Camps. Hankelet-based dynamical systems modeling for 3D action recognition. *Image and Vision Computing*, 44:29–43, 2015. 30

[183] Raymond Ptucha, Grigorios Tsagkatakis, and Andreas Savakis. Manifold based sparse representation for robust expression recognition without neutral subtraction. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2136–2143, 2011. 19

[184] Lishan Qiao, Songcan Chen, and Xiaoyang Tan. Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 43(1):331–341, 2010. 16

[185] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015. 76

[186] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 16

[187] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Raps: Robust and efficient automatic construction of person-specific deformable models. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, June 2014. 42, 79, 108

[188] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical face frontalization. In *Proceedings of IEEE Int'l Conf. on Computer Vision (ICCV 2015)*, Santiago, Chile, December 2015. 74

[189] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical frontalization of human and animal faces. *International Journal of Computer Vision, Special Issue on "Machine Vision Applications"*, June 2016. 74

[190] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 896–903. IEEE, 2013. 64

[191] Tomoya Sakai, Hayato Itoh, and Atsushi Imiya. Multi-label classification for image annotation via sparse similarity voting. In *Computer Vision–ACCV 2010 Workshops*, pages 344–353. Springer, 2011. 54

[192] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2015. 16

[193] Lawrence K Saul and Sam T Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, 4:119–155, 2003. 16

[194] Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017. 23

[195] Klaus R Scherer, Tanja Bänziger, and Etienne Roesch. *A Blueprint for Affective Computing: A sourcebook and manual*. Oxford University Press, 2010. 126

[196] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *INTERSPEECH*. Citeseer, 2013. 36, 37, 39, 82, 87

[197] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll. Using multiple databases for training in emotion recognition: To unite or to vote? In *INTERSPEECH*, pages 1553–1556. Citeseer, 2011. 82

[198] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing (ICIP)*, volume 2, pages II–370, 2005. 16

[199] Caifeng Shan, Shaogang Gong, and Peter W McOwan. A comprehensive empirical study on linear subspace methods for facial expression analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 153–153. IEEE, 2006. 16

[200] Linlin Shen and Li Bai. A review on gabor wavelets for face recognition. *Pattern analysis and applications*, 9(2-3):273–292, 2006. 16

[201] Yuan Shen, Zaiwen Wen, and Yin Zhang. Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, 29(2):239–263, 2014. 105

[202] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979. 83, 88, 119

[203] Marco Signoretto, Volkan Cevher, and Johan AK Suykens. An SVD-free approach to a class of structured low rank matrix optimization problems with application to system identification. In *IEEE Conference on Decision and Control (CDC)*, 2013. 30, 97, 99, 110

[204] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 67, 90, 133

[205] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004. 26, 81

[206] Patrick Snape, Yannis Panagakis, and Stefanos Zafeiriou. Automatic construction of robust spherical harmonic subspaces. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 218–233, 2015. 42

[207] Ian Sneddon, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, 2012. 36, 37

[208] Dennis L Sun and Cédric Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6201–6205. IEEE, 2014. 105

[209] Amit Surana, Arie Nakhmani, and Allen Tannenbaum. Anomaly detection in videos: A dynamical systems approach. In *IEEE Conference on Decision and Control (CDC)*, pages 6489–6495, 2013. 30, 31

[210] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 31

[211] Sima Taheri, Vishal M Patel, and Rama Chellappa. Component-Based Recognition of Faces and Facial Expressions. *IEEE Transactions on Affective Computing*, 4(4):360–371, 2013. 21, 22, 50, 53, 55, 57, 59

[212] Sima Taheri, Qiang Qiu, and Rama Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *IEEE Transactions on Image Processing*, 23(8):3590–3603, 2014. 19

[213] Pete C Trimmer, Elizabeth S Paul, Mike T Mendl, John M McNamara, and Alasdair I Houston. On the evolution and optimality of mood states. *Behavioral Sciences*, 3(3):501–521, 2013. 25

[214] Matthew Turk, Alex P Pentland, et al. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991. 16

[215] Georgios Tzimiropoulos and Maja Pantic. Optimization Problems for Fast AAM Fitting in-the-Wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 77, 118

[216] Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. Subspace learning from image gradient orientations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2454–2466, 2012. 17, 22

[217] Michel Valstar, Jonathan Gratch, Bjorn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Guiota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016-Depression, Mood, and Emotion Recognition Workshop and Challenge. *arXiv preprint arXiv:1605.01600*, 2016. 83

[218] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd*

ACM international workshop on Audio/visual emotion challenge, pages 3–10. ACM, 2013. 9, 92

[219] Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):966–979, 2012. 51, 54, 63

[220] Michel F Valstar and Maja Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(1):28–43, 2012. 60

[221] Peter Van Overschee and BL De Moor. *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media, 2012. 29, 93, 96, 97, 108, 109

[222] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM review*, 38(1):49–95, 1996. 100

[223] Vladimir Vapnik, Steven E Golowich, and Alex Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in neural information processing systems (NIPS)*. Citeseer, 1996. 82

[224] M Alex O Vasilescu and Demetri Terzopoulos. Multilinear subspace analysis of image ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, pages II–93, 2003. 21

[225] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009. 1

[226] Alessandro Vinciarelli. Capturing order in social interactions [social sciences]. *IEEE Signal Processing Magazine*, 26(5), 2009. 3

