Facial Landmark Detection with Local and Global Motion Modeling

Romain Belmonte

Université de Lille

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

Doctoral Dissertation

Computer Science and Applications

Facial Landmark Detection with Local and Global Motion Modeling

Romain Belmonte

Jury

Pr. Jean-Luc Dugelay	EURECOM	reviewer
Pr. Hichem Sahbi	Sorbonne University	reviewer
Pr. Nathalie Rolland	University of Lille	president
Pr. Karine Zeitouni	University of Versailles	examinator
Pr. Nicu Sebe	University of Trento	examinator
Pr. Chaabane Djeraba	University of Lille	supervisor
Dr. Nacim Ihaddadene	ISEN-Lille / Yncréa	co-supervisor
Dr. Pierre Tirilly	University of Lille	co-supervisor
Dr. Ioan Marius Bilasco	University of Lille	co-supervisor

Publicly defended at IRCICA on 11/26/2019

Université de Lille

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

Thèse de doctorat

Informatique et applications

Détection des points caractéristiques du visage par modélisation des mouvements locaux et globaux

Romain Belmonte

Jury

Pr. Jean-Luc Dugelay	EURECOM	rapporteur
Pr. Hichem Sahbi	Sorbonne Université	rapporteur
Pr. Nathalie Rolland	Université de Lille	président
Pr. Karine Zeitouni	Université de Versailles	examinateur
Pr. Nicu Sebe	Université de Trente	examinateur
Pr. Chaabane Djeraba	Université de Lille	directeur de thèse
Dr. Nacim Ihaddadene	ISEN-Lille / Yncréa	co-encadrant
Dr. Pierre Tirilly	Université de Lille	co-encadrant
Dr. Ioan Marius Bilasco	Université de Lille	co-encadrant

soutenue publiquement à l'IRCICA le 26/11/2019

Context

This thesis has been co-financed by the Institut Supérieur de l'Electronique et du Numérique de Lille (ISEN-Lille/Yncréa; Higher Institute for Electronics and Digital Training) and the Métropole Européenne de Lille (MEL; Lille European Metropolis).

The research was conducted at the Institut de Recherche sur les Composants logiciels et matériels pour l'Information et la Communication Avancée (IRCICA; Research Institute on software and hardware devices for information and Advanced communication) as well as at the Maison Intelligence (Smart Home), an R&D platform for experimentation and demonstration with a focus on ubiquitous and information and communication technologies.

Acknowledgements

First of all, I would like to thank my supervisor Chaabane Djeraba who trusted me to complete this thesis and who always remained very available. I also thank my co-supervisors Pierre Tirilly, Ioan Marius Bilasco, and Nacim Ihaddadene. I am sincerely grateful for the time they spent with me, their patience, and the many advices they gave me. These years as a Ph.D. student have been rich in many ways.

I would like to thank Jean-Luc Dugelay and Hichem Sahby for agreeing to be the reviewers of this work. I also thank Nathalie Rolland who gave me the honour of chairing the jury, as well as Karine Zeitouni and Nicu Sebe who agreed to be the examinators. I greatly appreciated the jury's comments and questions, which allowed me to take a new look at my work.

I would like to thank the FOX team for their hospitality. I am particularly grateful to Jean Martinet, José Mennesson and Benjamin Allaert for their help and support. My thanks also go to Delphine Poux and Veïs Oudjail who have always been ready to help me and with whom I have enjoyed sharing on each other's work.

Also, I would like to thank all the people at ISEN-Lille, especially the Department of Computer Science & Applied Mathematics and the ADICODE "Urbawood" who have always been very kind and supportive.

Many thanks to everyone I have been able to interact with during these years, colleagues and students, at the University of Lille and the Catholic University of Lille, especially at CRIStAL, IRCICA, ISEN, and IUT A, which made this experience even more valuable.

Finally, I would like to thank my family and friends for their constant support.

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse Chaabane Djeraba de m'avoir fait confiance pour la réalisation de cette thèse et d'avoir toujours fait preuve d'une grande disponibilité. Je remercie également mes co-encadrants Pierre Tirilly, Ioan Marius Bilasco, et Nacim Ihaddadene. Je leur suis sincèrement reconnaissant pour le temps qu'ils m'ont consacré, leur patience, et les nombreux conseils qu'ils ont pu me prodiguer. Ces années de doctorat ont été enrichisssantes à bien des égards.

Je tiens ensuite à remercier Jean-Luc Dugelay et Hichem Sahby d'avoir accepté d'être les rapporteurs de ces travaux. Je remercie également Nathalie Rolland qui m'a fait l'honneur de présider le jury, ainsi que Karine Zeitouni et Nicu Sebe qui ont accepté d'être les examinateurs. J'ai beaucoup apprécié les remarques et questions du jury qui m'ont permis de porter un regard nouveau sur mon travail.

J'adresse mes remerciements à l'ensemble de l'équipe FOX pour leur accueil. Je suis particulièrement reconnaissant envers Jean Martinet, José Mennesson et Benjamin Allaert pour leur aide et soutien. Mes remerciements vont également à Delphine Poux et Veïs Oudjail qui ont toujours été prêts à me rendre service et avec qui j'ai aimé échanger sur les travaux des uns et des autres.

Je tiens aussi à remercier l'ensemble du personnel de l'ISEN-Lille, en particulier le département Informatique & Mathématiques Appliquées et les ADICODE "Urbawood" qui m'ont accompagné avec beaucoup de gentillesse et m'ont constamment encouragé.

Un grand merci à tout ceux que j'ai pu cotoyer durant ces années, collègues et étudiants, à l'Université de Lille et à l'Université Catholique de Lille, notamment au CRIStAL, à l'IRCICA, à l'ISEN, et à l'IUT A, et qui ont rendu cette expérience encore plus enrichissante.

Mes remerciements vont enfin à mes proches, famille et amis, pour leur soutien inestimable.

Abstract

Facial landmark detection is an essential task for a large number of applications such as facial analysis (e.g., identification, expression, 3D reconstruction), human-computer interaction or even multimedia (e.g., content indexing and retrieval). Although many approaches have been proposed, performance under uncontrolled conditions is still not satisfactory. The variations that may impact facial appearance (e.g., pose, expression, illumination, occlusion, motion blur) make it a difficult problem to solve. In this thesis, a contribution to both the analysis of the performance of current approaches and the modeling of temporal information for video-based facial landmark detection is made. An experimental study is conducted using a video dataset to measure the impact of pose and expression variations on landmark detection. This evaluation highlights the most difficult poses and expressions to handle. It also illustrates the importance of a suitable temporal modeling to benefit from the dynamic nature of the face. A focus is then placed on improving temporal modeling to ensure consideration of local motion in addition to global motion. Several architectures are designed based on the two main models from the literature: coordinate regression networks and heatmap regression networks. Experiments on two datasets confirm that local motion modeling improves results (e.g. in the presence of expressions). These experiments are extended with a study on the complementarity between spatial and temporal information as well as local and global motion to improve the design of the proposed architectures. By leveraging these complementarities more effectively, competitive performance with current state-of-the-art approaches is achieved, despite the simplicity of the proposed models.

Keywords: computer vision; deep learning; facial analysis; landmark detection.

Résumé

La détection des points caractéristiques du visage est une tâche essentielle pour un grand nombre d'applications telles que l'analyse faciale (p. ex., identification, expression, reconstruction 3D), l'interaction homme-machine ou encore le multimédia (p. ex., recherche, indexation). Bien que de nombreuses approches aient été proposées, les performances en conditions non contrôlées ne sont toujours pas satisfaisantes. Les variations susceptibles d'impacter l'apparence du visage (p. ex., pose, expression, éclairage, occultation, flou cinétique) en font un problème encore difficile à résoudre. Dans cette thèse, une contribution est faite à la fois sur l'analyse des performances des approches actuelles mais aussi sur la modélisation de l'information temporelle pour la détection des points caractéristiques du visage basée sur la vidéo. Une étude expérimentale est réalisée à l'aide d'un jeu de données vidéo permettant d'évaluer l'impact des variations de pose et d'expression sur la détection des points caractéristiques. Cette évaluation permet notamment de mettre en évidence les poses et expressions posant le plus de difficultés. Elle permet également d'illustrer l'importance d'une modélisation temporelle capable de tenir compte efficacement de la nature dynamique du visage. L'accent est ensuite mis sur l'amélioration de la modélisation temporelle afin de considérer le mouvement local en plus du mouvement global. Plusieurs architectures sont conçues en s'appuyant sur les deux principaux modèles de la littérature : les réseaux de régression de coordonnées et les réseaux de régression de cartes de chaleur. Les expérimentations sur deux ensembles de données confirment que la modélisation du mouvement local améliore les résultats (p. ex. avec les expressions). Ces expérimentations sont étendues par une étude portant sur la complémentarité entre l'information spatiale et temporelle ainsi que le mouvement local et global dans le but d'améliorer la conception des architectures proposées. En exploitant davantage ces complémentarités, de meilleures performances, compétitives avec l'état de l'art, sont obtenues, et ce, malgré la simplicité des modèles proposés.

Mots-clés : vision par ordinateur; apprentissage profond; analyse faciale; détection de points caractéristiques.

Contents

List of Figures

List of Tables

List of Abbreviations

AAMs . . . . . Active Appearance Models

AL . . . . . Adversarial Learning

ANNs . . . . . Artifical Neural Networks

ASMs . . . . . Active Shape Models

AUC . . . . . Area Under the Curve

BoFs . . . . . Bag of Features

BPTT . . . . . BackPropagation Through Time

CED . . . . . Cumulative Error Distribution

CLMs . . . . . Constrained Local Models

CNNs . . . . . Convolutional Neural Networks

CSR . . . . . Cascaded Shape Regression

DL . . . . . Deep Learning

DNNs . . . . . Deep Neural Networks

DR . . . . . Direct Regression

DRL . . . . . Deep Reinforcement Learning

ELS . . . . . Exhaustive Local Search

FCN . . . . . Fully Convolutional Network

FVs . . . . . Fisher Vectors

FR . . . . . Failure Rate

GANs . . . . . Generative Adversarial Networks

GPA . . . . . Generalized Procrustes Analysis

GRU . . . . . Gated Recurrent Unit

HG . . . . . HourGlass

HOG . . . . . Histogram of Oriented Gradients

HSM . . . . . Hard Sample Mining

KLT . . . . . Kanade–Lucas–Tomasi

KDE . . . . . Kernel Density Estimate

LR . . . . . Learning Rate

LSTM . . . . . Long Short-Term Memory

ML . . . . . Machine Learning

MSE . . . . . Mean Square Error

RMSE . . . . . Root Mean Square Error

NRMSE . . . . Normalized Root Mean Square Error

PCA . . . . . Principal Component Analysis

PDM . . . . . Point Distribution Model

PSs . . . . . Pictorial Structures

RANSAC . . . RANdom SAmple Consensus

RNNs . . . . . Recurrent Neural Networks

SDM . . . . . Supervised Descend Method

SGD . . . . . Stochastic Gradient Descent

SIFT . . . . . Scale-Invariant Feature Transform

SURF . . . . . Speeded Up Robust Features

SVMs . . . . . Support Vector Machines

3DMM . . . . 3D Morphable Model

Chapter 1

Introduction

Contents

1.1 Background

Over recent years, the ubiquity of sensors (i.e., smartphones, computers, in public spaces) has led to an explosion of data, especially visual data. Each day, massive amounts of visual data are produced. Every minute, 500 hours of video and 66,000 images are uploaded on YouTube and Instagram [1]. Processing this amount of data is currently beyond the reach of humans and the quantity of data continues to increase. According to CISCO [2], video will represent 80% of all Internet traffic by 2021 [3]. Hence, it is critical to develop algorithms capable of understanding visual data, which is the purpose of computer vision. One of the most popular research topics within computer vision is human behaviour understanding since visual data represents a considerable source of information about people and their behaviour. The face provides the ability to recognize people, estimate their age, gender, and their emotional state. Hence, a large amount of research is focused on the automatic analysis of facial images, which aims to extract such information. Applications cover a wide variety of domains, e.g., education, transport, entertainment, security, surveillance, or medicine, just to name a few. Concrete exemples are engagement recognition in e-learning environments to improve teacher-student interaction and stimulate learning [82], driver inattention or drowsiness detection to prevent potential accidents [177], or mental health assessment for objective diagnosis [289]. Apart from the use of affective states, facial analysis also allows image editing, which can provide entertainment on social networks or improve the customer experience in retail [151].

Figure 1.1: Facial landmark detection [193]. The structure of the face is identified and its different components are characterized. This is achieved through the detection of facial landmarks, which are usually located around the eyes, nose and mouth.

[1] https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/

[2] CISCO is a leading provider of communications and Internet solutions and services.

[3] https://newsroom.cisco.com/press-release-content?type=webcontent&articleId=1853168

Facial landmark detection, also known as face alignment, is a fundamental task in facial analysis (e.g., face recognition, facial expression recognition, 3D face reconstruction, eye gaz tracking, face frontalization). As shown in Figure 1.1, this task defines the facial geometry, so that the structure of the face can be identified and its different components characterized. Facial landmark detection is useful for most facial analysis tasks and more broadly in human computer interaction, animation, video compression, image super resolution, indexing and retrieval, etc.

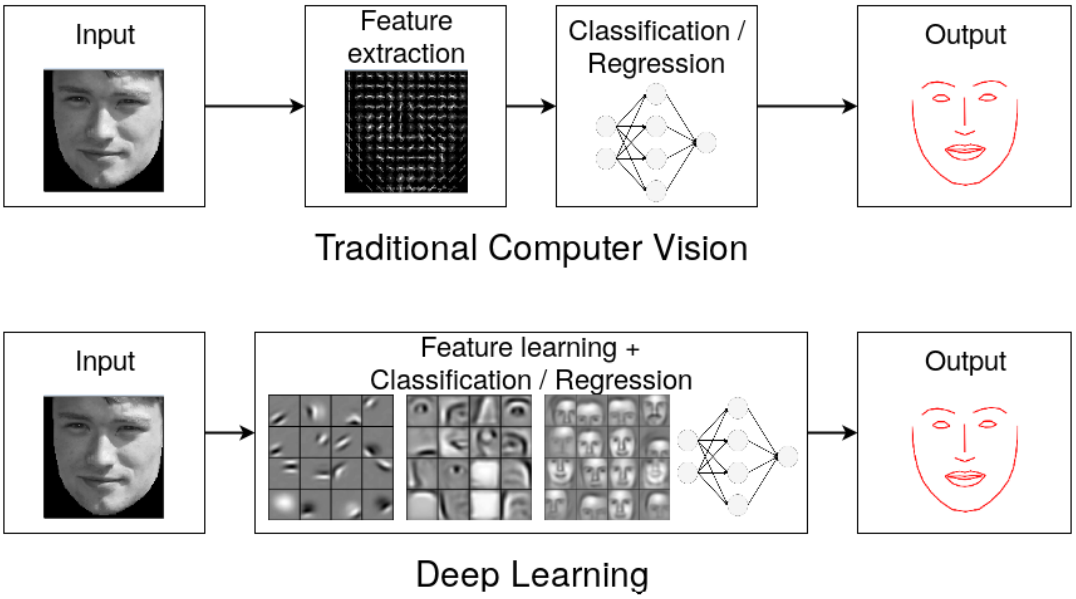

Figure 1.2: Comparison between traditional machine learning algorithms and deep learning algorithms. In the latter, hierarchical features are learned instead of using handcrafted ones such as histograms of oriented gradients [9, 126].

Identifying the facial structure is trivial for humans. From the perspective of an algorithm, an image is represented by an array of pixels. From these pixels, the aim is to provide enough abstraction to reach a semantic level that allows facial landmarks to be retrieved. However, it is difficult for an algorithm to have this kind of high level of understanding. To overcome this semantic gap, two solutions have been proposed (see Figure 1.2). The first one, the traditional approach, uses domain-specific knowledge to transform raw data into features (i.e., feature engineering). These features provide useful input for the prediction task, performed by Machine Learning (ML) algorithms. The second one, the Deep Learning (DL) approach, instead of using handcrafted features, lets the algorithm discover the features (i.e., feature learning) needed for the prediction task directly from the raw data. DL involves a series of computational layers, which generally extract low level features such as edges and curves, up to more abstract concepts. Such models are called Deep Neural Networks (DNNs). Thanks to the increase

in computational power and available data, DNNs have shown over the past few years impressive capabilities. They have completely disrupted the field of computer vision [122]. In most tasks, the community has shifted from feature engineering to DNNs architecture engineering.

1.2 Problem Formulation

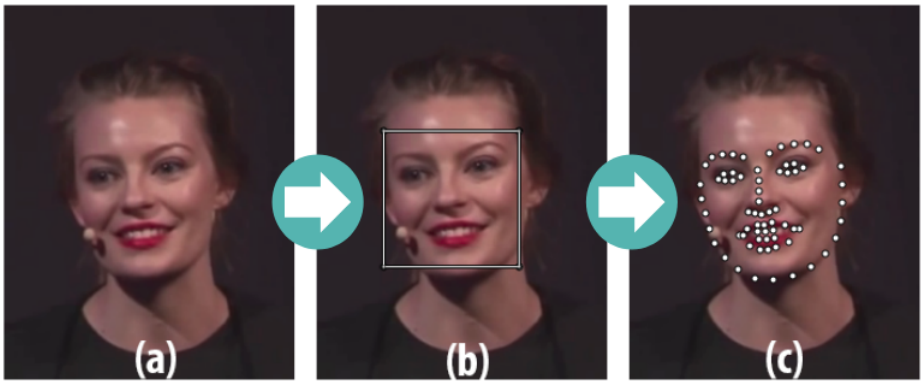

Figure 1.3: Facial landmark detection process: (a) original image, (b) face detection, and (c) landmark detection.

Given the position and size of face, the alignment process, illustrated in Figure 1.3 consists in modeling non-rigid facial structures. It can be done by identifying facial landmarks, which are usually located around the eyes, nose, and mouth. Landmarks correspond to corners (e.g., eyes, mouth) or edges that connect these corners. From the landmarks, it is then easier to remove transformations, using for example Generalized Procrustes Analysis (GPA) [75], to achieve the alignment of two or more faces. Work on face alignment is generally focused essentially on the detection of landmarks. The detection can be done in 2D, which means that each landmark is represented by Cartesian coordinates (x, y) on the 2D plane. Note that 3D detection, which involves an additional coordinate z corresponding to the depth, is also possible but is beyond the scope of this work. Three steps are generally necessary but not mandatory to perform landmark detection: face preprocessing, shape initialization, and an iterative process of feature extraction and shape prediction.

The purpose of this work is to model the relationship between an image or image sequence I and a facial shape $\hat{S}$. The mapping function can be expressed as follows:

$$M : I \mapsto \hat{S} \tag{1.1}$$

Two approaches can be found in the literature to achieve this mapping. They either regress the coordinates directly or compute heatmaps, one for each landmark. Due to their effectiveness and versatility, Convolutional Neural Networks (CNNs) are often used to learn these mappings in a supervised manner. The training uses N images annotated with the corresponding shapes $(I_1, S_1), ..., (I_N, S_N)$ where I can be a C-channel image $\in \mathbb{R}^{H \times W \times C}$ as well as a C-channel image sequence $\in \mathbb{R}^{T \times H \times W \times C}$ with H and W the height and width of the image and T the temporal window. Mean Squared Error (MSE) is used to measure the difference between the prediction and the associated ground truth over n samples. The MSE cost function is minimized according to the approach.

Coordinate Regression. In the case of coordinate regression, the ground truth is a vector $s = [x_1, y_1, x_2, y_2, ..., x_L, y_L]^T$ with $s \in \mathbb{R}^{2L}$, L the number of landmarks and x, y the Cartesian coordinates of a landmark. The MSE is:

$$\text{MSE}_{\text{coordinates}} = \frac{1}{n}\sum_{i=1}^{n}||s_i - \hat{s}_i||_2^2 \tag{1.2}$$

with $\hat{s} = \hat{S} = f(I)$, and $||.||_2$ the Euclidean norm.

Heatmap Regression. For heatmap regression, the ground truth $h \in \mathbb{R}^{L \times H \times W}$ is a set of L heatmaps of dimension $H \times W$ in which the coordinates of the maximum value correspond to the coordinates of a landmark. The MSE is:

$$\text{MSE}_{\text{heatmaps}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L}||h_{ij} - \hat{h}_{ij}||_F^2 \tag{1.3}$$

with $||.||_F$ the Frobenius norm.

1.3 Major Challenges

Under favorable conditions, a static frontal face with sufficient uniform lighting and no occlusion, current landmark detection approaches perform fairly well, with accuracy rates close to human performance. However, such conditions are unrealistic. When used in practical applications, the environment is generally not controlled and rarely at the advantage of the algorithm. Facial landmark detection depends on the face detection, so the quality of detection impacts landmark detection. Many variations may affect the appearance of the face. They can be related to the intrinsic dynamics of the face, the environment, and the capture conditions. The major challenges are described and illustrated (see Figure 1.4) below:

- Pose: Pose variations lead to significant changes in the appearance of the face. The difference between a frontal face and a profile face is considerable. The latter leads to self-occlusions causing a partial or complete disappearance of certain components. Moreover, some landmarks overlap each other.

- Expression: Facial expressions produce deformations on most facial components. So, significant changes in appearance (e.g., mouth open and closed) are noted, with landmarks that may overlap each other.

- Occlusion: Occlusions are very common. They can be related to objects (e.g., glasses, hat, make-up), or self-occlusions due to facial pilosity, hair, or hands. They lead to a more or less important loss of information.

- Illumination: Depending on its type of source and its intensity (i.e., non-uniform, low intensity), insufficient lighting causes the loss of certain details, or even make certain parts of the face disappear.

- Motion blur: Motion blur is the consequence of fast motion of the subject or of the camera. It causes trails in the direction of the motion with a loss of sharpness and visual details.

- Resolution: Due to some constraints, the capture may not be performed in suitable resolutions. The level of detail is then reduced. So, it becomes difficult to clearly distinguish the different components of the face.

Figure 1.4: Major challenges encountered under uncontrolled conditions: (a) occlusion, (b) pose, (c) illumination, (d) resolution, (e) motion blur, (f) expression [193].

The design of robust algorithms to tackle these different challenges is necessary. It is also important to consider the constraints of the applications. They typically include the ability to run in real time and limited resources, especially embedded applications (e.g., smartphone). Finding a trade-off between accuracy and efficiency is often unavoidable.

1.4 Motivation

A large number of methods have been proposed to achieve robust detection in such scenarios. Despite considerable progress in recent years, the performance of facial landmark detection under uncontrolled conditions is still not fully satisfactory. The theoretical minimum error achievable on one of the most popular dataset in the literature is still far from being reached [184]. Some landmarks (e.g., boundary of the face, mouth) are still difficult to detect accurately due possibly to a lack of discriminative information or ambiguity problems. Moreover, the degree of satisfaction of landmark detection is generally based on its overall accuracy (distance between the prediction and the ground truth), and does not really take into account the requirements of the applications (e.g., facial expression recognition, 3D face reconstruction). Many approaches have been developed with an emphasis on few challenges. Although they improve robustness, there is still a lack of approaches that can address all challenges at once. Besides, the majority of applications are built on video. Yet, most of the current approaches, when used in these applications, implement a tracking-by-detection strategy and are therefore unable to take advantage of the temporal coherence. We believe that the key to achieve such an approach could be to leverage the dynamic nature of the face.

Video-based approaches have recently generated interest [193] and have already shown their benefits [162]. They provide more robustness under uncontrolled conditions and have the potential to overcome most of the difficulties inherent to these conditions. One of the main reasons to include an additional dimension to the problem is to leverage the temporal coherence. However, temporal information and more specifically facial dynamics are currently under-exploited. The facial dynamics presents complex motion that can be observed at different scales. Facial dynamics includes global motion related to head movements and local motion, such as eye or lip movements, caused by facial deformations, for instance related to expressions. The approaches proposed so far are limited since they are unable to properly include local motion, which is important for accurate detections. In addition, they do not make a sufficient use of the possible synergy between motion, appearance, and shape.

Beyond the algorithms, many datasets have been made available. These datasets propose more and more complex data with multiple severe challenges. Data generally comes from the Web, which represents a seemingly unlimited source of diversified data captured under uncontrolled conditions. Until now, the focus of the dataset was static data. Efforts have been made to provide datasets to compare performance in controlled and uncontrolled conditions. However, the lack of contextual annotations that describes the acquisition conditions limits the level of analysis. Current datasets make it difficult to conduct comprehensive analyses, such as quantifying the impact of the different challenges on facial landmark detection. Yet, such analyses are crucial to effectively estimate the contribution of current approaches and to get relevant insights on how to improve these approaches. They may particularly highlight the role of temporal information under uncontrolled conditions.

1.5 Contributions

The work of this thesis is structured around the following objectives, each of which resulted in contributions:

- **providing a better analysis of current approaches**: A study is conducted using a recent, richly annotated, video dataset. It quantifies the impact of two major challenges, variations of pose and expression, on facial landmark detection. This analysis is performed at different levels: the overall shape and the landmarks, to identify the most critical situations and facial regions with regard to these challenges. In addition, recommendations are formulated, including the use of temporal information as a solution to address these challenges, all at once.

- **improving spatio-temporal modeling**: Local motion modeling is integrated into the two main models of the literature (i.e., coordinate and heatmap regression networks) through direct pixel connectivity. Experiments validate the assumption that local motion modeling improves performance, especially during complex localized motion such as expressions.

- **studying the relationships between spatial and temporal information, and local and global motion**: An investigation of DNN design is carried out and lead to better use of the complementarity between spatial and temporal information and local and global motion. This results in an improvement of the proposed architectures and additional performance gains.

In summary, the work in this thesis shows the importance and benefits of spatio-temporal modeling for facial landmark detection. Capturing facial dynamics provides valuable information in uncontrolled conditions, especially during pose and expression variations.

1.6 Outline

This dissertation is structured in three main parts. In Chapter 2, existing work on facial landmark detection is reviewed. Image-based approaches are first discussed. This mainly includes generative and discriminative approaches, which represent the two main ways to address this problem. Given the prominence of DL today, a particular attention is given to this type of approaches. Finally, various solutions are discussed. They have been proposed to specifically address one of the challenges that may be encountered under uncontrolled conditions (i.e., head pose, occlusions, expressions, or others). The focus then switches to video-based approaches which, despite being less popular, are attracting more and more attention. These approaches are organized into five categories. The most naive approaches, such as tracking by detection or box, landmark or pose tracking, are first presented. More sophisticated approaches, such as adaptive approaches, joint detection and tracking, and constrained approaches, are then featured.

Chapter 3 examines the effectiveness of current facial landmark detection approaches. Datasets and evaluation protocols are first presented. Image and video datasets captured under uncontrolled conditions are described. The pre-processing and augmentation techniques applied to the raw data are detailed, as well as the most commonly used evaluation metrics. Two benchmarks, image and video, are then conducted using a compilation of the results from the literature. In this way, a large number of approaches can be discussed. However, this does not allow to quantify the impact of the various challenges encountered under uncontrolled conditions. To go further, a comprehensive analysis is conducted to quantify the impact of poses and expressions on facial landmark detection. The analysis is performed up to the level of individual landmarks. It provides useful insights on how to further contribute towards solving the alignment problem, and highlights the key role of temporal information in addressing challenges such as head pose and expression variations.

Afterwards, efforts are concentrated on spatio-temporal modeling. After having identified the limitations of current video-based approaches, a solution to overcome these limitations is proposed in Chapter 4. It consists in improving the way temporal information is modeled. The main objective is to capture more subtle motion and not just

global motion related to head movements. For this purpose, existing spatio-temporal modelling solutions are first reviewed. Both handcrafted and DL approaches are considered. Once identified, the most relevant solutions for the problem, especially 3D convolutions, are used to propose new architectures for video-based landmark detection. Experiments on two datasets, SNaP-2DFe [2] and 300VW [193], are conducted. The results confirm the benefits of the proposed approach. To go further, questions about the complementarity between spatial and temporal information and local and global motion are investigated. Improvements to the proposed architectures are identified.

Chapter 2

Facial Landmark Detection

Contents

As stated in [99], the face corresponds to a deformable object that can vary in terms of shape and appearance. The first attempts to facial landmark detection can be traced back to the 1990s. The Active Shape Models (ASMs) [38] represents one of the seminal works on the subject. Such a generative approach consists of a parametric model that can be fitted to a given face by optimizing its parameters. This process of applying the model to new data is called inference. Rapid progress has been made through the development of Active Appearance Models (AAMs) [37], Constrained Local Models (CLMs) [40], and other extensions [216, 4, 9, 11], to the point where the problem is now considered well addressed for constrained faces [99]. The research has therefore shifted to unconstrained faces with multiple and complex challenges, e.g., occlusion, variations in pose, illumination, and expression. Faster and more robust discriminative methods such as Cascaded Shape Regression (CSR) [53] have been proposed to address these challenges. They differ from generative approaches as they directly learn a mapping function between images and facial shapes with better generalization ability. Due in particular to the phenomenon of data massification and the increase in the computational capacity of machines, a subset of these approaches, DL, has stood out. These approaches need larger datasets than traditional ML algorithms but do not require feature engineering. Above all, they currently outperform all previous approaches and most research now focuses on them.

Consequently, two main categories of approaches can be defined, generative and discriminative, not to mention the complementary solutions that can be applied to most of these approaches, e.g., multi-task learning, to explicitly address some selected challenges. The proposed nomenclature is close to the ones proposed in [240, 99]. Considering its wide range of applications and the persistent difficulties encountered under uncontrolled conditions, facial landmark localization remains a very active area of research. Recently, there has been a trend towards video-based solutions [193]. One of the main reasons to include this additional dimension to the problem is that temporal consistency provides a useful input to achieve robust detection under uncontrolled conditions. Different strategies have been proposed, from the most traditional tracking strategies, such as tracking by detection, to more complex strategies that can jointly extract features and perform tracking. Current work is mostly tracking-oriented and focuses on global head movements.

All these items are discussed in detail in the following sections. In Section 2.1, the groundbreaking work and advances in the two main categories of approaches are reviewed, with an emphasis on DL approaches. This allows to properly describe, in Section 2.1.4, the complementary solutions that can be used to handle major challenges.

In contrast to existing literature reviews [240, 99], the different strategies proposed in the literature to extend facial landmark localization to video are extensively discussed in Section 2.2. Finally, Section 2.3 concludes with a positioning of the work presented in this thesis.

2.1 Facial Landmark Detection in Still Images

Efforts to tackle the problem of facial landmark detection have long focused on still images and much work has been published. In Section 2.1.1 and 2.1.2, the two main categories of approaches, generative and discriminative, are detailed as well as their developments. Among the discriminative approaches, deep learning approaches have become very popular. Therefore, the latter is given close attention in Section 2.1.3. Finally, complementary solutions to handle the difficulties encountered under uncontrolled conditions are reviewed in Section 2.1.4.

