Computer vision for deciphering and generating faces

Antitza Dantcheva

Computer vision for deciphering and generating faces

Antitza Dantcheva

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

Inria

HDR

HABILITATION A DIRIGER DES RECHERCHES

COMPUTER VISION FOR DECIPHERING AND GENERATING FACES

Antitza DANTCHEVA

STARS Team, Inria

10/09/2021

Jury:

Reviewers: Prof. Patrick Flynn, University of Notre Dame, USA
Prof. Nicu Sebe, University of Trento, Italy
Prof. Albert Ali Shalah, Utrecht University, Netherlands

Examiners:
Prof. Mohammed Daoudi, Université de Lille, France
Prof. Anil Jain, Michigan State University, USA

Présentée en vue de l'obtention du grade de l'HDR
En Sciences et Technologies de l'Information et de la Communication
de l'Université Côte d'Azur
et de Inria

Table of Contents

Abstract

The main volume of the work presented here is in computer vision, and it aims to holistically decipher information enciphered in human faces.

Motivation originates from the emerging importance of automated face-analysis in our evolving society, be it for security or health applications, as well as from the practicality of such systems. Specifically, we have placed emphasis on *learning representations of human faces* concerning two main domains of application: *security* and *healthcare*. While seemingly different, these applications share the core processing-competence, which has proven to be beneficial as it has brought to the fore cross-fertilization of ideas across areas. With respect to security, we have designed algorithms, which extract soft biometrics attributes such as gender, age, ethnicity, height and weight. We have aimed at mitigating bias, when estimating such attributes. Prior, we have established the impact of facial cosmetics on automated face analysis systems and have then focused on the design of methods that reduce such impact and ensure for makeup-robust face recognition.

Results related to healthcare deal with facial behavioral analysis, as well as apathy analysis of Alzheimer's disease patients. In our current work with the STARS team of INRIA and the Cognition Behaviour Technology (CoBTeK) lab of the Université Nice Sophia Antipolis (UNS), we have developed a series of spatio-temporal methods for facial behavior, emotion and expression recognition.

Most recently, we have additionally focused on Generative Adversarial Networks (GANs), which have witnessed increasing attention due to their abilities to model complex visual data distributions. We have proposed a number of novel approaches towards conditional and unconditional *generation of realistic videos* and have additionally aimed at disentangling the latent space into appearance and motion, as well as interpreting it.

This mémoire d'habilitation manuscript revisits my work after the Ph.D. and about 10 years of research from January 2012 to date. Specifically, the results in Chapters 2 and 3, provide solutions to a series of problems relating to facial analysis in security and healthcare, respectively. Chapter 4 presents results on the novel and exciting topic of *face generation*. The manuscript is followed by information on my supervision, organization, funding, as well as the complete list of my publications since my Ph.D.. In the text, citations to my own work between 2012 and 2021 appear in alpha style as in [DBB+16], whereas citations to other works appear in plain style as in [35].

Chapter 1

Introduction

The topic of human facial analysis has engaged researchers in multiple fields including computer vision, biometrics, forensics, cognitive psychology and medicine. Interest in this topic has been fueled by scientific advances that offer insight into a *person's identity*, *intent*, *attitude* as well as *health* solely based on their face images.

The methodological breakthrough behind this success, lies in the fact that *computer vision* sees the human face as a natural object and aims to perform the tasks of detection, tracking, coding, and matching from images and videos and most recently the task of generation. The task of facial recognition, for the purpose of establishing human identity, is the central focus in *biometrics*, where face images have also been used to deduce soft biometric attributes such as an individual's age, gender and ethnicity. In *forensics*, local facial features such as moles, scars, tattoos and wrinkles have been used to validate identity in one-to-one matching cases involving photos. Real-time face tracking, coupled with the use of soft biometric features, has allowed for new applications, such as continuous user monitoring and authentication in work environments. In *cognitive vision* and *social psychology*, videos and images of faces have been analyzed to infer an individual's emotional state or to detect interpersonal deception. The neuropsychological processes pertaining to how humans recognize faces have also been actively studied over several decades. From a *medical* perspective, face images and videos may also offer information about an individual's health.

Motivated by the above some tantalizing questions emerged: “What is ciphered in human faces? How can we decipher faces? How can we generate faces?”

1.1 Goals

One main goal of my research has been to *learn representations of human faces* that are instrumental in deciphering and characterizing appearance and dynamics of faces, originating from the emerging importance of automated human-analysis in our evolving society, be it for security or health applications, as well as from the practicality of such systems. Specifically, focus has been placed on designing computer vision methods concerning two main domains of application: *security* and *healthcare*. While seemingly different, these applications share the core processing-competence, e.g., learning and classifying suitable representations, which has brought to the fore cross-fertilization of ideas across areas.

A more recent goal has been the design of *generative models* able to *generate* realistic *face videos*. In particular, in video generation, we have placed emphasis on naturalism, control of generated results, as well as most recently on interpretability.

This manuscript provides highlights of my work after the Ph.D., which I classify in three *research axes*, namely Axis I. Face analysis for *security*, Axis II. Face analysis for *healthcare* and Axis III. *Face Generation*. This chapter motivates these research directions (Section 1.2) and showcases related challenges (Section 1.3). The results in the following Chapters 2 and 3 provide solutions to a series of problems relating to Axis I and Axis II, respectively. The Chapter 4 presents results on the novel and exciting topic of face generation. Chapter 5 concludes the thesis and provides future directions. In the end Chapter 6 presents details on my supervising students and postdoctoral researchers, on my scientific engagement, as well as scientific projects.

1.2 Motivation

Axis I. Security The first axis is motivated by the *rapid evolution* in *volume*, *complexity* and *utility of biometrics* [36]. The past decade has witnessed significant technical progress in the field of biometrics partly due to convolutional neural networks (CNNs) and large datasets. While biometrics has

been traditionally involved person identification by employing simple matching of well-defined single biometric traits (such as face, iris or fingerprint), we are now witnessing a transition to large-scale biometric systems that reliably determine the identity of a person and that involve many traits and many tasks that often include deducing ancillary information beyond identity [DER15]. According to market value studies [DRD+20] the interest and investment in such evolved systems is large and rapidly growing and is fueled by applications ranging from border control to smartphones; from autonomous vehicles to e-voting; from crime scene investigation to personalization of customer service. Currently, the largest biometric system is operated by the Unique Identification Authority of India, whose national ID system (Aadhaar) accommodates almost the entire Indian population of 1.25 billion enrolled subjects at the time of this writing.

Axis II. Healthcare At a time of a rapid growth in the population of elderly individuals[1] and at a time of decreased/pressed availability of human healthcare-resources, automated face analysis has the potential to offer efficient and cost-effective methods for monitoring of a number of pathologies. Facial appearance is determined by *skull morphology, muscles, innervation of blood flow, fat deposit*. Abnormal dry skin, eye bags, facial asymmetry, as well as paleness can provide cues to internal health issues. In addition, many *psychological, genetic, and physical health disorders* can be directly diagnosed by facial analysis (*e.g., emotion, appearance, motion, symmetry, color, shape*), determining cause and course of diseases, as well as assessing efficiency of treatments for diseases. As stated by McKinsey Global Institute [53], it is envisioned that such analysis can reduce costs for the US health-care industry by $300 Billion per year. We here have focused on neurogenetic disorders (ND) and specifically on Alzheimer's disease, which is characterized by a decline in mental ability - severe enough to terminate independence in daily life. ND concern over 60 million people worldwide[2][3]. Consequently there is an urgent societal need and an equally certain scientific challenge to provide new solutions for *early detection, progress analysis and long-term analysis* of ND, which would improve intervention effects and decrease the global burden of ND [26].

[1] http://www.un.org/esa/population/publications/worldageing19502050/pdf/80chapterii.pdf

[2] https://www.alz.co.uk/research/statistics

[3] https://parkinsonsnewstoday.com/parkinsons-disease-statistics/

Axis III. Generation Human video generation has attracted an increased commercial and academic attention due to numerous real-world applications, as well as due to its applicability in computer vision for data augmentation. The former has mainly to do with *entertainment*, where video generation can greatly facilitate creating dynamic scenes. Current methods [10, 147, 148] have already been able to transfer motion from target video to input avatars, allowing users to move like Michael Jackson to even lipsync his songs. We envision that video generation will pave the way for customized movies, video games and law enforcement, where generation of realistic faces can be instrumental in cases of witness descriptions, where the descriptions are the only available evidence (*e.g.*, in the absence of facial images).

Data augmentation constitutes another pertinent application for GANs motivated by the inherently hungry deep CNNs. Real data is costly and cumbersome to obtain and inherently incorporates human biases. In addition, concerns related to privacy and usage rights have hindered recent data collection. Hence, synthetic data is now seen by some as a panacea [4], as it can be provided in abundance. Further, such data can be easily modified by walks in the latent space and labeled with respect to pose, illumination, scale, age, shape, and ethnicity, allowing for AI-systems to be unbiased across populations (diverse datasets). Each manipulated sample is considered as an *augmentation* (or another 'view') of the original data. A set of companies have developed their business models with such considerations, in sectors such as finance[4], insurance[5], and healthcare[6].

GAN generated and manipulated videos have become increasingly realistic. At the same time such *deepfake* techniques are now widespread via a number of websites and phone applications, and they pose without a doubt an imminent security threat to us all. To date, deepfakes are able to mislead face recognition systems, as well as humans. Unfortunately, existing methods for deepfake detection perform poorly when tested on unseen manipulation techniques, which is the most realistic scenario in practice.

1.3 Challenges

To understand some of the challenges that we are facing, we start by recalling that classical face analysis systems acquire an image from an individual, extract a set of features (e.g., representing edges and texture

[4] https://www.facteus.com/
[5] https://mostly.ai/synthetic-data-for-insurance/
[6] https://synthea.mitre.org/

descriptors) from the image, and proceed to compare this feature set with templates in the database in order to verify a claimed identity or to determine an identity. The comparison is being performed by machine learning tools such as principal component analysis (PCA) or support vector machines (SVM). Such classical systems have been challenged by a variety of co-variates including pose, illumination, expression (PIE), age.

Convolutional neural networks (CNNs) Starting with DeepFace in 2014 [119] we have witnessed the remarkable progress of convolutional neural networks (CNNs) in face analysis, rapidly replacing classic methods. CNNs can be trained with very large datasets, aimed at learning face representations that achieve high accuracy in downstream tasks. Such representations are robust to co-variates sufficiently represented in the training data. Deviating from classical systems, CNNs are data-driven, learning jointly in an end-to-end manner discriminative features, reducing dimensionality and classification. Therefore, CNNs intrinsically learn intra-class variations from training data and have overcome several related challenges, for example with respect to in-the-wild face recognition.

1.3.1 Axis I. Security

While CNNs have solved a series of challenges related to vulnerabilities in face recognition, there is a plethora of challenges that remains unsolved. Such remaining challenges include biasness, privacy preservation, anti-spoofing methods that generalize across different types of attacks, scalability, as well as the quantification of face-uniqueness and permanence thereof [35, 92].

Representation. Identifying a suitable representation scheme for a given biometric trait is essential. Such a representation should retain discriminative information that is distinctive to a person, and at the same time remain invariant to intra-subject variations.

Reliability. The next fundamental step in a face-based biometric algorithm is given a representation scheme, to designing a robust matcher. The desired matching algorithm must model the variations in the features belonging to the same individual, while accounting for variations between features of different individuals.

Face Analysis *in the Wild*. Face based biometrics has been challenged by Real-world data, referred to as “in the wild” that includes for example low-resolution images incorporating variations in pose,

illumination and expression, which is frequent in surveillance applications. Such challenges have been addressed in a number of works [49].

Biasness. An algorithm is considered to be biased, given that significant differences in its operation can be observed for different demographic groups (e.g., gender or ethnicity), thereby privileging and disadvantaging certain demographic groups. Biasness concerns all areas in artificial intelligence (AI), as it is rooted in factors such as skewed training data and sensors. In addition, humans are known to exhibit a broad range of biases [17, 73], which are transferred to AI. Biasness in face based biometrics [22] has sparked immense interest related to a set of concerns, which have raised questions with respect to algorithmic design, interactions with and use of biometric systems. In this context, we have introduced a multi-task algorithm aimed at mitigating biasness in esitmating gender, age and ethnicity [DDB18]. Recently, we revisited biasness in biometrics in an overview article [DRD$^+$20].

Privacy. Privacy preservation in biometrics has to do with respect and confidentiality of an individual's personal information or data, as well as transparency surrounding its use and storage. The General Data Protection Regulation (GDPR) for European Member States addresses biometric data and represents a significant step forward for data protection and privacy with a real international impact. Algorithmic data protection strategies include (a) the storage of data in a *de-centralized manner*, (b) the *elimination for storage of any identity* related information in the database, as well as (c) the transformation of raw images and videos into "*privacy - securing*" biometric templates. We note that similar privacy concerns have been discovered also in the context of GANs, and specifically identity leakage in image generation [122].

Presentation attacks. With recent advances in deep CNNs, biometric systems have become remarkably accurate, fast, and more resilient to environmental and user co-variates. However, biometric systems can be attacked and bypassed by "spoofs" or presentation attacks. Examples of presentation attacks include photographs, videos, masks, as well as facial makeup. Such attacks pose serious challenges in face recognition. This is why identification and mitigation of such vulnerabilities has received substantial attention.

In 2012 we were the first to establish the *impact of facial cosmetics* on automated face analysis systems [CDSR17, CDR14, DCR12b, RDB19] and have then focused on the design of methods that

reduce such impact [CDR13, CDR16]. Specifically, we have quantified such an impact on a number of face recognition, as well as age and gender-estimation systems, and have then designed methods for detecting facial cosmetics (based on appearance, texture and color features) and have proceeded to design an algorithm based on an ensemble of patch-based subspaces for makeup-robust face recognition. This work was awarded with the Best Tabula Rasa Spoofing Attack Award 2013, the Best Poster Award at the IAPR International Conference on Biometrics (ICB) 2013 and with the Best Paper Award (Runner up) at the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA) 2017. Most recent presentation attacks relate to GAN-generated attacks such as adversarial attacks, as well as deepfakes in the realm of behavior based authentication. We address deepfake detection in Section 4.4.

Scalability. The success of India's Aadhaar national ID system (based on fingerprints, face and iris) have proven that biometric recognition systems are highly scalable [104]. Despite the success, very few evaluations exist in the literature to show how biometric recognition systems operate at a scale the size of Aadhaar (an average of 35M biometric authentications per day[7]).

Related to scalability, soft biometrics are instrumental in search space reduction of large datasets, as they are able to prune or pre-filter such datasets. We have explored novel soft biometrics such as weight and height from facial images, as well as novel modalities such as smile-dynamics for gender estimation in Section 2.1.1.

Face uniqueness. The knowledge of distinctiveness of the human face is incomplete and often relegated to anecdotal interpretation of error rates rather than a systematic exploration of the biology of the characteristics [13, 38, 92]. Hence, we lack an estimate for the upper bound of the amount of discriminatory information contained in a face. We recently attempted to shed some light on facial uniqueness [BHBD21].

1.3.2 Axis II. Healthcare

Naturally, automated face-based security and health analysis have many intersections and hence share similar core algorithmic challenges. Both applications necessitate carefully designed face representation schemes, demand high levels of reliability and robustness, require the analysis of unconstrained data, as

[7] `https://uidai.gov.in/aadhaardashboard/authtrend.php`

well as privacy preservation mechanisms to secure the involved sensible data. These similarities motivate our joint exposition of these problems. However, we note that named challenges have particularities, when encountered in Axis I or II. Further contributing to particularities, we have that while in Axis I, we predominantly analyzed images, in Axis II, we analyzed *spatio-temperal* features of faces. Hence, differences in the challenge 'identification of representation' stem mainly from the additional time-dimension. Linked to this, addressing the time-dimension is a pertinent long term challenge of Axis II. At the same time, biometrics has already overcome some challenges that we still encounter in healthcare, such as data scarcity. Moreover, we might envision that presentation attacks might emerge as challenge in Axis II, when healthcare monitoring systems become ubiquitous. It is likely that some patients might aim to mislead a healthcare monitoring system, in order to avoid monitoring or to test related boundaries. Finally, while biasness is undesirable in both Axes, related personalization and specificity in Axis II is sought and might be a path forward.