[227] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. Canal9: A database of political debates for analysis of social interactions. In *International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*, pages 1–4, 2009. 35, 36, 37

[228] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012. 2, 3, 4, 26, 27, 34, 70, 82, 86

[229] M Vrigkas, C Nikou, and IA Kakadiaris. A Review of Human Activity Recognition Methods. *Front. Robot. AI 2: 28. doi: 10.3389/frobt*, 2015. 24

[230] Andrew Wagner, John Wright, Arvind Ganesh, Zihan Zhou, Hossein Mobahi, and Yi Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012. 16, 20

[231] Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block bregman admm for nonconvex composite problems. *arXiv preprint arXiv:1505.03063*, 2015. 105

[232] Fenghui Wang, Zongben Xu, and Hong-Kun Xu. Convergence of bregman alternating direction method with multipliers for nonconvex composite problems. *arXiv preprint arXiv:1410.8625*, 2014. 105

[233] Hongcheng Wang and Narendra Ahuja. Facial expression decomposition. In *IEEE International Conference on Computer Vision*, pages 958–965, 2003. 21

[234] Su-Jing Wang, Wen-Jing Yan, Guoying Zhao, Xiaolan Fu, and Chun-Guang Zhou. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In *Workshop at the European Conference on Computer Vision*, pages 325–338. Springer, 2014. 7, 17

[235] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2016. 105

[236] Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014. 100

[237] Chia-Po Wei, Chih-Fan Chen, and Yu-Chiang Frank Wang. Robust Face Recognition With Structurally Incoherent Low-Rank Matrix Decomposition. *IEEE Transactions on Image Processing*, 23(8), 2014. 20, 22, 54, 55, 57, 60, 61, 64

[238] F Weninger, F Eyben, BW Schuller, M Mortillaro, and KR Scherer. On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Frontiers in psychology*, 4:292–292, 2012. 76

[239] Felix Weninger, Johannes Bergmann, and Björn Schuller. Introducing currennt: The munich open-source cuda recurrent neural network toolkit. *The Journal of Machine Learning Research*, 16(1):547–551, 2015. 81

[240] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, pages 597–600. Citeseer, 2008. 82, 87, 88

[241] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn W Schuller, Cate Cox, Ellen Douglas-Cowie, Roddy Cowie, et al. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Interspeech*, volume 2008, pages 597–600, 2008. 26

[242] Britta Wrede and Elizabeth Shriberg. Spotting" hot spots" in meetings: human judgments and prosodic cues. In *INTERSPEECH*, 2003. 3

[243] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 7, 17, 18, 19, 54, 57, 60, 64, 66

[244] Yangyang Xu, Wotao Yin, Zaiwen Wen, and Yin Zhang. An alternating direction algorithm for matrix completion with nonnegative factors. *Frontiers of Mathematics in China*, 7(2):365–384, 2012. 105

[245] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007. 16

[246] Allen Y Yang, Shankar S Sastry, Arvind Ganesh, and Yi Ma. Fast ℓ 1-minimization algorithms and an application in robust face recognition: A review. In *IEEE International Conference on Image Processing (ICIP)*, pages 1849–1852, 2010. 55

[247] Bo Yang and Songcan Chen. A comparative study on local binary pattern (lbp) based face recognition: Lbp histogram versus lbp image. *Neurocomputing*, 120:365–379, 2013. 16

[248] Meng Yang, Lei Zhang, Simon CK Shiu, and David Zhang. Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary. *Pattern Recognition*, 46(7):1865–1878, 2013. 20

[249] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Robust sparse coding for face recognition. In *EEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 625–632. IEEE, 2011. 20, 22

[250] Ming-Hsuan Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *fgr*, volume 2, page 215, 2002. 16

[251] Zi-Lu Ying, Zhe-Wei Wang, and Ming-Wei Huang. Facial expression recognition based on fusion of sparse representation. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, pages 457–464. Springer, 2010. 19

[252] Adams Wei Yu, Wanli Ma, Yaoliang Yu, Jaime Carbonell, and Suvrit Sra. Efficient Structured Matrix Rank Minimization. In *Advances in Neural Information Processing Systems*, pages 1350–1358, 2014. 30, 97, 99

[253] Lazaros Zafeiriou, Mihalis A Nicolaou, Stefanos Zafeiriou, Symeon Nikitidis, and Maja Pantic. Probabilistic Slow Features for Behavior Analysis. *IEEE transactions on neural networks and learning systems*, 27(5):1034–1048, 2016. 72

[254] Stefanos Zafeiriou and Maria Petrou. Sparse representations for facial expressions recognition via l 1 optimization. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 32–39, 2010. 19

[255] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. 2, 3, 26, 70, 77, 81, 85, 87

[256] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010. 100

[257] D Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *IEEE International Conference on Computer Vision (ICCV)*, pages 471–478, 2011. 62

[258] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007. 65, 66

[259] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003. 14

[260] Zhonglong Zheng, Mudan Yu, Jiong Jia, Huawen Liu, Daohong Xiang, Xiaoqiao Huang, and Jie Yang. Fisher discrimination based low rank matrix recovery for face recognition. *Pattern Recognition*, 47(11):3502–3511, 2014. 20, 22

[261] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. 44, 49, 106