2.1.1 Generative Approaches

Generative methods typically build a statistical model for both shape and appearance. They are referred to as generative approaches since they provide a model of the joint probability distribution of the shape and appearance. These models are parametric and therefore have some degrees of freedom constrained during training so they are expected to match only possible facial configurations. By optimizing the model parameters, e.g., by minimizing the reconstruction error for a given image, the best possible instance of the model for that image can be generated. Initially only shape variations were modelled [38]. But, on the same line of thought, texture variations were also added to jointly apply constraints to variations in shape and texture [37]. Facial appearance can be represented in different ways. The entire face can be considered, which constitutes a holistic representation. The face can also be decomposed into different parts, e.g., patches centered at each landmark, which is known as a part-based representation. Generative approaches provide a good fitting accuracy with little training data but are sensitive to shape initialization and do not generalize well to new images. The optimization can be difficult under uncontrolled conditions due to the high dimensionality of the appearance space. They tend to get trapped in local minima as well. Note also that these methods are mainly based on Principal Component Analysis (PCA) [161] and therefore the distribution is assumed to be Gaussian, which does not fit the true distribution.

The AAMs [37] is probably the most representative method in this category and the most widely studied. In the following, AAMs modeling and fitting are presented as well

as some of their improvements. The modeling can be split into 3 parts: a shape model, an appearance model, and a motion model. The shape model, also called Point Distribution Model (PDM) [38], is common among deformable models [38, 37, 40, 9, 216]. It is built using the N training facial shapes. These shapes are first normalized (aligned) using a GPA [75] to remove affine transformations (i.e., rotation, scale, and translation variations) and keep only local, non-rigid shape deformation. Next, PCA is applied to obtain a set of orthogonal bases that capture the maximal variance. Only the n eigenvectors corresponding to the largest eigenvalues are kept as they summarize well the data. This reduces the dimensions while ensuring that the key information is maintained. The model can then be expressed as:

$$s_p = \overline{s} + U_s p \tag{2.1}$$

where $\overline{s} \in \mathbb{R}^{2L,1}$, $U_s \in \mathbb{R}^{2L,n}$, and $p \in \mathbb{R}^n$ are respectively the mean shape, the shape eigenvectors, and the vector of shape parameters. Four eigenvectors corresponding to the similarity transforms (scaling, in-plane rotation, and translation) are added to U_s (re-orthonormalization) to be able to fit the model on any image [145].

To build the appearance model, features from the N training images are first extracted using the function F. Initially, holistic pixel-based representation was used, which is sensitive to lighting and occlusions [37]. To improve robustness, feature-based representations such as Histogram of Oriented Gradients (HOG) [44] or Scale-Invariant Feature Transform (SIFT) [136] can also be used. In this way, relevant facial features are extracted, with a better ability to generalize to new faces. These features are then warped into the reference shape using the motion model. PCA is finally applied onto these vectorized warped feature-based images. The model can be expressed as:

$$a_q = \overline{a} + U_a q \tag{2.2}$$

where $\overline{a} \in \mathbb{R}^{M,1}$, $U_a \in \mathbb{R}^{M,m}$ and $q \in \mathbb{R}^m$ are respectively the mean appearance vector, the appearance eigenvectors, and the vector of appearance parameters.

The motion model refers typically to a warp function W such as a piece-wise affine warp or a thin-plate spline [145, 218]. Given a shape generated from parameters p, this function defines how to warp the texture into a reference shape, e.g., the mean shape $\overline{s}$. Figure 2.1 shows an example of AAMs instantiation.

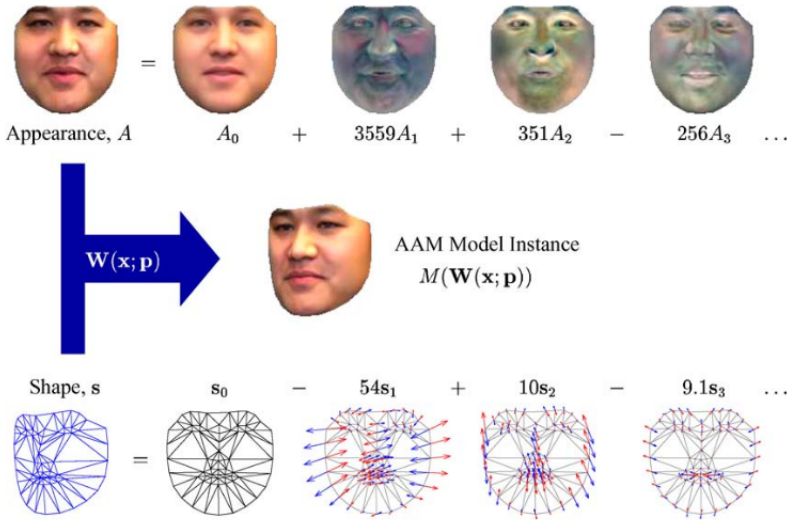

Figure 2.1: Example of AAMs instantiation. The appearance computed from the parameters q is warp into the shape computed from the parameters p [145].

Finally, the AMMs can be fit to a previously unseen image by optimizing the following cost function, which consists in the minimization of the appearance reconstruction error with respect to the shape parameters:

$$\arg\min_{p,q} \; ||F(I)(W(p)) - \overline{a} - U_a q||^2 \tag{2.3}$$

The optimization is an iterative process by which parameters p and q are found so that the best instance of the model is fitted to the given image. This can be solved by two main approaches: analytically or by learning. The first one typically uses gradient descent optimization algorithms [145, 77, 157]. The standard gradient descent being inefficient for this problem, other ways to update the parameters have been proposed. In the project-out inverse-compositional algorithm [145], shape and appearance are decoupled by projecting out the appearance variations. This results in a faster fitting but also with convergence issues that decrease robustness. With the simultaneous inverse compositional algorithm [77], shape and appearance parameters are optimized simultaneously to provide a more robust fitting, but at a higher computational cost. It is however pos-

sible to speed it up using the alternating inverse compositional algorithm [157], which optimizes the shape and appearance parameters in an alternated manner instead of simultaneously. The second optimization approach typically uses cascaded regression to learn a mapping function between the facial appearance and the shape parameters [37, 187]. The original AAMs [37] employs linear regression. Non-linear regression has also been proposed [187]. This approach may be efficient but makes the invalid assumption that there is a constant linear relationship between the image features and the parameter updates [145]. Despite various efforts in optimization strategies, the AAMs remains difficult to optimize under uncontrolled conditions due to the high dimensionality of the appearance space and the propensity of the optimizer to converge to local minima.

To overcome these problems, more robust, part-based representations can be used. In part-based generative models, the local appearance around each landmark is extracted and combined to build a model for the whole face [219, 216], which reduces the size of the appearance space. Active pictorial structures [4] go further by taking advantage of the tree structure used in Pictorial Structures (PSs) [65] to replace PCA. PSs are extended to model the appearance of the face using multiple graph-based pairwise distributions between the facial parts and prove to be more accurate than PCA. Note that ASMs is also considered as a part-based generative model, with the difference that an appearance model is used for each facial part rather than a single model for all parts. However, its evolution towards models such as CLMs [40, 188], considered as discriminative, will be discussed in the next section. Although holistic and parts-based models appear to have the same representational power, part-based models are easier to optimize and more robust to poor initializations, lighting, and occlusions since local features are usually not as sensitive as global features.

2.1.2 Discriminative Approaches

Unlike generative approaches, that rely on statistical parametric models of appearance and shape, discriminative approaches learn a mapping from the image to the facial shape. In this section two categories of approaches are distinguished. The first category refers to hybrid approaches such as CLMs [40, 188]. Such approaches are based on independent discriminative models of appearance, one for each facial part, constrained by a statistical parametric shape model, e.g., PDM. This means that there is still an optimization step during inference for this category. Besides, the ambiguity between the local appearance models of different landmarks can strongly impact the performance

under uncontrolled conditions. To address these difficulties, a second category of approaches has emerged. These approaches, the most notable of which is CSR [53, 248], infer the whole face shape by directly learning one or more regression functions which implicitly encode the shape constraint. It provides more freedom than a parametric shape model would as no explicit assumption on the distribution of the data is made. During inference, there is no optimization as well. The facial shape is directly estimated from the image features. Because of these distinctions, discriminative approaches are faster and more robust in comparison to generative ones. They generalize better to new unseen images as they can benefit from large datasets, which are omnipresent nowadays. However, it remains challenging to directly map the appearance to the facial shape when confronted to severe difficulties such as extreme poses or expressions. A third category could be added for DL approaches, e.g., CNNs. Most of the work today is based on DNNs, leaving traditional methods behind [1]. These approaches are capable of learning highly discriminative and task-specific features. They no longer need feature engineering. Given the breakthrough of DL in computer vision, a separate section (Section 2.1.3) is dedicated to these approaches.

Hybrid Approaches

CLMs is an extension of ASMs with the main difference that the appearance models are discriminative rather than generative. It is a part-based hybrid approach composed of two elements: discriminative local detectors or regressors to represent the appearance and a generative shape model to regularize the deformations and ensure a valid face. In the following, the modeling, fitting, and improvements related to this approach are described.

Local detectors or regressors are used to compute response maps providing the probability that a given landmark is located at a specific position. To build them for each landmark, a statistical model of the grey level structure along with the Mahalanobis distance as the response were initially proposed in ASMs [38]. However, more powerful discriminative approaches, e.g., binary classifiers such as logistic regression [188] or Support Vector Machines (SVMs) [138], have been used. Consider a linear SVMs that determines whether or not a given patch matches the description of the region of a landmark. From the training data, positive and negative examples are extracted for each landmark to train different SVMs, also called patch experts.

[1] https://jponttuset.cat/dl-lstm-gan-evolution/
https://medium.com/@karpathy/a-peek-at-trends-in-machine-learning-ab8a1085a106

The response can be expressed as follows:

$$R = -I(c + \delta c)^T \sum_{v=1}^{V} \alpha_v \lambda_v(c) \tag{2.4}$$

where I is the given image, c the initial coordinates, δc a displacement constrained by the PDM, $\lambda_v(c)$ the V support vectors and α_v the support weights. The output corresponds to the inverted classifier score, which is 1 if positive, -1 otherwise. By fitting a logistic regression function to the output, Platt Scaling [171], an approximate probabilistic output can be obtained. To refine the detection and maintain a valid shape, a shape constraint is then imposed using a statistical shape model such as PDM, already presented in Section 2.1.1.

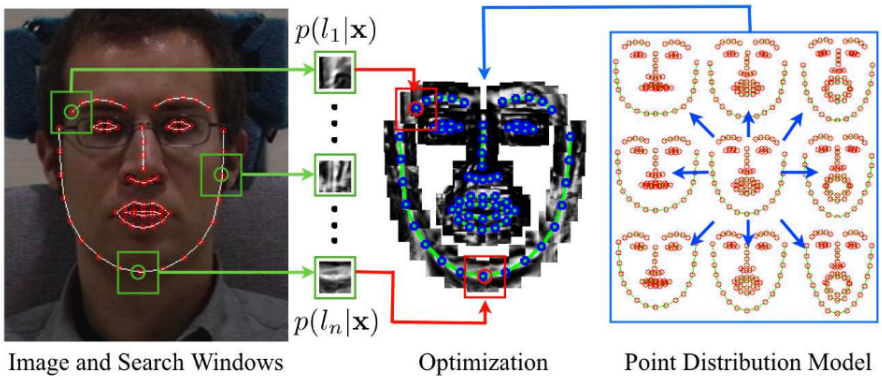

Figure 2.2: Overview of CLMs fitting: ELS is generally performed to get a response map for each landmark. To refine the detection and maintain a valid shape, a shape constraint is then imposed using a statistical shape model [188].

Hence, two steps, illustrated in Figure 2.2, can be differentiated when fitting CLM. An Exhaustive Local Search (ELS) [188] is generally performed first to get a response map for each landmark. This step is followed by a shape refinement over these response maps to find the shape parameters that maximize the probability that the landmarks are accurately detected given the appearance features. The cost function can be expressed as:

$$\arg\min_{p} \sum_{l=1}^{L} R_l(I(\overline{s}_l + U_l p)) \tag{2.5}$$

where I is the given image, L the number of landmarks, $\overline{s}_l$ the coordinate of the l-th landmark from the mean shape, U_l the shape eigenvectors, p the shape parameters, and R the response. The Gauss-Newton algorithm, although it suffers from local minima, is generally employed to solve this problem [188]. Also, the true response maps are not directly used due to performance issues; they are replaced by approximations. There are several approximation techniques, most of them parametric, yet a non-parametric representation known as regularized landmark mean-shift has proved to be a good balance between representational power and computational complexity [188]. It takes the form of a Gaussian Kernel Density Estimate [195]:

$$KDE = \sum_{y_i \in \psi_i} \pi_{y_i} \mathcal{N}(x_i; y_i, pI) \tag{2.6}$$

where y_i is a candidate location among ψ_i locations within a given region, π_{y_i} is the likelihood that the i-th landmark is aligned at location y_i, x_i is the location of the i-th landmark generated by the shape model, p is the variance of the noise on landmark locations, and I is the identity matrix. A regression-based fitting approach can also be used to learn mapping functions from response maps to the shape parameter updates. It is known as discriminative response map fitting [5]. Given its nature and ability to benefit from large amounts of data, this approach achieves a performance gain over RLMS fitting.

One of the limitations of part-based approaches is the ambiguity between local detectors due to the small size and large appearance variations of the training patches. Feature descriptors can be used to obtain a more robust appearance representation. As an example, HOG has proved to be effective [5]. To further reduce ambiguity, more reliable patch experts have been introduced. They are capable of learning non-linear and spatial relationships between pixels and responses, e.g., the minimum output sum of squared errors [16, 143] or the local neural field [9]. A second limitation of part-based approaches may be the ELS, which can be computationally expensive. However, it can be avoided by computing response maps using regression voting [36, 41]. A third limitation is related to the use of PDM as a shape model, which has restricted degrees of freedom due to PCA. Other shape models have been proposed, from separate statistical shape models for each facial component to preserve local deformations [91] to discriminative shape models based on restricted Boltzmann machines [237]. The

most distinctive one is the exemplar-based shape model [11, 98] which, implicitly from training data, imposes shape constraints using a consensus of non-parametric models. It has the advantage of being independent of the initialization. It should be noted that there are some other variants close to the CLMs but generally considered as independent. This includes the combination of local regressors based on boosted support vector regression with graph models such as Markov random fields [223, 142], or the use of tree-structured models [286, 222, 90]. However, globally optimizing Markov random fields appear to be difficult [223]. Tree-based models are easier to optimize thanks to dynamic programming but struggle to perform in real time [65, 286].

Regression-Based Approaches

Direct Regression (DR) learns the mapping from the image appearance to the facial shape in one iteration without any initialization nor parametric appearance or shape models. The landmark locations are jointly estimated while the shape constraints are implicitly encoded. The model can be expressed as follows:

$$M : F(I) \mapsto s \tag{2.7}$$

where M is the model, $F(I)$ is a function that extract appearance features from an image I, and $s \in \mathbb{R}^{2L,1}$ is the predicted facial shape. The regression function being central, its choice is decisive. Regression forests [19] are generally used to cast votes for the facial shape from randomly sampled local patches [46, 256, 257]. However, the random strategy to sample patches is sensitive to occlusions [257]. Additionally, the information provided by local patches is not sufficient to perform global shape estimation. More recently, holistic approaches performed DR using global facial images. In particular, DL approaches [205] is discussed in Section 2.1.3. This mapping is more difficult to learn due to the significant variations in appearance that can occur in the global facial image.

Since DR is challenging, CSR has been proposed to divide the regression process into stages, as shown in Figure 2.3. It usually starts with an initial shape, typically the mean shape of the training data. The early stages focus on large variations, e.g., yaw, roll, or scaling, while the later ones focus on subtle variations, e.g., face contours and motions of the mouth, nose, and eyes, to sequentially update and refine the prediction, following a coarse-to-fine strategy. Each stage corresponds to an independent regressor, which learns the mapping from the shape-indexed features to the shape update. Shape-indexed features correspond to local features extracted based on the current landmark location estimates.

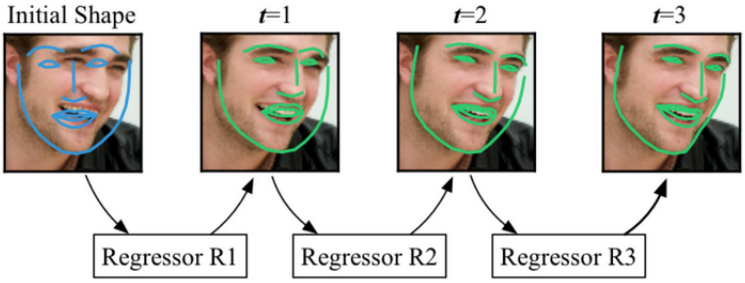

Figure 2.3: Coarse-to-fine prediction using CSR [230]. The regression process is divided into stages. The early stages focus on large variations while the later ones focus on subtle variations, to sequentially update and refine the prediction.

A shape increment can be expressed as follows:

$$\delta\hat{s}_t = r_t(\phi_t(I, s_{t-1})) \tag{2.8}$$

where δs_t is the shape increment at stage t, r_t the stage regressor, and ϕ_t a function that extracts shape-indexed features from image I using the landmark locations at stage $t-1$. The shape constraints are implicitly and adaptively encoded through the sequential training of the stage regressors r_t. During inference, the initial shape is refined by estimating shape increments δs to reduce errors. This results in fast and accurate predictions.

Explicit shape regression [27] represents one seminal work on CSR, using random ferns as stage regressors and pixel intensity differences as shape-indexed features. To train such a model, the mean shape of the training data is first calculated and normalized, i.e., rescaled and centered at the origin. Then, each stage regressor is trained using the ground truth transformed into the mean shape coordinates through GPA [75]. This ensures invariance in scale. The objective is to minimize the mean squared error between the target and the estimate shape increment. It can be expressed as follows:

$$\arg\min_{r} \sum_{n=1}^{N} ||\delta s_{t,n} - r_t(I_n, s_{t-1,n})||^2 \tag{2.9}$$

where r is the stage regressor, δs is the target shape increment at stage t, I is the n-th training image, and s is the shape estimate from the previous stage. The initial shape is generally perturbated multiple times, using data augmentation, to improve general-

ization. During inference, the same objective as during training is kept. The shape estimate is gradually updated using the shape increments transformed back into global coordinates. The complete process is a two-level boosted regression, which consists of a cascade of random ferns, i.e., 10 stages of 500 ferns with a depth of 5. Candidate features are proposed by generating a random set of normalized pixel coordinates indexed relatively to the nearest landmark on the mean shape. These shape-indexed features are extracted based on the current shape estimate at each stage by transforming the pixel locations back into global coordinates. Local pixel intensity differences provide discriminative features invariant to illumination. A correlation-based feature selection strategy is then applied to retain only the most discriminative and distinctive features using Pearson's correlation coefficient. The sum of the outputs of each fern constitutes the final output of the stage. With some adjustments, this approach can achieve impressive execution speeds and handle partial or uncertain labels during training [105].

CSR approaches differ mainly in terms of the shape-indexed features and regressors they use. An approach worth mentioning is the Supervised Descend Method (SDM) [248]. Initially developed to solve general non-linear least square problems, SDM uses linear regression with SIFT features to learn descent directions that minimize the objective. Random forests can also be used to learn discriminative local binary features for each landmark and to learn the mapping between the concatenated features and the facial shape [178]. Gaussian process regression trees [125] have been proposed to enhance generalization. A Gaussian process regression tree corresponds to a kernel function defined by a set of trees to measure the similarity between two inputs and uses differences of Gaussians as input features.

One of the main concerns of CSR may be the initialization with the mean shape, which is sub-optimal. Poor initialization can lead to local optima and impact robustness, e.g., for large head poses [198]. DR can be used to generate a better initial estimate [274]. The bounding box can be randomly shift and rescaled to generate multiple shape hypotheses and learn to rank or combine them to get the final result [252]. The coarse-to-fine strategy can also be employed to perform a progressive and adaptive shape searching [285]. Another concern related to the mapping from the face bounding box to the facial shape is the sensitivity to the face detector [184]. A mapping learned with bounding boxes from one detector could perform poorly if a different detector is used during inference, due to the possible differences in box sizes and locations. Lastly, a minor concern could be the fixed number of cascades. Good results or locally optimal solutions can be delivered early in the process, making the rest unnecessary. There is currently no way to stop the process or adjusting the number of cascades.

2.1.3 Deep Learning Approaches

Traditional ML techniques that have been used so far are not very efficient when dealing with raw image data directly, i.e., to work directly with pixels. Facial landmark detection is too difficult due to the variations in appearance (e.g., position, orientation, illumination) that can occur, especially under uncontrolled conditions. To overcome this issue, discriminative features as invariant as possible to variations are extracted (e.g., HOG, SIFT), which require engineering skills and domain expertise. This can be avoided by using DL approaches, which are capable of learning highly discriminative task-specific features. This subset of ML methods is based on Artificial Neural Networks (ANNs). These methods, although not new [123, 119], have recently led to breakthroughs in many fields, including computer vision [112]. This is associated to the increase in computational power and data availability currently required to leverage such approaches.

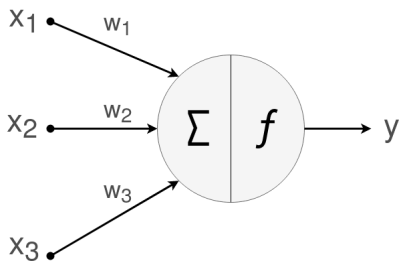

Figure 2.4: Structure of an artificial neuron. A weighted sum of the inputs is performed. The result is then passed through a non-linear activation function.

The most common type of ANNs is the feedforward network, in which information is forwarded in a single direction through multiple layers of computational units, without cycles or loops. There is a variety of architecture with a potentially high number of layers, hence the use of the term DL. The multilayer perceptron [181] can be considered as the groundwork of DL and is often referred as vanilla neural networks. Nowadays, CNN is the architecture commonly used to process images [124]. ANNs are often referred to as biologically inspired models [180]. An artificial neuron, illustrated in Figure 2.4, corresponds to a very simplified version of biological neuron. It can be formulated as follows:

$$y = f(\sum_{i=1}^{I} W_i x_i + b) \tag{2.10}$$

where x_i is the i-th input, W are the connection weights, b is the neuron bias, and f is called the activation function. The same applies to CNNs, which mimic the ventral pathway of the visual cortex by considering an image as a compositional hierarchy

using convolution and pooling layers inspired by simple cells, complex cells, and the LGN–V1–V2–V4–IT hierarchy in the brain [92, 64]. Each layer of a CNN represents the input at a specific level of abstraction. The first layers focus on low-level features such as colors, edges, and orientation, while high-level ones (parts of objects, objects) are obtained in the subsequent layers by composing these low-level features. Although this is an over-simplification from the actual visual cortex, CNNs allow to learn complex functions, which outperform traditional ML techniques in most tasks [3]. Among the popular CNN architectures are AlexNet [112], VGG [197], Inception [206], and ResNet [86], which are widely used in the literature, or at least they serve as a basis for more specific architectures. Today, much work is focused, not exclusively, on designing architectures: deep [197], wide [243], new layer [43], new connection [86], new layer composition [206], efficiency [93], structural diversity [278], combination with a projection model [288], etc. Many criteria need to be considered, including the amount, type, and quality of data available for the targeted problem.

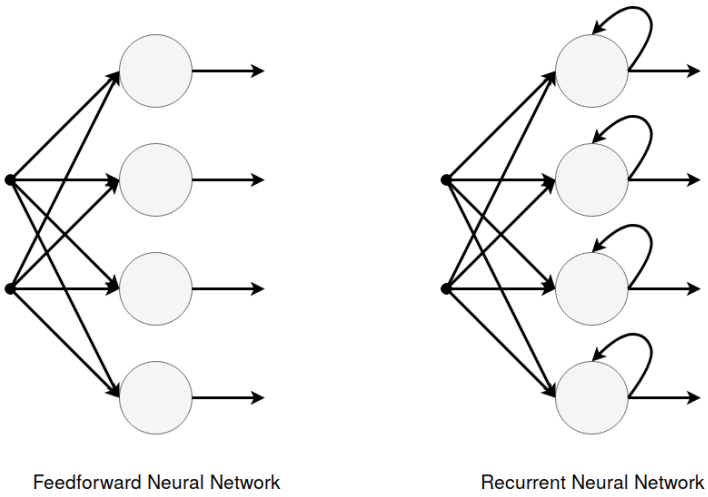

Figure 2.5: Comparison between feedforward and recurrent neural networks. In feedforward networks, the information flows in one direction only, from the input layer to the output layer. In recurrent neural networks, there are cycles or loops, which make them usueful for time series/sequential tasks.

Conventional feedforward networks have no persistence. To allow a reasoning about their past decisions, they can be extended by adding a feedback loop to them and thus making them recurrent (see Figure 2.5 [182]. This introduces a notion of sequence where information can be passed from one step to another. A common way to represent Recurrent Neural Networks (RNNs) is to unroll each step, which results in multiple copies of the same network interacting with each other. In the following the different layers used to build a DNN architecture, how to train and deploy such an architecture, and the specificities related to the landmark detection problem are described.

Theory

A Convolutional Neural Network is a composition of layers, each performing a (supposedly) differentiable function, all together resulting in an architecture that enables the learning of a non-linear mapping between input(s) and output(s). There are several types of layers, the most common ones being fully connected (*fc*) layers, convolution (*conv*) layers, activation layers, and pooling (*pool*) layers.

As illustrated in Figure 2.6, *fc* layers are composed of a fixed number of neurons that are connected to all the neurons in the adjacent layers. Each neuron output is a weighted sum of its inputs. This leads to a high level reasoning through the extraction of global relationship between neurons/features.

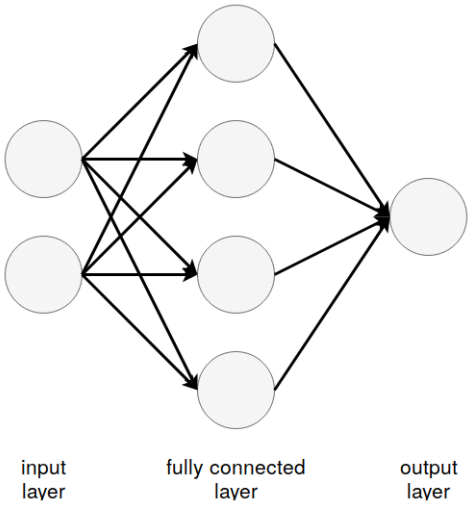

Figure 2.6: Illustration of an *fc* layer. Each neuron in the *fc* layer is connected to each neuron in the previous layer

Using *fc* layer for high-dimensional data being impracticable due to too many connections, *conv* layer was proposed [124]. CNNs make the explicit assumption that the inputs are images, which results in two key properties of the *conv* layer: local connections and weight sharing. A *conv* layer produces a set of 2-dimensional feature map. Each neuron of a feature map is connected to a local region, its receptive field, in the previous layer through a set of parameters called filter or kernel (see Figure 2.7). Local group of pixels are generally highly correlated. Since a feature should be useful to compute at different positions, all neurons in a feature map share the same filter. This provides a feature extractor with translation equivariance and far less parameters than with an *fc* layer.

2D convolution can be expressed as:

$$v_{ij}^{xy} = b_{ij} + \sum_{m} \sum_{p=0}^{Pi-1} \sum_{q=0}^{Qi-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)} \tag{2.11}$$

where v_{ij}^{xyz} is the value at position (x, y) on the j-th feature map in the i-th layer, b is the bias, m is the index of the feature map in the previous layer, P, Q are respectively the height and width of the kernel, and w is the value of the kernel. Note that a 1-dimensional *conv* layer where the filter is of the same size as the input is equivalent to an *fc* layer.

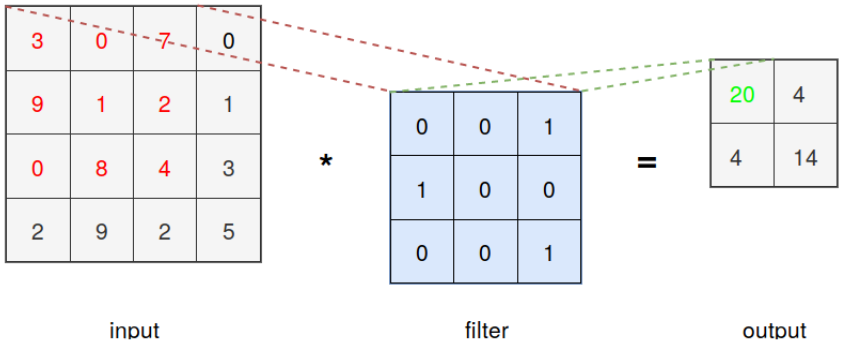

Figure 2.7: Example of a convolution operation with a 2x2 filter and stride 1. The dot products between the filter and a local region of the input is illustrated.

The activation layer is an elementwise function, which performs a non linear transformation, e.g., rectified linear unit (ReLu) [150]. Without this transformation, only linear operations are performed by the network. By increasing its non linear properties, this makes the network more suitable to solve complex problems. ReLu is one of the most commonly used activation. In comparison to the first used activations, e.g., tanh, sigmoid, it makes the training faster and alleviate optimization problems. It can be expressed as follow:

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \tag{2.12}$$

Finally, *pool* layer corresponds to a downsampling operation, which generally computes the maximum or average of a local region. Max pooling is depicted in Figure 2.8. It helps further reduce the amount of parameters and computation, allowing a better control of overfitting. It also adds invariance to translation (small shift) and distortion [124].

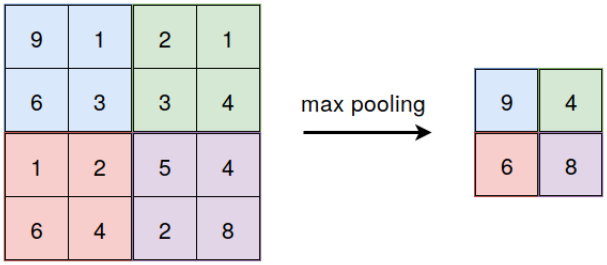

Figure 2.8: Example of a max pooling with a 2x2 filter and stride 2. The maximum of local regions is computed, which leads to a downsampling of the input.

Most of these layers have hyperparameters, which are not learned but must be defined manually. It is necessary to specify the capacity in terms of number of neurons for *fc* layers. The number, size and the displacement (stride) of filters are to be defined for *conv* layers. Similarly, the size and displacement of the window are also to be provided for *pool* layers. To control the size of the input and avoid losing information at the boundaries, *conv* and *pool* layers also require a certain amount of padding. Figure 2.9 depicts stride and padding. *conv* and *fc* layers also have parameters that need to be learned.