Motivated by the above, we proceed to elaborate on the challenges, we have encompassed in the context of *automated healthcare*.

Fine-Grained Representation. In face analysis, a key goal has been to obtain a discriminative appearance representation, which allows for precise classification of aspects such as expressions, mental states or neurogenetic diseases. To obtain such representation, one can employ various strategies that include (i) global feature maps (pertaining the full detected face), (ii) key local features (e.g., facial landmarks), (iii) attention mechanisms (placing the focus of the network on relevant key spatial, temporal regions or channels), as well as (iv) partitioning (where for example, the last convolutional layer of a network is partitioned into a number of horizontal stripes). Some of the limitations of the above strategies are described below. For the first strategy (i), a main limitation is that the errors incurred due to the massiveness of the contained information, while for the second strategy (ii) a key limitation is that the additional errors pertaining to landmark-detection are included in the extraction. In strategy (iii) infrequent (but potentially pertinent) information is removed, whereas in strategy (iv) key facial features might occur in different horizontal stripes. Our work on heartrate estimation from RGB-videos (Section 3.4) necessitated a fine-grained representation, which we designed with the means of channel and spatial-temporal attention mechanisms.

Face Analysis *in the Wild*. Towards developing algorithms that properly capture the right facial appearance, as well as behavior, we have worked with real-world data that encompasses continuous facial pose variations, and real expressions of different intensity. This has been particularly challenging, especially in settings that involve human analysis. An example of such a challenge has to do with the fact that real-world expressions of older adults are often subtle (modulated by saggy muscles or by pain), and can be occluded due to unfavorable poses or hands. Facial analysis systems face additional challenges with large intra- and inter-class variability of patients, as well as pose variations, as such systems are generally trained with constrained data and sometimes often with posed and rather "exaggerated" expressions. Moreover, noisy acquisition challenges automated healthcare methods, with a variety of ambient illuminations, camera movements and artefacts. Due to such factors, real-world facial analysis is still an open challenge in computer vision. We have faced above challenges in a set of works of Chapter 3 and more specifically in our works on music therapy (Section 3.2, as well as in the series of works on apathy classification (Section 3.3)).

Data Scarcity. A further challenge concerns healthcare data, where currently it is often the case that only limited data (e.g., containing a small number of patients) is available for analysis. Further, especially in a multimodal setting, missing input data of a modality (e.g., for a short period of time) poses a challenge. We proposed in the context of expression recognition weakly supervised (Section 3.1.1) and semi-supervised (Section 3.1.2) approaches. Along these lines, we can also envision to introduce active, few shot and cooperative learning. In addition, given the limited data, we built our work of apathy classification firstly in a hand-crafted manner, in order to rationalize learned features and avoid overfitting (Section 3.3). Finally, effective *data augmentation* is beneficial in the setting of limited available healthcare data. We explored different data augmentation methods in our work on heartrate estimation (Section 3.4). Further, we employed a similar attention mechanism in Axis III, Section 4.2.3 accounting for improved video quality in generated videos.

Addressing the Time-Dimension in DNNs. While discriminative models for perception tasks (such as object detection) have witnessed substantial progress with the resurgence of deep CNNs, effective methods for incorporation of temporal information into CNNs have proven more challenging [130, 143] and are still being actively explored. Such approaches are important and play a fundamental role in

modeling humans. Existing 3D convolutional networks such as I3D [9] and two-stream convolutional neural networks [18] suffer from (i) being limited to processing only a limited amount of frames at a time and hence lack the ability to model long-term dependencies, as well as suffer from (ii) limitations that enforce the processing of only fixed length videos. We envision that finding algorithms, which extract *high-level temporal features* would allow for processing of long videos, which is instrumental in hospital monitoring.

Privacy. Privacy here refers to assuring that healthcare data collected from a patient is not used to deduce any additional type of information about the individual (e.g., identity). In this context, we are anonymizing the data and are exploring a federated learning approach, in order to avoid for data to leave the hospital.

Deployment. While the majority of proposed algorithms have been evaluated and validated in constrained lab settings, the transition to *real life* deployment, and potentially real time analysis remains an open challenge. Deployment constraints such as *latency and interoperability* can only be tested during such deployment.

Personalization. A long term goal will be to design approaches for *intelligent personalized* monitoring and diagnosis based on *adaptive classification*. Such models will be trained on data associated to a single patient, individually. This can be highly instrumental for accounting for large inter-class variations. In an experiment of our music therapy work, we showed that such personalization can increase the accuracy of behavior estimation.

In addition to the the above highlighted open challenges, which are directly related to AI and computer vision techniques, there are few more challenges related to *data acquisition in unconstrained environment, publicly availability of datasets, as well as precise annotation*.

1.3.3 Axis III. Generation

GANs often suffer from following limitations. Firstly, model parameters may oscillate, destabilize and *fail to converge*. Further, in *mode collapse* the generator collapses, which leads to limited varieties of samples. In case of a too strong discriminator, the generator gradient can vanish and fail to learn, referred to as *diminished gradient*. An unbalance between generator and discriminator causes *overfitting*. We also

have that GANs are highly sensitive to the hyperparameter selections. Named challenges have been of concern in our work on image generation (Section 4.1).

Related to *video generation* we have identified a set of challenges that we proceed to list. Some of the main challenges of video generation include GAN design, and associated video representation. We elaborate on these challenges below.

GAN design. While video generation can be considered as the inverse problem of video understanding [WDB17], it constitutes a far more challenging problem due to its extra requirements such as stable training, high visual-quality and interpretability. While two widely-used architectures in video understanding, namely 3D ConvNets and 2D ConvNet+RNN, have been explored reversibly for generation, both entail notable limitations such as large complexity with more training parameters, which may render models difficult to be optimized (3D ConvNets), as well as unstable training owing to gradient vanishing and gradient explosion (RNN). *Discriminators* ensure that generated videos encompass *visual-quality*, as well as *temporal consistency*. For the latter, the transition between consecutive frames should be smooth, which is highly challenging. In Chapter 4, we present a two-stream discriminator, which combines 3D ConvNets and 2D ConvNets to learn spatio-temporal distribution (Section 4.2.3). In Section 4.2.4, we introduce a novel temporal pyramid discriminator equipped with only 2D ConvNets.

Entanglement of appearance and motion in videos. Appearance and motion are two major factors in videos. In the absence of additional information such as human keypoints or optical flow, learning to disentangle such factors is challenging, as it requires building specific model components to represent both factors, respectively. Due to lack of explicit formulation of the disentanglement of motion and appearance, designing model components, which disentangle these two factors remains challenging. In Section 4.2.3 and Section 4.2.4, we introduce two complementary disentangling approaches, as well as comparison thereof.

Interpretability. Deep neural networks have been widely used as black-boxes, and GANs are no exception. Due to the large amount of parameters, it is difficult to identify the types of knowledge GANs

have learned. In addition, given that features such as textures, concepts, semantics and objects are represented in a hierarchical manner in GANs [5], discovering and locating information of interest becomes difficult. To interpret different features, novel methods are necessitated. In the context of image generation for example, in an attempt to interpret attributes (e.g., gender, age) in StyleGANs [42,43], pretrained classifiers were used to provide scores for generated samples [105]. At the same time landmark detectors were necessitated for interpretation of pose. Deviating from that, we aimed at interpreting motion in video GANs. This is discussed in Chapter 4.2.4.

Evaluation. We note that lack of effective evaluation metrics is a major challenge in current GAN research. Often generated images and videos are evaluated for how realistic they are and in particular by user studies, which is though inefficient and time-consuming. Towards evaluating GANs in an objective manner, two quantitative evaluation metrics, namely Inception Score (IS) [99] and Fréchet Inception Distance (FID) [28], have been proposed. IS and FID use statistical methods, that rely on features extracted from pretrained models on large-scale datasets, in order to measure the distance between real and generated distributions. Due to large variability in space and time, evaluation in video generation remains very challenging.

Deepfakes. The access to large-scale public datasets, jointly with the fast progress of GANs, have led to the generation of very realistic generated images and videos, which entail corresponding implications towards society in this era of fake news. In our effort to detect manipulated images and videos, i.e., *deepfakes* we faced two fundamental challenges. Firstly we have a "cat-and-mouse-game", honing deepfake generation and deepfake detection, one against the other. In improving the detection mechanism, generation can be improved accordingly, which in turn can be beneficial in improving the detector. This results in this game, which can never be won. Secondly, we have that deep models are highly domain-specific and likely yield big performance degradation in cross-domain deployments, especially with large train-test domain gap. The second challenge indicates that detectors trained on known manipulation techniques generalize poorly to tampering methods outside of the training set, which we show in a very recent work, in which we compare 2D and 3D deepfake detection algorithms (Section 2).

Chapter 2

Axis I. Facial Analysis for Security

Biometrics is the science of *recognizing* individuals *based on* their physical, behavioral, and physiological *characteristics* [33, 35, 36, 39, 91] such as fingerprint [39], face [37], and iris [14]. Biometrics is aimed at ensuring an accurate person recognition, which ensures high recognition rates and low error rates (e.g., False Reject Rate (FRR) and False Accept Rate (FAR)). Despite fingerprint and iris entailing generally higher recognition accuracy than face, face has catapulted itself as the most compelling modality commercially for its user-friendliness. However, its susceptibility to change due to factors such as expression or aging, have brought to the fore a number of challenges, see Section 1.3.

Automating face recognition dates back to 1964, when Woody Bledsoe, Helen Chan Wolf, and Charles Bisson [6] firstly studied the problem. This laid the foundation for *geometric-based* face recognition, which is based on distances between pre-defined facial landmarks. In contrast, the first approach considering the face *holistically*, namely Eigenfaces dates back to 1991 and was proposed by Turk and Pentland [129]. The novelty herein was a compact face representation, obtained by mapping the high-dimensional face image into a lower dimensional sub-space, which is aimed at reducing the intra-class distances on the features and increasing the inter-class distances. In 2004 Viola and Jones revolutionized the field by introducing a real-time face detector based on Haar filters [134]. A large number of hand-crafted algorithms such as local binary pattern (LBP), Gabor Wavelet, scale invariant feature transform (SIFT), histogram of oriented gradients (HoG), and also sparsity-based representations continued the progress in face recognition. Notably, the progress of sensors and cameras, e.g., the first digital cameras

in the early 1990s, further pushed forward face-based biometrics by allowing for 3D face recognition or face recognition beyond the visible spectrum. The most recent and prominent milestone was set by Taigman et al. with DeepFace [119], entailing a CNN for face recognition, which notably approached human performance on the unconstrained condition for the first time ever (DeepFace 97.35% vs. Human: 97.53%). Interestingly, Wang and Deng [145] drew attention that starting from the first few layers of a deep neural network, patterns similar to the Gabor Wavelets are encoded (i.e., oriented edges with different scales). Following layers learn increasingly complex features such as 'high-bridged nose' and 'big eyes', whereas the last few layers incorporate facial attributes such as smile and eye color. Notable CNN-approaches include VGGface [74], COCO algorithm [51], SphereFace [50], CosFace [140], ArcFace [16]. Most recently, algorithms are concerned with bias-mitigation [20], as well as with privacy, introducing federated learning in the context of face recognition, which allows for data not to be transferred, nor stored [1]. Models after 2017 have predominantly focused on developing new loss functions for more discriminative feature learning. The remarkable progress of CNNs for face-based biometrics [59] is on account of large-scale datasets such as MegaFace, incorporating more than 690.000 identities, or Google's private dataset including over 10 Million individuals. This progress has enabled current biometric systems to reliably recognize cooperative individuals in controlled environments [38].

Given the increasing reliability of biometrics, in the last decade researcher have studied forgery in biometrics [19, 54, 81], as well as methods to counter-act and distinguish fake biometric features from authentic ones [15].

The National Institute of Standards and Technology (NIST) has been organizing frequent evaluations and challenges, reporting on the progress of face recognition accuracy[1]. The interest and investment into biometric technologies is large and rapidly growing according to various market value studies [DRD+20].

While biometric data is typically used to recognize individuals, it is possible to deduce other types of attributes of an individual from the same data. For example, attributes such as age, gender, ethnicity, height, hair color and eye color can be deduced from data collected for biometric recognition purposes. These additionally deduced attributes, while not necessarily unique to an individual, can be used in a variety of applications. For example they can be used in conjunction with primary biometric traits in

[1] https://www.nist.gov/biometrics

order to improve or expedite recognition performance. It is perhaps this latter application that has led to these attributes being referred to as *soft biometrics* [DER15]. In this context, soft biometrics can be traced back to Bertillon [86]. We note that soft biometrics are instrumental in bridging the semantic gap between human and machine descriptions of biometric data.

In this context, we present here work on (a) facial *behavior* in estimating soft biometrics (Section 2.1.1), (b) extracting novel soft biometrics such as height and weight *from face* (Section 2.1.2), (c) algorithmic *bias* in biometric systems w.r.t. soft biometrics (Section 2.3), as well as (d) vulnerabilities in biometric systems (Section 2.2). In (a) and (b) we have designed algorithms aiming at exploring the relatively novel modality - behavior, as well as the novel for the face height, weight and BMI. In (c) one notable result comprises a multi-task algorithm for analyzing gender, age and ethnicity in an unbiased manner that won a European Conference of Computer Vision 2018 challenge on unbiased face analysis. In (d) we established the impact of facial cosmetics on automated face analysis systems and have then focused on the design of methods that reduce such impact and more specifically designed an ensemble algorithm of patch-based subspaces for makeup-robust face recognition. This work was awarded with the Best Tabula Rasa Spoofing Attack Award 2013, the Best Poster Award at the IAPR International Conference on Biometrics (ICB) 2013 and with the Best Paper Award (Runner up) at the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA) 2017.

2.1 Soft Biometrics or 'What else is there in your biometric data'

Soft biometrics [DER15,34] constitute personal attributes, such as gender, age, ethnicity, hair color, body weight, which find applications that include video surveillance or human-computer interaction. Such attributes are particularly useful in bridging the *semantic* gap between human and machine descriptions of the biometric data. Soft biometrics are seen as a crucial stepping stone towards the evolution of human analysis into a science that reflects the dynamic nature and uncertainty that exists in decision making of real world pertinent problems.

2.1.1 Gender estimation based on smile dynamics

One pertinent soft biometric attribute is gender, for related numerous applications, including video surveillance, human-computer interaction, anonymous customized advertisement, and image retrieval. Most commonly, the underlying algorithms analyzed the facial appearance for clues of gender. We proposed novel methods for gender estimation [BDB16, DB16], which exploited dynamic features gleaned from smiles and we proceed to show that: (a) facial dynamics incorporate clues for gender dimorphism and (b) while for adult individuals appearance features are more accurate than dynamic features, for subjects under 18 years facial dynamics can outperform appearance features. In addition, we fused proposed dynamics-based approach with state-of-the-art appearance-based algorithms, predominantly improving performance of the latter. Results showed that smile-dynamics include pertinent and complementary to appearance gender information.

2.1.2 Show me your face and I will tell you your height, weight and body mass index

Staying in the realm of soft biometrics, we explored novel attributes such as body height, weight, as well as the associated and composite body mass index (BMI). Such attributes are of pertinence in biometrics, as well as in healthcare. Previous work on automated estimation of height, weight and BMI had predominantly focused on 2D and 3D full-*body* images and videos. Little attention had been given to the use of face for estimating body height and weight. Motivated by this, we explored the possibility of estimating height, weight and BMI from single-shot facial images by proposing a regression method based on the 50-layers ResNet-architecture [DBB18]. In addition, we introduced a novel dataset consisting of 1026 subjects and showed results, which suggested that facial images contain discriminatory information pertaining to height, weight and BMI, comparable to that of body-images and videos. In conclusion, we performed a gender-based analysis of the prediction of height, weight and BMI.