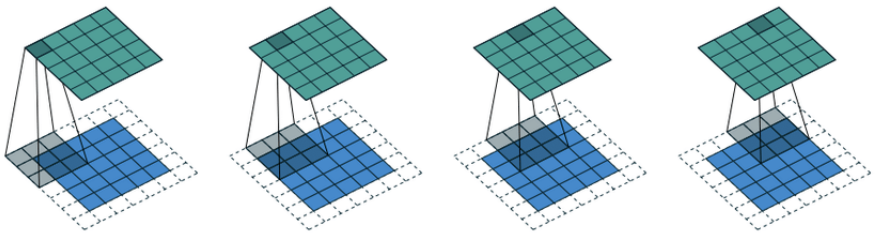

Figure 2.9: Behaviour of a filter of size 3x3 with padding and stride 1.

Once the architecture is defined, the training phase aims to adjust the layer parameters to obtain the best possible predictions. An objective or loss function is applied at the output of the network to measure the error between predictions and ground truth. Mean Squared Error (MSE) is one of the most common regression loss functions:

$$MSE = \frac{1}{N}\sum_{n=1}^{N} ||\overline{s}_i - s_i||^2 \tag{2.13}$$

where N is the number of training samples, $\overline{s}_i$ the i-th ground truth shape, and s_i the i-th predicted shape. The (mini-batch) Stochastic Gradient Descent (SGD) algorithm [179]

is used to find the minimum of this function. The error is computed and distributed backwards (backpropagation). A gradient vector for each parameter is computed from the last layer to the first using the chain rule. It determines the amount by which the error would increase/decrease if the parameter was changed. The parameter is then adjusted in the direction opposite to the gradient vector multiplied by an additional hyperparameter called the Learning Rate (LR). The latter defines by how much the parameters are changed according to the error, which helps to control the speed at which the model learns. It is called stochastic because batches of training samples are selected randomly to compute gradient at each iteration instead of the full training set. It helps prevent the algorithm from reaching local minima. A parameter update can be expressed as follows:

$$W \leftarrow W - \lambda(\frac{\partial MSE}{\partial W}) \tag{2.14}$$

where W corresponds to the parameters, λ the LR and $\frac{\partial MSE}{\partial W}$ the partial derivative of the lost function with respect to the parameters. To avoid gradient problems when performing gradient descent optimization, the parameters are initialized with random values in such a way that they are not too small nor too large, e.g., Glorot initialization. Going through all the training data corresponds to an epoch; it is done multiple times. This iterative process leads to a convergence towards a minimum generally close to the global solution. It is crucial to carefully tune hyperparameters, some as the LR being critical to ensure a good convergence. Variants of SGD such as Adam [107] can be used to maintain an LR for each parameter and to adapt it dynamically during training. Once training is completed and the parameters are set, the network can then return the corresponding landmarks from an input image by passing it through the stack of layers.

RNNs rely on an extension of backpropagation called BackPropagation Through Time (BPTT), which unrolls the network and propagates the error backward over the entire input sequence. The weights are then updated with the accumulated gradients. BPTT can be computationally expensive as the number of timesteps increases. To avoid computational burden, truncated BPTT is generally used. The sequence is split into smaller sequences that are treated as separate training cases. Vanilla RNNs have fundamental optimization limitations as the gradients tend to either vanish or explode during optimization, resulting in an unstable network unable to learn long term dependencies [12]. Gradient explosion can be addressed through gradient clipping [160]. Gradient vanishing can be solved using Long Short-Term Memory (LSTM) [88] illustrated in

Figure 2.10. By adopting a memory cell coupled with gating functions to control the flow of information, the network is better able to learn over long sequences. LSTM can be expressed as:

$$
\begin{aligned}
i_t &= \text{sigmoid}(W_{xi}X_t + W_{hi}H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\
f_t &= \text{sigmoid}(W_{xf}X_t + W_{hf}H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\
C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc}X_t + W_{hc}H_{t-1} + b_c) \\
o_t &= \text{sigmoid}(W_{xo}X_t + W_{ho}H_{t-1} + W_{co} \circ C_t + b_o) \\
H_t &= o_t \circ \tanh(C_t)
\end{aligned}
\tag{2.15}
$$

where X_t and H_t are respectively the input and the hidden state at time t. W denotes the weights, b the bias, and $\circ$ the Hadamard product. C_t is the memory cell of the LSTM, which is updated by the input gate i_t, the forget gate f_t, and the output gate o_t. Theses gates help reduce the vanishing gradient effect. Gated Recurrent Unit (GRU) [31] is an increasingly popular alternative with a simplified structure, which combines the forget and input gates into a single update gate and merges the hidden state and memory cell.

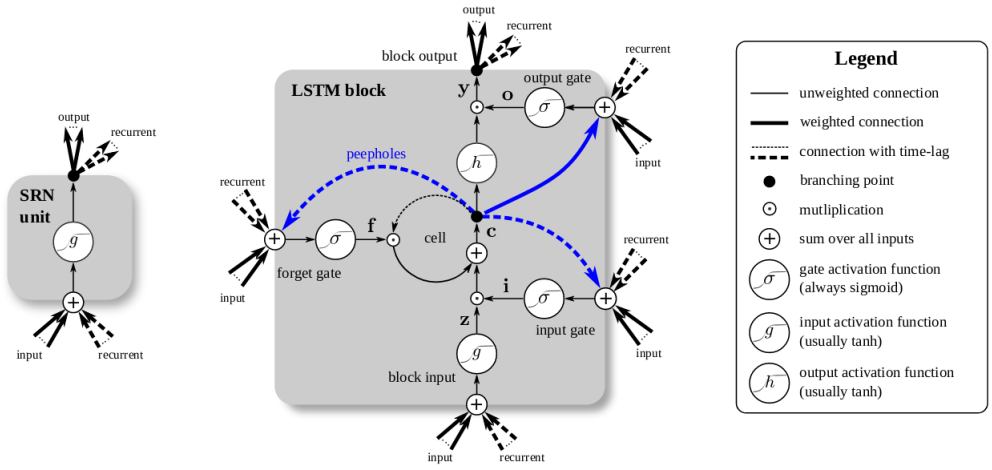

Figure 2.10: Comparison of a vanilla RNN unit (left) and a LSTM block (right) [76]. Compared to vanilla rnn, LSTM has a memory cell updated by different gates to control the flow of information and reduce the vanishing gradient effect.

Beyond the standard supervised learning framework, other techniques can be used. Adversarial Learning (AL) is an increasingly popular framework, which consists of replacing or supplementing the explicit loss function with one or several DNNs. This is

often presented as a game theory scenario where several networks are used to compete against each other. Training is modeled as a zero-sum game. The gain of one necessarily constitutes a loss for the other. As shown in Figure 2.11, one neural network, called the generator (G), generates new data (e.g. an heatmap), while its opponent, the discriminator (D), tries to detect if the data is real or if it is the result of G. The aim of G is thus to fool D. The gradients accumulated from the loss of D are used to update G. In this way D provides feedback to G so that it may improve. This strategy is also known as Generative Adversarial Networks (GANs) [73]. The original GANs loss function can be expressed as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \quad (2.16)$$

where D maximize the probability it correctly classifies reals and fakes ($logD(x)$) and G minimize the probability that D will predict its outputs are fake ($log(1-D(G(z)))$).

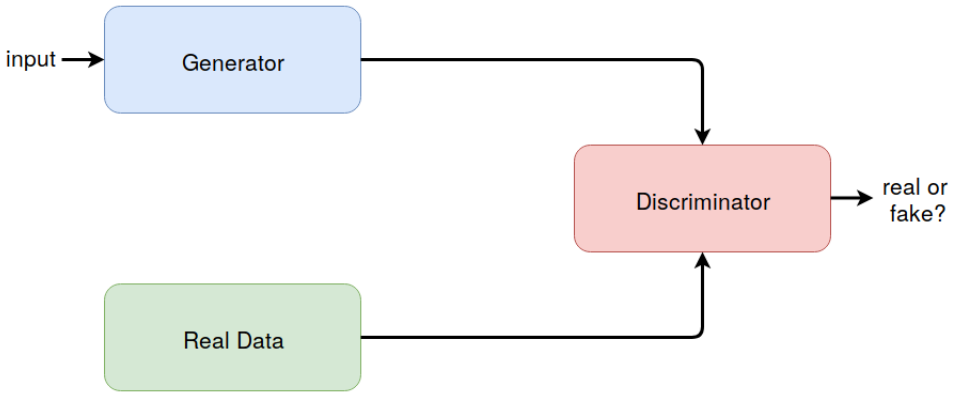

Figure 2.11: Overview of GANs. The generator generates new data while its opponent, the discriminator, tries to detect if the data is real or if it is the result of the generator. The discriminator then provides feedback to the generator so that it may improve.

Advances

In the recent Menpo challenge [271], which is a good indicator of trends on facial landmarks detection problem, only DL approaches have been proposed. Currently, two strategies to design architectures for landmark detection can be distinguished: coordinate regression and heatmap regression. They respectively regress the coordinates directly using a fully connected layer [205] or compute heatmaps, one for each landmark with the same size as the input, using a FCN [23]. The coordinates of the maximum value in a heatmap are generally considered as the coordinates of a landmark.

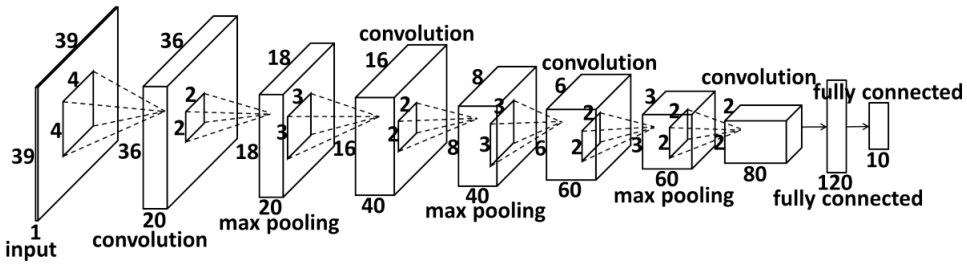

Figure 2.12: Early CNN architecture for facial landmark detection [205]. It consists of 4 *conv* and *pool* layers and 2 *fc* layers to produce landmark coordinates.

One of the first DNNs designed for landmark detection is based on the CSR approach presented in Section 2.1.2 [205, 282]. It consists of multiple levels of CNN, the first level processing the whole face to capture texture information and geometric constraints, while the subsequent levels perform a local refinement. Figure 2.12 illustrates the CNN architecture composed of 4 conv and pool layers and 2 fc layers allowing the direct regression of landmark coordinates. There is now a trend towards Fully Convolutional Network (FCN) architectures. It has some advantages [135], including no input size limit, better spatial information handling and a better computational cost/representation power ratio. The HourGlass (HG) network shown in Figure 2.13 is currently very popular and considered as the state-of-the-art facial landmark detection method [23]. Initially designed for human pose estimation [152], HG consists of a symmetrical encoder-decoder with skip connections between the two to consolidate information at different scales. It is generally stacked several times (up to 8) with intermediate supervision to follow the CSR approach. A compromise between these two approaches is also possible, by regressing coordinates and using landmark heatmaps to transfer spatial information across stages [111]. HG is the most popular model, yet it struggles to run in real time due to its high computational complexity. Neural network compression such as weight binarization can be applied to improve speed and reduce the size of the model, but at the expense of accuracy [22].

CSR remains popular within DL approaches and can be implemented in different ways, either by stacking networks [274, 87, 80], as seen previously, or by formulating it as a recurrent process by combining CNNs and RNNs [215, 230, 116]. In the latter, each recurrent step corresponds to a shape increment; shape increments are jointly learned relying on an explicit memory shared across steps. CSR can be improved by adding attention mechanisms to the network to enable it to focus on relevant parts of the data. Such a mechanism, combined with a RNN, can identify reliable landmarks and use them as attention centers to refine primarily landmarks close to them [245]. It can also be incorporated into *conv* layers using intermediate supervision to generate top-down

attention maps [267]. This allows the network to focus on regions around the landmarks and thus capture more discriminative and powerful features for facial landmark detection. This approach suggests that CSR is not a necessity and that DR remains a valid solution. Deep Reinforcement Learning (DRL) has also proven to be a promising alternative to CSR through the use of a shape searching policy, which maximizes a shape evaluation function [134].

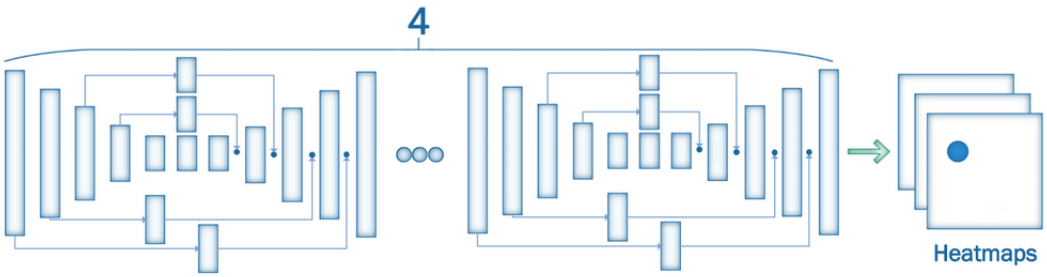

Figure 2.13: Hourglass network [23]. It consists of a symmetrical encoder-decoder with skip connections between the two to consolidate information at different scales. It is generally stacked several times to produce heatmaps, one for each landmark, in a coarse-to-fine manner.

Other improvements have been proposed. A comparative analysis of different loss function (L1, L2) [66] suggests that more attention should be paid to small and medium range errors and the authors proposed a new loss function called Wing loss to address this problem. The loss function can also be used to apply more specific constraints during training. A contextual loss has been proposed to capture the visual context of each landmark [273]. It can be combined with a structural loss to add an explicit geometric constraint in order to exploit the hierarchical structure of the face. This constraint can be incorporated into the model in different ways. A first alternative involves the use of a second model as a discriminator to distinguish between real and fake shapes with the aim of providing a feedback to the first model [30]. A second alternative is to integrate other information related to landmarks such as boundary of the face and use them as a representation of geometric structure [235].

2.1.4 Handling Challenges

Under uncontrolled conditions, the approaches presented in Section 2.1.1, 2.1.2, and 2.1.3 continue to encounter difficulties, e.g., head pose, occlusions, facial expressions, illumination. Some of the approaches are capable of implicitly dealing with these challenges, but they remain most often ineffective when extreme cases occur. A number

of authors propose to explicitly address some selected challenges and the proposed solutions are generally complementary to most approach. These include the use of 3D models or multiple specialized models (e.g., multiview) to treat profile faces more consistently, multitask learning with expression recognition as a related task to better handle facial expressions, or extended dataset annotations to explicitly detect occlusions. In this section a focus is placed on three challenges considered to be the most complex to address (i.e., head pose, occlusions, expressions) and review the proposed solutions to tackle them.

Head pose

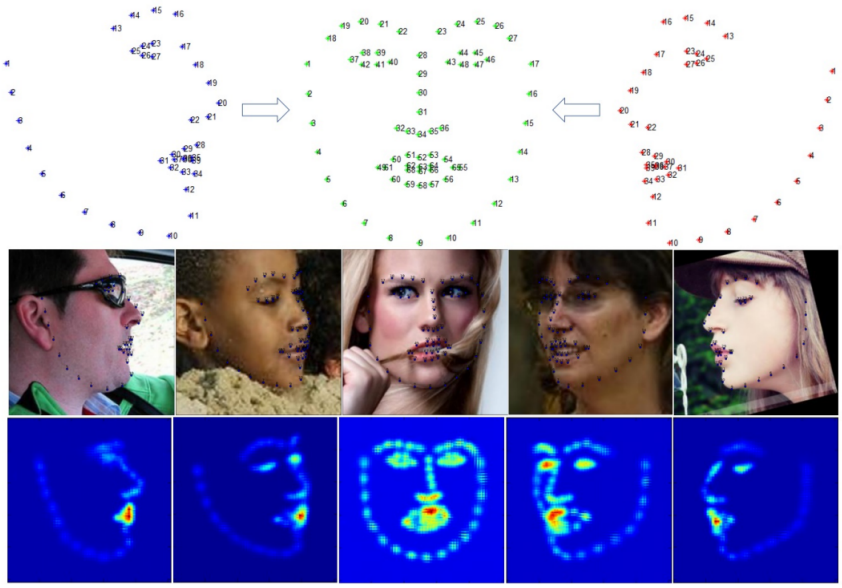

Figure 2.14: Joint multiview by capitalizing on the correspondences between views [50]. This helps achieve better accuracy under extreme poses.

Head pose variations are one of the main sources of error. They lead to significant changes in appearance and shape especially when switching from frontal to profile views, which hides an entire part of the face. The approaches presented so far assume that the landmarks to be detected are visible. This is not true with profile faces. Moreover, there is little training data for extreme poses due to the difficulty of annotating it. To address these problems, apart from the few multi-tasking or conditioning solutions, which condition the measurement of 2D coordinates on 3D pose [251, 114], two solutions currently stand out: multiple view-specific models and dense 3D model.

Multiple view-specific models can be used to achieve a better accuracy under extreme poses, one for each view, e.g., left profile, frontal view, and right profile. The best

model is then selected depending on the scenario. It can be done using head pose based on a few landmarks [253, 264, 48] or the confidence score of the models [39, 286]. To reduce selection failures and improve fault tolerance, one possibility, instead of splitting the training data by pose, is to use the whole training data for each model and weight more heavily the training samples of the specific pose domain, with overlap between domains [67]. An alternative is to merge the results of different pose-dependent models [46]. Still, model selection and model fusion are not trivial tasks as it requires view estimation or the use and ranking of different models. More recently, with the introduction of a specific landmark configuration for profile faces [271], a joint multi-view model has been proposed to avoid using distinct models, by capitalizing on the correspondences between views (see Figure 2.14) [50].

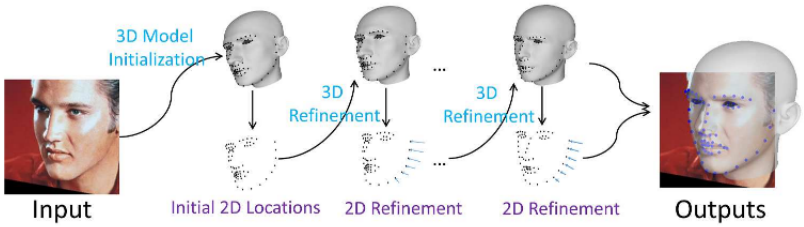

Figure 2.15: 2D-3D dual learning to mutually reinforce each task [244]. Considering the 3D structure of the face is particularly useful to address more effectively self-occluded landmarks.

Pose variations can also be handled by fitting a dense 3D model, e.g., a 3D Morphable Model (3DMM) [15], to a 2D image. Considering the 3D structure of the face is particularly useful to reliably reason about self-occluded landmarks. The projection matrix and model parameters are generally estimated using cascaded regression, then a 3D-to-2D projection is performed to infer the 2D landmark locations from the 3D face [100, 287]. Using a single model has also been proposed to improve training and inference times [101, 14]. The main limitation with these methods is that the shape is restricted by their linear parametric 3D model, which may lead to a lack of precision in detection. This can be mitigated by jointly refining the 3D face model and 2D landmark locations through a recurrent 2D-3D double learning process to mutually reinforce each other (see Figure 2.15) [244].

Occlusions

Among other causes of error are occlusions, which lead to a loss of information since they cover a more or less important part of the face. Occlusions that are locally consis-

tent, meaning that they do not hide all the landmarks, are considered. This challenge is particularly complex to deal with due to the variety of occlusions, which can be caused by objects with arbitrary appearances and shape, or self occlusions, e.g., due to head pose. It requires to rely on non-occluded parts and, therefore, determine which parts are hidden. Occlusions can be detected explicitly to improve the robustness to outliers (i.e., misdetected landmarks).

A first approach is to use occlusion-dependent models. The facial image is generally divided into several regions in which some are assumed occluded. Different models are trained with each of these regions. The resulting predictions are merged according to certain criteria such as the amount of occlusion present in the regions [25, 265]. The main drawback of this approach is that facial occlusions are totally arbitrary and therefore it cannot cover the variety of occlusions. Moreover, relying on a single region provides insufficient appearance information. Another strategy for dividing and processing the facial image is to segment the image into non-overlapping regions and predict for each region a heatmap, which indicates how useful is the information contained in it. This heatmap is used together with the facial image to perform landmark detection [255].

A second approach helps further alleviate the mentioned drawbacks by using unified framework to jointly model the appearance around each landmark, landmark locations and visibility, and occlusion pattern [71]. CSR can be used to predict landmarks location and visibility with an explicit occlusion pattern constraint and more weight assigned to visible landmarks [238]. With recent DL approaches such as FCN, occlusion information can be embedded into heatmaps labels [269]. Broadly speaking, these approaches require ground truth annotations for occlusions or synthetically occluded data, which makes the annotation of datasets more laborious.

Expressions

Facial expressions also cause important changes of facial appearance and shape. There are 7 universal expressions [57]: happiness, sadness, fear, disgust, anger, contempt, and surprise. All over the world they are used to convey the same emotion. The various state transitions and intensity are source of localized and complex motions. Beyond these universal expressions, more complex expressions are likely to be encountered in natural environments [183, 172]. Recent approaches, i.e. DL, deals reasonably well with expressions in terms of Euclidean distance. However, it is unclear at this time whether

this is suited to applications since Euclidean distance is a specific evaluation metric for landmarks detection and does not reflect some cases such as instability on videos. A few methods have been specifically designed to be robust to expressions by using multi-state local shape models or face shape prior models to deal with the shape variations of facial components [210, 237]. Considering the correlation between shape and expression, a more common solution is not to treat facial landmark detection as an independent problem but to jointly learn various related tasks in order to achieve individual performance gains. Among the related task is facial action unit detection [130, 191]. The relationship between facial action units and landmarks can serve as a constraint when iteratively updating landmarks during CSR [239]. Facial expression recognition is also widely used [281, 280, 175, 241]. An example is depicted in Figure 2.16. While such a solution may be straightforward to implement, by adding as many outputs as auxiliary tasks, it can make the training stage much more complex because the optimal convergence rates may vary from one task to another [280].

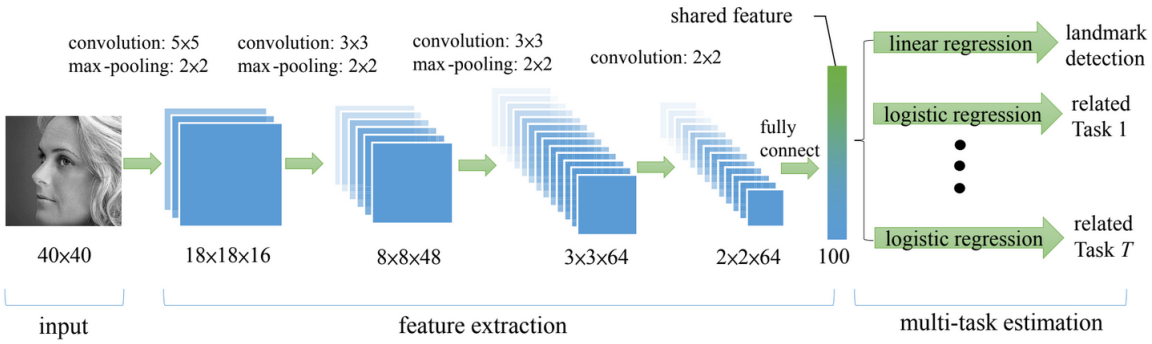

Figure 2.16: Landmark detection with auxiliary tasks based on multitask learning [279]. Facial landmark detection is not treated as an independent problem but jointly learned with various related tasks in order to achieve individual performance gains.

Other

Some challenges, mostly related to the quality of the image, e.g., illumination, resolution, blur, have attracted less interest, either because they are easier to handle or because they do not have major impacts in terms of occurrence frequency or error. Recently, work based on GANs has been carried out to detect landmarks more accurately on low-resolution images [24]. As shown in Figure 2.17, it involves three models to improve both super-resolution and facial landmark detection: a generator for synthesizing high-resolution images from low-resolution ones, a first discriminator to distinguish between synthesized and original high-resolution images, and a second discriminator for the detection of landmarks on synthesized high-resolution images.

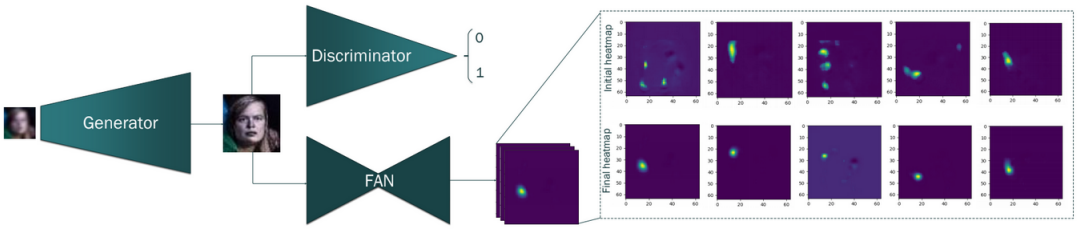

Figure 2.17: Adversarial learning involving super-resolution to improve landmark detection on low resolution images [24]. A generator synthesizes high-resolution images from low-resolution ones while a first discriminators distinguishes between synthesized and original high-resolution images, and a second discriminator detects landmarks on synthesized high-resolution images.

The datasets used to solve the landmark detection problem present significant intra/inter-dataset variations, which can lead to poor generalization. By coupling two models, it is possible to leverage these variations [236]. The first model takes advantage of intra-dataset variations by using an implicit feature sharing mechanism. It takes an image and its flipped version as input. The second model, which focuses on inter-dataset variations, decides which candidate should be considered. Some authors have rather focused their work on simplifying the landmark detection task by normalizing the input to a canonical pose. The most advanced solutions employ supervised face transformations such as spatial transformer networks to remove the translation, scale, and rotation variations of the face [139, 111, 259, 61, 29]. While most authors focus on the variance of faces, the intrinsic variance of image styles can also be handled to improve performance using a style-aggregated network, which takes as input two complementary images of the same face, the original style image and an aggregated style image generated by GANs [55].

2.1.5 Summary

Generative approaches can achieve interesting performance under uncontrolled conditions using only few data, with suitable representation and optimization strategy [218, 217]. However, they can be computationally intensive and due to the difference between the model and the true distribution of data, their ability to generalize is often restricted [13]. Besides, a small error in the prediction of the model coefficients may lead to a large error regarding the landmarks. This is not true discriminative approaches, which are becoming more and more popular. Instead of predicting the model coefficients, landmarks are directly predicted resulting in a better accuracy.

Part-based approaches such as CLMs, although more robust to partial occlusions and lighting, suffer from an ambiguity problem along with performance issues due to their local appearance representation coupled with a global shape optimization. To overcome the limitations of these approaches, another popular approach is to estimate the whole facial shape directly from the image features while implicitly exploiting spatial constraints. This is less restrictive than modeling them explicitly with a shape model. Two solutions can be found: DR and CSR, the latter being the most commonly used and a state-of-the-art approach. By providing a mapping from the image to the facial shape without the use of any parametric appearance or shape model, faster and more accurate predictions can be obtained. The coarse-to-fine strategy of CSR offers a progressive and flexible way to encode the shape constraints while increasing robustness. The availability of large datasets further improves the generalization capacity of discriminative approaches. However, it is still not easy to perform this mapping in some scenarios, such as extreme expression and pose variations. Recently, a shift from traditional approaches to DL-based approaches has occurred. The latter can model more effectively the nonlinear relationship between images and shapes and has already shown impressive performance.

Facial landmark detection problem has benefited from the advances in DL. With such approaches, the need for feature engineering is no longer necessary. The research has shifted to architecture design where domain expertise can still be useful but not as crucial as it used to be. Although CSR continues to predominate, DR appears to be practicable with DNNs, as some authors showed it [146, 148]. However, DNNs require a significant amount of computational power and annotated data, which is not easily affordable. It is also worth noting that landmark detection is a structural prediction problem that can hardly benefit from models pre-trained on massive amounts of data such as ImageNet [47] for object detection and localization. On the other hand, overfitting can be an issue when building a deep model from scratch due to the lack of data available. Despite these challenges, DL approaches have helped take a further step towards solving the problem of facial landmark detection, but there is still much to be done.

Complementary solutions have been developed, which can be integrated into both the DL and the more traditional approaches. This include the use of 3D model, multi view-specific models, multitask learning, GANs, explicit occlusions modeling. They can help solve the remaining challenges and further increase generalization. By integrating such solutions, it is possible to significantly improve the robustness to specific challenges. However, it is often a combination of these challenges that are encountered

under uncontrolled conditions. Although these solutions are not mutually exclusive, it may be too complex and costly to apply them together. Another kind of solution based on temporal consistency and allowing several challenges to be addressed at the same time is possible.

All the approaches presented so far, when applied to a video stream, are not able to use temporal information. Yet, with the ubiquity of video sensors, the vast majority of applications rely on videos. Moreover, there is a need for an approach that can deal with all challenges at once. Recent work has proved that taking into account video consistency help to deal with the variability in facial appearance and ambient environment encountered under uncontrolled conditions. These works are reviewed in the following section.

2.2 Extending Facial Landmark Detection to Videos

Despite the quantity of facial landmark detection approaches proposed in the literature and the recent major advances, the difficulties encountered under uncontrolled conditions are still far from being solved (see Section 2.1). As already seen in Section 1.3, variations in pose and expression along with occlusions are among the most difficult challenges due to their significant influence on facial appearance. Moreover, applications of face analysis are mainly based on image sequences. So far, image-based methods cannot use the motion information from image sequences nor adapt their models online to make it person-specific. In the following sections, it is shown how temporal information can be beneficial to landmark detection, especially under uncontrolled conditions, by addressing many of its challenges. To this end, temporal approaches are reviewed from the most naive but popular tracking by detection to the most recent and more complex ones based on DNNs.

2.2.1 Tracking by Detection

Recently, a comparative analysis of video-based facial landmark detection approaches showed that the most popular strategy for this problem is tracking by detection [193]. As illustrated in Figure 2.18 tracking by detection consists in applying face detection followed by facial landmark detection, independently on each frame, without taking into account the coherence of adjacent frames.

Face detection can possibly return false positives. To prevent these, predictions with the highest confidence are usually retained. The detector may also fail to detect any

faces on an image, making facial landmark detection no longer possible. In this case, the bounding boxes from the previous image can possibly be used. This is a reasonable assumption if the frame rate is high enough to ensure that there is no significant variation between the current and the previous images.