2.2 Vulnerabilities in Biometrics

A contribution that I am particularly proud of relates to our pioneering work on the topic of facial cosmetics and automated facial analysis systems, which we proposed during my 2 PostDocs in the USA. Specifically, we showed that automated face analysis not only can be substantially impacted by the application of facial makeup but also can be spoofed, surpassing all current spoofing countermeasures. The problem has been of particular interest, since (a) face recognition systems have been increasingly deployed in security and commercial applications and (b) facial cosmetics are a simple, non-permanent, cost-efficient method to substantially alter facial appearance. Hence facial cosmetics has the potential to compromise the accuracy of biometric systems. We proposed research investigating this and I was the PI of the project "Impact of Cosmetics on the Performance and Security of Face Recognition" from National Science Foundation (NSF) Industry / University Cooperative Research Center CITeR (Center for Identification Technology Research with affiliates such as the Federal Bureau of Investigation). Results of this work were published a series of successful, well cited publications covering following topics.

2.2.1 Establishing the impact of facial cosmetics on automated face analysis algorithms

We verified the significant impact of facial cosmetics by evaluating the matching accuracy of multiple face recognition algorithms and the accuracy of multiple age and gender-estimation algorithms on four different datasets [CDSR17, CDR14, DCR12b].

2.2.2 Mitigation of impact

Motivated by the above, we designed a *makeup detection-algorithm* [CDR13] extracting a feature vector that captures the shape, texture and color characteristics of the input face, and employs a classifier to determine the presence or absence of makeup. Further, we proposed an adaptive pre-processing scheme that exploits knowledge of the presence or absence of facial makeup to improve the matching accuracy of a face matcher.

Towards finding an *algorithm for cosmetics-robust face recognition [CDR16]* we introduced an ensemble learning scheme to generate multiple common semi-random subspaces for before-makeup and after-makeup samples, instead of two separate subspaces. We illustrate the scheme in Figure 2.1. In random subspace methods, a set of multiple low-dimensional subspaces are generated by randomly sampling feature vectors in the original high-dimensional space. Specifically, multiple texture descriptors were used to describe a face-patch. A combination of sparse and collaborative classifiers were used in these subspaces. A random subspace method can be used to generate multiple common subspaces, where each subspace contains a small portion of discriminative information pertaining to the identity. At the same time, by randomly selecting different patches as the input to each subspace-based classifier, the overfitting issue is avoided.

2.3 Bias in Biometrics

Biometric systems have become ubiquitous in personal, commercial, and governmental identity management applications. Both cooperative (e.g., access control) and non-cooperative (e.g., surveillance and forensics) systems have benefited from biometrics. Recently, however, automated decision systems (including biometrics) faced public and academic concerns related to systemic bias. Most prominently, face recognition algorithms have been labelled as "racist" or "biased" by the media, non-governmental organisations, as well as researchers.

We revisited algorithmic bias in the context of biometrics [DRD+20], providing a comprehensive survey of existing literature on biometric bias estimation and mitigation, as well as discussing pertinent technical and social matters, and outlining remaining challenges.

In addition, we explored joint classification of gender, age and race [DDB18], where we proposed a Multi-Task Convolution Neural Network (MTCNN) employing joint dynamic loss weight adjustment towards classification of named soft biometrics, as well as towards mitigation of soft biometrics related bias. The proposed algorithm achieved promising results on the UTKFace and the Bias Estimation in Face Analytics (BEFA) datasets and was ranked first in the the BEFA Challenge of the European Conference of Computer Vision (ECCV) 2018.

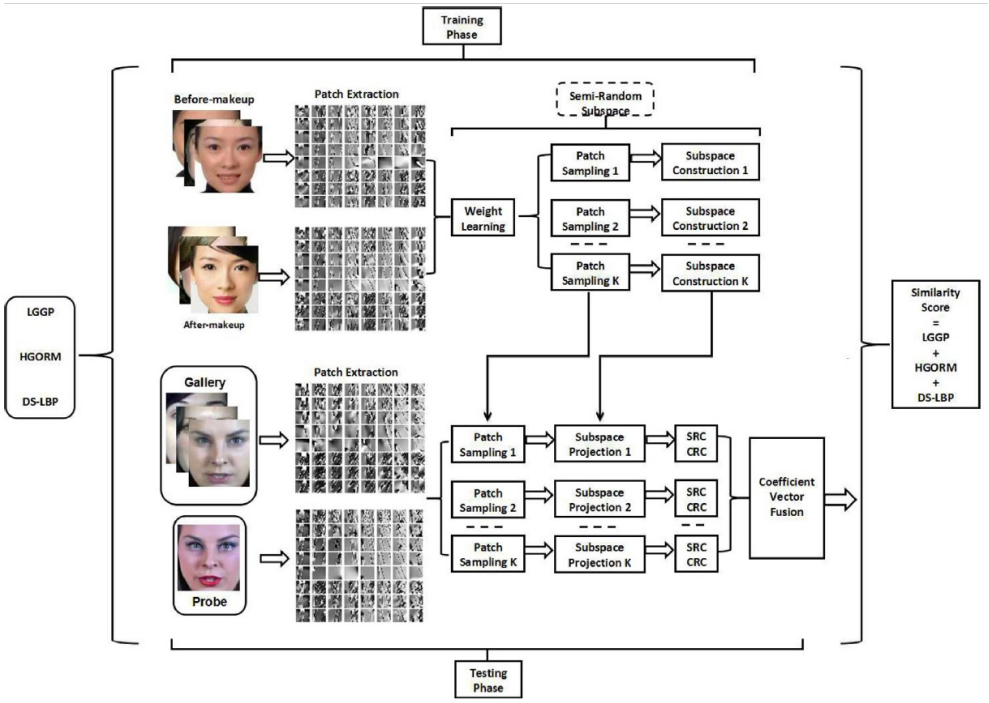

Figure 2.1: Proposed framework for matching after-makeup images with before-makeup images. During the training phase, for each feature descriptor, a pool of patches is extracted, followed by weight learning, patch sampling and random subspace construction. In the testing phase, patches from an input image are projected onto the learned random subspace. A combination of sparse representation based classification (SRC) and collaborative representation based classification (CRC) classifiers are used to compare feature vectors in these subspaces and generate a match score. This process is repeated for each descriptor and the matching scores corresponding to individual feature descriptors are fused to generate the final similarity score.

We note that DeepFakes, representing maliciously modified images/videos is a recent challenge in biometrics, which we elaborate on in Chapter 4.

Selected Publications

1. P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, C. Busch. Demographic Bias in Biometrics: A Survey on an Emerging Challenge, *IEEE Transactions on Technology and Society (T-TS)*, vol. 1, no. 2, pp. 89–103, 2020.

   arXiv preprint arXiv:2003.02488, 2020.

2. A. Dantcheva and F. Brémond. Gender estimation based on smile-dynamics. *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 12, no. 3, pp. 719-729, March 2017.

3. C. Chen, A. Dantcheva, A. Ross. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, vol. 32, pp. 80-92, November 2016.

4. A. Dantcheva, P. Elia, A. Ross. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 11, no. 3, pp. 441-467, March 2016.

5. D. Anghelone, C. Chen, A. Ross, P. Faure, A. Dantcheva. Explainable Thermal to Visible Face Recognition Using Latent-Guided Generative Adversarial Network, In *FG'21, 16th IEEE International Conference on Automatic Face and Gesture Recognition*, December 15 - 18, 2021, Jodhpur, India (Virtual Event).

6. A. Das, A. Dantcheva, F. Brémond Mitigating bias in gender, age, and ethnicity: a multi-task convolution neural network approach In *ECCVW'18, International Workshop on Bias Estimation in Face Analytics (BEFA) in conjunction with the European Conference on Computer Vision*, September 9, 2018, Munich, Germany.

7. A. Das, C. Galdi, H. Han, R. Ramachandra, J.-L. Dugelay, A. Dantcheva. Recent advances in biometric technology for mobile devices In *BTAS'18, IEEE International Conference on Biometrics: Theory, Applications and Systems*, October 22-25, 2018, Los Angeles, USA.

8. P. Bilinski, A. Dantcheva, F. Brémond. Show me your face and I will tell you your height, weight and body mass index. In *IAPR ICPR'18, International Conference on Pattern Recognition*, August 20-24, 2018, Beijing, China.

9. C. Chen, A. Dantcheva, A. Ross. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, vol. 32, pp. 80-92, November 2016.

10. C. Chen, A. Dantcheva, T. Swearingen, A. Ross. Spoofing Faces Using Makeup: An Investigative Study. In *ISBA'17, 3rd IEEE International Conference on Identity, Security and Behavior Analysis*, February 2017, New Delhi, India.

11. C. Chen, A. Dantcheva, A. Ross. Impact of facial cosmetics on automatic gender and age estimation algorithms In *VISAPP'14, 9th International Conference on Computer Vision Theory and Applications*, January 5-8, 2014, Lisbon, Portugal.

12. C. Chen, A. Dantcheva, A. Ross. Automatic facial makeup detection with application in face recognition In *IAPR ICB'13, 6th IAPR International Conference on Biometrics*, June 4-7, 2013, Madrid, Spain.

13. A. Dantcheva, C. Chen, A. Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *IEEE BTAS'12, 5th IEEE International Conference on Biometrics: Theory, Applications and Systems*, September 23-26, 2012, Washington DC, USA.

Chapter 3

Axis II. Facial Analysis for Healthcare

It is interesting to note that automated face-based security and health analysis have naturally many intersections and have effectively similar underlying principles. Specifically, both applications demand high levels of reliability and robustness and hence coerce carefully designed face representation schemes. Both applications require the analysis of unconstrained data, captured in an non-intrusive and efficient manner. Both security and healthcare share high societal impact and require privacy preservation mechanisms to secure the involved sensible data. These similarities motivate our joint exposition of these problems. We believe that these applications - despite their differences in interpretation of the common extracted facial information - share similar core algorithmic challenges and similar core algorithmic solutions. It is this joint analysis and empirical experimentation that may reveal deeper connections between seemingly unrelated challenges. One example has to do with our result suggesting that smile-dynamics encode the gender of an individual [DB16,BDB16]. Having this in mind, when encoding facial expressions in a framework analyzing apathy, we built separate gender-specific models, which showed beneficial in the accuracy.

In addition, joint exposition of security and healthcare has allowed for cross-fertilization of ideas across areas, where we proposed an algorithm for gender estimation [BDB16] and were able to adopt the algorithm in the setting of facial behavior recognition in Alzheimer's disease patients [DBN$^+$17a].

Face-based health diagnosis and treatments have advanced rapidly, associated to the rapid progress in computer vision and machine learning, and carry the premise to constitute a fundamental part of future

assisted living frameworks [25, 98]. Specific applications include *clinical diagnosis*, where symptoms or specific health conditions are evaluated, *prognosis*, referring to the monitoring of a patient for a specific health condition and predicting how this health condition will evolve in the future, *assertive techniques* providing assistance directly to patients, *personal level health monitoring* and *healthcare management*.

During the past decade, notable computer vision based approaches have focused on depression detection [41, 44, 146], pain detection [7], assessment of neurological disorder [11, 87, 127], as well as phenotypes of genetic disorders [23, 24]. In the context of healthcare also stress [151] and affect [47, 62, 71, 103, 114] have been studied.

Motivated by the medical need, as well as by the great progress in CNNs, the topic of automated healthcare monitoring has sparked high interest in the community and consequently several scientific events, such as special sessions (Kinect-based Kinematic Data Analysis and Evaluation for Clinical Applications) and scientific event workshops (Face and Gesture Analysis for Health Informatics) and special session (Human Heath Monitoring based On Computer Vision) have been organized associated to the IEEE FG 2018[1], IEEE FG 2019[2], IEEE FG 2020[3], IEEE FG 2021[4], respectively. Previous overview articles such as [107] and [120] have revisited technical challenges pertaining to automated healthcare monitoring, witnessing an increased number of publications associated to this research area.

This scientific attention has been further fueled by the prevalent commercial deployment of automated health-care systems, by dedicated projects based on health monitoring such as Patient@home[5] and Smarthome [52].

In the context of healthcare, we have sought to predict the needs of patients with Alzheimer's disease (AD), in order to better address them. At a time of an increased elderly population growth[6] and decreased availability of human healthcare-resources, *computer vision aided face analysis* has the potential to offer efficient and cost-effective methods for monitoring of AD-patients.

[1] https://fg2018.cse.sc.edu
[2] http://fg2019.org/participate/special-sessions/hhmbcv/
[3] https://fg2020.sunai.uoc.edu/program/
[4] http://iab-rubric.org/fg2021/session.html
[5] http://www.en.patientathome.dk/projects/computer-vision-for-in-home-medical-diagnosis-and-monitoring.aspx
[6] http://www.un.org/esa/population/publications/worldageing19502050/pdf/80chapterii.pdf

In this context we proposed (a) novel computer vision and machine learning methods for automated *recognition of facial expressions and dynamics in severely demented patients* (Section 3.1). Recognizing behavior and expression of AD-patients is essential, because AD patients usually lose a substantial amount of their cognitive capacity, and some even their verbal communication ability (*e.g.*, aphasia). This means that clinicians often need to interpret the patients' verbal and non-verbal messages; these messages can be important, and they must be properly assessed because they might convey discomfort or pain. Such assessment classically requires the patients' presence in a clinic, and it usually involves time consuming examinations involving medical personnel. Thus, (non-automated) expression assessment is costly and logistically inconvenient, and as a result can severely hinder large-scale monitoring. Our work constitutes a large improvement over the majority of previous approaches on emotion and expression recognition, which have focused on posed expressions or spontaneous expressions in highly constrained settings [2] and which limits the validity and generalizability of their models to more complex concepts, such as mental health.

Building upon proposed facial emotion classification methods, we proposed (b) models, which explore complex concepts of mental states by employing behavior algorithms (Section 3.3). In particular, we introduced novel machine learning frameworks to classify apathetic and non-apathetic patients based on analysis of facial dynamics, entailing both *emotion and facial movement*. In addition, we explored regression models to predict the clinical scores related to the *mini-mental state examination (MMSE)* and the *neuropsychiatric apathy inventory NPI* using motion and emotion features, and successfully augmented the accuracy of apathy classification. We note that assessment of mental health is a significant rising problem, with reports showing an astonishing 46% of subjects meet criteria for symptoms such as anxiety disorders, mood disorders (including depression and bipolar disorders, impulse-control disorders, and substance use disorders (including alcohol and drug abuse)) at least once in their lives [12]. Automatic assessment of mental health on a large scale is a new big challenge, which I am interested in. A further work in this Axis includes (c) a face-based algorithm for remote *heartrate detection* [NZH+19], capturing remote photoplethysmography (rPPG), which constitutes a pulse triggered perceivable chromatic variation (Section 3.4).

We note that, despite these challenges, it is imperative to work with such data, as it is representative for current (vast amount of) video-documentation of medical doctors, requiring automated analysis.

3.1 Emotion Analysis

The *universal hypothesis* suggests that the six basic emotions - *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* - are being expressed by similar facial expressions by all humans. Recognizing such expressions remains highly challenging, predominantly due to (a) lack of sufficient data, (b) subtle emotion intensity, (c) subjective and inconsistent annotation, as well as due to (d) *in-the-wild* data containing variations in pose, intensity, and occlusion.

3.1.1 A Weakly Supervised Learning Technique for Classifying Facial Expressions

While existing datasets support the universal hypothesis and comprise of images and videos with discrete disjoint labels of profound emotions, real-life data contains jointly occurring emotions and expressions of different intensities. Models, which are trained using categorical one-hot vectors often over-fit and fail to recognize low or moderate expression intensities. Motivated by the above, we aimed to tackle challenge (a) described in Section 3.1, namely lack of sufficient data by a weakly supervised learning technique for expression classification [HDB19], which leveraged the information of not annotated data. Crucial in our approach (see Figure 3.1) was that we firstly trained a CNN with label smoothing in a supervised manner and proceeded to tune the CNN-weights with both labelled and unlabelled data simultaneously. Experiments on four datasets demonstrated large performance gains in cross-database performance, as well as show that the proposed method achieves to learn different expression intensities (above described challenge (b) in Section 3.1), even when trained with categorical samples.

3.1.2 Semi-supervised Emotion Recognition using Inconsistently Annotated Data

Seeking to address above described challenges in a unified framework, we proposed a self-training based semi-supervised CNN-framework [HDB20b], which directly addressed the challenge of (a) limited data

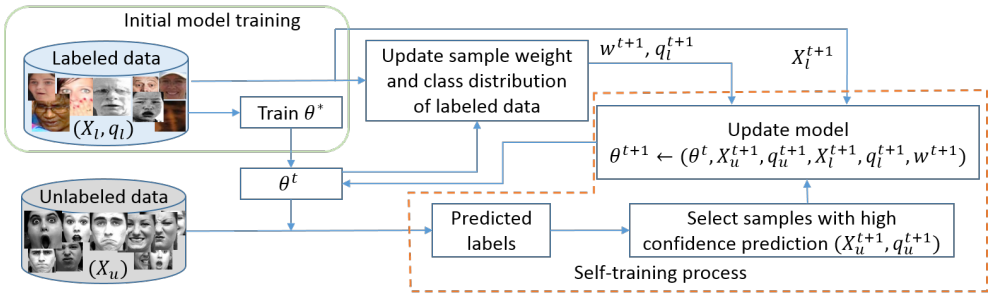

Figure 3.1: Workflow of the proposed expression recognition method.

by leveraging information from unannotated samples. Our method used 'successive label smoothing' to adapt to the subtle expressions and improve the model performance for (b) low-intensity expression samples. Further, we addressed challenge (c) inconsistent annotations by assigning sample weights during loss computation, thereby ignoring the effect of incorrect ground-truth. We observed significant performance improvement in *in-the-wild* datasets by leveraging the information from the *in-the-lab* datasets, related to challenge (d). Associated to that, experiments on four publicly available datasets demonstrated large performance gains in cross-database performance, as well as showcased that the proposed method achieves to learn different expression intensities, even when trained with categorical samples.

3.2 Music Therapy

We assessed facial dynamics in patients AD. Such assessment was challenging, but of high impact, as such patients have lost a substantial amount of their cognitive capacity, and hence communication ability (e.g., to indicate discomfort or pain). We proposed an initial handcrafted approach based on the extension of Improved Fisher Vectors (IFV) [141] for videos, representing a video-sequence using both, local, as well as the related spatio-temporal features [DBN$^+$17b, DBB$^+$16]. Later, we compared CNN-methods for assessing facial dynamics such as *talking*, *singing*, *neutral* and *smiling* captured during music mnemotherapy sessions [WDB$^+$18b]. Specifically, we compared 3D ConvNets [125], Very Deep Neural Network based Two-Stream ConvNets [109, 144], as well as Improved Dense Trajectories. We adapted these methods from prominent action recognition methods and our promising results suggest

that the methods generalize well to the context of facial dynamics. The Two-Stream ConvNets in combination with ResNet-152 obtained the best performance on our dataset, capturing well even minor facial dynamics and has thus sparked high interest in the medical community. In personalizing the former approach, we obtained higher accuracy [HDB+20a].

3.3 Apathy Analysis

This section stands out, as it has the goal to go beyond expression and behavior analysis and focuses on a more complex, mental state. Analysis thereof is substantially more challenging, as it constituted a novel topic in computer vision and it was to be explored of how to deduce an internal characteristic state from external facial clues. We tackled the challenge on classifying apathy, which is defined by symptoms including reduced emotional response, lack of motivation and limited social interaction. Current methods for apathy diagnosis require the patient's presence in a clinic, and time consuming clinical interviews and questionnaires involving medical personnel, which are costly and logistically inconvenient for patients and clinical staff, hindering among other large scale diagnostics.

3.3.1 Initial Framework

In this context, we introduced a novel machine learning framework to classify apathetic and non-apathetic patients based on analysis of facial dynamics, entailing both emotion and facial movement [HDD+19]. Our approach catered to the challenging setting of current apathy assessment interviews, which include short video clips with wide face pose variations, very low-intensity expressions, and insignificant inter-class variations. We tested our algorithm on a dataset consisting of 90 video sequences acquired from 45 subjects and obtained an accuracy of 84% in apathy classification. Based on extensive experiments, we showed that the fusion of emotion and facial local motion produces the best feature set for apathy classification. In addition, we trained regression models to predict the clinical scores related to the mini-mental state examination (MMSE) and the neuropsychiatric apathy inventory (NPI) using the motion and emotion features. Our results suggested that the performance can be further improved by appending the predicted clinical scores to the video-based feature representation.

3.3.2 Multi-task learning for apathy classification

Leveraging on findings from the above framework, we proceeded to propose a multi-task learning (MTL) framework for apathy classification based on facial analysis, entailing both *emotion* and *facial movements* [HDD+20]. In addition, it leveraged information from other auxiliary tasks (i.e., clinical scores), which might be closely or distantly related to the main task of apathy classification. Our proposed MTL approach (termed MTL+) improved apathy classification by jointly learning model weights and the relatedness of the auxiliary tasks to the main task in an iterative manner. Our results on 90 video sequences acquired from 45 subjects obtained an apathy classification accuracy of up to 80%, using the concatenated emotion and motion features. Our results further demonstrated the improved performance of MTL+ over MTL. We improved this algorithm by considering spatio-temporal relations in a GRU-based model [DND+21] obtaining 90% of accuracy.

3.4 Heartrate Estimation

Adding upon the above emotion and behavior analysis, we here aimed to analyze the physiological trait heart rate (HR) from RGB-facial-videos [NZH+19]. Towards this, we present an end-to-end approach for robust remote HR-measurement based on remote photoplethysmography (rPPG), depicted in Figure 3.2, which constitutes a pulse triggered perceivable chromatic variation, sensed in RGB-face videos. Challenging in this setting has been that rPPGs can be affected in less-constrained settings. To unpin the shortcoming, the proposed algorithm utilizes a spatio-temporal attention mechanism, which places emphasis on the salient features included in rPPG-signals. In addition, we investigate an effective rPPG augmentation approach, generating multiple rPPG signals with varying HRs from a single face video (see Figure 3.3). Experimental results on the public datasets VIPL-HR and MMSE-HR show that the proposed method outperforms state-of-the-art algorithms in remote HR estimation.

Selected Publications

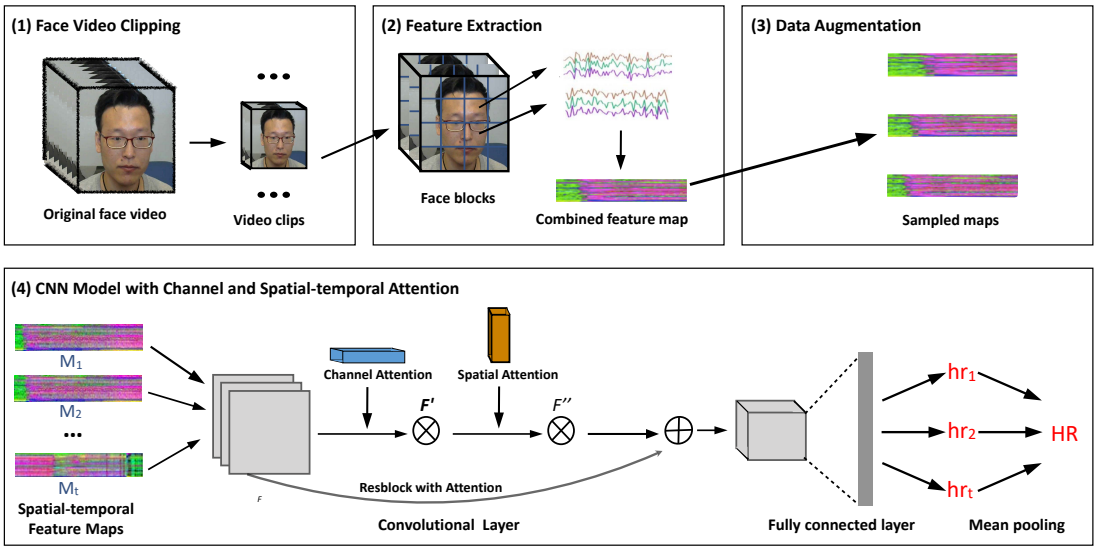

Figure 3.2: Overview of the proposed end-to-end trainable approach for rPPG based remote HR measurement via representation learning with spatial-temporal attention.

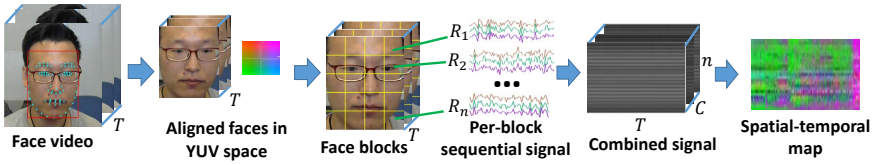

Figure 3.3: Illustration of the spatial-temporal map generation procedure.

1. A. Das, X. Niu, A. Dantcheva, SL Happy, H. Han, R. Zeghari, P. Robert, S. Shan, F. Bremond, X. Chen. A Spatio-temporal Approach for Apathy Classification, IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), doi: 10.1109/TCSVT.2021.3082857, 2021.

2. SL Happy, A. Dantcheva, A. Das, F. Bremond, R. Zeghari, P. Robert Apathy classification by exploiting task relatedness In *FG'20, IEEE International Conference on Automatic Face and Gesture Recognition*, May 18-22, 2020, Buenos Aires, Argentina.

3. SL Happy, A. Dantcheva, F. Bremond Semi-supervised emotion recognition using inconsistently annotated data In *FG'20, IEEE International Conference on Automatic Face and Gesture Recognition*, May 18-22, 2020, Buenos Aires, Argentina.

4. SL Happy, A. Dantcheva, A. Das, R. Zeghari, P. Robert, F. Bremond Characterizing the state of apathy with facial expression and motion analysis In *FG'20, IEEE International Conference on Automatic Face and Gesture Recognition*, May 14-18, 2019, Lille, France.

5. SL Happy, A. Dantcheva, F. Bremond. A weakly supervised learning technique for classifying facial expressions, Pattern Reognition Letters, 2019.

6. SL Happy, A. Dantcheva, A. Routray Dual-threshold based local path construction method for manifold approximation and its application to facial expression analysis In *EUSIPCO'19, European Signal Processing Conference*, September 2-6, 2019, A Coruna, Spain.

7. X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, X. Chen Robust remote heart rate estimation from facce utilizing spatial-temporal attention In *FG'19, IEEE International Conference on Automatic Face and Gesture Recognition*, May 14-18, 2019, Lille, France.

8. A. Dantcheva, P. Bilinski, H. Nguyen, J. C. Broutart, F. Bremond Expression Recognition for Severely Demented Patients in Music Reminiscence - Therapy In *EUSIPCO'17, 25th European Signal Processing Conference*, August 28 - September 2, 2017, Kos island, Greece.

9. A. Dantcheva, P. Bilinski, J. C. Broutart, P. Robert, F. Brémond. Emotion facial recognition by the means of automatic video analysis. *Gerontechnology*, vol. 15, pp. 12s, September 2016.

Chapter 4

Axis III. Face Generation

Given an image or a short video sequence, humans are able to forecast future potential events. Such ability is instrumental in making decisions. In this chapter, we intend to tackle the question: *can we endow machines with a similar ability to forecast the future?* We formulate this problem as a conditional generation task and pursue it incorporating *generative adversarial networks* (GANs).

GANs as introduced by Goodfellow *et al.* [21] incorporate two networks, a *Generator*, which generates new data instances and a *Discriminator*, which evaluates them for authenticity. The generator accepts noise as input and generates new samples of data in line with the observed training data. GANs have succeeded in applications such as image generation, image translation, super-resolution imaging, as well as face image synthesis.

Conditional GANs enhance the GAN-concept by providing both, the discriminator and generator with additional class information, in order to generate samples conditioned on different classes. It has been beneficial in domain transfer, super-resolution imaging, as well as image editing. Notable approaches include the conditional generative adversarial networks (cGANs) work by Mirza and Osindero [60] and Isola [31].

Recently GANs [21] have witnessed increased attention, attributed to the associated abilities to model complex data distributions, which allow them to generate and translate images. A number of tasks have benefited from this ability including *image generation* [118], *image translation* [31, 153], *super-resolution imaging* [48], as well as face *image synthesis* [106]. In contrast to images, *videos* represent

richer representations of the visual world and hence related *video generation* encompasses challenges w.r.t. complexity and computation, associated to the simultaneous modeling of appearance, as well as motion. Specifically, in inferring and modeling the distribution of human videos, generative models face three main challenges: (i) generating uncertain motion and retaining of human appearance, (ii) modeling spatio-temporal consistency, and (iii) understanding of latent representation. Finding suitable model architectures and representation learning methods, which are able to address these challenges, are critical to visual quality and plausibility of rendered novel video sequences.

Recent works in this context aim to generate actions [46], as well as to *predict videos*, i.e. anticipating what will happen in a video - essential to automate decision making [69, 113]. Particularly *video prediction* refers to the generation of future frames, given past observations by learning dynamic visual patterns from videos [40]. In this context, emphasis has been predominantly placed on predicting high-level semantics including action [135], event [29] and motion [137]. Such approaches have alleviated the challenges by conditioning the generation on potent priors such as input images, human keypoints and optical flow. This relates to learning to sample from conditional distributions, assuming access to the marginal distributions instead of learning joint distributions.

With respect to face generation, we proposed (a) a 2D model for conditional image generation based on attribute labels (Section 4.1), which we extend into (b) a 3D conditional approach for video generation based on attribute labels (Section 4.2). We then explored the challenging *video generation* that entails the mapping of a prior distribution (e.g., Gaussian distribution) and video distribution. In a first step, we designed (c) ImaGINator (Section 4.2.2), an architecture, which preserved the appearance information learned from an input image and animates this image using a motion label. Slightly altering the scenario of interest, we then introduced (d) a spatio-temporal generative model (Section 4.2.3), which sought to capture the distribution of high dimensional video data and to model appearance and motion in disentangled manner. As opposed to ImaGINator, this new model, referred to as G^3AN was streamlined to generate videos in an unconditional manner and hence did not require an input image. Our associated results showed that G^3AN indeed disentangled appearance and motion and hence both, appearance and motion can be manipulated in generated videos. The follow up model, (e) MintGAN (Section 4.2.4) was

targeted to allow for *interpretation of the latent space*. Specifically, we decomposed motion into semantic sub-spaces, which allowed for motion in generated videos to be easily manipulated. The sub-spaces were implemented by a motion dictionary, whose atoms form an orthogonal basis in the latent space. We showcased (f) the efficacy of generated data for data augmentation in a contrastive learning approach for person re-identification in Section 4.3. We proceeded to (g) detect *deepfakes* in Section 4.4, which constitute videos created or manipulated by generative models such as GANs.

We note that GANs have been instrumental in face-based biometrics related to Axis I, where we introduced a GAN, which improved face sketch recognition via adversarial sketch-photo transformation [YHS+19]. Most recently, we improved thermal to visible face recognition by a novel GAN [ACR+21]. It is aimed at explicitly decomposing an input image into identity code that is spectral-invariant and style code that is spectral-dependent. By using such a disentanglement, we were able to analyze the identity preservation by interpreting and visualizing the identity code. Similarly, we envision GANs being beneficial in Axis II for data augmentation, given the challenge concerning 'data scarcity'.

4.1 Image generation (2D model)

In generating still face images based on attribute-labels [WDB18a], we aimed to fit the conditional probability $P(x|z, y)$ in a conditional GAN, as depicted in Figure 4.1. We let z be the noise vector sampled from $\mathcal{N}(0, 1)$ with dimension 100, y be the vector representing attribute-labels (with $y_i \in \{\pm 1\}$, where i corresponds to the i^{th} attribute). We trained a GAN, adding attribute-labels in both, generator and discriminator. While the generator accepted as input the combination of prior noise $p(z)$ and attributes vector y, the discriminator accepted both, real or generated images, as well as the attribute-labels. Generated samples are shown in Figure 4.2.