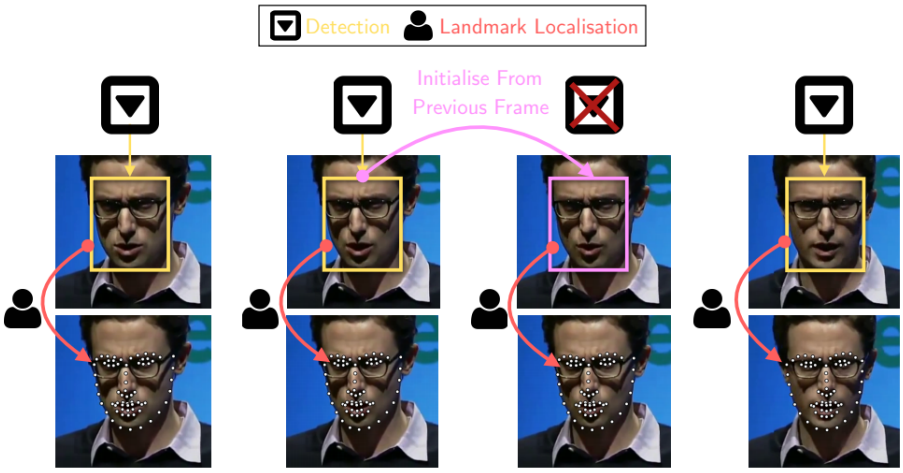

Figure 2.18: Tracking-by-detection over a few frames. It consists in applying face detection followed by facial landmark detection, independently on each frame, without taking into account the coherence of adjacent frames. Detection from previous frame is used in case of failure. [32].

Although tracking by detection is naive, a recent study has demonstrated its viability for deformable face tracking [32] since current detectors are effective enough to operate under uncontrolled conditions [272]. However, this strategy has severe limitations. It relies on two distinct static models that are individually trained, which is time consuming. Neither of these two models are able to integrate motion information and personalize the models online to the targeted person. Additionally, facial landmark detection being sensitive to initialization, it is very likely to obtain poor predictions due to bad initializations far from the ground truth, especially during large variations, e.g., pose or expression.

2.2.2 Box, Landmark and Pose Tracking

To avoid using face detection at each frame with the related initialization issues, an alternative is to use a substitute for the detection in order to adapt to the temporal domain. Three main solutions have been proposed to establish a correlation between the previous and current frames [258]: box, landmark, and pose tracking. In each of these solutions, the aim is generally to take advantage of the previous prediction to provide a better initialization for the current frame.

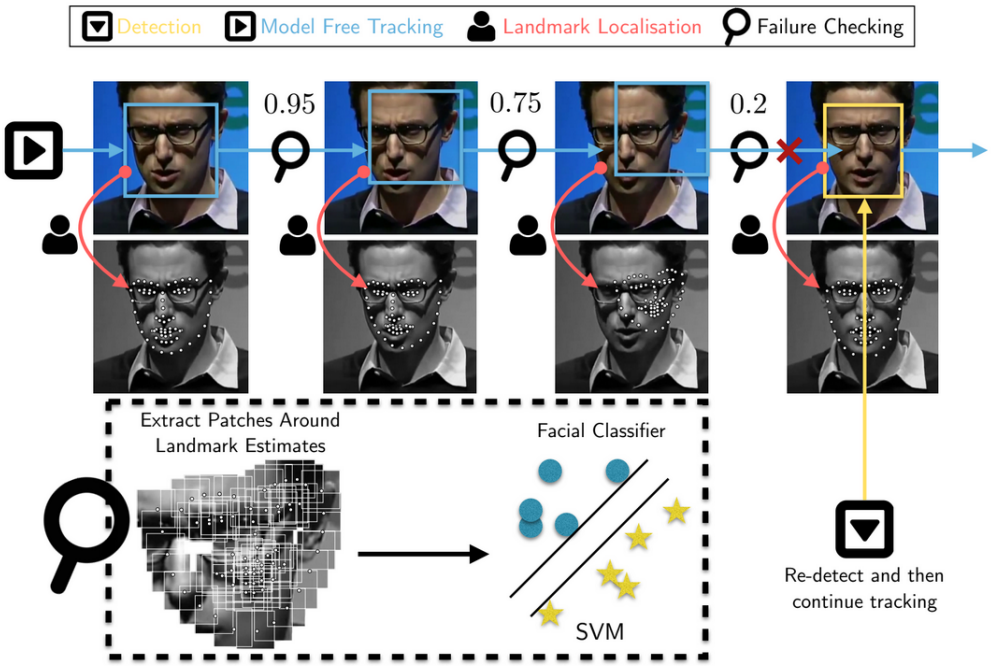

Figure 2.19: Box tracking over a few frames based on a rigid tracking algorithm instead of a face detector in each frame. A reset mechanism is used in case of drift [32].

Box tracking consists in using a rigid tracking algorithm instead of a face detector in each frame (see Figure 2.19). Such algorithms are able to be robust to some variations in the facial appearance during tracking [258, 32]. However, it can be as much time consuming as detection, and more importantly, it can easily drift over time, especially under uncontrolled conditions. This solution is found to be equivalent to tracking by detection with the same drawbacks [32]. The second solution, landmark tracking, involves the use of the previous shape as an initialization for the current frame [258]. It provides the ability to inherit information about rotation, translation, and scale, and thus provide an initialization closer to the ground truth and potentially a better convergence and more accurate predictions. It might, however, lead to a poor convergence due to cumulative errors in video, when the initialization is not the same as in the training set. A third solution is to keep only the similarity transformation parameters from the previous prediction, i.e., pose tracking [258, 194] as shown in Figure 2.20. These parameters are used to adjust the mean shape, which then serves as an initialization. By tracking the pose and not the full shape, the noise from the previous prediction is smoothed, making the following predictions more stable over time. It can greatly reduce the failure rate while improving overall performance [193].

In the same vein but less popular, ensemble learning algorithms can be used to generate multiple initializations for an image, based on the shape parameters from previous predictions [166, 127]. This can be summarized as applying facial landmark detection using the different initializations and then recovering the best result. The obvious disadvantage of this approach is its efficiency; a balance between computational efficiency and accuracy has to be found. Beyond initialization, the shape parameters of the previous prediction can also be used to select a pose dependent model to obtain a better and more efficient convergence, as explained in Section 2.1.4 [258, 249].

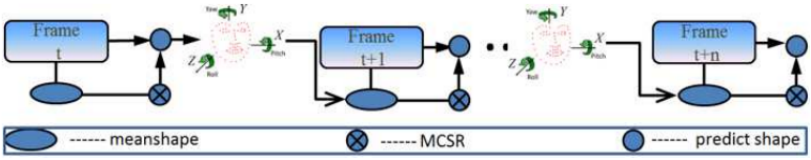

Figure 2.20: Overview of pose tracking. The mean shape is adjusted based on the similarity transformation parameters from the previous frame [258].

Regardless of the tracking strategy, it is necessary to incorporate a reinitialization mechanism in order to avoid any drift over time. Such mechanism is illustrated in Figure 2.19. It is generally based on the quality of the prediction, using SVMs [258, 32] or more recently DNNs [164] to decide if the shape and appearance still match those of a face. The classifier is trained with patches extracted from the ground truth for positive examples, and negative cases are randomly sampled from the region around the ground truth. It can also be done in a holistic way by concatenating the face image and the landmark heatmap. The score of the classifier is then used as a fitting score to judge the quality of the predictions. If the score is below a certain threshold, face detection is applied instead of tracking.

Non-rigid tracking produces a more accurate fitting than tracking by detection. It takes advantage of adjacent frames to improve initialization. Facial landmark detection is therefore no longer dependent on the detection window, but rather takes advantage of the previous prediction to improve initialization. These solutions remain trivial and still do not benefit from person-specific information.

2.2.3 Adaptive Approaches

In contrast to the work presented in Section 2.2.2, which retains one generic model after learning, a more advanced approach consists in integrating an on-line model update, also called incremental learning, in order to make it person-specific and thus more accurate

[186, 166, 6, 33]. This is close to rigid tracking algorithms but, instead of considering the object as a single bounding box, its shape and deformations are taken into account along with its appearance. As shown in Figure 2.21, the pre-trained model is updated over time to learn a person-specific representation without re-training from scratch. Incremental learning can be achieved using generative approaches as proposed with the incremental PSs [34, 33] but also apply to discriminative approaches as with incremental SDM [6]. These approaches are generally costly. Therefore, efficient methods based on continuous regression, as opposed to standard discrete regression, have been proposed and are able to reach real-time performance [186]. To avoid compromising the model during drifting, an update criterion similar to the reinitialization mechanisms presented in Section 2.2.2 is generally used.

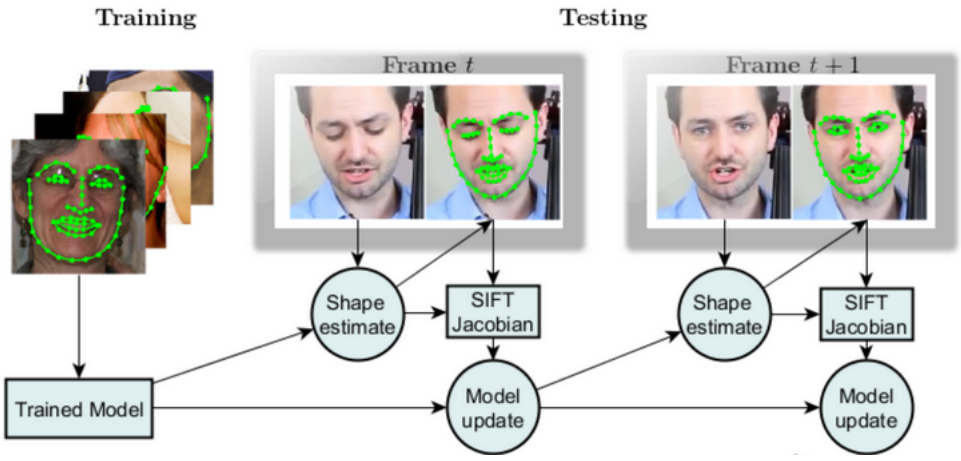

Figure 2.21: CSR with on-line model update [186]. The pre-trained model is updated over time to learn a person-specific representation without re-training from scratch.

Adaptive approaches provide more accurate predictions but, as with generic tracking, they are prone to drifting, while tracking-by-detection is drift-free but less accurate. It is possible to establish a synergy between both detection and tracking in order to limit the weaknesses of these two approaches. This is known as joint detection and tracking.

2.2.4 Joint Approaches

To avoid drift issues and provide accurate predictions, both the tracking by detection approach and the conventional tracking approach may be combined to complement each other [106, 81]. One of the first solutions to be proposed relies on the use of different detection and tracking initializations [106]. Each of these initializations define different optimization sub-problems. Independent shape updates for each sub-problem are computed and then coupled using global variable consensus optimization [17].

Using Equation 2.3 of the generative model (Section 2.1.1, Page 16), it can be formulated as follows:

$$\underset{p_1 \ldots p_M, q_1 \ldots q_M}{\arg\min} \sum_{m=1}^{M} f_m(p_m, q_m) \tag{2.17}$$

where M, f, p and q correspond respectively to the different initializations, sub-problems, shape, and appearance parameters. This strategy can be integrated within any state-of-the-art approach. However, the synergy between both tasks is limited due to the use of separate models, one for landmark detection and another for tracking. The two tasks are not optimized together and are actually applied subsequently.

DRL [132] has been proposed to explicitly exploit the synergy between bounding box generation and landmark detection [81]. Two agents, a tracking agent and an alignment agent, are trained interactively (see Figure 2.22). The aim of the tracking agent is to adjust the bounding box based on a set of actions corresponding to displacements or changes in scale. The role of the alignment agent is essentially to determine whether or not it is necessary to continue the search process using continue/stop actions. Both tasks are jointly conducted through the sequence of action defined by the two agents with an emphasis on the quality of the alignment.

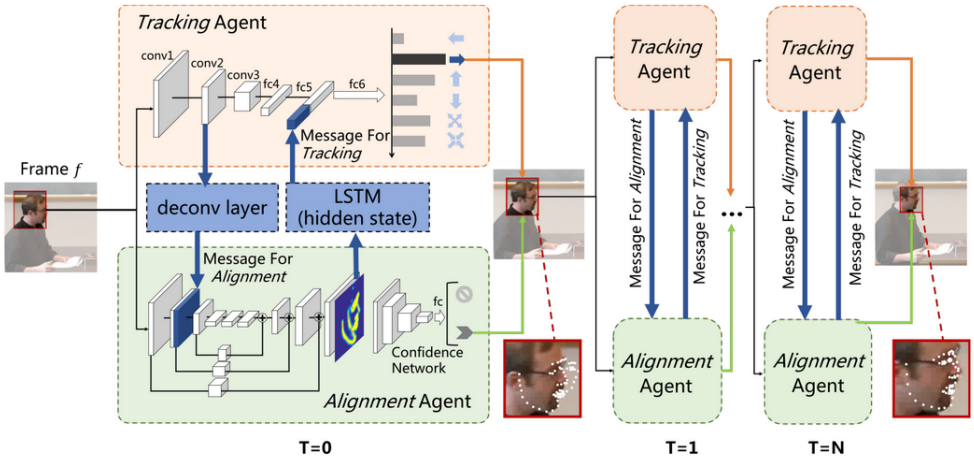

Figure 2.22: Joint detection and tracking using Deep Reinforcement Learning [81]. A tracking agent adjusts the bounding box based on a set of actions corresponding to displacements or changes in scale. An alignment agent determines whether or not it is necessary to continue the search process using continue/stop actions. The two agents are trained interactively.

Joint approaches are able to benefit from both detection and tracking while avoiding their downsides. However, they do not take advantage of the dynamic nature of the face. Other information, such as the trajectories of the landmarks through the image sequence, seems relevant to consider.

2.2.5 Temporal Constrained Approaches

Whether explicit or implicit, the shape constraints present in most facial landmark detection methods are crucial to obtain good performance under uncontrolled conditions. In image sequences, an additional constraint may be applied to the trajectories of the landmarks. One simple way to achieve this is to smooth predictions between frames to avoid jittering and increase stability. Bayesian filters such as Kalman filters and their extensions (e.g., particle filters) can be used for this purpose. However, they require a complex design and tuning phase and are generally restricted to global rigid motion tracking, i.e., tracking the detected bounding box [221, 173], ignoring local non-rigid facial deformations. It has also been demonstrated that theses filters only provide a marginal gain for video-based facial landmark detection [79, 32].

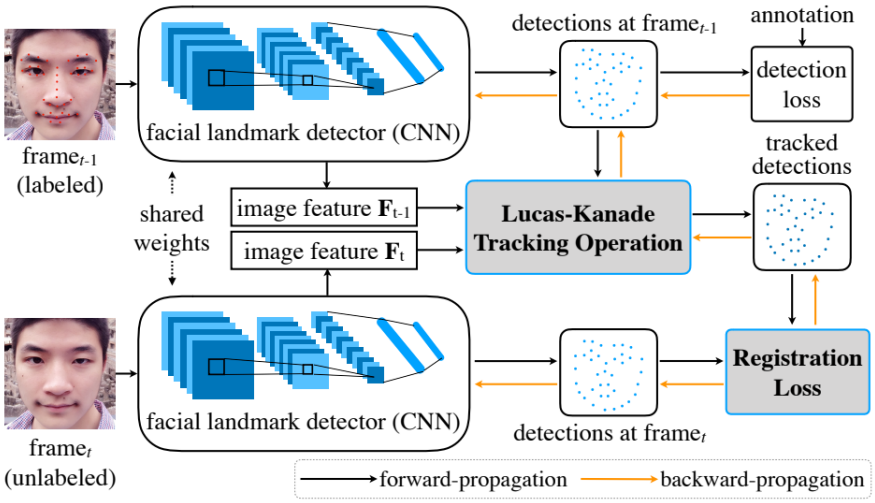

Figure 2.23: Semi-supervised training using the coherency of optical flow as a source of supervision [56]. Lucas-Kanade tracker tracks landmarks on future frames based on past predictions from the detector. The registration loss computes the distance between these predictions and the ones from the detector. This allows to train a detector with unlabeled videos to provide temporally consistent predictions.

To consider the whole shape, a stabilization model based on a Gaussian mixture has been proposed [207]. The model is trained on a set of videos to learn the statistics of different kinds of movements. A loss function with two terms to address time delay and non-smooth issues is used to estimate its parameters. The model is then applied

after landmark detection; it takes as an input all predictions from previous images and produces a more accurate and stable result for the current image. A balance has to be found between accuracy and stability.

Another way to reduce jittering in video is to take advantage of the coherency of optical flow by using it as a source of supervision [56]. The main advantage is that no manual labelling is required. The Lucas-Kanade tracker coupled with a registration loss function can be used to train a detector with unlabeled videos to provide temporally consistent predictions. The latter tracks landmarks on future frames based on past predictions from the detector. The registration loss computes the distance between these predictions and the ones from the detector. The whole training procedure is illustrated in Figure 2.23. Although optical flow registration can encourage temporal consistency to improve existing facial landmark detectors, they still remain image-based detectors with limited motion modeling capability.

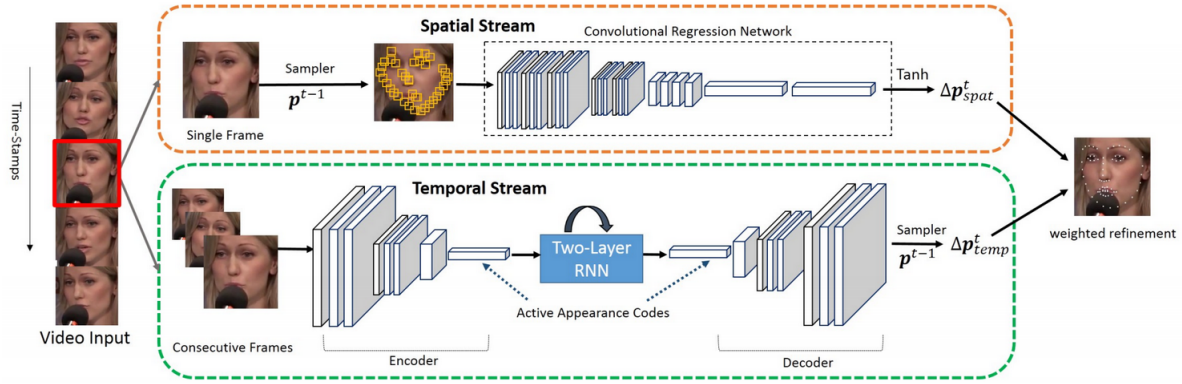

Figure 2.24: Two-stream network which decomposes the video input to the spatial and temporal streams [133]. The spatial stream aims to preserve the holistic facial shape structure while the temporal stream focuses on temporal consistency to improve shape refinements. A weighted fusion of both streams produces the final prediction.

RNNs are particularly suitable for time series analysis. They have proved to be effective for facial landmark detection using a CNN as a backbone to jointly estimate and track visual features over time [162, 79, 89]. When training is done in conjunction with a CNN, optimal features can be extracted and temporal dependencies at different scales can be captured. As a result, RNNs help stabilize predictions over time and improve robustness under uncontrolled conditions [79]. It is possible to start from a pre-trained CNN and transform any pre-trained feed-forward layer into a recurrent layer [262]. As depicted in Figure 2.24, spatial and temporal information processing can also be decoupled in order to explicitly exploit their complementarity [133]. The spatial stream aims to preserve the holistic facial shape structure while the temporal stream focuses on

temporal consistency to improve shape refinements. A weighted fusion of both streams produces the final prediction.

However, these approaches only provide late temporal connectivity on small feature maps with a possibly high level of abstraction at the level of the recurrent layer. This layer is then able to compute global motion features (e.g., head motion) but may not be capable of accurately detecting local motion (e.g., eye and lip movements) due to the fact that CNNs cannot model any motion.

2.2.6 Summary

Existing solutions for video-based facial landmark detection have been reviewed in this section. The first observation is that, even today, one of the most popular strategies for this problem is tracking by detection. Despite being considered viable, this strategy remains limited since it does not leverage temporal information at all.

More sophisticated solutions have been proposed. They take advantage of conventional tracking and adaptive learning to improve accuracy. Although these solutions are more accurate, they are likely to drift over time and thus require a reset mechanism. Moreover, they do not sufficiently make use of the temporal information as they generally only use the prediction on the adjacent frames.

Since tracking is more accurate than detection but prone to drift and detection is drift-free, they have been coupled in a joint approach. This solution allows the advantages of both approaches to be exploited more effectively but it still makes limited use of temporal information.

Similar to the shape constraint present in most image-based approaches, several solutions have also been proposed to add temporals constraint in order to increase stability over time: either using 2D models with specific training strategy or using RNNs. Due to their nature, temporal-constrained 2D models are reduced to smoothing only. Regarding RNNs, the most recent and advanced methods suffer especially from limited temporal connectivity since it relies on the use of a CNN as a backbone. However, the CNN is unable to capture temporal information, which only allows to track high level features (i.e., global motion). Fine motion cannot be easily modelled while they are just as relevant for landmark detection.

Most of the proposed approaches are based on tracking or temporal smoothing. The main weakness of all these methods is that they do not fully exploit the dynamic nature of the face. Taking this dynamic into account appears crucial to ensure high accuracy and robustness under uncontrolled conditions, especially when the appearance is no longer reliable.

There are many open questions. How can motion be modeled with more detail and integrated into current approaches based on DL? How to combine local and global motion while taking advantage of their complementarity? The same question also arises regarding spatial and temporal information. Finally, which level of time dependency is necessary for a problem like facial landmarks detection?

2.3 Discussion

Throughout the years, research has always been focused on static images. The approaches proposed so far can be grouped into two major categories: generative and discriminative. The fundamental differences between these two categories of approach include:

- model fitting based on coefficient prediction or mapping between the image and the shape;
- explicit shape model or implicitly embedded shape constraints;
- independent landmark prediction or joint prediction;
- holistic or part-based representation;
- one step prediction or coarse-to-fine strategy;
- the need of a shape initialization.

Many solutions have also been developed to address specific challenges and can be integrated into any type of approach. They are based on techniques including 3D model [287], multiview model(s) [50], multitask learning [280], GANs [24], explicit occlusions modeling [269]. However, these solutions can make dataset annotation and model training much more complex.

Discriminative approaches perform landmark prediction through a mapping between the image and the shape. This makes them faster and more accurate than the approaches

based on generative models [105, 178]. Landmarks are jointly predicted, which implicitly integrates the shape constraints. The mapping is generally implemented using a cascade of regressors, coarse-to-fine strategy, since direct mapping is hard to achieve. Among the discriminative approaches, DL ones has gained increasing popularity [23]. They allow discriminative and problem-specific feature learning, as opposed to handcrafted features used in traditional approaches. The latter are generic and likely to be suboptimal for facial landmark detection [122]. With DNNs, feature extraction and regression are trained jointly. The relationship between images and shapes is modeled more effectively. DL has notably been successful with direct mapping, showing that DR remains a relevant solution [146, 148].

Current approaches still encounter difficulties under uncontrolled conditions [184]. Although most of the challenges have been addressed, this is often achieved with solutions specific to one or a few challenges. There is a lack of approaches that can address all the difficulties at once. With regard to the literature, temporal approaches appear to be the most promising direction. Today, due to the increase in video data and their potential benefits, the interest in temporal approaches is growing [193]. Currently, most landmark detection approaches are unable to take advantage of the temporal information. They are often applied in tracking-by-detection manner. More advanced strategies have recently been proposed [32]. They mainly rely on tracking (e.g., box, landmark, pose) along with incremental learning and for a smaller part on temporal smoothing [56, 207]. However, these approaches are subject to drift and so far do not fully exploit facial dynamics. The most advanced ones based on RNNs appear to be limited to the global movements of the head [133, 89, 79]. The CNNs currently used as a backbone cannot model any motion.

Moreover, given the overall progress accomplished, more attention should be paid to evaluation protocols. There is a lack of comprehensive analysis to quantify the performance of current approaches according to the different difficulties [271]. Regarding temporal approaches, the contribution of the temporal information according to the difficulties, and compared to static approaches, is not well quantified. This would help to better identify the remaining efforts to be made but could also be useful for anyone who uses landmarks as a pre-processing in an application (e.g., expression recognition).

Consequently, the issues addressed in this work are related to the evaluation protocols and temporal approaches. The first goal is to quantify the impact of major challenges on landmark detection on the one hand, and on the other, to quantify the contribution of temporal approaches. The recent availability of datasets such as SNaP-2DFe

[2] presents an opportunity to address these issues. The second goal is to better exploit the dynamic nature of the face in order to obtain more stable predictions over time and more robustness to the full set of variations that considerably impact facial appearance. One hypothesis is that early temporal connectivity could help to model motion more finely as well as complement current approaches. Among the open questions that are also of interest are how to effectively combine facial motion information with the facial appearance. To this end, CNNs appears to be an effective and versatile baseline tool.

Chapter 3

Effectiveness of Facial Landmark Detection

Contents

Landmark detection is a common and often crucial pre-processing step in the context of facial analysis. Although its overall performance continues to improve, its impact on subsequent tasks must be considered. Let us consider, for example, a typical expression recognition process (see Figure 3.1). It relies on landmarks to normalize the face and extract appearance, geometry or motion features. Poor landmark detections may lead to confusion between expressions, decreasing the accuracy of the recognition process. Hence, it is necessary to ensure the robustness and stability of facial landmark detection under uncontrolled conditions and the suitability of the detected landmarks to the subsequent task (here, expression recognition).

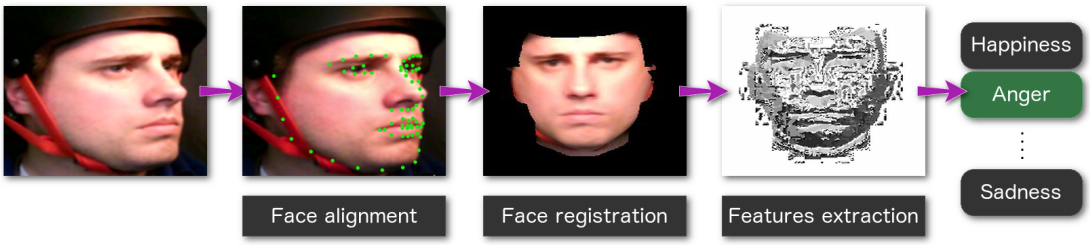

Figure 3.1: Typical facial expression recognition process. It relies on landmarks to normalize the face and extract appearance, geometry or motion features.

To the light of the review in Chapter 2, we can observe a large number and variety of approaches for facial landmark detection. Besides, these approaches are no longer limited to images but also include video-based approaches. Nowadays, it may be difficult to clearly understand the current state of the problem. Hence, there is a need for benchmarking to better identify the benefits and limits of the various approaches proposed so far, especially deep learning ones. Beyond overall performance, there is also a need to quantify the accuracy of these approaches according to the different challenges mentioned in Section 1.3. This would help to address the problem more effectively.

An overview of each of the benchmarks conducted is provided in Section 3.1. In Section 3.2, the datasets and evaluation metrics used in the literature to solve the facial landmark detection problem are first covered. In Section 3.3, the overall performance of image-based and video-based approaches is then reviewed. A comprehensive analysis of a selection of approaches is also proposed in Section 3.4, with a focus on some critical challenges for facial analysis, head pose and facial expressions, through a cross-dataset evaluation. The results are finally discussed in Section 3.5.

3.1 Overview

To get useful insights about the current state of the problem, figures that are available in the literature have been collected. We focus on two widely used datasets, 300W for still images and 300VW for videos. In this way, the figures of a large collection of approaches can be gathered. It also gives the possibility to link these results to the conclusions of the 300W [184] and 300VW [193] competitions. State-of-the-art approaches that perform fairly well on these datasets have then been selected and a third benchmark has been conducted based on SNaP-2DFe. This video dataset emphasizes two challenges, head pose and facial expressions, with annotations that allows an in-depth analysis of the performance in the presence of these challenges, independently and simultaneously. These measurements are especially relevant as landmark detection is critical for many facial analysis tasks, which are now shifting to uncontrolled conditions. Overall, the results obtained through each of these benchmarks provide an opportunity to discuss about:

- the respective value of the different types of approaches;
- the difference in performance between controlled and uncontrolled conditions;
- the most critical difficulties;
- the most impacted facial regions.

In order to properly interpret these benchmarks, a review of the datasets and evaluation metrics used to solve the problem of facial landmark detection is first provided in the following section.

3.2 Datasets and Evaluation Metrics

Datasets and evaluation metrics are fundamental to train and demonstrate the validity of algorithms. In Section 3.2.1, the major datasets captured under uncontrolled conditions to address the problem of facial landmark detection are first reviewed. The evolution of the data sets, their specificities, and the annotation process are covered. The common preprocessing and augmentation techniques are then presented in Section 3.2.2. The latter has become more and more important to limit the dataset imbalance problem and to obtain the amount of data needed by DL approaches. Evaluation metrics are similarly reviewed in Section 3.2.3.

3.2.1 Image and Video Datasets

In recent years, many datasets for facial landmark detection have been made available to the scientific community (see Table 3.1). The images included in these datasets are collected on social networks or image search services such as Google, Flickr, or Facebook, bringing more realism to the data. The annotation is performed manually, which is time consuming, but it can be alleviated with the help of the Amazon Mechanical Turk platform [1] or using semi-automatic methods [193]. The quality of the annotations, however, may vary [184, 23]. The annotation scheme used, i.e., the positions and the number of landmarks, may also differ from one dataset to another [184]. Currently, the scheme composed of 68 landmarks [78, 184] illustrated in Figure 3.2 is the most widely used. Since this scheme does not fit extreme poses (typically, profile poses) due to self-occlusions, as landmarks get stacked up, a 39-landmark scheme has also been proposed for profile faces [78, 271]. Other schemes with either fewer landmarks (e.g., only the rigid ones for each facial component) or with over a hundred landmark (e.g., including the contours of the face and of each component) can also be useful depending of the application (e.g., human-computer interaction, motion capture). Nonetheless, the 68-landmark-based scheme might be suitable for many applications with a reasonable trade-off between annotation time and information capture.

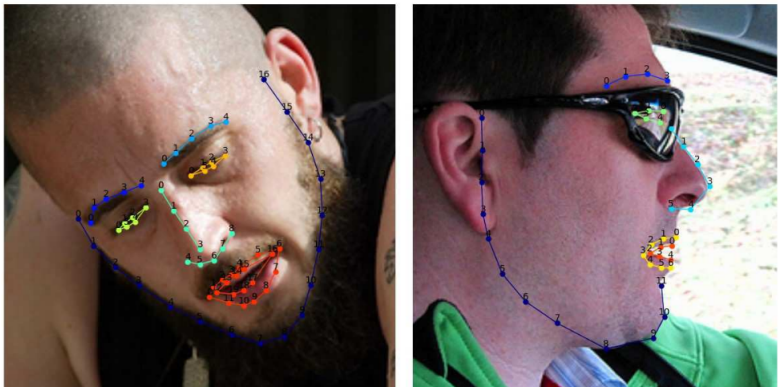

Figure 3.2: 68-landmark (left) and 39-landmark (right) schemes [271]. These schemes might be suitable for many applications with a reasonable trade-off between annotation time and information capture.