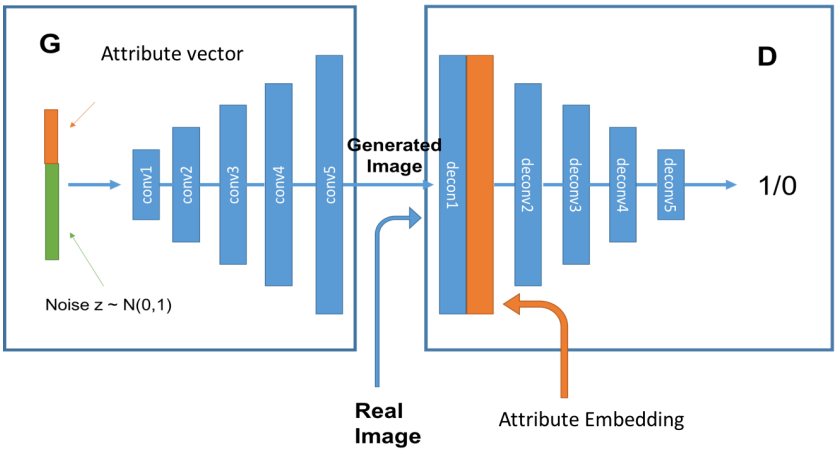

Figure 4.1: Architecture of proposed 2D method consisting of two modules, a discriminator D and a generator G. While D learns to distinguish between real and fake images, classifying based on attribute-labels, G accepts as input both, noise and attribute-labels in order to generate realistic face images.

(a) no glasses, female, black hair, smiling, young

Figure 4.2: Example images generated by the proposed 2D model.

4.2 Video generation (3D models)

4.2.1 Attributes guided video generation

We expanded the above presented 2D-model onto the spatio-temporal dimension, in order to create a conditional 3D-GAN [WDB17] (Figure 4.3) for generating videos. In both, generator and discriminator, the convolutional kernels have been expanded onto three dimensions (H, W, C, T), where H, W, C and T denote the height, width, channel and temporal step of the receptive fields in each kernel.

We feed the attribute vectors into the 3D model in a similar manner as in the 2D model. Specifically, in the generator we concatenate the attribute vector with the noise vector. In the discriminator, the feature map after the first layer has the dimension of (H, W, C, T), each $(H, W, C, t), t \in T$ containing spatio-temporal features of a certain time period. Our goal is to generate face videos based on attributes, hence

we proceed to provide each spatial-temporal feature map with the same attribute embedding. Based on this, we insert an attribute embedding into the spatio-temporal feature map from the first layer of the discriminator, creating a new feature map with the dimension $(H, W, C+y, T)$, where y is the dimension of the attribute vector.

We proposed a conditional video generation framework to generate smiling face sequences by providing facial attributes, as well as proposed a new method to insert attributes labels into spatio-temporal feature maps.

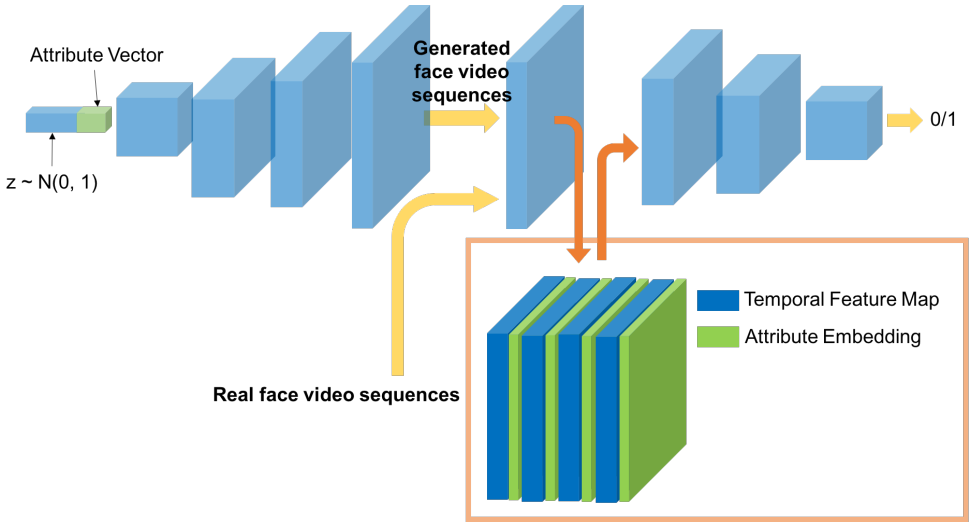

Figure 4.3: Architecture of proposed 3D model for face video generation

4.2.2 Video generation from a single image: ImaGINator

Here our goal has been to generate a video sequence, given an appearance information (i.e., a single image frame) and a motion class-label [WBBD20b]. We assumed that a video y can be decomposed into appearance c_a (originating from the input-image) and motion c_m (originating from the category-label), based on which we proceeded to generate videos. Hence, we formulated our task as learning a conditional mapping G: $\{z, c_a, c_m\} \rightarrow y$, where $z \sim \mathcal{N}(0,\ 1)$ denotes the random noise.

In this context, we proposed a framework that consists of the following 3 main components: (i) *Generator* G, that accepts c_a, c_m and noise as inputs, and seeks to generate realistic video sequences,

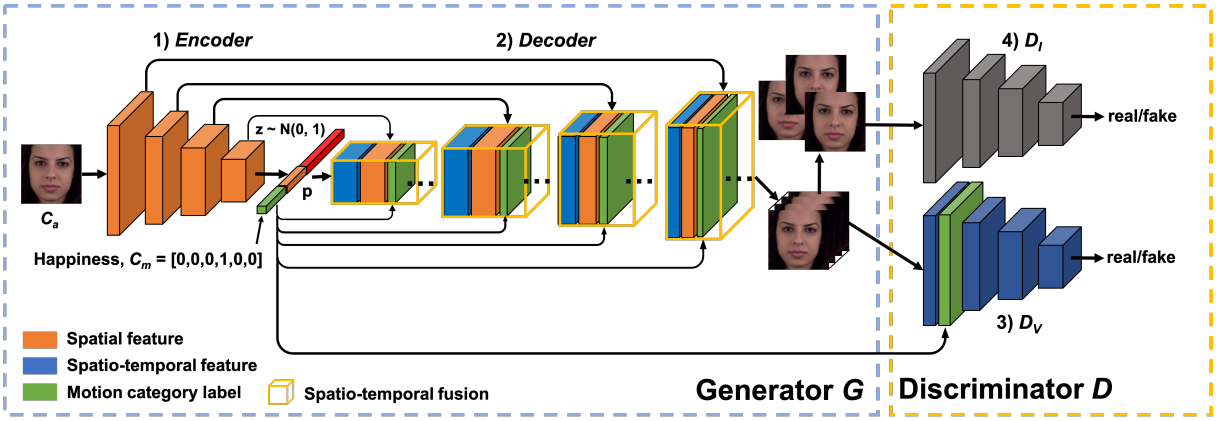

Figure 4.4: **Overview of the proposed ImaGINator.**

(ii) *image Discriminator* D_I that determines the frame-level based appearance quality, and (iii) *video Discriminator* D_V, which additionally discriminates, whether the generated video sequences contain authentic motion, see Figure 4.4. Generated frames are shown in Figure 4.5.

Related to that, our **contributions** included the design of a novel generative model incorporating (a) a *novel spatio-temporal fusion* mechanism, aiming at *retaining the appearance* by enforcing G to employ the spatial information in both, low and high feature levels, as well as (b) a *novel transposed (1+2)D convolution*, factorizing the transposed 3D convolutional filters into separate temporal and spatial components.

4.2.3 Unconditional Generation

In continuation of the above work, we aimed to generate videos given merely a motion class-label. To tackle this challenge, we proposed the novel spatio-temporal GAN-architecture G^3AN [WBBD19b] (see Figure 4.6), which sought to capture the distribution of high dimensional video data and to model appearance and motion in disentangled manner. The latter was achieved by decomposing appearance and motion in a three-stream Generator, where the main stream aims to model spatio-temporal consistency, whereas the two auxiliary streams augment the main stream with multi-scale appearance and motion features, respectively. An extensive quantitative and qualitative analysis showed that our model systematically and significantly outperformed state-of-the-art methods on the facial expression datasets MUG

Figure 4.5: **Example generated video frames by ImaGINator**

and UvA-NEMO, as well as the Weizmann and UCF101 datasets on human action. Additional analysis on the learned latent representations confirmed the successful decomposition of appearance and motion.

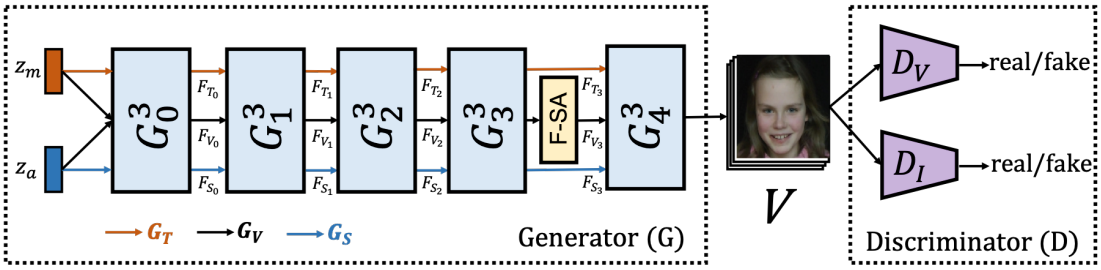

Figure 4.6: **Overview of our G^{3}AN architecture,** a fully convolutional GAN aimed at generating realistic video sequences. It consists of a three-stream Generator and a two-stream Discriminator. The Generator has 5 stacked G^3 modules, a factorized self-attention (F-SA) mechanism, and takes as input two random noise-vectors, z_a and z_m, aiming at controlling appearance and motion, respectively.

4.2.4 Interpretable Generation

Following the above, we presented an unconditional video generative model, MintGAN [WBD21], targeted to allow for interpretation of the latent space. Towards this, we designed a model (see Figure 4.7) that generates high quality videos, placing emphasis on the interpretation and manipulation of *motion*.

Video quality of generated videos was ensured given that they entail (a) naturalism, as well as (b) diversity. Naturalism and related video quality are generally impacted by a number of factors such as model architecture, objective function, and regularization, our major interest has been in the *model architecture*.

We decomposed motion into semantic sub-spaces, which allowed for control of generated samples. We designed the generator of MintGAN in accordance to proposed Linear Motion Decomposition, which carried the assumption that motion can be represented by a dictionary, whose atoms form an orthogonal basis in the latent space. Each vector in the basis represented a semantic sub-space. In addition, a Temporal Pyramid Discriminator analyzed videos at different temporal resolutions. Extensive quantitative and qualitative analysis showed that our model systematically and significantly outperformed state-of-the-art methods on the VoxCeleb2-mini, BAIR-robot and UCF101 datasets with respect to video quality, as well as confirmed that decomposed sub-spaces were interpretable and moreover, generated motion was controllable.

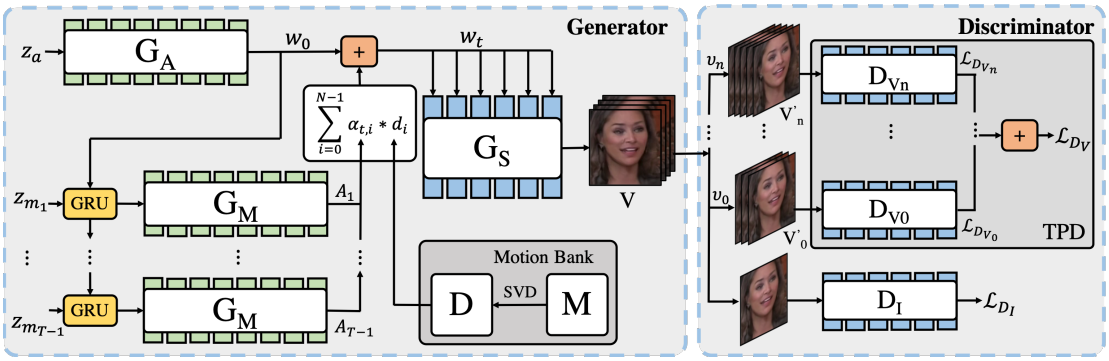

Figure 4.7: **MintGAN-architecture.** MintGAN comprises of a Generator and a two-stream Discriminator. We design the architecture of the Generator based on proposed Linear Motion Decomposition. Specifically, a motion bank is incorporated in the Generator to learn and store a motion dictionary D, which contains motion-directions $[d_0, d_1, .., d_{N-1}]$. We use an appearance net G_A to map appearance noise z_a into a latent code w_0, which serves as the initial latent code of a generated video. A motion net G_M maps a sequence of motion noises $\{z_{m_t}\}_{t=1}^{T-1}$ into a sequence $\{A_t\}_{t=1}^{T-1}$, which represent motion magnitudes. Each latent code w_t is computed based on Linear Motion Decomposition using w_0, D and A_t. Generated video V is obtained by a synthesis net G_S that maps the sequence of latent codes $\{w_t\}_{t=0}^{T-1}$ into an image sequence $\{x_t\}_{t=0}^{T-1}$. Our discriminator comprises an image discriminator D_I and a Temporal Pyramid Discriminator (TPD) that contains several video discriminators D_{V_i}, leveraging different temporal speeds v_i to improve generated video quality. While D_I accepts as input a randomly sampled image per video, each D_{V_i} is accountable for one temporal resolution.

4.3 Generated data for data augmentation in a contrastive learning approach for person re-identification

Recent self-supervised contrastive learning has provided an effective approach for unsupervised person re-identification (ReID) by learning invariance from different views (transformed versions) of an input. We incorporated a GAN and a contrastive learning module into one joint training framework [CWL+21]. While the GAN provided online data augmentation for contrastive learning, the contrastive module learned view-invariant features for generation. In this context, we proposed a mesh-based view generator. Specifically, mesh projections served as references towards generating novel views of a person. In addition, we proposed a view-invariant loss to facilitate contrastive learning between original and generated views. Deviating from previous GAN-based unsupervised ReID methods involving domain adaptation, we did not rely on a labeled source dataset, which made our method more flexible. Extensive experimental results showed that our method significantly outperformed state-of-the-art methods under both, fully unsupervised and unsupervised domain adaptive settings on several large scale ReID datsets.

4.4 Deepfake Detection

While technically intriguing, progress in generating realistic images and videos raises a number of social concerns related to the advent and spread of fake information and fake news. Such concerns necessitate the introduction of robust and reliable methods for fake image and video detection. Towards this, we studied the ability of state of the art *video* CNNs including 3D ResNet [27], 3D ResNeXt [152], and I3D [9] in detecting manipulated videos [WD20]. We presented related experimental results on videos tampered by four manipulation techniques, as included in the FaceForensics++ dataset [93]. We investigated three scenarios, where the networks were trained to detect (a) *all* manipulated videos, as well as (b) separately *each* manipulation technique individually. Finally and deviating from previous works, we conducted cross-manipulation results, where we (c) detected the veracity of videos pertaining to manipulation-techniques not included in the test set. Our findings clearly indicated the need for a better understanding of manipulation methods and the importance of designing algorithms that can successfully generalize onto unknown manipulations.

Selected Publications

1. H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, F. Bremond Joint Generative and Contrastive Learning for Unsupervised Person Re-identification In *CVPR, IEEE Conference on Computer Vision and Pattern Recognition*, June 19-25, 2021, virtual.

2. Y. Wang, P. Bilinski, F. Bremond, A. Dantcheva G$\hat{3}$AN: Disentangling motion and appearance for video generation In *CVPR, IEEE Conference on Computer Vision and Pattern Recognition*, June 16-18, 2020, Seattle, USA.

3. Y. Wang, A. Dantcheva A video is worth more than 1000 lies. Comparing 3D CNN approaches for detecting deepfakes In *FG, IEEE International Conference on Automatic Face and Gesture Recognition*, May 18-22, 2020, Buenos Aires, Argentina.

4. Y. Wang, P. Bilinski, F. Bremond, A. Dantcheva ImaGINator: conditional spatio-temporal GAN for video generation In *WACV, Winter Conference on Applications of Computer Vision*, March 2-5, 2020, Aspen, USA.

5. S. Yu, H. Han, S. Shan, A. Dantcheva, X. Chen, Improving face sketch recognition via adversarial sketch-photo transformation In *FG, IEEE International Conference on Automatic Face and Gesture Recognition*, May 14-18, 2019, Lille, France.

6. Y. Wang, A. Dantcheva, F. Brémond From attribute-labels to faces: face generation using a conditional generative adversarial network In *ECCVW Women in Computer Vision (WiCV) Workshop in conjunction with the European Conference on Computer Vision*, September 9, 2018, Munich, Germany.

7. Y. Wang, A. Dantcheva, J. C. Broutart, P. Robert, F. Brémond, Bilinski, Piotr Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders In *ECCVW, 6th International Workshop on Assistive Computer Vision and Robotics (ACVR) in conjunction with the European Conference on Computer Vision*, September 9, 2018, Munich, Germany.

8. Y. Wang, A. Dantcheva, F. Brémond From attributes to faces: a conditional generative adversarial network for face generation In *BIOSIG, 17th International Conference of the Biometrics Special Interest Group*, September 26-28, 2018, Darmstadt, Germany.

Chapter 5

Conclusions

This HDR presented a number of scientific works related to computer vision and specifically face analysis for security and healthcare, as well as to face generation. During the past 10 years I consolidated my knowledge in the domain of pattern recognition and computer vision, designing presented works.

With respect to *security*, our work extended prior state-of-the art, identifying bias and vulnerabilities in face analysis algorithms and proceeding to mitigate such by proposing novel ensemble-learning, as well as multi-task algorithms. Finding new face representations allowed for new insights into face, allowing to deduce novel characteristics such as height and weight.

Deepfake detection. We aim at designing heterogeneous strategies for deepfake detection that successfully generalize onto unknown manipulations. One strategy involves the learning of behavioural-signatures (e.g., talking-signature) representing enrolled subjects. Based on such signatures we will examine the integrity of videos, regardless of manipulation techniques. Few-shot learning will allow to transfer learned behaviour-patterns onto unseen subjects. Further, we intend to explore domain adaptation, transfer learning, metric learning, as well as one-class learning. Domain adaptation and transfer learning has been able to mitigate dataset bias and to increase cross-dataset accuracy. Metric learning enables the maximization of feature-wise distances between real and manipulated frames, while minimizing the feature-wise distances between frames obtained from different manipulation methods, focusing on features that generally occur in manipulation artefacts. In one-class learning, deepfake detection can

be formulated as an anomaly detection problem. In this context, the distribution of non-manipulated face images would be modelled using deep convolutional neural networks (CNN), aiming to identify manipulated face images as anomalies. Additional directions of investigation include the fusion of multiple detection strategies, as well as a multi-asset analysis, which incorporates diverse contextual information such as IP address, accompanying text, audio, as well origin of data. Finally, increasing the interpretability of deep CNNs will allow for the improvement of a reliable detection, providing higher robustness with respect to unseen malicious attacks.

In terms of *healthcare*, our work presented methods catering to the substantial gap between state-of-the-art behavior recognition systems, where related algorithms were trained and tested on small-scale datasets, and the colossal amounts of unconstrained real-life data. While the majority of research on automated recognition using biological and behavioral characteristics had focused on individual highly constrained settings, and had taken a rather microscopic view to the problem, we designed more macroscopic spatio-temporal algorithms that were efficiently applied in real-life healthcare settings.

We explored these topics from different perspectives and contributed to tools directly usable in security, as well as in clinical practice for the assessment of facial appearance and behavior. These works were developed in collaboration with an international network of scientific, clinical, as well as industrial collaborators, resulting in a set of projects and grants in the domain of computer vision for facial analysis. Despite these achievements, the state-of-the-art remains challenged in several settings, which I intend to overcome in our future work.

Multimodal analysis. In the context of healthcare, which involves processing data prone to operational randomness and uncertainty, I intend to explore additional sensors, which capture aspects such as depth or spectra beyond the visible ones. Specifically depth sensing entails a set of benefits such as (a) the determination of an absolute size of observed subject, (b) robustness to noise (as opposed to RGB) and presentation attacks (i.e., spoofing), (c) sensing that is possible in night environments, (d) an improved performance of 2D data for face analysis with applicability in pose estimation and 3D reconstruction. However depth maps are of poor quality due to occlusions and artefacts, and furthermore the related

depth map resolution is orders of magnitude below that of embedded RGB cameras. Additional *multi-modal sensors*, which have the ability to support face analysis include devices measuring physiological signals such as pulse and vein monitoring.

Towards dealing with the scarcity of training data, we will investigate GANs for data augmentation, which has been shown to improve algorithmic performance when used as additional training data [CWL+21].

One long term goal in the area of healthcare constitutes the design of a holistic framework, which analyzes human emotions and mental states, predicts emergencies, monitors the progress of potential diseases and proceeds to suggest interventions such as serious games and music therapy. Such a framework should be streamlined to extract salient features from long videos, deducing a semantic summary.

Finally our work on face generation allowed us to learn meaningful generative models, entailing the challenge of being able to simultaneously generate both appearance, as well as motion. Towards this and hence towards modeling the distribution of high dimensional video data, we disentangled appearance and motion, allowing to tackle the two bottlenecks in video generation, namely ensuring spatio-temporal consistency, as well as preserving appearance throughout generated videos. While we introduced a set of image and video generation methods with increasingly realistic generation results, associated results remain limited in *resolution* and far from perfect w.r.t. *video quality*. The main reason for these limitations has to do with the challenging training of video GANs, due to large model-complexity, training instability and optimization issues. These can be addressed in more sophisticated architectures, involving more robust loss functions and stable training procedures. We believe that a further research direction in video generation constitutes developing simpler, lower complexity and memory-efficient architectures that require less training as shown in the context of image processing [3].

Explainability of Video GANs. In view of a most recently financed project on explainability of video GANs (XGAN, Action exploratoire, Inria), we intend to pierce the black box of GANs for video generation by proposing strategies to interpret the latent space in (a) design of interpretable architectures, and by (b) analysis of symmetric functions in input and output of patch-based generation.

We also intend to design frameworks, which allow for video generation based on an audio and text-input. A long term goal in generation will be to design a system able to generate personalized complex videos incorporating interaction, solely based on written descriptions.

Chapter 6

Supervision, Responsibilities and Other Research Activities

I am currently the main advisor of 1 PostDoc and 2 Ph.D. students. The funding for these has been secured by the ANR JCJC, the ANR RESPECT projects, as well as Thales. One PostDoc and one Ph.D. student will join in the next months, whose research will be funded by Inria, Action Exploratoire `https://www.inria.fr/fr/actions-exploratoires-inria-prendre-des-risques`.

6.1 Current Supervision

- *Student Name:* **Yaohui Wang**

  Subject: "Automatic Holistic Analysis of Humans"

  Institution: INRIA, Sophia Antipolis, France

  Dates: December 2017 - November 2020.

  Supervisors: Antitza Dantcheva

- *Student Name:* **David Anghelone**

  Subject: "Facial analysis beyond the visible spectrum"

  Institution: INRIA, Sophia Antipolis, France

Dates: April 2019 - March 2022.

Supervisor: Antitza Dantcheva

In collaboration with Thales

Post-doctoral fellow Supervision

- *PostDoc Name:* **Indu Joshi**

  Subject: "Deepfake detection"

  Institution: INRIA, Sophia Antipolis, France

  Dates: July 2021 - February 2022.

  Supervisor: Antitza Dantcheva

6.2 Past supervision

- *PostDoc Name:* **S L Happy**

  Subject: "Automatic Holistic Analysis of Humans"

  Institution: INRIA, Sophia Antipolis, France

  Dates: February 2018 - July 2019.

  Supervisor: Antitza Dantcheva (80%) François Brémond (20%)

- *PostDoc Name:* **Abhijit Das**

  Subject: "Facial expression recognition with application in health monitoring"

  Institution: INRIA, Sophia Antipolis, France

  Dates: February 2018 - Mai 2019, December 2020 - February 2021.

  Supervisor: Antitza Dantcheva (80%) François Brémond (20%)

- *PostDoc Name:* **Michal Balazia**

  Subject: "Computer Vision for Neurodegenerative Disorders"

Institution: INRIA, Sophia Antipolis, France

Dates: September 2019 - March 2021.

Supervisor: Antitza Dantcheva

- *Research engineer:* **Thanh Hung Nguyen**

  Subject: "Head and Face Detection"

  Dates: 2016 - 2018

  Institution: INRIA, STARS team

  Supervisors: Antitza Dantcheva, François Brémond

  Significance of the work: The outcomes of this work will be published in a pending conference and a pending journal paper.

Previous Ph.D. Student mentoring:

- *Student Name:* **Cunjian Chen**

  Subject: "Facial Cosmetics and automated Face Analysis Systems"

  Dates: 2012 - 2014

  Institution: West Virginia and Michigan State University

  Supervisors: Arun Ross

  Mentor: Antitza Dantcheva

  Significance of the work: The outcomes of this work have been featured in a series of publications (see the full list of contributions) and were a major part of the student's Ph.D. thesis. He successfully defended middle of 2014.

- *Student Name:* **Ester Gonzalez Sanchez**

  Subject: "Fusion of Facial and Body Soft Biometrics"

  Dates: 2016 - 2017, during Ester's visit at Eurecom, France

  Institution: Universidad Autónoma de Madrid

  Supervisors: Ruben Vera, Julian Fierrez

Mentor: Antitza Dantcheva

Significance of the work: The outcomes of this work is a conference paper (ICPR'16) and a submitted journal publication (submitted November 2017 to Trans. Image Processing).

Masters final project student supervision:

- *Student Name:* **Leah Dankovcik**

  Subject: "Interrelation of Soft Biometrics"

  Dates: February 2012 - May 2012

  Institution: West Virginia University

  Supervisors: Arun Ross

  Mentor: Antitza Dantcheva

  Significance of the work: The outcomes of this work have been used in my publication in the IEEE Transactions on Information Forensics and Security (TIFS 2016).

- *Student Name:* **Leila Zangar**

  Subject: "Soft Biometrics: Eye Colors"

  Dates: Summer Semester 2009

  Institution: Eurecom

  Supervisors: Antitza Dantcheva, Jean-Luc Dugelay

  Significance of the work: The outcomes of this work have been used for a publication in the IEEE Winter Conference on Applications of Computer Vision 2011 (WACV'11).

6.3 Scientific Engagement

I am continuously being invited to serve in different committees and to serve as reviewer for major conferences and journals that I proceed to list below.

- Member of the **ELLIS Society** (`https://ellis.eu/members`)

- **Program Co-chair** at the International Conference of the Biometrics Special Interest Group (BIOSIG) 2017-2021 (http://fg-biosig.gi.de/biosig-2019.html)
- Member of the **Evaluation Committee of ANR** AAPG 2020 - Comité Sécurité Globale et Cybersécurité
- **Associate Editor** of Pattern Recognition, since 2020 (https://www.journals.elsevier.com/pattern-recognition/editorial-board)
- **Editorial Board** of Journal Multimedia Tools and Applications, since 2017 (http://www.springer.com/computer/information+systems+and+applications/journal/11042?detailsPage=editorialBoard)
- Serving in the Technical Activities Committee of **IEEE Biometrics Council** (2017-2021) (http://ieee-biometrics.org/index.php/homepage/committees) comprising of a chair and 4 members, responsible to review tutorials, support award nominations, cooperate with the publications, conferences and education committees, review and create expert web content, as well as actively shape the technical future of biometrics within the IEEE.
- Served in the Education Committee of the **IEEE Biometrics Council** (2013 - 2016) comprising of a chair and 4 members, who manage the Distinguished Lecturers Program (DLP); organize biometric seminars; evaluate scholarship requests from students; facilitate the development of biometric tutorials; and interface with other entities to promote biometric education across IEEE as well as society-at-large.
- Serving in the **EURASIP Biomedical Image & Signal Analytics (BISA) special area team** (2018-2021)
  (https://www.eurasip.org/index.php?option=com_content&view=article&id=151&Itemid=1151)
  comprising of a chair and 23 members, aiming to strengthen the biomedical activities at EURASIP in terms of workshops, tutorials, and special sessions, in particular, at EUSIPCO, initiating a Best Paper Award at EUSIPCO for papers in the area of biomedical image and signal processing, as well as providing expertise to the EURASIP Board of Directors.

- Tutorial chair at the IEEE International Joint Conference on Biomtrics (IJCB) 2021.
- Publication chair at the IEEE International Joint Conference on Biomtrics (IJCB) 2020.
- Publication chair at the IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2019.
- Media Co-chair at the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS) 2013.
- Member of the **European Association of Biometrics** (EAB) since 2018 (`https://www.eab.org/membership/members.html?ts=1518609713652`)
- I have been jury member for Ph.D. committees of several students of Universite Cote d'Azur.
- A selection of invited lectures include: workshop at Winter Conference on Applications of Computer Vision (WACV'20); Information Sciences Institute (ISI) at the University of Southern California (USC), Los Angeles, USA (2018); Institute of Computing Technology (ICT) at the Chinese Academy of Sciences (CAS), Beijing, China (2018); Summer School Brain Innovation Generation @ UCA (Big@UCA, Sophia Antipolis (2018); Institute of Biomedical Engineering (IBME), University of Oxford, UK (2017); Biometrics Congress, London, UK (2017); Day of Biometrics, Caen, France (2017); World Conference on Gerontechnology, Nice, France (2016); Research center for Information Security (SBA), Vienna, Austria (2015); University of Cyprus, Nicosia, Cyprus (2015); INRIA, Sophia Antipolis, France (2013); Workshop on 3D and 2D face analysis and recognition, Lyon, France (2011); Lane Department of Computer Science and Electrical Engineering, West Virginia University, USA (2011).
- I have been session chair in a number of conferences including IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2020), International Conference of the Biometrics Special Interest Group (BIOSIG) 2017, 2018, European Conference on Computer Vision (ECCV) Workshop "What's in a Face?" (2012), SPIE Defense, Security + Sensing: Biometric Technology for Human Identification IX (2012)

I have been in the Technical Program Committee of the IEEE International Conference on Multimedia and Expo (ICME) (2019), the International Joint Conference on Biometrics (IJCB) (2014, 2017), the IAPR International Conference on Biometrics (ICB) (2013, 2016, 2018), the IEEE Conference on Computer Vision and Pattern Recognition Workshop ChaLearn Looking at People 2016 the Workshop on Affective Interaction with Virtual Assistants within the Healthcare Context in conjunction with the EAI International Conference on Pervasive Computing Technologies for Healthcare Conference (2016), the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS) (2013, 2015, 2016), the Cairo International Biomedical Engineering Conference (CIBEC) (2012).

6.4 Collaborative Projects and Funding

I proposed research directions and ideas in several successful international projects listed below. Particular attention is placed on the ANR Jeunes Chercheuses et Jeunes Chercheurs (JCJC), which is a personal funding with 12.6% acceptance rate for 2017, with the aim to support young researchers and enable them to autonomously develop a specific research topic theme and to give them the possibility to form their team and express their capacity for exploratory research and innovation.

More recent, we were also awarded the selective ANR PRCI (French-German) with Eurecom, France and the Hochschule Darmstadt, Germany.

Further, I have initiated collaboration with the two companies Thales[1], and the startup Blu Manta[2] in the area of 2D and 3D face authentication, focusing on face analysis beyond the visible spectrum with challenges such as robust depth map reconstruction from structured light dot pattern.

- PI of INRIA Action Exploratoire (AEx) 2021 "**XGAN. Interpretable Representation Learning for Video GANs**", September 2021 - August 2024.
- INRIA PI of ANR PRCI (French-German) "**RESPECT. Reliable, Secure and Privacy preserving multibiometric Person Authentication**", April 2019 - March 2022.

[1] `https://www.thalesgroup.com/en`
[2] `https://www.societe.com/societe/blu-manta-832907471.html`

- PI of ANR JCJC "**ENVISION. Computer Vision for Automated Holistic Analysis of Humans**", (Project duration: November 2017 - October 2021).

- Co-PI of INRIA - CAS "**FER4HM. Facial expression recognition with application health monitoring**", (project duration: November 2017 - October 2019).

- PI of Labex Post Doctoral Fellowship "**Big Data and Biometrics**". Project funded my research at INRIA (project duration: from March 2016 to February 2017).

- PI of ERCIM ABCDE Project nr. 246016 of the European Commission "**Facial Analysis for Health Monitoring**". Project funded my research at INRIA (project duration: from March 2014 to July 2015).

- PI of the project "**Impact of Cosmetics on the Performance and Security of Face Recognition**" Funded by the National Science Foundation (NSF) Industry / University Cooperative Research Center CITeR (Center for Identification Technology Research) (project duration: July 2012 - June 2013).

- Participant of the EIT Digital European Project on "**Cross - linguistic comparison of speech features in older adults with Alzheimer's Disease and related disorders**" (project duration January 2017 - December 2018).