300W [184] is one of the most used datasets in the literature to train and evaluate facial landmark detection algorithms under uncontrolled conditions. This is the product of a competition of the same name, which has federated several datasets, i.e., LFPW [11], HELEN [121], AFW [286]. All these datasets have been re-annotated using the

[1] Crowdsourcing marketplace allowing individuals and companies to perform tasks, in exchange for remuneration, that computers are currently unable to perform.

68-landmark scheme and new images have been added for training and evaluation with a total of 4350 images. Some challenges such as pose variations being under-represented, an extension called 300W-LP [288] has been proposed in order to provide more images with large pose variations. However, these new data, synthesized using the 3DMM [15] contain artifacts that might affect the accuracy of landmarks.

Type	Datasets	Year	#Images	#Landmarks
Static	AFLW [109]	2011	25,993	21
	AFW [286]	2012	205	6
	HELEN [121]	2012	2,330	194
	LFPW [11]	2013	1,432	29
	COFW [25]	2013	1,007/507	29
	IBUG [185]	2013	135	68
	MTFL [279]	2014	12,995	5
	MAFL [280]	2016	20,000	5
	300W [184]	2016	600	68
	300W-LP [288]	2016	61,225	68
	MENPO [271]	2017	10,993/3852	68/39
	WFLW [235]	2018	10,000	98
Dynamic	300VW [193]	2015	218,595	68
	SNaP-2DFe [2]	2018	93,240	68

Table 3.1: Datasets captured under unconstrained conditions. The 68-landmark scheme is the most widely used. There is only limited video data available.

More specific datasets have also been proposed. COFW [25] gives more emphasis to occlusions. Initially annotated with 29 landmarks, it has recently been updated to the 68-landmark scheme. MTFL [279] and MAFL [280] focus on multitask learning and facial attributes, but with only 5 annotated landmarks. In an effort to increase the number of challenging data, especially for DL approaches, MENPO [271] has been developed, in conjunction with the competition of the same name. It contains 10,993 images for semi-frontal faces and 3852 for profile faces, obtained from large scale datasets, AFLW [109] and FDDB [96], and respectively annotated with the 68-landmark and 39-landmark schemes. Samples can be seen in Figure 3.3. However, these datasets make it difficult to determine the sources of error. Recently, a new dataset, WFLW [235], based on WIDER Face [260], has been published with rich annotations: 98 landmarks, occlusions, pose, make-up, illumination, blur, and expression. It aims to enable a comprehensive analysis of existing algorithms.

In practical applications, facial landmark detection algorithms are commonly applied to videos. It appears crucial to have data that is representative of the problem in

order to answer it. Static datasets do not cover all the difficulties encountered by applications; especially, the ones related to the movements of persons or cameras are not currently considered. 300VW [193] is the only dataset developed for video-based landmarks detection under uncontrolled conditions to be published. It contains 114 videos of about 1 minute each and featuring one person annotated with 68 landmarks. 50 videos are intended for training and 64 for testing. The test set is divided into 3 categories of increasing difficulty: category 1 presents videos recorded in well-lit conditions with various head poses, category 2 contains additional lighting variations, and category 3 includes severe difficulties such as lighting, occlusions, expression, and head pose (see Figure 3.4). However, the challenging data are not as diverse and extreme as those found in static datasets.

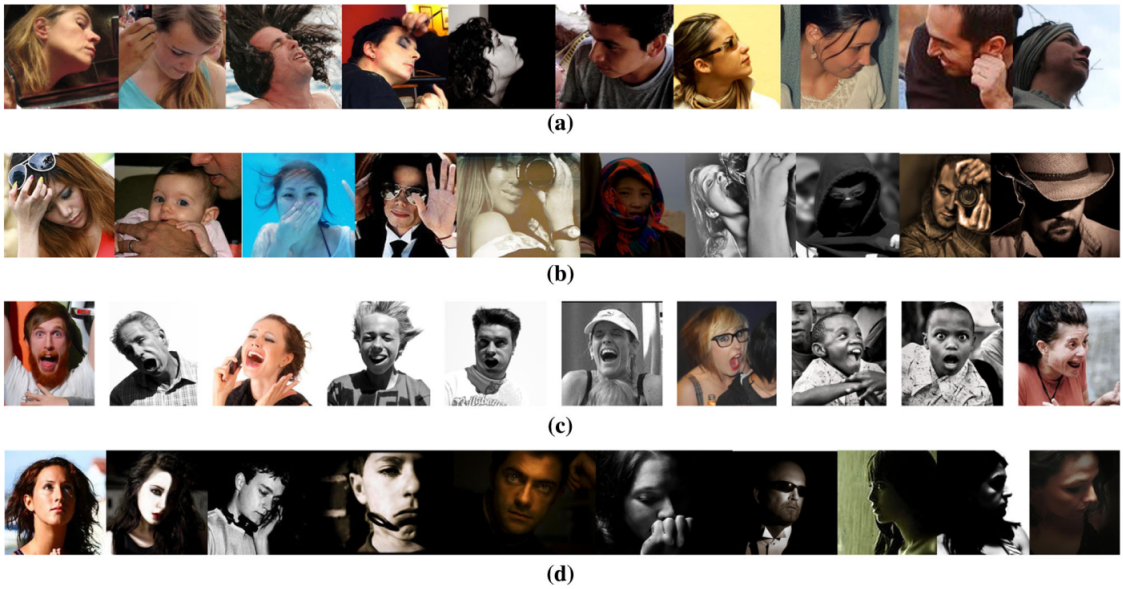

Figure 3.3: Illustration of the images contained in the MENPO dataset (a = pose; b = occlusion; c = expression; d = illumination) [49].

Given the limited video data available, other video datasets have been reviewed, and one of them has shown valuable specifications. SNaP-2DFe [2] is a video dataset recently developed to quantify the impact of head movements on expression recognition performance. As it contains landmark annotations, it also provides rich annotations, movement and expression, to study video-based facial landmark detection. It consists of 6 movements composed of a horizontal translation and/or a rotation (roll, pitch, yaw), each associated with 7 acted expressions (neutral, happiness, fear, anger, disgust, sadness, surprise) corresponding to the universal expressions suggested by [58]. The temporal patterns of expression activation (i.e., neutral-onset-apex-offset-neutral) are also provided, where apex refers to the highest intensity of an expression. Data from 15

participants have been collected by two synchronized cameras (i.e., helmet and static camera), with a total of 1260 videos. Samples with pitch movement and surprise expression are illustrated in Figure 3.5.

Figure 3.4: Illustration of each category of the 300VW dataset: (a) well-lit conditions, (b) lighting variations, and (c) severe difficulties [32]. These difficulties are not as extreme as in still image datasets.

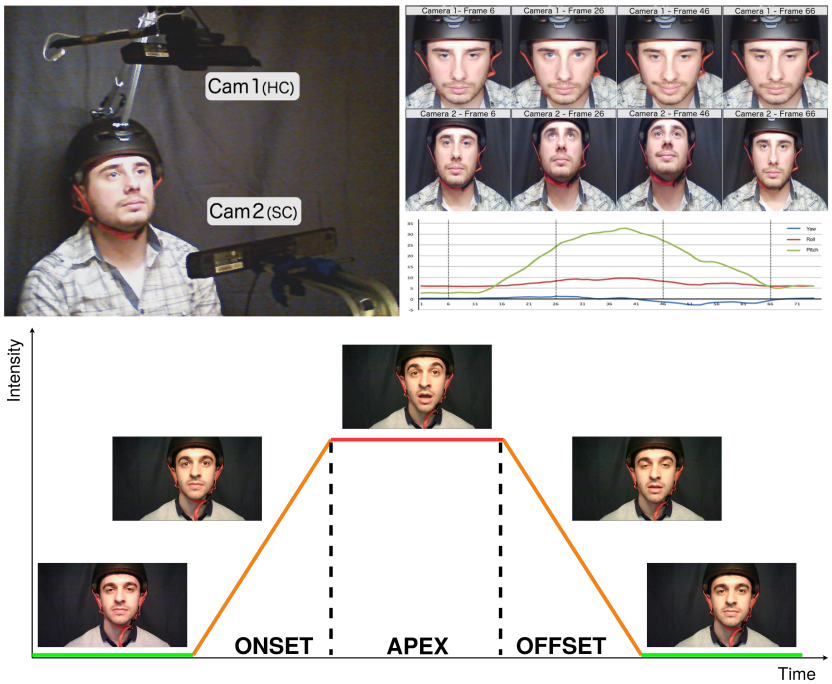

Figure 3.5: Sample images of facial expressions recorded under pitch movements from the SNaP-2DFe dataset (row 1: helmet camera, only expression movement; row 2: static camera, expression and head movement). The bottom plot shows the different patterns of a facial expression: neutral (green), onset (orange), apex (red), and offset (orange).

3.2.2 Face Preprocessing and Data Augmentation

Various preprocessings are generally applied to facial images before being used for facial landmark detection. The most common and relevant are related to scale variations and face region size. To reduce the variance of the regression target, scale variations are removed by rescaling the bounding box produced by the face detector, e.g., to 256x256 pixels for the state-of-the-art Hourglass Network [23]. Depending on how the face region is obtained (e.g., different detectors give different bounding boxes), some landmarks may be missing. It is useful to enlarge the region, e.g., by 30% [46], to ensure that all landmarks are included. Histogram Equalization can be used to improve contrast by adjusting the distribution of intensities. Normalization is also applied to ensure similar data distribution. It is generally calculated by subtracting the mean from each pixel followed by a division by the standard deviation. The resulting distribution is expected to be a Gaussian centered at zero. When using ANNs, it can be necessary to scale the data depending of the activation function to prevent saturation.

Today, due in particular to the emergence of DL approaches, it may be necessary to have a large number of annotated images to reduce overfitting and improve robustness, which existing datasets do not necessarily provide. In the literature, various augmentation methods are used to circumvent this problem. Some simple operations (e.g., rotation, mirroring, blur, jittering of the position and size of the detection window, jittering of the initial landmarks) can be applied to images to generate new training samples while preserving the ground truth (see Figure 3.6).

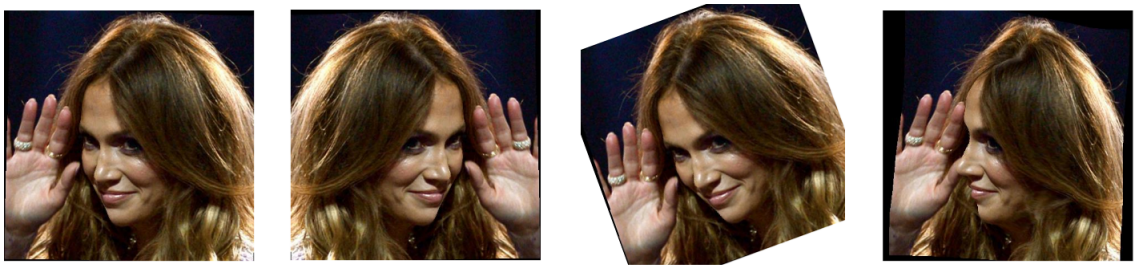

Figure 3.6: Illustration of different augmentation operations: (a) original image, (b) mirroring, (c) rotation, (d) synthesized profile view.

Other more complex processes may also be used to provide new relevant samples based on the initial training data. As mentioned in Section 3.2.1 about the 300W-LP dataset, 3D approaches have proved [288] to be suitable for certain specific augmentation. Inspired by face frontalization [85], it consists, once the depth of a face image is predicted, in synthesizing the corresponding face image in profile view with a 3D

rotation. The result is depicted in Figure 3.6-(d). However, current solutions may be affected by undesirable artifacts. Besides, Hard Sample Mining (HSM) has become a popular strategy to increase the number of training data while explicitly addressing the data imbalance problem. This problem can result in overfitting on well-represented samples and a poor adaptation to under-represented ones. HSM can be performed very easily by looking at the error value during training. By following an offline strategy, training samples with high error values are kept for a second training phase [114]. The same process can be done online with DNNs by selecting about 70% of training samples with the highest loss values during the forward pass, and only compute gradient values for these selected samples in the backward pass to ignore the easiest cases [275]. More advanced offline techniques that rely on PCA have also been used [66]. By projecting the original shapes onto the space defined by the shape eigenvector controlling pose variations, under-represented samples are highlighted and privileged for augmentation.

More recently, with the emergence of GANs, the joint optimization of data augmentation and model training has also been proposed [165]. An augmentation network (A) is used to generate augmentation operations online to improve the training of the target network (T). The two networks being in competition with each other, A aims to exploit T's weaknesses by generating hard augmentations while T learns from these hard augmentations.

3.2.3 Evaluation Metrics

To evaluate the predictions of multiple landmarks on an image, the Root Mean Square Error (RMSE) normalized by the interocular (or interpupil) distance (NRMSE) is generally used. The purpose of the normalization is to ensure the invariance to scale, to avoid an undesirable impact of face size on the error. NRMSE can be expressed as:

$$NRMSE = \frac{1}{L} \frac{\sum_{i=1}^{L} \sqrt{(x_i^p - x_i^g)^2 + (y_i^p - y_i^g)^2}}{\sqrt{(x^{le} - x^{re})^2 + (y^{le} - y^{re})^2}} \tag{3.1}$$

where L is the number of landmarks, p_i the i-th predicted landmark, g_i the corresponding ground truth landmark, and le and re the left and right pupil centers from the ground truth. The normalization by the interocular distance, although not very robust to extreme poses due to non visible eyes, is the most widespread. Other normalization factors are sometimes used, such as the face size obtained from the bounding box, or a distance

computed on one side of the face, e.g., between an eye and the mouth [23, 32, 238]. On a set of images, average NRMSE is the simplest and most intuitive evaluation metric. It should be noted that it can be strongly affected by a few outliers. Figure 3.7 shows that beyond an error value of 8%, landmarks are mostly not located correctly. Hence, the prediction is generally considered as a failure.

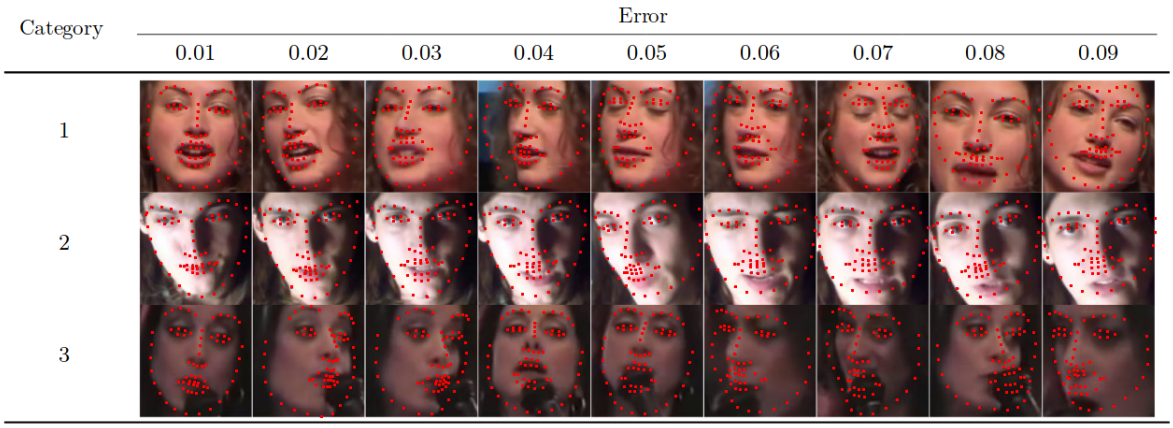

Figure 3.7: Predictions quality for each error value [32]. Beyond an error value of 8%, landmarks are mostly not located correctly.

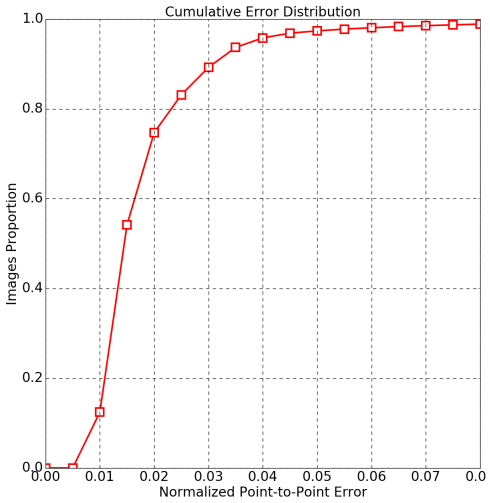

Figure 3.8: Graphical representation of the CED function.

A graphical representation of the Cumulative Error Distribution (CED) function is also widely used (see Figure 3.8). It corresponds to the proportion of images as a function of the NRMSE up to a certain threshold, e.g., 8%. The Area Under the Curve (AUC) and the Failure Rate (FR), that is, the percentage of images for which the error is greater than the threshold, are sometimes calculated from this representation.

The AUC and FR are expressed as:

$$\mathrm{AUC}_\alpha = \int_0^\alpha f(e)de \tag{3.2}$$

$$\mathrm{FR}_\alpha = 1 - f(\alpha). \tag{3.3}$$

where f is the cumulative error distribution (CED) function and α the threshold. Another important evaluation criterion is the efficiency in terms of speed, which is calculated in Frames Per Second (FPS) using regular CPU/GPU hardware, and size of the model.

3.2.4 Summary

Over the past few years, there has been an evolution towards datasets that are increasingly complex and diversified to best represent natural conditions. Major challenges are nowadays well represented along with extreme cases. The data are generally annotated manually or using semi-automatic methods. There is a harmonization of the landmark scheme with the most common used being the 68-landmark, along with a more restricted version based on 39-landmark for profile faces. For extreme poses and profile faces, half of the face being self-occluded, the 68-landmark scheme no longer makes sense as many landmarks overlap. To limit dataset imbalance, various HSM techniques have also been developed. Regarding evaluation, NRMSE in terms of accuracy and FPS in terms of efficiency are the metrics typically used. The normalization of the RMSE by the diagonal of the bounding box appears to be a good choice as it is more robust to extreme pose variations.

Due to the shift towards DL, there is a need for large scale datasets. Although there have been efforts in this direction, it is not yet satisfactory. The difficulty of building large datasets is probably related to the annotation of the landmarks, which can be very time-consuming. Currently, the preferred solution to obtain more data is to use augmentation techniques, which are becoming more and more sophisticated. As the community focuses on static data, there is also only limited video data available. Besides, with the continuous improvement of landmark detection approaches, richer annotations are also required to allow more in-depth analysis of algorithm capabilities. Datasets that were not originally designed for the landmark detection problem can still be useful.

3.3 Image and Video Benchmarks

In the following, the two benchmarks obtained by compiling the results from the literature are introduced. The first benchmark covers image-based algorithms and can notably be used to confront traditional and deep learning approaches along with fine and challenging conditions. In the second benchmark, video-based algorithms are included to provide an additional comparison between static and dynamic approaches. Full details on the datasets and approaches used can be found in Section 2.1, 2.2, and 3.2.

3.3.1 Compiled Results on 300W

Figure 3.9: Samples of ESR, SDM, LBF and TCDCN on 300W [280]. Landmarks of non-rigid facial structures, especially on the outline of the face, are found to be the most difficult to detect.

The performances of leading methods for image-based facial landmark detection are reported in Table 3.2. These methods were evaluated on the 300W dataset [184], which is the most widely used dataset in the literature. As a reminder, it composed of 2 categories of variable difficulty. Category A corresponds to images that do not include severe difficulties. Category B contains more complex facial images with large variations in pose, illumination, and expressions, as well as occlusions. First, note that CSR based on various DNN architectures is the predominant approach (see for example, CFAN, MDM, DAN, LAB). Secondly, for category B, the average error is more than twice the one obtained on category A. Despite the quantity of methods proposed in the literature and the recent major advances, these results show that the problems encountered under uncontrolled conditions are still far from being solved.

Approach	Method	Year	Category A	Category B	Full set
Traditional	DRMF [5]	2013	6.65	19.79	9.22/-/-
	RCPR [25]	2013	6.18	17.26	8.35/-/-
	SDM [248]	2013	5.57	15.40	7.52/42.94/10.89
	ESR [27]	2014	5.28	17.00	7.58/43.12/10.45
	ERT [105]	2014	-	-	6.40/-/-
	LBF [178]	2014	4.95	11.98	6.32/-/-
	CFSS [285]	2015	4.73	9.98	5.76/49.87/5.08
Deep learning	CFAN [274]	2014	5.50	16.78	7.69/-/-
	TCDCN [279]	2014	4.80	8.60	5.54/-/-
	MDM [215]	2016	4.83	10.14	5.88/52.12/4.21
	3DDFA [288]	2016	6.15	10.59	7.01/-/-
	RAR [245]	2016	4.12	8.35	4.94/-/-
	TSR [139]	2017	4.36	7.42	4.96/-/-
	DRR [116]	2017	-	-	4.90/-/-
	DAN [111]	2017	-	-	3.59/55.33/1.16
	DVLN [236]	2017	3.94	7.62	4.66/-/-
	FARN [89]	2018	4.23	7.53	4.88/-/-
	TCD [114]	2018	3.67	7.62	4.44/-/-
	RDN [134]	2018	3.31	7.04	4.23/-/-
	LAB [235]	2018	3.42	6.98	4.12/-/-
	SBR [56]	2018	3.28	7.58	4.10/-/-

Table 3.2: Performances of recent image-based methods on 300W. NRMSE is reported for categories A and B. NRMSE/AUC/FR are reported for the full set. Numbers are taken from the literature. CSR based on various DNN architectures is the predominant approach. The average error on category B is more than twice the one obtained on category A. The problems encountered under uncontrolled conditions are still far from being solved.

By looking at qualitative results such as the ones in Figure 3.9, interesting observations can also be made. Because of their significant influence on facial appearance, variations in pose and expression along with occlusions seem to be among the most difficult challenges. Landmarks of non-rigid facial structures (e.g., those around the mouth) are found to be the most difficult to detect as they can be strongly impacted by deformations related to expressions. Landmarks on the outline of the face are even more challenging, probably due to the lack of distinctive features. Deep learning approaches are now very popular and state-of-the-art. They demonstrate striking results compared to traditional approaches, especially in category B. This is most likely related to their ability to represent complex non-linear functions and to learn discriminative task-specific features while benefiting from data augmentation. Nevertheless, they have difficulty running in real-time, even on GPU. Traditional CSR approaches are much faster, reaching up to 1000 to 3000 FPS [105, 178]. It should be noted that the training

procedure (pre-training, training data, augmentation) is not always clearly provided and may vary from one method to another, which can impact the results and make them difficult to compare. However, similar observations with additional details are reported in competition reports, which provide original evaluations with identical datasets and protocols [184, 271].

3.3.2 Compiled Results on 300VW

Approach	Method	Year	Category 1	Category 2	Category 3	Full set
Static	SDM [248]	2013	7.41	6.18	13.04	7.25
	ESR [27]	2014	-	-	-	7.09
	CFSS [285]	2015	7.68	6.42	13.67	6.13
	CFAN [274]	2014	-	-	-	6.64
	TCDCN [279]	2014	7.66	6.77	14.98	7.59
Dynamic	RED-NET [162]	2016	-	-	-	5.15
	TSTN [133]	2018	5.36	4.51	12.84	-

Table 3.3: Performances of recent image-based and video-based methods on 300VW. NRMSE is reported for categories 1, 2, and 3 and the full set. Numbers are taken from the literature. DNNs based on RNN are the predominant approach. The results demonstrate better robustness of temporal approaches under uncontrolled conditions. Temporal constraints appear to be just as important as spatial constraints.

In the same way as image-based approaches, the performance of leading video-based facial landmark detection methods are referenced in Table 3.3. These methods were evaluated on the 300VW dataset [193] which, as a reminder, is composed of three categories of increasing difficulty. It is currently the only video dataset available. Some image-based methods ("Static" row in Table 3.3) have been included for comparison [248, 27, 285, 274, 279]. Since the interest in video-based approaches is more recent, possibly due to data availability, much fewer results are available.

DNNs are also the predominant approach. They tend to provide more stable predictions without large errors since they take advantage of the temporal coherence (see Figure 3.10). Temporal constraints appear to be just as important as spatial constraints, as proved by the results that demonstrate better robustness under uncontrolled conditions. RNN-based approaches are currently very popular [162, 79], especially two-stream architectures [133]. The lack of data being more important when considering videos (more complex model, redundancy in data), one of the advantages of such architectures is that fine-tuning can be performed for the spatial stream from models trained on static datasets. Although the results demonstrate the usefulness of temporal infor-

mation, the trend to use RNNs turns out to be limited. Combining CNNs and RNNs provides late temporal connectivity, most certainly capable of detecting only global motion from high level feature maps, e.g., head movements, since the CNNs used cannot detect any motion.

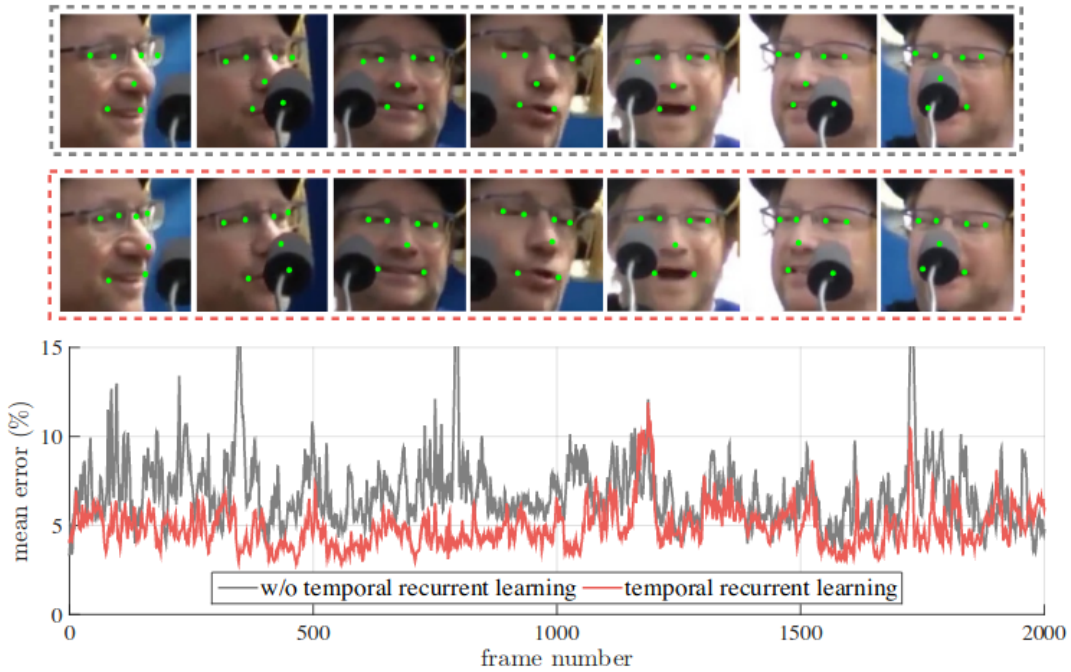

Figure 3.10: Mean error of RED-NET with and without recurrent learning on 300VW [162]. Recurrent learning tends to provide more stable predictions without large errors.

3.4 Cross-Dataset Benchmark

Relying on popular datasets does not allow for a comprehensive analysis of the results. In recent years, many datasets have been published and have helped to improve the robustness of facial landmark detection algorithms. Although these datasets currently exhibit a wide range of difficulties, they do not allow a straightforward identification and quantification of the weaknesses of algorithms with regard to these difficulties. This is actually difficult to determine the origins of failures as the data are not annotated by challenge. They lack suitable annotations as well as similar data captured in the absence of challenges in order to provide a rigorous evaluation. The performance of the approaches cannot therefore be effectively measured in the presence of head pose, facial expression, or any other difficulty. However, new datasets with richer annotations, such as SNaP-2DFe [2], have been introduced, allowing for a comprehensive analysis regarding major challenges. SNaP-2DFe represents a great candidate for our third and

last benchmark. In this section, the evaluation protocol is first presented, and then an in-depth analysis is conducted to investigate both the impact of head pose and facial expressions on facial landmark detection.

3.4.1 Evaluation Protocol

The aim of the following benchmark is not to make a comparison between approaches in order to find the best one but rather to evaluate their aptitudes in the presence of head pose and facial expression. Note that we chose not to do any fine-tuning on SNaP-2DFe. One reason is that the generalization capacity of current landmark detection approaches can also be assessed by doing so. As the models used are trained on uncontrolled datasets where data are more challenging, they should provide equivalent performance on SNaP-2DFe.

Selected Dataset

As already mentioned, SNaP-2DFe has been selected to conduct the third benchmark. Unlike other datasets, annotations of 6 head movements and 7 facial expressions along with the temporal pattern of their activation (i.e., neutral-onset-apex-offset-neutral) are provided. Only the data from the static camera (sc) is used, which represents a total of 630 videos. The 6 head movements include horizontal translation and/or rotation (i.e., Static, Translation-x, Roll, Yaw, Pitch, Diagonal). The 7 facial expressions correspond to the universal expressions (i.e., Neutral, Happiness, Fear, Anger, Disgust, Sadness, Surprise).

Selected Approaches

Given the large number of facial landmark detection approaches in the literature, we select a subset of recent approaches representative of the current state of the problem. Deep learning approaches based on state-of-the-art models for coordinate (DAN [111]) and heatmap (HG, SAN [23, 55]) regression are included. Complementary multi-task (TCDCN [280]) and dynamic solutions (SBR, FHR [56, 207]) are considered.

Unfortunately, the temporal component of the selected dynamic approaches was not provided by the authors. Due to the lack of code and critical details for implementation, we were not able to successfully integrate it. These approaches have still been kept as image-based algorithms and therefore the focus has been narrowed down to a static 2D landmark evaluation.

Since the code and the pre-trained models provided by the respective authors are used, it is important to consider the datasets used to train these approaches. As shown in Table 3.4, most approaches are trained mainly on 300W, HELEN, and AFLW. HG, TCDCN, SBR, and FHR made use of additional datasets, either to improve performance by limiting overfitting, or to provide the necessary multi-task component.

Table 3.4: The different datasets used to train the selected approaches. Most approaches are trained mainly on 300W, HELEN, and AFLW.