- Participant of the "**ACTIBIO - Unobtrusive Authentication using Activity Related and Soft BIOmetrics**" European Commisision under FP7-215372 (project duration: March 2008 - March 2011).

6.4.1 Publications after the Ph.D.

International journals (10)

1. A. Das, X. Niu, A. Dantcheva, SL Happy, H. Han, R. Zeghari, P. Robert, S. Shan, F. Bremond, X. Chen, A spatio-temporal approach for apathy classification *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2021.

2. SL Happy, A. Dantcheva, F. Bremond, Expression recognition with deep features extracted from holistic and part-based models Download pdf *Image and Vision Computing (IMAVIS)*, 2020.

3. P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, C. Busch. Demographic Bias in Biometrics: A Survey on an Emerging Challenge, *IEEE Transactions on Technology and Society (T-TS)*, vol. 2, no. 2, pp. 1-15, 2020.

4. C. Rathgeb, A. Dantcheva, C. Busch. Impact and detection of facial beautification in face recognition: An overview, *IEEE Access,* vol. 7, no. 1, December, 2019.

5. SL Happy, A. Dantcheva, F. Bremond. A weakly supervised learning technique for classifying facial expressions, *Pattern Reognition Letters*, 2019.

6. A. Dantcheva and F. Brémond. Gender estimation based on smile-dynamics. *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 12, no. 3, pp. 719-729, March 2017.

7. C. Chen, A. Dantcheva, A. Ross. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, vol. 32, pp. 80-92, November 2016.

8. A. Dantcheva, P. Bilinski, J. C. Broutart, P. Robert, F. Brémond. Emotion facial recognition by the means of automatic video analysis. *Gerontechnology*, vol. 15, pp. 12s, September 2016.

9. A. Dantcheva, P. Elia, A. Ross. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 11, no. 3, pp. 441-467, March 2016.

10. A. Dantcheva, J.-L. Dugelay. Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics. *Mutimedia Tools and Applications (MTAP)*, vol. 74, no. 24, pp. 11331-11355, 2014.

Reviewed international conferences (33)

1. D. Anghelone; C. Chen; A. Ross; P. Faure; A. Dantcheva Explainable thermal to visible face recognition using latent-guided generative adversarial network In *FG'21, 16th IEEE International*

Conference on Automatic Face and Gesture Recognition, December 15-18, 2021, Jodhpur, India (hybrid).

2. D. Yang; Y. Wang; A. Dantcheva; L. Garattoni; G. Francesca; F. Bremond Self-supervised video pose representation learning for occlusion-robust action recognition In *FG'21, 16th IEEE International Conference on Automatic Face and Gesture Recognition*, December 15-18, 2021, Jodhpur, India (hybrid).

3. H. Chen*; Y. Wang*; B. Lagadec; A. Dantcheva; F. Bremond Joint generative and contrastive learning for unsupervised person re-identification In *CVPR'21, IEEE Conference on Computer Vision and Pattern Recognition*, June 19-25, 2021, virtual. arXiv:2012.09071

4. I. Joshi; A. Utkarsh; R. Kothari; V. Kurmi; A. Dantcheva; S. Roy; P. Kalra On learning sensor-invariant features for fingerprint ROI segmentation In *IJCNN'21, The International Joint Conference on Neural Networks*, July 18-22, 2021, virtual.

5. I. Joshi; A. Utkarsh; R. Kothari; V. Kurmi; A. Dantcheva; S. Roy; P. Kalra Learning noise-aware preprocessing of fingerprints In *IJCNN'21, The International Joint Conference on Neural Networks*, July 18-22, 2021, virtual.

6. Y. Wang; F. Bremond; A. Dantcheva InMoDeGAN: Interpretable motion decomposition generative adversarial network for video generation arXiv:2101.03049

7. I. Joshi; R. Kothari; A. Utkarsh; V. K. Kurmi; A. Dantcheva; S. D. Roy; P. Kalra Explainable fingerprint ROI segmentation using Monte Carlo dropout In *WACVW'21, Workshops of the Winter Conference on Applications of Computer Vision*, January 5-9, virtual.

8. M. Balazia; SL Happy; F. Bremond; A. Dantcheva How unique is a face: An investigative study In *ICPR'20, 25th International Conference on Pattern Recognition*, January 10-15, 2021, Milan, Italy (virtual).

9. Y. Wang, P. Bilinski, F. Bremond, A. Dantcheva G$\hat{3}$AN: Disentangling motion and appearance for video generation In *CVPR, IEEE Conference on Computer Vision and Pattern Recognition*, June 16-18, 2020, Seattle, USA.

10. X. Li, H. Han, H. Lu, X. Niu, Z. Yu, A. Dantcheva, G. Zhao, S. Shan The 1st Challenge on Remote Physiological Signal Sensing (RePSS) In *CVPRW, Workshops of the IEEE Conference on Computer Vision and Pattern Recognition*, June 16-18, 2020, Seattle, USA.

11. Y. Wang, A. Dantcheva A video is worth more than 1000 lies. Comparing 3D CNN approaches for detecting deepfakes In *FG, IEEE International Conference on Automatic Face and Gesture Recognition*, May 18-22, 2020, Buenos Aires, Argentina.

12. SL Happy, A. Dantcheva, A. Das, F. Bremond, R. Zeghari, P. Robert Apathy classification by exploiting task relatedness In *FG, IEEE International Conference on Automatic Face and Gesture Recognition*, May 18-22, 2020, Buenos Aires, Argentina.

13. SL Happy, A. Dantcheva, F. Bremond Semi-supervised emotion recognition using inconsistently annotated data In *FG, IEEE International Conference on Automatic Face and Gesture Recognition*, May 18-22, 2020, Buenos Aires, Argentina.

14. Y. Wang, P. Bilinski, F. Bremond, A. Dantcheva ImaGINator: conditional spatio-temporal GAN for video generation In *WACV, Winter Conference on Applications of Computer Vision*, March 2-5, 2020, Aspen, USA.

15. Y. Wang, P. Bilinski, F. Bremond, A. Dantcheva G$\hat{3}$AN: This video does not exist. Disentangling motion and appearance for video generation arXiv preprint arXiv:1912.05523, 2019.

16. SL Happy, A. Dantcheva, A. Routray Dual-threshold based local path construction method for manifold approximation and its application to facial expression analysis In *EUSIPCO, European Signal Processing Conference*, September 2-6, 2019, A Coruna, Spain.

17. SL Happy, A. Dantcheva, A. Das, R. Zeghari, P. Robert, F. Bremond Characterizing the state of apathy with facial expression and motion analysis In *FG, IEEE International Conference on Automatic Face and Gesture Recognition*, May 14-18, 2019, Lille, France.

18. X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, X. Chen Robust remote heart rate estimation from facce utilizing spatial-temporal attention In *FG, IEEE International Conference*

on Automatic Face and Gesture Recognition, May 14-18, 2019, Lille, France. award Best Poster Award

19. S. Yu, H. Han, S. Shan, A. Dantcheva, X. Chen, Improving face sketch recognition via adversarial sketch-photo transformation In *FG, IEEE International Conference on Automatic Face and Gesture Recognition*, May 14-18, 2019, Lille, France.

20. A. Das, C. Galdi, H. Han, R. Ramachandra, J.-L. Dugelay, A. Dantcheva. Recent advances in biometric technology for mobile devices In *BTAS, IEEE International Conference on Biometrics: Theory, Applications and Systems*, October 22-25, 2018, Los Angeles, USA.

21. A. Das, A. Dantcheva, F. Brémond Mitigating bias in gender, age, and ethnicity: a multi-task convolution neural network approach In *ECCVW International Workshop on Bias Estimation in Face Analytics (BEFA) in conjunction with the European Conference on Computer Vision*, September 9, 2018, Munich, Germany.

22. Y. Wang, A. Dantcheva, F. Brémond From attribute-labels to faces: face generation using a conditional generative adversarial network In *ECCVW Women in Computer Vision (WiCV) Workshop in conjunction with the European Conference on Computer Vision*, September 9, 2018, Munich, Germany.

23. Y. Wang, A. Dantcheva, J. C. Broutart, P. Robert, F. Brémond, Bilinski, Piotr Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders In *ECCVW, 6th International Workshop on Assistive Computer Vision and Robotics (ACVR) in conjunction with the European Conference on Computer Vision*, September 9, 2018, Munich, Germany.

24. Y. Wang, A. Dantcheva, F. Brémond From attributes to faces: a conditional generative adversarial network for face generation In *BIOSIG, 17th International Conference of the Biometrics Special Interest Group*, September 26-28, 2018, Darmstadt, Germany.

25. P. Bilinski, A. Dantcheva, F. Brémond. Show me your face and I will tell you your height, weight and body mass index. In *IAPR ICPR International Conference on Pattern Recognition*, August 20-24, 2018, Beijing, China.

26. A. Dantcheva, P. Bilinski, H. Nguyen, J. C. Broutart, F. Bremond Expression Recognition for Severely Demented Patients in Music Reminiscence - Therapy In *EUSIPCO'17, 25th European Signal Processing Conference*, August 28 - September 2, 2017, Kos island, Greece.

27. C. Chen, A. Dantcheva, T. Swearingen, A. Ross. Spoofing Faces Using Makeup: An Investigative Study. In *ISBA'17, 3rd IEEE International Conference on Identity, Security and Behavior Analysis*, February 2017, New Delhi, India.

28. Gonzalez-Sosa, Ester; A. Dantcheva, Vera-Rodriguez, Ruben; J.-L. Dugelay,; Bremond, Francois; Fierrez, Julian. Image-based Gender Estimation from Body and Face across Distances In *ICPR'16, 23rd International Conference on Pattern Recognition*, December 4-8, 2016, Cancun, Mexico.

29. P. Bilinski, A. Dantcheva, F. Brémond. Can a smile reveal your gender? In *BIOSIG'16, 15th International Conference of the Biometrics Special Interest Group*, September 21-23, 2016, Darmstadt, Germany.

30. C. Chen, A. Dantcheva, A. Ross. Impact of facial cosmetics on automatic gender and age estimation algorithms In *VISAPP'14, 9th International Conference on Computer Vision Theory and Applications*, January 5-8, 2014, Lisbon, Portugal.

31. A. Dantcheva, A. Ross, C. Chen. Makeup challenges automated face recognition systems. In *SPIE Newsroom 2013, Defense and Security.* DOI: 10.1117/2.1201303.004795.

32. C. Chen, A. Dantcheva, A. Ross. Automatic facial makeup detection with application in face recognition In *IAPR ICB'13, 6th IAPR International Conference on Biometrics*, June 4-7, 2013, Madrid, Spain.

33. A. Dantcheva, C. Chen, A. Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *IEEE BTAS'12, 5th IEEE International Conference on Biometrics: Theory, Applications and Systems*, September 23-26, 2012, Washington DC, USA.

Books and book chapters (2)

1. A. Dantcheva, P. Elia, J.-L. Dugelay. Facial Soft Biometrics for Person Recognition. In A. Naït-Ali and R. Fournier, editors, *Signal and Image Processing for Biometrics*. Wiley, 2012.

2. A. Dantcheva, C. Yemdji, P. Elia, J.-L. Dugelay. Biométrie faciale douce pour l'identification des individus. In A. Naït-Ali and R. Fournier, editors, *Traitement du signal et de l'image pour la biométrie*. Traité IC2, série Signal et Image dirigée par Henri Maître et Francis Castanié, Hermes Science, 2012.

References pertaining to the list of my publications 2012-2021

[ACR+21] David Anghelone, Cunjian Chen, Arun Ross, Philippe Faure, and Antitza Dantcheva. Explainable thermal to visible face recognition using latent-guided generative adversarial network. In *FG'21, 16th IEEE International Conference on Automatic Face and Gesture Recognition*, 2021.

[BDB16] Piotr Bilinski, Antitza Dantcheva, and Francois Brémond. Can a smile reveal your gender? In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, volume 15, 2016.

[BHBD21] Michal Balazia, SL Happy, François Brémond, and Antitza Dantcheva. How unique is a face: An investigative study. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7066–7071. IEEE, 2021.

[CDR13] Cunjian Chen, Antitza Dantcheva, and Arun Ross. Automatic facial makeup detection with application in face recognition. In *Proc. of IAPR International Conference on Biometrics (ICB)*, 2013.

[CDR14] Cunjian Chen, Antitza Dantcheva, and Arun Ross. Impact of facial cosmetics on automatic gender and age estimation algorithms. In *Proc. of International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014.

[CDR16] Cunjian Chen, Antitza Dantcheva, and Arun Ross. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, 32:80–92, 2016.

[CDSR17] Cunjian Chen, Antitza Dantcheva, Thomas Swearingen, and Arun Ross. Spoofing faces using makeup: An investigative study. In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2017.

[CWL+21] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2004–2013, 2021.

[DB16] Antitza Dantcheva and François Brémond. Gender estimation based on smile-dynamics. *IEEE Transactions on Information Forensics and Security*, 12(3):719–729, 2016.

[DBB+16] Antitza Dantcheva, Piotr Bilinski, Jean-Claude Broutart, Philippe Robert, and Francois Brémond. Emotion facial recognition by the means of automatic video analysis. *Gerontechnology*, 15(suppl):12s, 2016.

[DBB18] Antitza Dantcheva, François Bremond, and Piotr Bilinski. Show me your face and i will tell you your height, weight and body mass index. In *International Coference on Pattern Recognition (ICPR)*, 2018.

[DBN+17a] Antitza Dantcheva, Piotr Bilinski, Hung Thanh Nguyen, Jean-Claude Broutart, and Francois Bremond. Expression recognition for severely demented patients in music reminiscence-therapy. In *Signal Processing Conference (EUSIPCO), 2017 25th European*, pages 783–787. IEEE, 2017.

[DBN+17b] Antitza Dantcheva, Piotr Bilinski, Hung Thanh Nguyen, Jean-Claude Broutart, and Francois Bremond. Expression recognition for severely demented patients in music reminiscence-therapy. In *25th European Signal Processing Conference (EUSIPCO)*, pages 783–787, Aug 2017.

[DCR12a] Antitza Dantcheva, Cunjian Chen, and Arun Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *Proc. of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 391–398. IEEE, 2012.

[DCR12b] Antitza Dantcheva, Cunjian Chen, and Arun Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *Proc. of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2012.

[DCR13] Antitza Dantcheva, Cunjian Chen, and Arun Ross. Makeup challenges automated face recognition systems. *SPIE Newsroom*, pages 1–4, 2013.

[DD15] Antitza Dantcheva and Jean-Luc Dugelay. Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics. *Multimedia Tools and Applications*, 74(24):11331–11355, 2015.

[DDB18] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age, and ethnicity: a multi-task convolution neural network approach. In *European Conference on Computer Vision - Workshops (ECCVW)*, 2018.

[DER15] Antitza Dantcheva, Petros Elia, and Arun Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, pages 1–26, 2015.

[DGH+18] Abhijit Das, Chiara Galdi, Hu Han, Raghavendra Ramachandra, Jean-Luc Dugelay, and Antitza Dantcheva. Recent advances in biometric technology for mobile devices. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, volume 9, 2018.

[DHD+20] Abhijit Das, SL Happy, Antitza Dantcheva, Hu Han, Francois Bremond, and Xiling Chen. Computer vision-based human health monitoring: Recent advancement. In *Pending Submission*, 2020.

[DND+21] Abhijit Das, Xuesong Niu, Antitza Dantcheva, SL Happy, Hu Han, Radia Zeghari, Philippe Robert, Shiguang Shan, Francois Bremond, and Xilin Chen. A spatio-temporal approach for apathy classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[DRD+20] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[DYED12a] Antitza Dantcheva, Christelle Yemdji, Petros Elia, and Jean-Luc Dugelay. Biométrie faciale douce pour l'identification des individus. In A. Naït-Ali and Régis Fournier, editors, *Traitement du signal et de l'image pour la biométrie*. Hermes Science, 2012.

[DYED12b] Antitza Dantcheva, Christelle Yemdji, Petros Elia, and Jean-Luc Dugelay. Facial soft biometrics for person recognition. In A. Naït-Ali and Régis Fournier, editors, *Signal and Image Processing for Biometrics*. Wiley, 2012.