Dataset	HG	TCDCN	DAN	FHR	SBR	SAN
300W [184]	✓	✓	✓	✓	✓	✓
HELEN [121]	✓	✓	✓	✓	✓	✓
AFLW [109]	✓	✓	✓	✓	✓	✓
COFW [25]	-	✓	-	-	-	-
MAFL [280]	-	✓	-	-	-	-
300W-LP [288]	✓	-	-	-	-	-
Mempo [271]	✓	-	✓	-	-	-
300VW [193]	✓	-	-	-	-	-

Evaluation Metrics

The RMSE normalized by the diagonal of the ground truth bounding box, $\sqrt{width^2 + height^2}$, is used as evaluation metric for its robustness to head pose variations [32]. From this metric, the AUC and FR with a 4% threshold are computed. Above this threshold, the prediction is considered as a failure, since a facial component can be mismatched. In some experiments, the difference in performance between a static neutral face and a face with the different variations, $performance_{variations} - performance_{static\ neutral}$, is also used.

3.4.2 Comparison of Selected Approaches

In the following experiments, the robustness of the selected facial landmark detection approaches in the presence of head pose variations and facial expressions is investigated. An overall performance analysis is first performed in order to identify which factor of variation challenges the approaches studied the most. The analysis then focuses on each facial landmark individually in order to identify more precisely which facial regions are more difficult to characterize due to facial expressions and head pose variations.

Overall Performance Analysis

In this experiment, the objective is to determine which movement, from the one caused by head pose variations or the one produced by facial expressions, has the greatest impact on facial landmark detection. Table 3.5 shows the performance of the selected approaches in the presence of head pose variations only (i.e., neutral face with the 6 head movements), facial expressions only (i.e., static face with the 6 facial expressions), and head pose variations combined with facial expressions (i.e., the 6 facial expressions along with the 6 head movements). For each approach, AUC and FR are computed over the entire sequences.

Table 3.5: AUC/FR by head pose variation and facial expression on SNaP-2DFe. AUC is lower and FR higher in the presence of head pose variations than in the presence of facial expressions for all approaches. Head pose variations appears to be more challenging than facial expressions for facial landmark detection.

	Head pose only	Expressions only	Head pose & expressions	Average
HG	54.30 / 0.42	55.35 / 0.05	52.94 / 0.58	54.19 / 0.35
TCDCN	54.86 / 4.49	59.73 / 0.00	52.97 / 5.01	55.85 / 3.16
DAN	68.73 / 8.44	71.77 / 6.41	67.88 / 8.34	69.46 / 7.73
FHR	69.90 / 0.17	71.84 / 0.00	68.52 / 0.64	70.08 / 0.27
SBR	70.44 / 1.02	73.76 / 0.57	68.86 / 2.02	71.02 / 1.20
SAN	70.97 / 0.21	73.80 / 0.00	70.27 / 0.44	71.68 / 0.21
Average	64.87 / 2.46	67.71 / 1.17	63.57 / 2.84	65.38 / 2.15

Table 3.5 indicates that the AUC is lower and the FR higher in the presence of head pose variations than in the presence of facial expressions for all approaches. These results suggest that head pose variations are more challenging than facial expressions for facial landmark detection. This is not surprising since the appearance is generally much more affected by a change in pose than a change in expression. In the simultaneous presence of both challenges, the performance tends to decrease further.

In order to obtain more accurate performance measurements of each approach according to head pose variations and facial expressions, the previous results on the full set (i.e., head pose variations and expressions) are detailed by activation pattern in Table 3.6.

As shown in Table 3.6, the accuracy of all approaches decreases the most for images adjacent to the apex state. More specifically, it corresponds to the moment where the expression and most head pose variations are at their highest intensity. As soon as the face gets closer to a neutral expression and a frontal pose, the accuracy improves. These

results suggest that facial expressions with head pose variations remain a major difficulty for facial landmark detection and show that not only the presence of an expression, but also its intensity, impact landmark detection.

Table 3.6: AUC/FR by activation pattern on SNaP-2DFe in the presence of facial expressions and head pose variations. These results suggest that facial expressions with head pose variations remain a major difficulty for facial landmark detection. Not only the presence of an expression, but also its intensity, impact landmark detection.

	Neutral / Onset	Onset / Apex	Apex / Offset	Offset / Neutral	All
HG	55.21 / 0.06	49.30 / 1.73	48.58 / 1.56	51.74 / 0.65	52.94 / 0.58
TCDCN	55.55 / 3.17	44.70 / 11.77	45.99 / 9.64	51.84 / 4.92	52.97 / 5.01
DAN	69.87 / 7.91	62.35 / 9.95	64.07 / 8.99	67.51 / 8.34	67.88 / 8.34
FHR	71.57 / 0.04	64.14 / 2.37	63.24 / 1.67	67.15 / 0.62	68.52 / 0.64
SBR	71.74 / 1.04	62.22 / 5.12	63.75 / 3.42	67.35 / 2.83	68.86 / 2.02
SAN	72.79 / 0.00	65.65 / 1.22	66.41 / 1.08	68.86 / 0.56	70.27 / 0.44
Average	66.12 / 2.04	58.06 / 5.36	58.67 / 4.39	62.41 / 2.99	63.57 / 2.84

Robustness to Head Pose Variation

In this experiment, we place the emphasis on head pose variations, by highlighting which type of head pose presents the most difficulties. In Table 3.7, the results are split by head pose variations. AUC and FR are computed on all sequences where the face is neutral between the onset and the offset phase. The results (ΔAUC and ΔFR) correspond to the difference in performance between a static neutral face and a neutral face in the presence of the different head poses. Negative (resp. positive) values for ΔAUC (resp. ΔFR) indicate difficulties in the presence of the given head pose.

According to the results in Table 3.7, some head pose variations seem more difficult to handle than others. Diagonal and Pitch lead to a severe drop in the AUC and a considerable increase in the FR, suggesting that these pose variations are the most challenging. They involve out-of-plane rotations, some of them on several rotation axes, which has a significant impact on the face appearance. This may also be related to an insufficient amount of training data for these movements. There is a drop in the AUC for Yaw, Roll, and Translation-x as well, but with an almost zero failure rate, showing a better ability of the approaches to manage these variations. The errors generated are fairly small and do not result in landmark detection failures. Even for SAN, which has the best overall performance, the average ΔAUC remains high with a value of -5.48 with, however, a average ΔFR of 0.

Table 3.7: ΔAUC / ΔFR between a static neutral face and the different head pose variations on SNaP-2DFe. Diagonal and Pitch lead to a severe drop in the AUC and a considerable increase in the FR. These pose variations are the most challenging.

	Translation-x	Roll	Yaw	Pitch	Diagonal	Average
HG	-0.89 / 0.0	+0.67 / 0.0	-4.17 / 0.0	-1.19 / +0.67	-10.18 / +3.67	-3.15 / +0.86
TCDCN	-0.36 / 0.0	-1.36 / 0.0	-21.33 / +24.0	-9.39 / 0.0	-30.65 / +33.67	-12.61 / +11.53
DAN	-4.0 / +2.0	-13.15 / +15.0	-1.76 / -1.0	-8.8 / +0.66	-15.79 / +4.33	-8.7 / +4.19
FHR	-0.97 / 0.0	-1.69 / 0.0	-0.66 / 0.0	-10.08 / 0.0	-7.5 / +2.0	-4.18 / +0.4
SBR	-4.21 / -0.33	-2.82 / -1.0	-7.15 / -0.33	-12.36 / +0.0	-16.71 / +1.67	-8.65 / 0.0
SAN	-0.75 / 0.0	-5.22 / 0.0	-2.36 / 0.0	-8.02 / 0.0	-11.08 / 0.0	-5.48 / 0.0
Average	-1.86 / +0.28	-3.93 / +2.33	-6.24 / +3.78	-8.31 / +0.22	-15.32 / +7.56	-7.13 / +2.83

Robustness to Facial Expressions

In this experiment, we emphasize facial expressions, by highlighting which type of facial expression presents the most difficulties. In Table 3.8 the results are split by facial expression. AUC and FR are computed on all sequences where the face is static from the onset to the offset of the facial expression activation. The results (ΔAUC and ΔFR) correspond to the difference in performance between a static neutral face and a static face with the different facial expressions. Negative (resp. positive) values for ΔAUC (resp. ΔFR) indicate difficulties in the presence of the given facial expression.

Table 3.8: ΔAUC / ΔFR of between a static neutral face and the different facial expressions on SNaP-2DFe. Disgust and Sadness lead to a significant drop in the AUC. These facial expressions are the most challenging.

	Happiness	Anger	Disgust	Fear	Surprise	Sadness	Average
HG	-4.85 / +0.36	-4.41 / 0.0	-8.68 / 0.0	-1.75 / 0.0	-1.35 / 0.0	-3.39 / 0.0	-4.07 / 0.06
TCDCN	-3.12 / 0.0	+1.51 / 0.0	-8.1 / 0.0	-0.02 / 0.0	+0.55 / 0.0	-4.57 / 0.0	-2.29 / 0.0
DAN	+0.48 / -4.30	-0.86 / -0.22	-7.04 / +0.74	-1.29 / -2.16	-1.21 / -0.12	-3.11 / -2.67	-2.17 / -1.45
FHR	-1.44 / 0.0	-2.28 / 0.0	-6.62 / 0.0	-0.96 / 0.0	+0.93 / 0.0	-4.09 / 0.0	-2.41 / 0.0
SBR	+0.04 / -0.64	-0.37 / -1.00	-7.27 / -0.63	-3.49 / +2.01	-0.37 / -1.00	-4.03 / -0.38	-2.58 / -0.27
SAN	+0.34 / 0.0	+1.0 / 0.0	-7.41 / 0.0	-1.52 / 0.0	-1.61 / 0.0	-4.12 / 0.0	-2.22 / 0.0
Average	-1.43 / -0.76	-0.90 / -0.20	-7.52 / +0.02	-1.51 / -0.03	-0.51 / -0.19	-3.89 / -0.51	-2.62 / -0.28

The results reported in Table 3.8 show that some facial expressions seem more difficult to handle than others. Disgust and Sadness lead to a significant drop in the AUC. These expressions involve complex mouth motions with significant changes in appearance, which may explain why the different approaches have more difficulties handling them. Besides, they also present a wider range of activation patterns and intensities, and may be less present in the datasets considered for training. The decrease in the AUC is less marked for happiness, anger, fear, and surprise, which seems to be handled better. As previously, if the focus is on the best selected approach in terms of overall performance (SAN), a large value of the average ΔAUC, -2.22, is observed.

Analysis at the Landmark Level

To better identify the strengths and weaknesses of facial landmark detection approaches, a more detailed analysis at the level of each landmark is proposed. Figure 3.11 illustrates the landmark error rates of the different approaches according to the head pose variations (Figure 3.11-A), the facial expressions (Figure 3.11-B) and the combination of both (Figure 3.11-C).

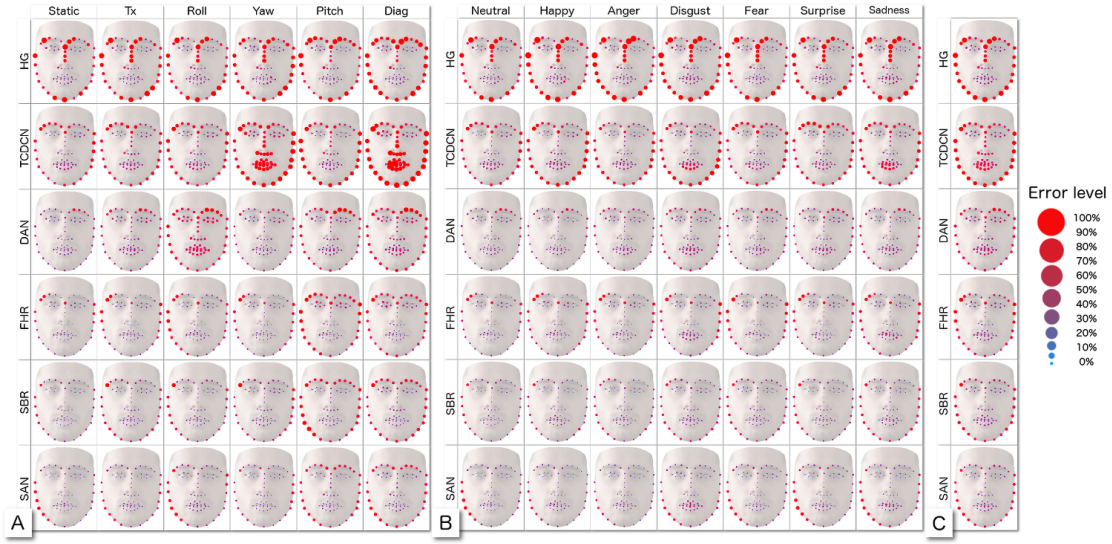

Figure 3.11: Heatmaps of landmark detection error (A: per head pose, B: per facial expression, C: in all). Eyebrows, facial edges, and the mouth are among the regions that are the most affected by head pose and facial expression.

Regarding head pose variations (Figure 3.11-A), in-plane transformations (Static, Translation-x, and Roll) do not have a strong impact on most approaches. Eyebrows, nose, and facial edges are among the regions where landmark detection lacks precision, especially for TCDCN, HG, and DAN. Out-of-plane (Yaw, Pitch and Diag) variations, however, result in much more significant decreases of the accuracy. These transformations tend to stack up landmarks due to self-occlusions of certain parts of the face. Pitch and Diagonal movements challenge all approaches; during them, eyebrows and facial edges are generally poorly detected. In the presence of facial expressions (Figure 3.11-B), eyebrows and facial edges are also the most impacted regions for all approaches. Disgust and sadness are among the most challenging expressions as they produce complex inner motions, which also lead to poor detection around the mouth. When both challenges occur (Figure 3.11-C), eyebrows, facial edges and mouth are therefore heavily affected. Note however that inner landmarks present less noise, especially around the mouth. This could simply be related to the training data or an implicitly stronger shape constraint during simultaneous variations.

3.5 Discussion

The first image-based benchmark has shown that the average error under uncontrolled conditions is more than twice the one obtained in good condition. It was also noted that the trend towards deep learning approaches has spread to this problem. While large datasets are not currently available, performance already tends to show the positive contribution of these approaches. However, regarding computation times, traditional approaches seem better as long as DNN optimization techniques (as presented in Section 2.1.3) are excluded. The second video-based benchmark highlighted that temporal consistency could help to address some of the difficulties encountered under uncontrolled conditions.

The third and last cross-dataset benchmark showed that the accuracy of current facial landmark detection approaches is still not satisfactory in presence of head pose variations and/or facial expressions. The difficulties encountered are mainly related to certain expressions (disgust and sadness), due to their complex motion, and to certain head pose variations (Pitch, Diag), due to out-of-plane transformations. Another reason could be their limited representation in the training datasets. The combination of both, which is closer to the data captured under uncontrolled conditions, naturally increases the difficulty of the detection. Eyebrows, facial edges, and the mouth are among the regions that are the most affected by these difficulties. This may be explained by a potential lack of distinctive features, especially for outline landmarks, and the propensity of landmarks to overlap in these regions. The landmark-wise analysis could be extended. Depending on the application, it may not be necessary to accurately detect all facial landmark to properly characterize a face.

Facial landmark detection approaches tend to become increasingly robust in the presence of head pose variations and facial expressions but are not yet robust enough to detect facial landmarks under uncontrolled conditions. It is reasonable to believe that the development and use of training datasets such as SNaP-2DFe will help to design algorithms that meet the expectations of the underlying tasks. While some suggest adding more data and/or building deeper models, a clever design of the model and/or the training protocol could greatly contribute to improve the robustness and generalization capacity of DNNs. SAN, which relies on the use of style-aggregated images, is a good illustration and has shown its effectiveness. Unfortunately, video-based approaches were unable to be included in the third benchmark but given the performance of FHR [207] and SBR [56] without their temporal component, one can still imagine the potential benefit it could have brought. For instance, the authors of FHR report a

decrease in the average error close to 20% on 300VW category 3 by adding the temporal component. However, these methods only rely on image-based models. We have seen that more advanced temporal approaches based on RNNs have also been proposed. Although their performance in the video-based benchmark appears promising, these approaches still suffer from some limitations, e.g., may only handle global head movements (see Section 2.2).

We can go further by considering local motion through the learning of spatio-temporal features. Dynamic features such as motion patterns, guided by appearance or not, could be used to add a constraint on the landmark trajectories in addition to the spatial constraint. As illustrated in Figure 3.12, the temporal evolution of the texture in each landmark during an expression variation shows specific patterns that could be exploited, especially for the mouth, eyes, and outline of the face. Capturing such information could help to avoid errors such as mismatches or drifts that can be observed with image-based and tracking approaches. This could provide more robustness and accuracy in uncontrolled conditions, which strongly encourages us to investigate this direction.

Figure 3.12: Temporal evolution of the texture in each landmark during an expression variation. The sequence is from 300VW, category 3. Specific patterns can be observed and could be exploited, especially for the mouth, eyes, and outline of the face.

Chapter 4

Facial Landmark Detection with Improved Spatio-Temporal Modeling

Contents

Despite considerable progress in recent years, the performance of facial landmark detection under uncontrolled conditions is still not fully satisfactory (see Section 3.3) and even today, this problem continues to be studied largely from still images. Yet, with the ubiquity of video sensors, the vast majority of applications rely on videos. Current approaches, when applied to videos, usually track landmarks by detecting them and are therefore not able to leverage the temporal dimension (see Section 2.2). Recent work has proved that taking into account video consistency help to deal with the variability in facial appearance and ambient environment encountered under uncontrolled conditions. It generally involves a CNN coupled to a RNN, which provides only limited temporal connectivity on feature maps with a high level of abstraction. Such architectures can model global motion (e.g., head motion) but not as easily local motion like the movements of the eyes or the lips, which are important to detect facial landmarks accurately.

Video analysis has been studied for a long time to tackle a variety of problems including human behaviour understanding. The interest for video has been growing today due to the explosion in the number of videos shared on the Internet, media, surveillance, medicine, and the need to understand them. The diversity and complexity of videos make their analysis very challenging. In addition to the issues related to video quality and changes in the visual context (e.g., illumination, occlusions), which are also found in still images, it requires to handle motion-related issues, e.g., displacements of the camera and the targeted object, motion blur, duration. All these works can provide useful insights to improve video-based facial landmark detection.

An overview of the objectives of this work is provided in the next section (Section 4.1). In Section 4.2, current possibilities for spatio-temporal modeling are described in a broader context. Both hand-crafted features and DL approaches are reviewed. A solution proposal is then introduced in Section 4.3. Our 2D baseline architectures are covered as well as their extension to the temporal domain. The experimental protocol, implementation details, and results with their analysis are presented in Section 4.4. Experiments on two datasets, 300VW [193] and SNaP-2DFe [2], are conducted in order to evaluate the results obtained and to compare them with state-of-the-art approaches. A more in-depth study of temporal integration is also performed in Section 4.5 through design investigations. Finally, the chapter ends with Section 4.6 by discussing the benefits and limits of the proposed approach.

4.1 Overview

Facial landmark detection remains difficult due to the many variations encountered under uncontrolled conditions. The purpose of this work is to better exploit the dynamic nature of the face in order to obtain more stable predictions over time and more robustness to variations that considerably impact facial appearance. We have seen that recent video-based approaches have begun to show their benefits. However, popular approaches do not sufficiently take advantage of the temporal information. They suffer from limited temporal connectivity, allowing only global motion to be modeled. We believe that with a better temporal modeling more robust predictions to the various challenges encountered in uncontrolled conditions can be achieved.

Even today the gap between handcrafted and DL approaches for video analysis is not as wide as it was with still images when AlexNet [112] was proposed. The design of a compact and efficient descriptor for video remains challenging. For this reason, it is still worthwhile to consider handcrafted approaches, to have a better understanding of the development of video analysis. Through this work, a particular interest is given to the following aspects:

- integration of local motion;
- complementarity between both local and global motion;
- complementarity between both spatial and temporal information.
- application to both predominant models of the literature, coordinate regression networks and heatmap regression networks.

4.2 Spatio-Temporal Modeling Review

Spatio-temporal modeling is generally built upon image-based solutions that are extended to video. Many handcrafted spatio-temporal feature detectors and descriptors have been proposed [117, 52, 190, 232, 263, 108]. A detector identifies sparse or dense spatio-temporal locations and scales. Spatial or spatio-temporal gradient or motion based features capture shape or motion information around the selected locations. Following the breakthrough of DL, the interest has shifted to the design of deep architectures for video [83]. Features from individual frames can be extracted using CNNs trained on images followed by their temporal integration using *pool* layers or RNNs.

More advanced approaches involve the use of 3D convolutions, which allow spatio-temporal feature learning, and two-stream architectures, which fuse the features from RGB and optical flow images. In this section, both handcrafted and DL approaches are reviewed.

4.2.1 Hand-Crafted Approaches

To obtain a compact representation of video without any complex processing such as object tracking or motion segmentation, a commonly used solution is to extend the notion of spatial interest points into the spatio-temporal domain. Different kinds of detectors with variable sparsities have been proposed. Space-time interest points [117] is an extension of the Harris detector [84] in which the detected points are characterized by large variation of the image values in all three directions altogether, i.e., spatio-temporal corners (see Figure 4.1). Corners being sparse, cuboids [52, 18] let more features to be detected using Gabor filters. In these methods, global information is ignored. Yet, it can be integrated by extracting structural information, allowing more relevant interest points to be detected [233]. Features that are scale-invariant, both spatially and temporally, can also be obtained by combining a point detector and scale selection using the Hessian matrix [232]. However, all these detectors remain sparse and only detect salient motion. Dense sampling of local features has proved to be more effective [227]. It consists of the extraction of video blocks at regular positions and scales in space and time, usually with overlap.

Figure 4.1: Space-time interest points detection [117]. The detected points are characterized by large variation of the image values in all three directions altogether, i.e., spatio-temporal corners.

A different approach is to work with trajectories instead of cuboids. Trajectory and cuboid are shown in Figure 4.2. The spatial and temporal domains having different properties, they should be handled differently. Points are no longer detected at certain positions in space and time, but are tracked over time using an optical flow algorithm. The motion information of trajectories can be leveraged in the process. To extract trajectories, the Kanade–Lucas–Tomasi (KLT) tracker [137] can be used along with the extended Harris detector [144] or simply by relying on ”good features to track” [147].

They can also be computed by matching SIFT descriptors between two consecutive frames [203]. The KLT tracker and SIFT descriptors can be combined as well [202]. The use of dense trajectories has also been proposed [224], as dense sampling was found to give superior performance. This involves densely sampled points tracked using a dense optical flow field, ensuring good coverage and tracking. When the focus is on a specific subject, e.g., in action recognition, dense trajectories can benefit from explicit camera motion estimation [220, 234, 158, 95]. If camera motion is known, unwanted trajectories in the background can be removed to only keep the trajectories of the object of interest. The optical flow is then corrected, which improves the performance of motion descriptors. To do so, Speeded Up Robust Features (SURF) [10] descriptors and dense optical flow can be used to match feature points between frames. Theses matches serve for homography estimation with RANdom SAmple Consensus (RANSAC) [69] and an object detector to remove inconsistent matches [226].

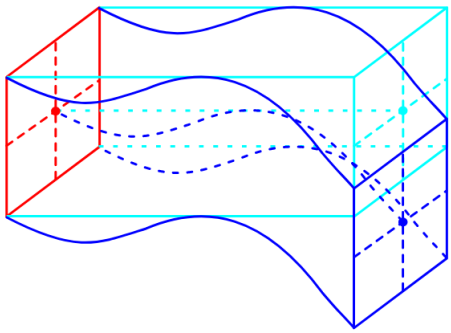

Figure 4.2: Difference between a cuboid and a trajectory [225]. Cuboid corresponds to a straight 3D bloc while trajectory follows the optical flow.

For each given sample point, a feature descriptor is computed to describe its neighbourhood (see Figure 4.3). HOG and HOF [118] are among the descriptors that have been found to perform well [118, 227]. The success of image-based descriptors led to their generalization to videos. Spatio-temporal extensions such as 3D SIFT [190], extended SURF [232], local trinary patterns [263] and HOG 3D [108] have been introduced. For the latter, integral images have been extended to integral videos for the efficient computation of 3D gradients. New motion descriptors have also been proposed. Motion boundary histograms [45] are based on spatial derivatives of optical flow calculated separately in x and y, which are then quantized into a histogram. The use of derivatives reduces the influence of the camera movements by removing constant motion. Features are generally combined to capture shape, appearence, and motion information and improve performance [45]. Once the features have been extracted, they are encoded into Bag of Features (BoFs) [167] or Fisher Vectors (FVs) [169], generally pooled using a spatio-temporal pyramid, and then fed to an SVMs classifier.

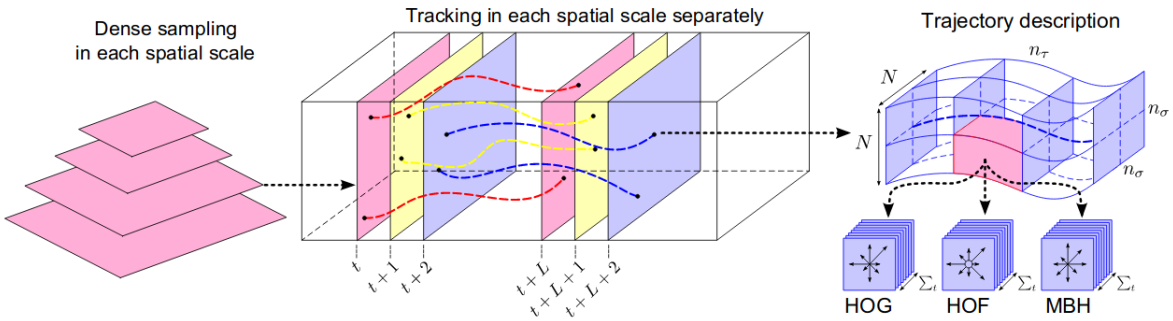

Figure 4.3: Extraction and characterization of dense trajectories [225]. For each given sampled and tracked point, a feature descriptor is computed to describe the trajectory of the point.

4.2.2 Deep Learning Approaches

Following the breakthrough of DNNs on still images [112], state-of-the-art handcrafted approaches have been adapted to DL or even replaced and new video representations have been proposed. Trajectory-pooled deep convolutional descriptors [228] use CNNs to learn discriminative feature maps from which local trajectory-aligned descriptors are computed. Deep Fisher networks [168] are composed of multiple layers of FVs and allow a refined hierarchical representation of information. Attention quickly turned to the design of CNN architectures to best capture temporal patterns with a good balance between computational cost and accuracy. Many attempts to DL for spatio-temporal learning have been made, relying on different modalities: RGB, optical flow [94], and different kinds of aggregation (pooling, RNN [268]). As early work has shown [104], expecting the model to learn spatio-temporal motion-dependent features can be a difficult task.

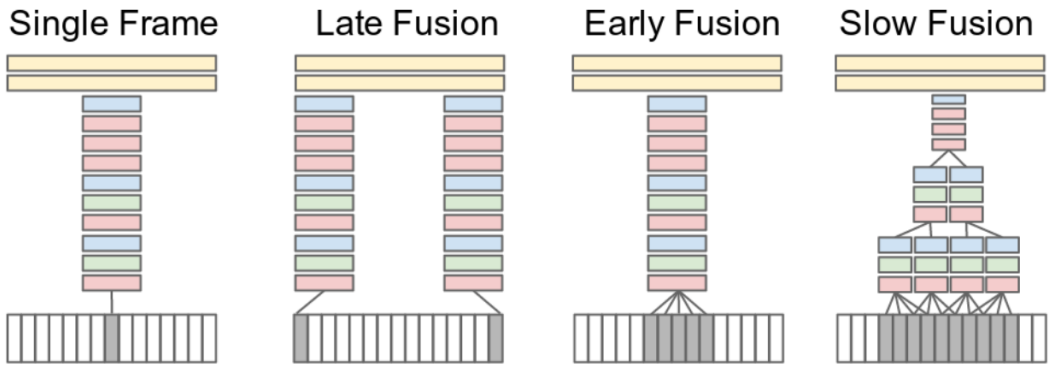

Figure 4.4: Strategies to extend CNNs to the temporal domain (red = *conv* layer; green = normalization layer; blue = *pool* layer) [104]. Slow fusion was found to be the best strategy.

Several ways to incorporate and fuse temporal information from multiple frames have been investigated, either at the first layers or later in the architecture [104]. The different strategies are illustrated in Figure 4.4. Early fusion operates at the pixel level by extending the first convolution to the temporal domain and taking a stack of images as input. Late fusion is based on two streams that share their parameters and take images spaced apart in time as input. Neither of the two streams is able to detect any motion. The first *fc* layer computes global motion features by comparing the outputs of each stream. Slow fusion is a combination of the two previous strategies. Temporal information is merged using 3D convolutions so that the last layers of the network gradually access more global information in both the spatial and temporal domains. The latter was found to be the best strategy but does not seem to capture the motion well as it performed similarly to the network operating on individual frames.

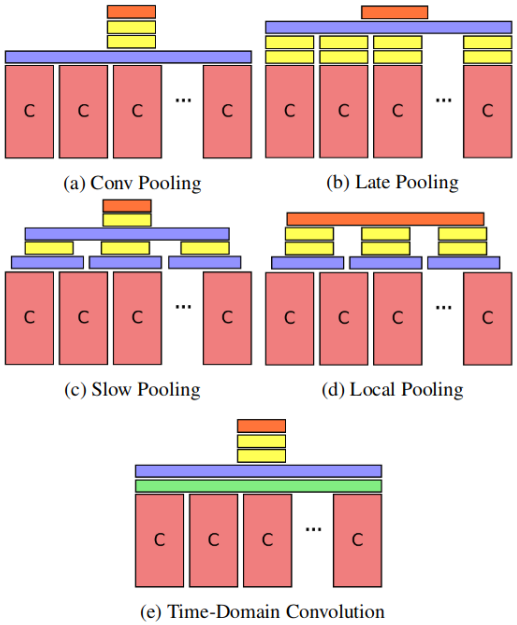

Figure 4.5: Feature pooling architectures (blue = *pool* layer; green = time-domain *conv* layer; yellow = *fc* layer; orange = softmax layer) [268]. Temporal pooling of *conv* layers was found to be the best strategy.

This kind of approach focusing only on small periods of time, temporal aggregation has also been studied [268] through the use of a shared CNNs along with pooling or RNN in order to process long sequences. Many pooling strategies have been considered: over *conv* layers, late (after the *fc* layer), slow (hierarchically combining information), local (combining only frame level information), and using 3D convolutions. They are all shown in Figure 4.5. Temporal pooling of *conv* layers performs better than the other four. However, RNNs provide slightly better results than pooling and may have better

modeling capacities as they consider the temporal structure. It is also shown that optical flow is not always helpful, especially when videos are captured under uncontrolled conditions.

Over the last years, 3 approaches have emerged and continue to widen the gap with image-based and handcrafted video-based solutions:

- explicit models of motion using two-stream architectures (spatial and temporal streams);
- embedding the temporal dimension into CNNs using 3D convolutions;
- video understanding as sequence modeling by combining CNNs and RNNs.

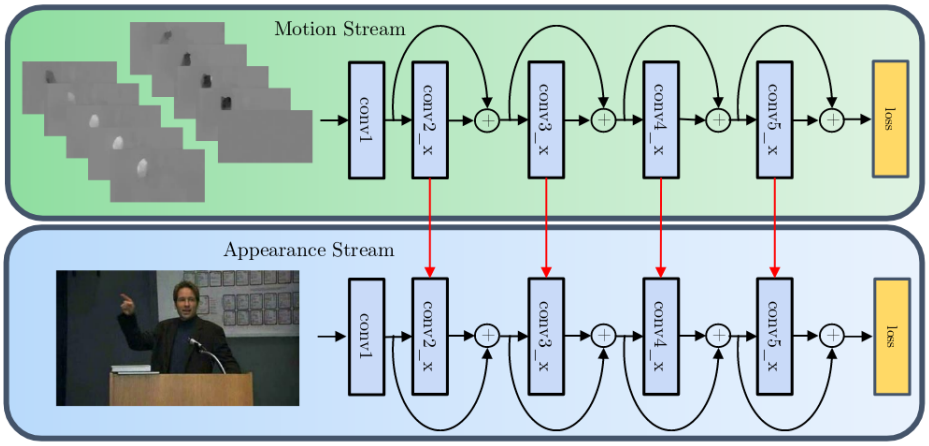

Figure 4.6: Two-stream network. Residual connections between the spatial stream and the temporal stream enforce the relationship between what is captured by the two streams [62].

Explicit models of motion using two-stream architectures draw their inspiration from neurosciences, in which the hypothesis that humans possess separate visual pathways is made [140]. The ventral pathway responds to spatial features while the dorsal pathway is sensitive to object transformations and their spatial relationships. HMAX [113] is one of the first two-stream biologically-inspired model that has been proposed and it has been naturally generalized to DL with CNNs [196, 63, 231, 229]. In such an architecture, a video is decomposed into spatial and temporal components (see Figure 4.6). The spatial stream extracts spatial features from RGB images while the temporal stream extracts motion features from the optical flow, which is stacked on multiple frames. This way appearance and motion across frames are captured. Two types of motion representations can be used, by sampling the optical flow either at the same

locations across several frames or along the motion trajectories, as with handcrafted descriptors [196]. To enable an interaction between the two streams and thus enforce the relationship between what is captured by the spatial stream and the motion captured by the temporal stream, residual connections between the two streams can be established [62].

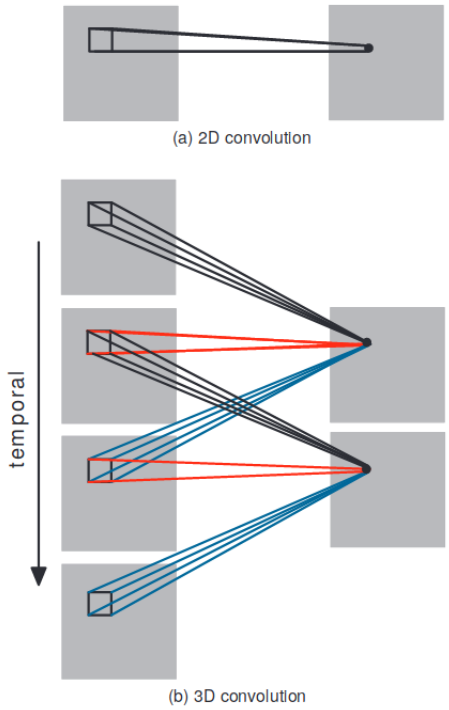

Figure 4.7: Comparison of 2D (a) and 3D (b) convolutions with a temporal depth of 3 [97]. A new dimension to the filters in order to obtain a hierarchical representation of spatio-temporal data.

3D convolution is probably the most natural way to extend CNNs to the temporal domain, allowing both appearance and motion to be modeled simultaneously. It was first used in human behavior-related problems [97, 7], while also studied for unsupervised learning of spatio-temporal features [208, 120]. As illustrated in Figure 4.7, a new dimension corresponding to the temporal depth is added to the filters in order to obtain a hierarchical representation of spatio-temporal data. A 3D convolution can be expressed as:

$$v_{ij}^{xyz} = b_{ij} + \sum_{m} \sum_{p=0}^{Pi-1} \sum_{q=0}^{Qi-1} \sum_{r=0}^{Ri-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)} \tag{4.1}$$

where v_{ij}^{xyz} is the value at position (x, y, z) on the j-th feature map in the i-th layer, b is the bias, m is the index of the feature map in the previous layer, P, Q, R are respectively the height, width and depth of the kernel, and w is the value of the kernel

(weight). 3D CNNs have rapidly achieved superior performance over 2D ones. They have become competitive with other approaches [211], and have demonstrated a good generalization capacity [149, 156]. A few studies have been conducted to better understand how to design and parameterize such an architecture and what kind of features they learn [211]. Shallow and very deep architectures have been investigated on currently available datasets with the idea that 3D CNNs have the potential to follow the same achievement as their 2D counterpart on still images [83]. However, video datasets equivalent to ImageNet [47] do not seem to be available yet, while 3D CNNs imply more parameters resulting in more difficulty to train from scratch, i.e., time consuming, overfitting. To address these issues, several ways to initialize 3D weights using pre-trained 2D ones have been proposed: supervision transfer [51], averaging by temporal size, initializing the first timestep with the 2D weights and others to zero, as well as variants of these [141]. Besides reducing overfitting and speeding up training, they also lead to improvements in accuracy.

Another growing solution to facilitate the training of 3D CNNs is to factorize 3D convolutional filters into separate spatial and temporal components [204, 213]. While being easier to optimize, it also allows for more nonlinearities, which yields a significant gain in accuracy. Various strategies to decompose 3D convolutions have been formulated, the most successful being to use a 2D spatial convolution followed directly by a 1D temporal convolution [174]. Moreover, motion may not be extracted across all the network. A trade-off between speed and accuracy can be found by replacing many of the 3D convolutions by 2D ones [213], depending on the application. Evolutionary algorithms have also been used to explore models with different types and combinations of *conv* layers to discover new and potentially optimal video architectures [170].

As already mentioned in Section 2.1.3, RNNs [182], historically FC-RNNs, use their past inputs to compute their output, by maintaining an internal state, which accumulates contextual information. Beyond this state, RNNs have the ability to capture temporal orderings and long-term dependencies. For the purpose of visual understanding, a CNN is usually used as a backbone to extract discriminative features, which are then fed into the RNN [261, 283]. Following the philosophy of DL, a stack of RNNs is often used to obtain a more efficient representation with a high level of abstraction and potentially operate at different time scales [54, 268]. RNNs have also been extended to convolutional architectures instead of *fc* ones, for both the input-to-state and state-to-state connections, which are therefore 3D tensors [8, 247]. As a result, ConvRNNs turn out to be better than FC-RNNs at handling spatio-temporal correlations. A visual comparison of both architectures is shown in Figure 4.8.

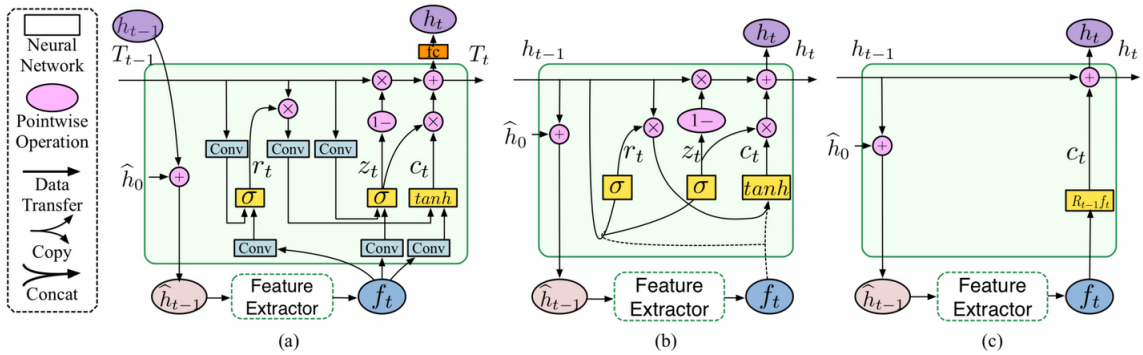

Figure 4.8: Comparison of different RNN architectures: (a) ConvGRU, (b) GRU, and (c) traditional RNN [230]. Compared to traditional RNN, GRU adopts a memory cell coupled with gating functions to control the flow of information. Using ConvGRU, the input and state are 3D tensors and convolutions are used for both input-to-state and state-to-state connections instead of matrix multiplications.

RNNs have also been used for unsupervised learning [200]. To better understand how a RNN works, studies have been conducted on visualization and the importance of the gates, structure, and depth [76, 102, 103, 159]. However, it only concerns language and audio modeling and only few studies focus on visual sequences. Yet, they have specific properties such as redundancy and variable temporal dependencies depending on the applications [189]. Recently, a solution has been proposed to leverage a pre-trained CNN by transforming any of its layers, either *conv* or *fc*, into a recurrent one [262]. An analysis of the distribution of gate activations has shown that pre-trained RNNs make better use of the temporal context. Also, one RNN is enough since when two RNNs are stacked, the second one behaves like an *fc* layer and does not take advantage of the previous hidden state.

4.2.3 Summary

Although the immediate purpose of the reviewed work is different from ours, they offer interesting insights for video-based facial landmark detection. There is also a need to encode spatio-temporal information. Handcrafted approaches rely on the description of cuboid or trajectories obtained through sparse or dense points sampling. Different kinds of descriptors, gradient-based or motion-based descriptors, are used and generally combined. They are then encoded in BoFs or FVs and classified using SVMs. Such solution can also benefit from explicit camera motion estimation. Although dense trajectories improved by camera estimation provide interesting performance [226], it is at a high computational cost, which makes them intractable on large-scale datasets.

Video analysis based on DL approaches has undergone considerable development in recent years. While early work has shown that learning spatio-temporal motion dependent features can be a difficult task for such approaches [104], recent work continues to suggest otherwise. They are mainly based on three architectures: two-stream networks, 3D CNNs, and RNNs combined with a CNN as a backbone. All these solutions are not mutually exclusive and potentially complementary. Two-stream networks and 3D convolutions may have facilities in capturing low-level motion, as opposed to RNNs, which may be better with higher-level variations. Additionally, explicit motion integration using an optical flow combined with raw images may still help to improve the temporal representation. There have already been attempts to combine two-stream networks and 3D convolutions [28], 3D convolutions and RNNs [149, 7, 276], and two-stream networks and RNNs [209, 284] with promising results.

The work discussed in the following sections focuses on CNNs for facial landmark detection, as they provide an effective and versatile baseline tool to solve this problem. In Section 2.2, we have seen that current video-based facial landmark detection approaches [133, 162, 79, 89] rely mainly on late temporal connectivity based on RNNs, which encode motion at a large scale. In contrast to this work, the proposed solution aims to improve the temporal connectivity of deep CNNs for video-based facial landmark detection, to include local motion to the model. To the best of our knowledge, this is the first work to focus on local motion and on the learning of low-level spatio-temporal features for this problem. To do so, 3D convolutions are used to perform spatio-temporal convolutions. This provides early temporal connectivity directly at the pixel level, allowing both appearance and motion to be modeled within all layers of the same network. 3D convolution has already been used for other tasks such as action recognition, and has shown its ability to learn relevant spatio-temporal features [213]. This makes it a natural candidate to model local motion in facial landmark detection, which can be easily applied to existing work.

4.3 Architecture Design

This section describes the architectures developed in this work to extend the connectivity of CNN-based landmark detectors to include local motion through early connectivity. We consider lightweight architectures to isolate the contribution of local motions from other factors.

4.3.1 Coordinate Regression Networks

2D Baseline. The proposed 2D baseline coordinate regression model (coord-2D), in Figure 4.9-(a), follows the philosophy of VGG network [197] and is inspired by recent work on facial landmark detection [66]. It contains only convolutions with a kernel size of 3×3, a stride of 1 and a padding of 2. The number of filters increases by a factor of 2 at each of the 5 convolutions, from 32 to 512, to accommodate the increase in abstraction. Each convolution is followed by a batch normalization layer, an ELU activation layer and a max-pooling layer with a pool size and stride of 2. ELU activation [35] was chosen as activation function as it provides better performance than the commonly used ReLu [150]. Batch normalization layers further improve the training procedures. This model produces feature maps of size 2x2x512 from an input image of size 64x64x3. These maps are vectorized and passed on to two *fc* layers to aggregate the most relevant information from the *conv* layers and thus obtain the landmark predictions. The first *fc* layer has a capacity of 1,024 units and is followed by a batch normalization layer and an ELU activation layer. We add a dropout layer [199] with a rate of 0.2 for regularization. The second and last *fc* layer outputs 136 values corresponding to the 68 landmark coordinates. As already stated, we limited the complexity of the model to facilitate experiments and to clearly highlight the contribution of temporal information, especially, no cascaded regression is used.

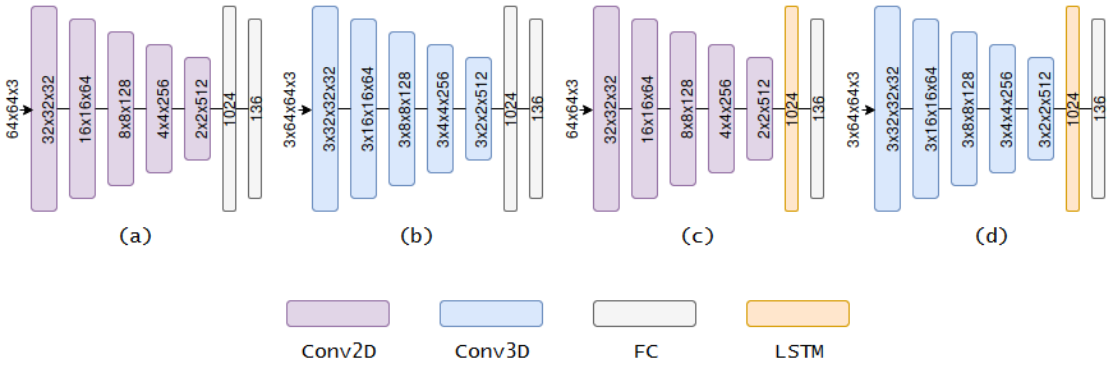

Figure 4.9: The proposed coordinate regression networks. (a; coord-2D) Baseline 2D architecture, (b; coord-3D) early temporal connectivity based on 3D convolutions, (c; coord-2RNN) late temporal connectivity based on RNNs, and (d; coord-3DRNN) both connectivity levels. These four lightweight architectures are used to evaluate the contribution of local motion to coordinate regression.

Convolutional Neural Network with Temporal Connectivity. We experiment with different types of connectivities:

- early connectivity (coord-3D), which is direct at the pixel level and based on 3D convolutions;

- late connectivity(coord-2DRNN), which occurs after a series of convolution and is based on RNNs;
- combination of early and late connectivity (coord-3DRNN).

Each architecture with temporal connectivity takes as an input a temporal window of 3 frames as this value seems suitable for facial landmark detection (see Appendix A.1). The use of a greater depth was investigated without any benefit. Long-term temporal dependencies may not be necessary for facial landmark detection.

As shown in Figure 4.9-(b), local motion information is added through early temporal connectivity by extending the *conv* and *pool* layers by one dimension [97]. This allows the network to detect local motion patterns and learn a hierarchy of motion features. We use the same parameters as for spatial dimensions, as suggested in [211], except for pooling. Since the temporal depth of the input is only 3, we do not apply any pooling along this dimension to preserve the temporal information.

Finally, we integrate late connectivity by replacing the first *fc* layer by a long short-term memory (LSTM) [247, 79] recurrent layer with identical capacity (see Figure 4.9-(c) and 4.9-(d)). The 2D CNN model is distributed in time with shared parameters. By processing the output of the streams, the recurrent layer is able to compute global motion features. Dropout is also performed on recurrent connections, with the same rate as for feed forward connections. We keep only the last timestep of the LSTM, which should encode relevant spatio-temporal context to predict landmarks on the last frame of the sequence.

4.3.2 Heatmap Regression Networks

2D Baseline. The proposed 2D heatmap regression model is designed by adapting the 2D base architecture from Section 4.3.1 to a fully-convolutional auto-encoder architecture. This model has the potential to better handle spatial correlations since there is no *fc* layer and therefore no vectorization of the features maps. Inspired by recent advances in the literature [152, 23], we adopt an hourglass-like architecture. At the encoder level, we remove the spatial pooling and replace it by convolutions with strides to obtain a fully-convolutional structure. We use a constant number of 256 filters and a stride of 2 along spatial dimensions. Note that we are not necessarily looking for optimal values but rather values that facilitate our experiments. The decoder has an equivalent number of transposed convolutions with identical parameters. Two *conv* layers with a kernel size of 1×1, a stride of 1 and a padding of 2 are applied after the decoder in order to

generate heatmap predictions. The first one has 512 filters and is followed by a batch normalization layer and an ELU activation layer. The second one outputs 68 heatmaps, one for each facial landmark. Regularization techniques such as dropout works well on coordinate regression models especially on *fc* layers. However, they are not as effective on fully convolutional models [128]. Therefore, no dropout is applied to the *conv* layers. Also, we did not integrate residual connections between encoder and decoder, since in some experiments these connections caused a drop in performance (see Appendix A.2).

Fully-Convolutional Network (FCN) with Temporal Connectivity. As before, the transition from 2D to 3D is done by extending the convolutions to the temporal domain with the same parameters as for spatial dimensions and a temporal depth of 3. Late connectivity is integrated through the use of a convolutional LSTM layer between the encoder and the decoder, with the same parameters as the other convolutions, except the stride which is set to 1 to preserve the spatial information. ConvLSTM turns out to be better than FC-LSTM at handling spatio-temporal correlations [247] and allows to preserve a fully-convolutional architecture. Unlike FC-LSTM, the input and state are 3D tensors and convolutions are used for both input-to-state and state-to-state connections instead of matrix multiplications. Decoding is done in 2D by keeping only the feature maps of the last timestep of the convolutional LSTM layer. We still used dropout on recurrent connections, with a rate of 0.2. Models with late temporal connectivity (heat-2DFCRNN) and both connectivity levels (heat-3DFCRNN) are illustrated in Figure 4.10-(a) and Figure 4.10-(b).

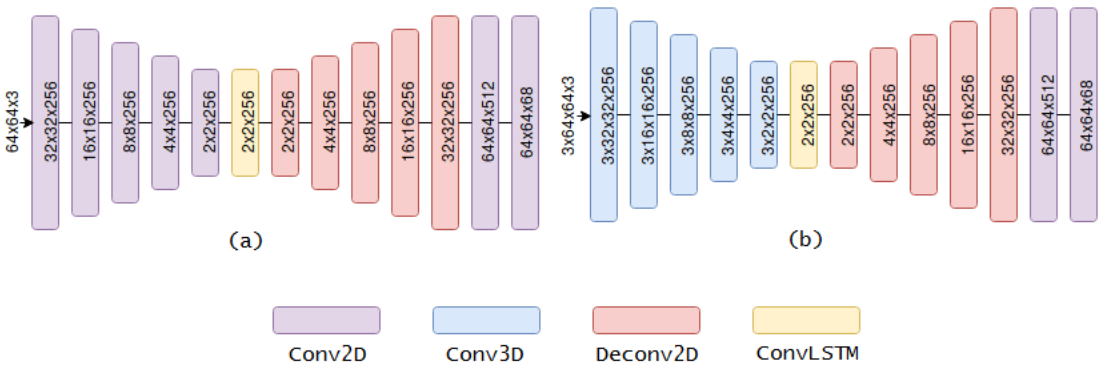

Figure 4.10: The proposed heatmap regression networks. (a; heat-2DFCRNN) Late temporal connectivity based on RNNs, and (b; heat-3DFCRNN) both connectivity levels. These two lightweight architectures are used to evaluate the contribution of local motion for heatmap regression.

4.4 Experiments

We conduct experiments on two datasets, 300VW [193] and SNaP-2DFe [2]. 300VW is the reference dataset for video-based facial landmark detection. SNaP-2DFe allows the study of the impact of head movements and facial expressions on facial landmark detection and to identify the respective benefits of early and late connectivities. Qualitative results are provided, especially to illustrate what the proposed models learn and their limits through failure cases. We also analyze the performance of each model in terms of speed, size and number of parameters.

4.4.1 Datasets and Evaluation Protocols

300VW. As already described in Section 3.2.1, 300VW [193] contains 114 videos. 50 videos are intended for training and 64 for testing. The test set is divided into 3 categories of increasing difficulty. We keep the same split as the 300VW challenge [193] for training and testing, without adding any external data. 20% of randomly selected training data was used for validation.

SNaP-2DFe. SNaP-2DFe [2] consists of 6 movements composed of a horizontal translation and/or a rotation (roll, pitch, yaw), each associated with 7 acted expressions (neutral, joy, fear, anger, disgust, sadness, surprise). Data from 12 participants has been collected at the time of the experiments. We use data from subjects 1 to 6 for fine-tuning, 7 to 8 for validation and 9 to 12 for testing.

Pre-processing. For 300VW, all models are trained from scratch without any pre-training. For SNaP-2DFe, we fine-tune the models trained on 300VW due to the limited number of images present in SNaP-2DFe. Training data is augmented by reversing each image sequence and flipping each image horizontally. Motion augmentation using re-sampling has been evaluated but found to be harmful (see Appendix A.3). The same data is used for each training session. The input of each 2D network is an RGB image cropped from the face bounding box, resized to a size of 64x64x3 and normalized as explained in Section 3.2.2. For their 3D counterpart, we opt for an input window of size 3x64x64x3. Each video of the training set is then split into several non-overlapping image sequences of size 3. This ensures a reasonable amount of training data. Ground truth coordinates are normalized between [-0.5; 0.5]. Ground truth heatmaps are generated by computing a bivariate Gaussian of bandwidth 3×3 centered at the location of the landmarks.

Evaluation metrics. In the experiments, we use the metrics introduced in [32], i.e., RMSE normalized by the diagonal of the ground truth bounding box. We compute the AUC and failure rate at thresholds of 8% and 4%. We choose this metric for its robustness to pose variations. The 8% threshold is mainly used in the literature. As already mentioned in Section 3.2.3, beyond this threshold, landmarks are mostly not located correctly. AUC can be high using a 8% threshold, especially on SNaP-2DFe, because the accepted error is large. We also provide some results with a 4% threshold. This threshold corresponds to predictions that are mostly acceptable. For a fair comparison with other existing approaches, we also use the average RMSE normalized by the interocular distance as in the 300VW challenge.

Note that we encountered difficulties (overfitting) in fine-tuning heatmap regression models on SNaP-2DFe, regularization using dropout not being helpful. For the latter, we only provide results on 300VW which is anyway the reference dataset. The comprehensive analysis using SNaP-2DFe is therefore only performed on coordinate regression models.

4.4.2 Implementation Details

Network weights are initialized with the Xavier uniform initializer. The Adam optimizer with L2 loss is used for all models, following the original paper parameters except for the learning rate that we set to 10^{-4}. During training, the learning rate is reduced by a factor of 0.1 when the error on the validation set has not improved for 10 epochs. We performed several training sessions of 100 epochs for each model and report the best run. Batch sizes of 8 and 4 are used for the 2D and 3D models, respectively. These parameters have been found to be optimal based on empirical tuning. We observed the same convergence behavior for both types of models, i.e., coordinate regression and heatmap regression.

4.4.3 Evaluation on SNaP-2DFe

Table 4.1 summarizes the performance of coordinate regression models over the entire SNaP-2DFe test set in terms of AUC and FR with thresholds at 8% and 4%. We observe a performance gain between coord-2D and coord-3D approaches; it is larger when the threshold is 4%, showing a better robustness in a more challenging evaluation context. This underlines the effectiveness of 3D convolutions for landmark localization. Early temporal connectivity provides a significant gain in accuracy while reducing the failure rate. The results of models with RNN suggest that early and late temporal connectivities may be complementary.

Table 4.1: Comparison of the different architectures presented in Section 4.3.1 on SNaP-2DFe (subjects 9 to 12). AUCs and failure rates at thresholds of 8% and 4% are reported. Early temporal connectivity provides a significant gain in accuracy while reducing the failure rate.

Method	AUC @8%	FR @8%	AUC @4%	FR @4%
coord-2D	78.53	0.27	57.44	0.79
coord-3D	**79.57**	**0.15**	**59.44**	**0.64**
coord-2DRNN	83.32	**0.08**	66.83	0.35
coord-3DRNN	**84.12**	0.09	**68.41**	**0.28**

In order to better understand the strengths and weaknesses of these models, we compute the error for each type of head movement (i.e., Static, Translation-x, Roll, Yaw, Pitch, Diagonal) and facial expression (i.e., Neutral, Happiness, Fear, Anger, Disgust, Sadness, Surprise). Table 4.2 presents the results regarding head movements. A notable performance gain is observed for all movements by applying early temporal connectivity only. However, when no movement occurs, there is a drop in performance, which may indicate a weakness of 3D convolution in static conditions. It could also be due to the *fc* layer since this weakness is not found when a recurrent layer is used instead. With the addition of late temporal connectivity, there is also an improvement in performance, but it is more modest. This can be explained by the fact that the RNN is already capable of capturing global motion, as shown by the significant improvement from coord-2D to coord-2DRNN.

Table 4.3 presents the results with respect to facial expressions. The performance gain for this kind of motion is more significant than with head movements, even with the addition of late temporal connectivity. Facial expressions consist of local motion patterns that RNNs cannot capture, unlike 3D convolution. In the absence of any expression (i.e., neutral sequences), with early connectivity alone, we obtain lower performances potentially due to the lack of motion, as observed earlier. These results clearly illustrate the value of local motion for facial landmark detection. They also demonstrate the complementarity of local motion modeling and global motion modeling with RNNs.

Table 4.2: Comparison of the different architectures presented in Section 4.3.1 on SNaP-2DFe (subjects 9 to 12; motion only). AUCs at thresholds of 8% and 4% are reported. Early temporal connectivity provides performance gain for all movements. Note, however, the drop in performance when no movement occurs. These results also suggest a complementarity between early and late temporal connectivity.

Method	coord-2D		coord-3D		coord-2DRNN		coord-3DRNN	
	@8%	@4%	@8%	@4%	@8%	@4%	@8%	@4%
Nothing	**82.11**	**64.22**	81.19	62.39	86.20	72.41	**86.84**	**73.68**
Diagonal	82.29	64.58	**82.77**	**65.54**	85.28	70.55	**85.40**	**70.81**
Pitch	82.69	65.37	**83.56**	**67.12**	84.76	69.51	**85.26**	**70.52**
Roll	79.72	59.43	**80.90**	**61.80**	83.38	66.75	**84.01**	**68.02**
Translation-x	80.43	60.87	**82.14**	**64.27**	85.28	70.55	**85.33**	**70.66**
Yaw	82.64	65.29	**84.23**	**68.47**	**85.42**	**70.83**	85.23	70.47

Table 4.3: Comparison of the different architectures presented in Section 4.3.1 on SNaP-2DFe (subjects 9 to 12; emotion only). AUCs at thresholds of 8% and 4% are reported. Early temporal connectivity provides performance gain for all facial expressions. Note, however, the drop in performance when no movement occurs. These results also suggest a complementarity between early and late temporal connectivity.

Method	coord-2D		coord-3D		coord-2DRNN		coord-3DRNN	
	@8%	@4%	@8%	@4%	@8%	@4%	@8%	@4%
Neutral	**82.11**	**64.22**	81.19	62.39	86.20	72.41	**86.84**	**73.68**
Anger	77.97	55.94	**79.27**	**58.53**	83.54	67.09	**84.16**	**68.33**
Disgust	76.35	52.70	**78.76**	**57.52**	83.57	67.15	**84.49**	**68.98**
Fear	77.51	55.01	**78.48**	**56.95**	83.92	67.85	**84.32**	**68.64**
Happiness	80.69	61.37	**83.80**	**67.60**	85.60	71.20	**86.42**	**72.83**
Sadness	76.75	53.49	**79.15**	**58.31**	82.33	64.67	**84.63**	**69.26**
Surprise	77.29	54.59	**80.93**	**61.85**	84.33	68.67	**85.36**	**70.72**

4.4.4 Evaluation on 300VW

Table 4.4: Comparison of the different architectures presented in Section 4.3.1 on the three categories of the 300VW test set. AUC and FR at a threshold of 8% are reported. By applying early temporal connectivity, more accurate predictions are obtained in categories 1 and 2, while in category 3 a lower failure rate is observed.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
coord-2D	71.82	**1.73**	67.17	**0.61**	**65.48**	4.54
coord-3D	**72.26**	1.88	**67.62**	2.21	64.23	**4.22**
coord-2DRNN	76.04	**1.52**	**73.45**	**0.14**	**71.25**	2.83
coord-3DRNN	**76.21**	1.60	73.14	0.17	70.86	**2.70**

Coordinate Regression Networks. Table 4.4 shows the performance of coordinate regression models in terms of AUC and FR with an 8% threshold on the 3 categories of 300VW. Compared to the results obtained on SNaP-2DFe, the contribution of early temporal connectivity is not as clear. However, a gain is systematically observed for models without late temporal connectivity, either in terms of AUC or FR. In categories 1 and 2, we obtain more accurate predictions, while in category 3 we observe a lower failure rate. According to these results and considering the observations made on SNaP-2DFe, it seems that the 3D convolution is affected by the lack of motion and by occlusions, which are common in 300VW.

Table 4.5: Comparison of the different architectures presented in Section 4.3.2 on the three categories of the 300VW test set. AUC and FR at a threshold of 8% are reported. Early temporal connectivity applied to heatmap regression networks brings significant and consistent performance gains over the 3 categories of 300VW.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
2DFCRNN	73.99	3.79	73.73	**0.30**	70.55	4.91
3DFCRNN	**75.25**	**2.50**	**74.23**	**0.30**	**71.60**	**4.40**

Heatmap Regression Networks. The performance of heatmap regression networks in terms of AUC and FR with an 8% threshold is reported in Table 4.5. Early temporal connectivity brings here significant and consistent performance gains over the 3 categories of 300VW. Despite the decoding from 2D feature maps, local motion information is preserved well in the reconstruction, which may explain the variance of the results with

coordinate regression models. As previously reported, the latter might be mostly due to their *fc* layers. Due to their fully-convolutional structure, heatmap regression networks seem more suitable to handle spatio-temporal correlations.