[GSDVR+16] Ester Gonzalez-Sosa, Antitza Dantcheva, Ruben Vera-Rodriguez, Jean-Luc Dugelay, Francois Brémond, and Julian Fierrez. Image-based gender estimation from body and face across distances. In *IAPR International Conference on Pattern Recognition (ICPR)*, volume 23, 2016.

[HDB19] SL Happy, Antitza Dantcheva, and Francois Bremond. A weakly supervised learning technique for classifying facial expressions. *Pattern Recognition Letters*, 128:162–168, 2019.

[HDB+20a] SL Happy, Antitza Dantcheva, Piotr Bilinski, Philip Robert, and Francois Bremond. Personalized facial expression and activity recognition in mnemotherapy for patients with major neurocognitive disorders. *Pending submission*, 2020.

[HDB20b] SL Happy, Antitza Dantcheva, and Francois Bremond. Semi-supervised emotion recognition using inconsistently annotated data. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 15, 2020.

[HDD+19] SL Happy, Antitza Dantcheva, Abhijit Das, Radia Zeghari, Philippe Robert, and Francois Bremond. Characterizing the state of apathy with facial expression and motion analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 14, 2019.

[HDD+20] SL Happy, Antitza Dantcheva, Abhijit Das, Francois Bremond, Radia Zeghari, and Philippe Robert. Apathy classification by exploiting task relatedness. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 15, 2020.

[HDR19] SL Happy, Antitza Dantcheva, and Aurobinda Routray. Dual-threshold based local patch construction method for manifold approximation and its application to facial expression analysis. In *European Signal Processing Conference (EUSIPCO)*, volume 27, 2019.

[NZH+19] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Shiguang Shan, Antitza Dantcheva, and Xilin Chen. A robust remote heart rate estimation technique from face analysis utilizing spatial-temproal attention learning. In *14th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2019.

[RDB19] Christian Rathgeb, Antitza Dantcheva, and Christoph Busch. Impact and detection of facial beautification in face recognition: An overview. *IEEE Access*, 7:152667–152678, 2019.

[RJDD21] Ritaban Roy, Indu Joshi, Abhijit Das, and Antitza Dantcheva. 3DCNN architectures and attention mechanisms for deepfake detection. In *Handbook of Digital Face Manipulation and Detection*, 2021.

[WBBD19a] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3AN: This video does not exist. Disentangling motion and appearance for video generation. *arXiv preprint arXiv:1912.05523*, 2019.

[WBBD19b] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: This video does not exist. disentangling motion and appearance for video generation. *arXiv preprint arXiv:1912.05523*, 2019.

[WBBD20a] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Gˆ3an: Disentangling motion and appearance for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[WBBD20b] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[WBBD20c] Yaohui Wang, Piotr Bilinski, Francois F Bremond, and Antitza Dantcheva. ImaGINator: Conditional Spatio-Temporal GAN for Video Generation. In *WACV*, 2020.

[WBD21] Yaohui Wang, Francois Bremond, and Antitza Dantcheva. Inmodegan: Interpretable motion decomposition generative adversarial network for video generation. *arXiv preprint arXiv:2101.03049*, 2021.

[WD20] Yaohui Wang and Antitza Dantcheva. A video is worth more than 1000 lies. comparing 3d cnn approaches for detecting deepfakes. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 15, 2020.

[WDB17] Yaohui Wang, Antitza Dantcheva, and Francois Bremond. From attributes to faces: a conditional generative adversarial network for face generation. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2017.

[WDB18a] Yaohui Wang, Antitza Dantcheva, and Francois Bremond. From attribute-labels to faces: face generation using a conditional generative adversarial network. In *European Conference on Computer Vision Workshops (ECCVW)*, 2018.

[WDB+18b] Yaohui Wang, Antitza Dantcheva, Jean-Claude Broutart, Philippe Robert, Francois Bremond, and Piotr Bilinski. Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 144–157. Springer, 2018.

[YHS+19] Shikang Yu, Hu Han, Shiguan Shan, Antitza Dantcheva, and Xilin Chen. Improving face sketch recognition via adversarial sketch-photo transformation. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2019.

References

[1] Divyansh Aggarwal, Jiayu Zhou, and Anil K Jain. Fedface: Collaborative learning of face recognition model. *arXiv preprint arXiv:2104.03008*, 2021.

[2] L. Ballihi, A. Lablack, B. B. Amor, I. M. Bilasco, and M. Daoudi. Positive/negative emotion detection from rgb-d upper body images. In *Face and Facial Expression Recognition from Real World Videos*, pages 109–120. Springer, 2015.

[3] Sandipan Banerjee, Walter J. Scheirer, Kevin W. Bowyer, and Patrick J. Flynn. Fast face image synthesis with minimal training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2126–2136, 2019.

[4] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *arXiv preprint arXiv:2106.05963*, 2021.

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019.

[6] Woodrow Wilson Bledsoe. The model method in facial recognition. *Panoramic Research Inc., Palo Alto, CA, Rep. PR1*, 15(47):2, 1966.

[7] S. Brahnam, L. Nanni, and R. Sexton. Introduction to neonatal facial pain detection using common and advanced face classification techniques. In *ACIPH–1*, pages 225–253. Springer, 2007.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[10] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942, 2019.

[11] Elvan Çiftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. The turkish audio-visual bipolar disorder corpus. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.

[12] G. G. Cole, D. T. Smith, and M. A. Atkinson. Mental state attribution and the gaze cueing effect. *Attention, Perception, & Psychophysics*, 77(4):1105–1115, 2015.

[13] National Research Council et al. *Strengthening forensic science in the United States: a path forward*. National Academies Press, 2009.

[14] John Daugman. How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier, 2009.

[15] Debayan Deb. *Towards Robust and Secure Face Recognition: Defense against Physical and Digital Attacks*. PhD thesis, Michigan State University, 2021.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[17] Jonathan St BT Evans. *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc, 1989.

[18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[19] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[20] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3414–3424, 2021.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[22] Patrick Grother. Bias in face recognition: What does that even mean? and is it serious? In *Biometrics Congress*, 2017.

[23] Yaron Gurovich, Yair Hanani, Omri Bar, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter Krawitz, Susanne B Kamphausen, Martin Zenker, Lynne M Bird, et al. Deepgestalt-identifying rare genetic syndromes using deep learning. *arXiv preprint arXiv:1801.07637*, 2018.

[24] Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M Krawitz, Susanne B Kamphausen, Martin Zenker, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine*, 25(1):60–64, 2019.

[25] Zakia Hammal, Di Huang, Kévin Bailly, Liming Chen, and Mohamed Daoudi. Face and gesture analysis for health informatics. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 874–875, 2020.

[26] Harald Hampel, Richard Frank, Karl Broich, Stefan J Teipel, Russell G Katz, John Hardy, Karl Herholz, Arun LW Bokde, Frank Jessen, Yvonne C Hoessler, et al. Biomarkers for alzheimer's disease: academic, industry and regulatory perspectives. *Nature reviews Drug discovery*, 9(7):560–574, 2010.

[27] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *CVPR*, 2018.

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017.

[29] Minh Hoai and Fernando De la Torre. Max-margin early event detectors. *IJCV*, 107(2):191–202, 2014.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 2017.

[32] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Proc. of ICBA*, 2004.

[33] Anil Jain, Lin Hong, and Sharath Pankanti. Biometric identification. *Communications of the ACM*, 43(2):90–98, 2000.

[34] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Can soft biometric traits assist user recognition? In *Biometric technology for human identification*, volume 5404, pages 561–572. International Society for Optics and Photonics, 2004.

[35] Anil K Jain, Debayan Deb, and Joshua J Engelsma. Biometrics: Trust, but verify. *arXiv preprint arXiv:2105.06625*, 2021.

[36] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007.

[37] Anil K Jain and Stan Z Li. *Handbook of face recognition*, volume 1. Springer, 2011.

[38] Anil K Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern recognition letters*, 79:80–105, 2016.

[39] Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20, 2004.

[40] Yunseok Jang, Gunhee Kim, and Yale Song. Video Prediction with Appearance and Motion Conditions. In *ICML*, 2018.

[41] Anis Kacem, Zakia Hammal, Mohamed Daoudi, and Jeffrey Cohn. Detecting depression severity by interpretable representations of motion dynamics. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 739–745. IEEE, 2018.

[42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[44] Heysem Kaya and Albert Ali Salah. Eyes whisper depression: A cca based multimodal approach. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 961–964, 2014.

[45] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[46] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *TPAMI*, 38(1):14–29, 2016.

[47] Jean Kossaifi, Antoine Toisoul, Adrian Bulat, Yannis Panagakis, Timothy M Hospedales, and Maja Pantic. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6060–6069, 2020.

[48] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[49] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J. Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019.

[50] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[51] Yu Liu, Hongyang Li, and Xiaogang Wang. Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv preprint arXiv:1710.00870*, 2017.

[52] S. Majumder, E. Aghayi, M. Noferesti, H. Memarzadeh-Tehran, T. Mondal, Z. Pang, and M Deen. Smart homes for elderly healthcare—recent advances and research challenges. *Sensors*, 17(11):2496, 2017.

[53] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, et al. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.

[54] Sébastien Marcel, Mark S Nixon, Julian Fierrez, and Nicholas Evans. *Handbook of biometric anti-spoofing: Presentation attack detection*. Springer, 2019.

[55] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017.

[56] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep Multi-Scale Video Prediction Beyond Mean Square Error. In *ICLR*, 2016.

[57] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, 2016.

[58] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

[59] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. Biometrics recognition using deep learning: A survey. *arXiv preprint arXiv:1912.00271*, 2019.

[60] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[61] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[62] Anna Mitenkova, Jean Kossaifi, Yannis Panagakis, and Maja Pantic. Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[63] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

[64] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *ICLR*, 2018.

[65] Mark S Nixon, Paulo L Correia, Kamal Nasrollahi, Thomas B Moeslund, Abdenour Hadid, and Massimo Tistarelli. On soft biometrics. *Pattern Recognition Letters*, 68:218–230, 2015.

[66] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. *arXiv preprint arXiv:1610.09585*, 2016.

[67] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional Image Synthesis With Auxiliary Classifier GANs. In *ICML*, 2017.

[68] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, 2017.

[69] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *NIPS*, 2015.

[70] Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. Realistic dynamic facial textures from a single image using gans. In *ICCV*, pages 5429–5438, 2017.

[71] Naima Otberdout, Anis Kacem, Mohamed Daoudi, Lahoucine Ballihi, and Stefano Berretti. Automatic analysis of facial expressions based on deep covariance trajectories. *IEEE transactions on neural networks and learning systems*, 31(10):3892–3905, 2019.

[72] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. *arXiv preprint arXiv:1903.04480*, 2019.

[73] Raja Parasuraman and Dietrich H Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

[74] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[75] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[76] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible Conditional GANs for image editing. In *NIPS Workshop on Adversarial Training*, 2016.

[77] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Prof. of European Conference on Computer Vision (ECCV)*, pages 143–156, 2010.

[78] Silvia L Pintea, Jan C van Gemert, and Arnold WM Smeulders. Déja vu. In *ECCV*, 2014.

[79] Yunchen Pu, Shuyang Dai, Zhe Gan, Weiyao Wang, Guoyin Wang, Yizhe Zhang, Ricardo Henao, and Lawrence Carin. Jointgan: Multi-domain joint distribution learning with generative adversarial nets. *arXiv preprint arXiv:1806.02978*, 2018.

[80] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[81] Raghavendra Ramachandra and Christoph Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 50(1):1–37, 2017.

[82] P. Ramachandran, P. J. Liu, and Q. V. Le. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*, 2016.

[83] Fitsum A. Reda, Guilin Liu, Kevin J. Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. sdc-net: Video prediction using spatially-displaced convolution.

[84] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *ECCV*, 2018.

[85] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[86] Henry Taylor Fowkes Rhodes. *Alphonse Bertillon, father of scientific detection*. Greenwood, 1968.

[87] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Summary for avec 2018: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2111–2112, 2018.

[88] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9. ACM, 2017.

[89] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *ICCV Workshops*, 2019.

[90] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[91] A Ross, A Jain, and K Nandakumar. Introduction to biometrics: A textbook, 2011.

[92] Arun Ross, Sudipta Banerjee, Cunjian Chen, Anurag Chowdhury, Vahid Mirjalili, Renu Sharma, Thomas Swearingen, and Shivangi Yadav. Some research problems in biometrics: The future beckons. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[93] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971*, 2019.

[94] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vision*, 126(2-4):144–157, April 2018.

[95] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011.

[96] Alexander Sage, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Logo Synthesis and Manipulation with Clustered Generative Adversarial Networks. In *CVPR*, 2018.

[97] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.

[98] Albert Ali Salah. Designing computational tools for behavioral and clinical science. In *Companion of the 2021 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pages 1–4, 2021.

[99] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

[100] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In *NIPS*. 2016.

[101] 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.

[102] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, 26(4):1965–1978, 2017.

[103] Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.

[104] Ram Sewak Sharma. *The Making of Aadhaar: World's Largest Identity Platform*. Rupa, 2020.

[105] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

[106] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis. In *CVPR*, 2018.

[107] O. R. Shishvan, D.-S. Zois, and T. Soyata. Machine intelligence in healthcare and medical cyber physical systems: A survey. *IEEE Access*, 6:46419–46494, 2018.

[108] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *ICCV*, 2019.

[109] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS*. 2014.

[110] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019.

[111] Yale Song, David Demirdjian, and Randall Davis. Tracking Body and Hands For Gesture Recognition: NATOPS Aircraft Handling Signals Database. In *FG*, 2011.

[112] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. Technical report, CRCV-TR-12-01, November 2012.

[113] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015.

[114] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016.

[115] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[116] Carly CG Sweegers, Atsuko Takashima, Guillén Fernández, and Lucia M Talamini. Neural mechanisms supporting the extraction of general knowledge across episodic memories. *Neuroimage*, 87:138–146, 2014.

[117] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[118] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[119] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[120] J. Thevenot, M. B. Lopez, and A. Hadid. A survey on computer vision for assistive medical diagnosis from faces. *IEEE J of BHI*, 22(5):1497–1511, 2018.

[121] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016.

[122] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1320–1328, January 2021.

[123] John Ronald Reuel Tolkien. *The Lord of the Rings: One Volume*. Houghton Mifflin Harcourt, 2012.

[124] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[125] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[126] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[127] M. Tsiknakis, L. Koumakis, M. Karachaliou, S. Voutoufianakis, and P. Vorgia. Vision-based absence seizure detection. In *Proc. Eng. Med. Biol. Soc.*, pages 65–68. IEEE, 2012.

[128] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018.

[129] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[130] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[131] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, 2017.

[132] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017.

[133] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *arXiv preprint arXiv:1704.05831*, 2017.

[134] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[135] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NIPS*, 2016.

[136] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Variable-state latent conditional random field models for facial expression analysis. *Image and Vision Computing*, 58:25 – 37, 2017.

[137] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, 2016.

[138] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, 2014.

[139] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017.

[140] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[141] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. Research Report RR-8050, INRIA, August 2012.

[142] J.-G. Wang, J. Li, W.-Y. Yau, and E. Sung. Boosting dense sift descriptors and shape contexts of face images for gender recognition. In *Proc. of IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 96–102, 2010.

[143] Le Wang, Jinliang Zang, Qilin Zhang, Zhenxing Niu, Gang Hua, and Nanning Zheng. Action recognition by an attention-aware temporal weighted convolutional neural network. *Sensors*, 18(7):1979, 2018.

[144] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *CoRR*, abs/1507.02159, 2015.

[145] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

[146] P. Wang, C. Kohler, and R. Verma. Estimating cluster overlap on manifolds and its application to neuropsychiatric disorders. In *CVPR*, pages 1–6. IEEE, 2007.

[147] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019.

[148] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.

[149] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[150] Nevan wichers, Ruben Villegas, Dumitru Erhan, and Honglak Lee. Hierarchical long-term video prediction without supervision. In *ICML*, 2018.

[151] Yujin Wu, Mohamed Daoudi, Ali Amad, Laurent Sparrow, and Fabien D'Hondt. Unsupervised learning method for exploring students' mental stress in medical simulation training. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 165–170, 2020.

[152] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[153] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*, 2017.