4.4.5 Comparison with Existing Models

Table 4.6 shows the average error of the best models with major models from the literature on the full 300VW test set. Numbers are reported from [133]. The comparison includes static methods with handcrafted [248, 285] and learned features [280], and approaches that leverage temporal information [133]. The results show the effectiveness of the proposed models, especially on Category 3, which is the most challenging one. Despite their simplicity, coordinate regression (coord-3DRNN) and heatmap regression model (heat-3DFCRNN) are among the top-performing methods. Thanks to their ability to capture both local and global motion, they outperform static methods and show competitive performance with other, more complex, video-based approaches.

Table 4.6: Comparison of the best proposed models, coord-3DRNN and heat-3DFCRNN, with existing models on the three categories of the 300VW test set. Mean error is reported. Despite their simplicity, coord-3DRNN and heat-3DFCRNN are among the top-performing methods.

Method	Category 1	Category 2	Category 3
SDM [248]	7.41	6.18	13.04
CFSS [285]	7.68	6.42	13.67
TCDCN [280]	7.66	6.77	14.98
TSTN [133]	5.36	**4.51**	12.84
coord-3DRNN	**5.34**	5.01	**8.14**
heat-3DFCRNN	5.73	4.83	8.70

4.4.6 Qualitative Results

Figure 4.11 shows some failure case and qualitative results from 300VW data set. We observe that under static conditions (a), 3D model has issues with accuracy. However, during expression variations (b), it shows the best performance especially around the mouth area. Notice the lack of influence of the RNN in this type of situation. Early connectivity proves to be more suitable for modeling local motions. In the presence of large head movements (c), coord-3DRNN provides more robustness than 3D alone. The RNN shows its ability to model global motions. The complementarity between early and late connectivities is also highlighted.

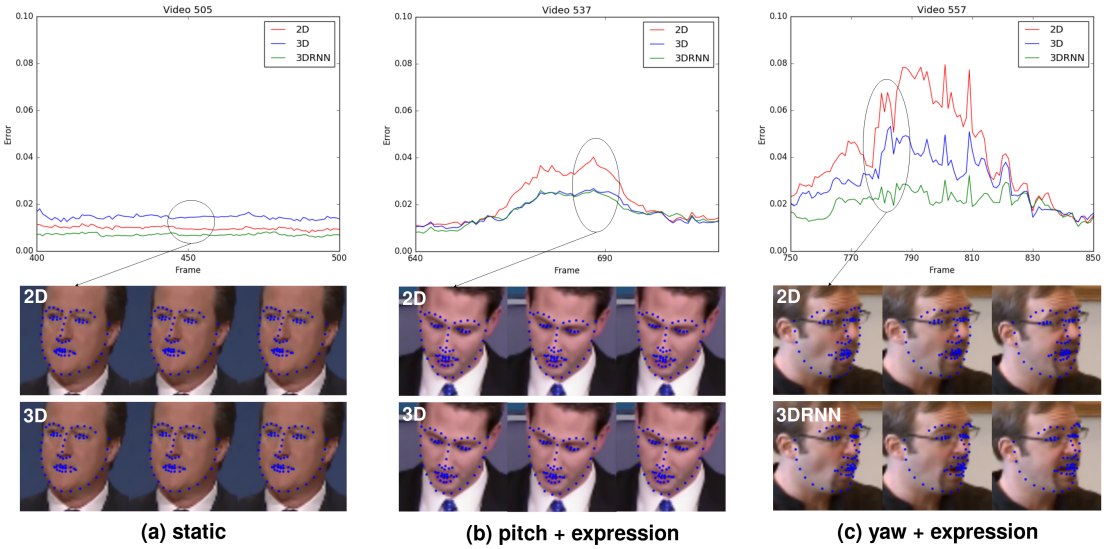

Figure 4.11: Failure case and qualitative results from 300VW data set. coord-3D model is less accurate under static conditions (a) but handles local motions better (b). When combined with a RNN, the robustness to large motions is improved (c).

In Figure 4.12, we show the inputs that activate the filters and the filter weights in the first layer for the coord-2D, coord-3D and coord-3DRNN models. It helps to understand what kind of patterns activate the filters. The coord-2D model encodes local patterns and color while the coord-3D and coord-3DRNN models encode variations of local patterns and color.

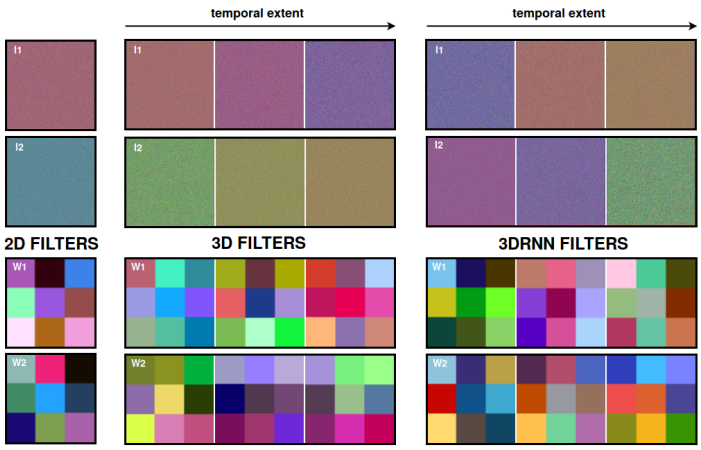

Figure 4.12: Activation maximization of two filters of the first layer (top) and their weights (bottom), for the coord-2D (left), coord-3D (center) and coord-3DRNN (right) models. The top images are sub-sampled while the bottom images are upsampled to facilitate viewing. The coord-2D model encodes local patterns and color while the coord-3D and coord-3DRNN models encode variations of local patterns and color.

4.4.7 Properties of the Networks

Table 4.7 presents different properties of the architectures proposed in this paper: number of parameters, size of the model, and prediction speed (in frames per second – FPS) on CPU and GPU. Runs are performed on an Intel Xeon E5 3.50GHz CPU and a Nvidia GeForce GTX 1080 Ti GPU. Although the number of parameters and the size increase considerably when the temporal dimension is added, the models remain light with less than 20M parameters and 70 MB. As a comparison, an architecture such as VGG16 [1], used for instance in [79, 163], has more than 130M parameters and a size of more than 500 MB. We can also observe that all models run in real time on GPU, and all coordinate regression models on CPU. Moreover, we believe that, by revising the architectures or by applying techniques such as binarization [22], real time could be reachable on CPU for FCN models too.

Table 4.7: Comparison of the proposed architectures regarding their numbers of parameters, model sizes and speeds. All models run in real time on GPU, and all coordinate regression models on CPU. VGG16 properties are also provided for comparison purposes.

Method	#params (M)	Size (MB)	Speed (FPS)	
			CPU	GPU
coord-2D	3.8	15	178	285
coord-3D	11.1	43	40	205
coord-2DRNN	14.2	55	60	252
coord-3DRNN	17.4	67	36	205
heat-2DFCRNN	10.2	40	14	104
heat-3DFCRNN	14.9	58	11	99
VGG16	130	500	0.3	24

4.5 Design Investigations

From the results of the first experiments, heatmap regression models may handle better spatial correlations than coordinates regression models due to their fully convolutional structure. The latter rely on fully connected layers which involve feature map vectorization with spatial information loss. The same may also be true for spatio-temporal correlations. Hence, we focus the rest of the experiments on heatmap regression models. Efficient 3D hourglasses are not trivial to design, as experienced in Section 4.3. The

[1] https://github.com/jcjohnson/cnn-benchmarks

architecture initially proposed is first revised in Section 4.5.1 with an emphasis on decoding. Two different approaches to better leverage the complementarity between spatial and temporal information are then investigated in Section 4.5.2. Either through the use of a two-stream architecture or by decoupling convolutions within a single stream architecture. Finally, in Section 4.5.3, the complementarity between local and global motion has also been subject to new experiments.

4.5.1 Encoder-Decoder

Given the lack of hindsight regarding the design and parameterization of hourglass-like architectures for facial landmark detection, we investigated:

- different strategies to encode image or image sequence into feature maps, i.e., convolutions with stride or max-pooling;
- different strategies to decode these feature maps into heatmaps, i.e., transposed convolution or upsampling;
- studied the influence of specific layers and connections, i.e., connections between encoder and decoder;
- studied the impact of regularization techniques, i.e., dropout.

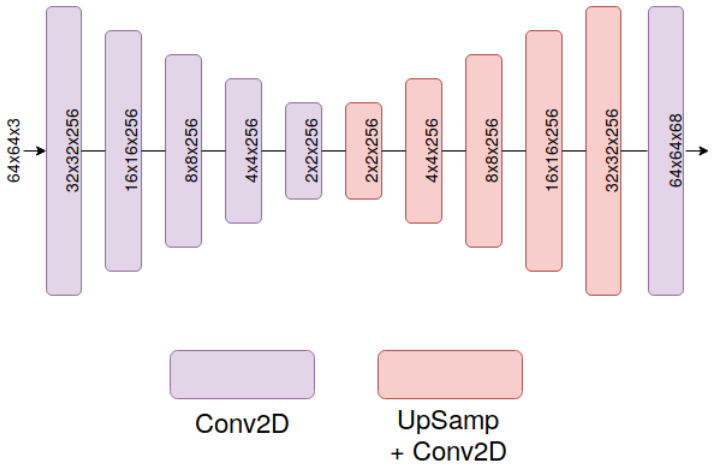

Figure 4.13: The proposed baseline network for heatmap regression. Compared to the models of Section 4.3.2, the encoder and decoder use max-pooling and upsampling rather than convolutions with stride.

We did not notice any significant difference between the aforementioned encoding and decoding strategies in terms of performance. The convolution with stride reduces process and the transposed convolution does not use a predefined interpolation method

and has learnable parameters. However, we opted for the use of max-pooling and upsampling to avoid potential unwanted artifacts (checkerboard pattern) [154], which could potentially lead to a loss of accuracy. The addition of dropout, at different positions and with different rates, as well as residual connections between encoder and decoder systematically resulted in no or even negative outcome. Dropout can lead to a decrease in performance when combined with batch normalization [128]. Since we use shallow networks, residual connections may be unnecessary. The results of these experiments are available in Appendix A.2. Figure 4.13 shows the new proposed 2D baseline for heatmap regression (2DFCN).

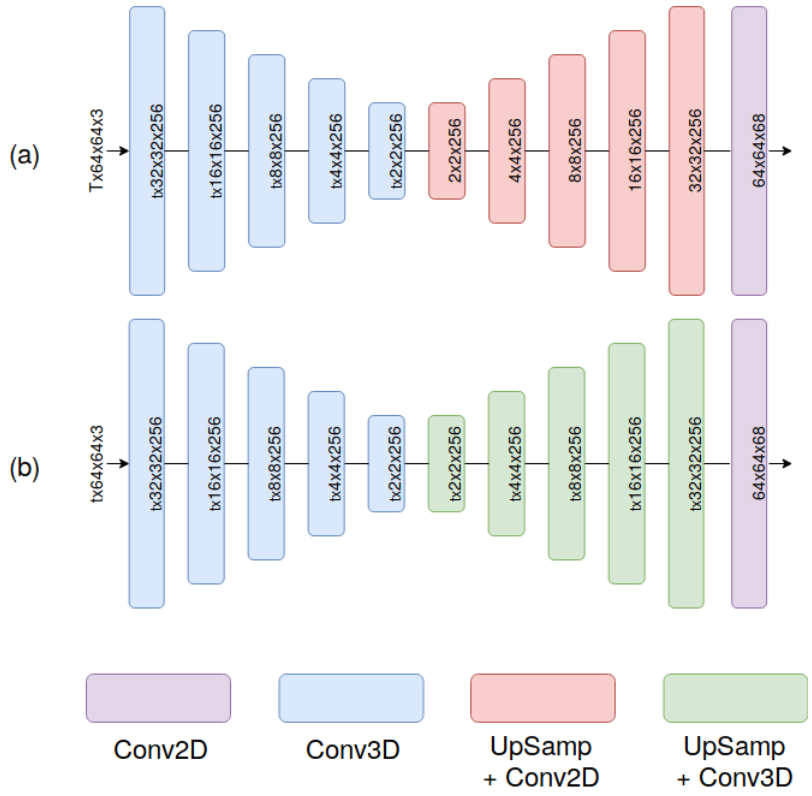

Figure 4.14: The proposed heatmap regression networks with early temporal connectivity based on 3D convolutions. (a) 2D decoding with early global temporal pooling and (b) 3D decoding with late global temporal pooling.

When it comes to decoding, the question arises as to whether it should be done in 2D or in 3D. 2D decoding appears to be appropriate and lightweight. As we want to perform the prediction on a single image, the last one of the sequence, the temporal dimension may not be necessary during decoding. However, to better consider motion information, it can be relevant to consider 3D decoding. Both strategies have been set up to assess the need of a 2D or a 3D decoding. Figure 4.14-(a) shows the 2D decoding (3DFCN-a). A global temporal pooling operation (max pooling with pool size equals

to the size of the input) is performed between the encoder and the decoder. 2D feature maps are then obtained, which encode the spatio-temporal context. Then 2D decoding is performed. Figure 4.14-(b) shows the 3D decoding (3DFCN-b). In the same way as for the encoder, the convolutions were extended by one dimension to also reconstruct the temporal dimension. Then follows a step of global temporal pooling in order to perform the prediction on a single image using the last convolution, which has been kept in 2D.

Table 4.8 shows the performance of 2D and 3D heatmap regression models in terms of AUC and FR with a 4% threshold on the three categories of 300VW. The 3D models produce better results than the 2D baseline in the three categories of the dataset and for the two metrics. 3D decoding in this configuration does not seem to have any major importance considering the marginal gain in performance obtained. Max and average pooling have also been evaluated for global temporal pooling without showing any significant difference (see Appendix A.2).

Table 4.8: Comparison of the different architectures presented in Section 4.5.1 on the three categories of the 300VW test set. AUC and FR at a threshold of 4% are reported. 3D models produce better results. 2D decoding performs on par with 3D decoding.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
2DFCN	55.11	5.67	49.77	2.93	49.17	7.32
3DFCN-a	56.28	**4.24**	50.52	**2.45**	**50.57**	6.96
3DFCN-b	**56.55**	4.53	**51.10**	3.24	**50.57**	**6.76**

4.5.2 Complementarity between Spatial and Temporal Information

In the first experiments, we observed a weakness of 3D convolution in static conditions. When designing a spatio-temporal model, decoupling/factorizing the processing of spatial and temporal information can help to better leverage the complementarity between both types of information. We propose two approaches which we present below.

Two-stream Fully Convolutional Network

The first approach is an architecture with a two-stream encoder, i.e., spatial and spatio-temporal stream (2SFCN). The goal of the spatial stream is to extract static features from still images while the temporal stream aims to extract the dynamic features of the face from an image sequence. The outputs of the two streams are concatenated and decoded in 2D. The same 2D and 3D encoders from Section 4.5.1 are used, as well as the 2D decoder.

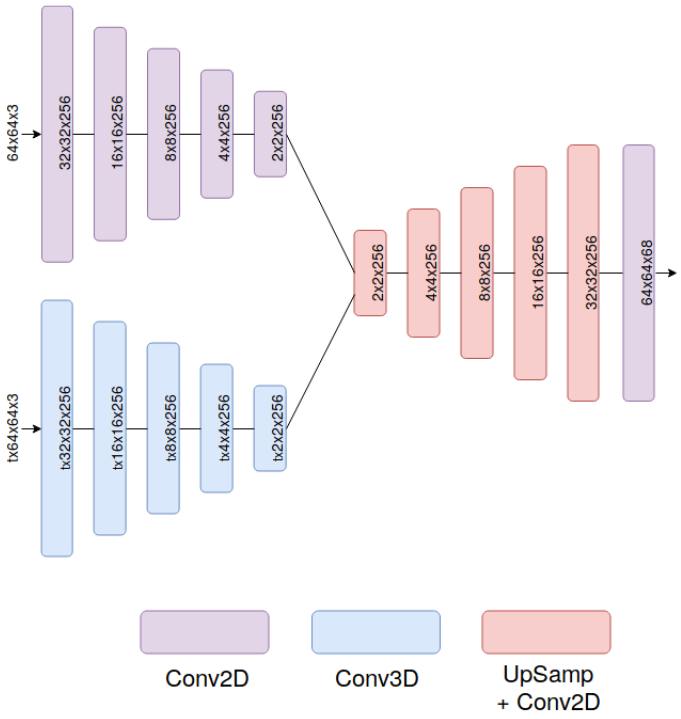

Figure 4.15: The proposed two-stream heatmap regression network. The same 2D and 3D encoders from Section 4.5.1 are used, as well as the 2D decoder.

Fully Convolutional Network with Decoupled Convolutions

In the second approach (D3DFCN), we decouple the 3x3x3 convolution kernel into a spatial kernel and a temporal kernel of dimensions 1x3x3 and 3x1x1 respectively. The complementarity between both types of information can therefore be exploited within a single stream. As seen in Section 4.2.2, the decomposition also facilitates the optimization and allows more complex functions to be represented by doubling the number of activation [213], i.e., additional non linear activation layer between spatial and temporal convolution. Several ways are proposed to apply these convolutions, as illustrated in Figure 4.16. The difference lies in the influence that the two types of filters can have on each other and on the output.

It should be noted that unlike [174] we do not use residual connections in the proposed architecture given its shallow depth (as in Section 4.5.1). In Figure 4.16-(a), the two convolutions are stacked by first applying spatial filters and then temporal ones. We therefore have a direct influence between the two types of information, and only the second convolution, i.e., the temporal one, is connected to the output. In Figure 4.16-(b), the two convolutions are applied in parallel, each along a different path. We therefore have an indirect influence between them. The outputs of the two convolutions are then cumulated. Finally, in Figure 4.16-(c), a balance between the two previous options is

made. We have a direct influence between the two types of convolutions, and both are connected to the output.

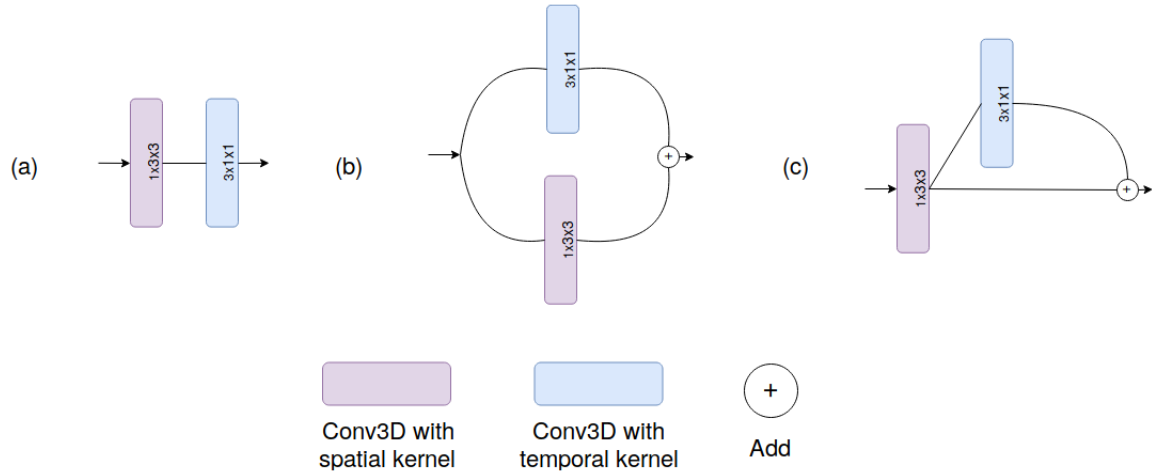

Figure 4.16: The proposed 3D convolution factorization strategies. The difference lies in the influence that the two types of filters can have on each other and on the output.

Table 4.9 presents the results of both two-stream and decoupled 3D convolutions on the three categories of 300VW. AUC and FR with a 4% threshold are reported. Overall, both approaches provide a performance gain over the use of a single vanilla 3D stream. Among the factorization strategies, (a) performed best, followed by (c) and then (b). Note that the latter results in a negative outcome. Inspired by the idea of structural diversity [278], we have also evaluated a random mix of the proposed factorization strategies: a-b-c-b-a. It was however unsuccessful. Compared to two-stream networks, factorization (a) shows slightly better results while being more advantageous since it requires a single stream.

Table 4.9: Comparison of the different architectures presented in Section 4.5.2 on the three categories of the 300VW test set. AUC and FR at a threshold of 4% are reported. Two-stream and factorization approaches provide a performance gain over the use of a single vanilla 3D stream. Among the factorization strategies, (a) performed best.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
3DFCN-a	56.28	4.24	50.52	2.45	50.57	6.96
2SFCN	56.24	4.43	52.16	**1.66**	51.43	**6.69**
D3DFCN-a	**57.61**	**3.84**	**52.08**	2.56	**52.11**	7.06
D3DFCN-b	55.32	5.03	49.81	2.91	49.55	7.23
D3DFCN-c	57.07	4.81	51.32	2.28	51.95	6.89
D3DFCN-mix (a-b-c-b-a)	56.63	4.24	50.59	2.58	50.72	7.05

4.5.3 Complementarity between Local and Global Motion

The main purpose of the proposed architecture is to model local motions through early temporal connectivity, i.e., directly at the pixel level. This connectivity provides a better robustness when local motions such as expression variations occur and complements existing work that focuses on global head movements, i.e., pose variations. In the following, different architectures are proposed to revise the combination of these two approaches and to better exploit their complementarity. We are especially interested in the necessity of having a ConvRNN layer in a fully convolutional architecture as 3D convolutions alone could be enough for short-term dependencies such as in facial landmark detection.

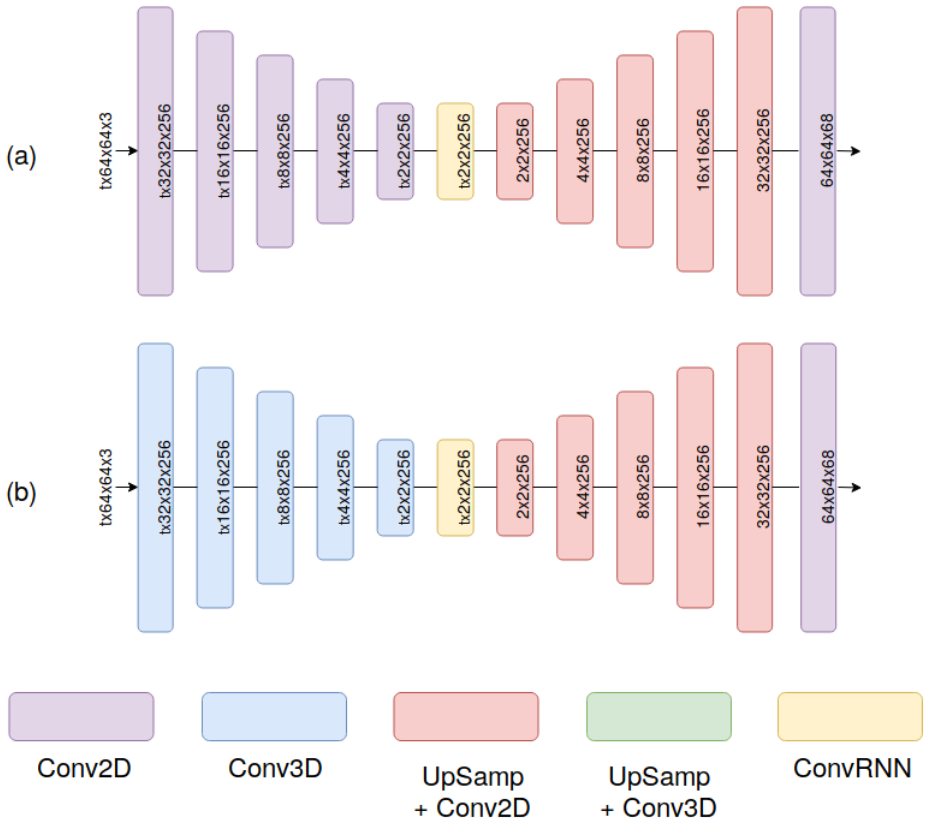

Figure 4.17: The proposed 2D (a) and 3D (b) recurrent networks for heatmap regression. The late temporal connectivity approach is followed using a ConvLSTM as it allows to preserve a fully-convolutional architecture.

Fully Convolutional Recurrent Network

Most architectures dealing with global motions perform coordinate regression and integrate a FC-RNN (LSTM, GRU) layer after the *conv* layers, in order to capture temporal information on high level vectorized feature maps. When designing the recur-

rent encoder-decoder network, we followed the late temporal connectivity approach but opted for the use of a ConvLSTM layer between the encoder and the decoder. ConvLSTM turns out to be better than FC-LSTM at handling spatio-temporal correlations [247] as it allows to preserve a fully-convolutional architecture.

Figure 4.17 shows the proposed baseline, which consists of the FCN from Section 4.5.1 to which we add a ConvLSTM with the same parameters as the other convolutions (2DFCRNN). The encoder is distributed in time with shared parameters. By processing the output of the streams, the recurrent layer is able to compute global motion features. Decoding is done in 2D by keeping only the feature maps of the last timestep of the ConvLSTM layer. To integrate both local and global motions, the same changes have been made to the FCN with temporal connectivity from Section 4.5.1 (3DFCRNN). We add a ConvLSTM between the 3D encoder and the 2D decoder with the same parameters as the other convolutions. No dropout and no recurrent dropout have been used. By combining 3D convolutions and ConvLSTM, the complementarity between local and global motions might be better exploited.

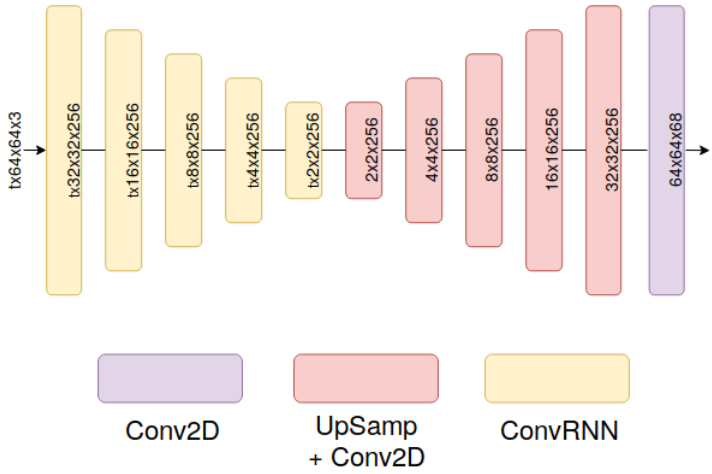

Figure 4.18: The proposed fully recurrent network for heatmap regression. The encoder is composed exclusively of ConvLSTM as it also enables temporal connectivity at the pixel level.

In addition, we also designed a model whose encoder is composed exclusively of ConvLSTM (FCRNN, see Figure 4.18). Due to its convolutional structure, ConvLSTM can also provide temporal connectivity at the pixel level. By doing so, we can determine which of 3D convolution or ConvLSTM is the most appropriate to deal with both local and global motions and get better insights on their complementarity.

Table 4.10: Comparison of the different architectures presented in Section 4.5.3 on the three categories of the 300VW test set. AUC and FR at a threshold of 4% are reported. Only a small gain is observed between static and recurrent models. The use of an encoder composed only of ConvRNN demonstrates performance in favour of early temporal connectivity.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
2DFCN	55.11	5.67	49.77	2.93	49.17	7.32
2DFCRNN	55.46	5.66	51.25	**1.81**	49.55	7.84
3DFCN-a	56.28	4.24	50.52	2.45	50.57	6.96
3DFCRNN	56.49	4.40	51.31	2.49	50.26	7.87
FCRNN	**57.37**	**3.99**	**52.19**	1.97	**52.14**	**6.41**

Table 4.10 summarizes the performance of recurrent models for heatmap regression over the three categories of 300VW in terms of AUC and FR with thresholds at 4%. The contribution of the recurrent layer for both 2D and 3D models is less significant than with coordinate regression models. Only a small gain is observed between static and recurrent models. However, the use of an encoder composed only of ConvRNN has demonstrated better performances than 3D convolutions except when the 3D convolutions are factorized. This demonstrates the importance of early temporal connectivity. Still, this type of architecture is much more costly to train.

4.6 Discussion

Experiments on two datasets confirm that modeling local motion improves the results (e.g. especially with expressions, see Table 4.3 in Section 4.4.3), and that it is complementary to RNNs, which model global motion. In Chapter 3, the difficulty of current approaches in the presence of pose and expression variations has been highlighted. By considering temporal information, a performance gain up to 30% for critical expressions (disgust and sadness) and 10% for critical poses (Pitch and Diagonal) can be observed on SNaP-2DFe for coordinate regression models. In light of the results obtained on 300VW, this gain is likely to be even greater for heatmap regression models.

In Table 4.11, a summary of the performance of all the proposed approaches is provided. By decoupling spatial and temporal information processing, either via a two-stream architecture or via 3D convolution factorization, accuracy can be improved. Factorization has been found to perform better while being lightweight and easier to train.

Table 4.11: Comparison of the different architectures presented in Section 4.5.1, 4.5.2 and 4.5.3 on the full 300VW testset. AUCs and failure rates at thresholds of 8% and 4% are reported. 3D convolution factorization was found to perform the best.

Method	AUC @8%	FR @8%	AUC @4%	FR @4%
2DFCN	74.88	1.55	52.35	5.31
2DFCRNN	75.19	1.71	53.01	5.12
3DFCN-a	75.57	1.48	53.46	4.37
2SFCN	75.88	1.47	54.07	4.20
D3DFCN-a	**76.37**	**1.23**	**54.90**	4.22
3DFCRNN	75.57	1.65	53.71	4.67
FCRNN	76.28	1.50	54.82	**4.00**

Given the fully convolutional structure of the proposed architectures, the interest and usage of a RNN for this type of architecture was also questioned. One of the advantages of a ConvRNN is that it can be applied directly to the first layers of a network, allowing early temporal connectivity. Although it is more difficult to train, it gives interesting performances close to the ones of factorized 3D convolution. This tends to show the importance of temporal connectivity at the pixel level. However, 3D convolution alone, if properly exploited, might be sufficient to handle both local and global motions through hierarchical learning.

In summary, this work has demonstrated that:

- capturing temporal information significantly contributes to facial landmark detection, especially during pose and expression variations;
- fully convolutional networks seem more inclined to handle spatio-temporal correlations;
- it is equally important to consider both local and global motion;
- it is crucial to decouple the processing of spatial and temporal information.

There is some room for improvement. Due to overfitting problems, heatmap regression models were unable to be evaluated on SNaP-2DFe. Although the results on 300VW confirmed the validity of the approach, quantifying the impact of head pose and expression would provide more insights regarding these models. Besides, our analysis by challenge is limited only to head pose and expression. Given the difficulties encountered on 300VW and the temporal nature of our approach, the impact of occlusions should be evaluated. In Section 4.5, we performed a design investigation based

mainly on the figures obtained on 300VW. To improve the understanding of the proposed architectures, it could be worth working on feature visualization aspects. Lastly, our approach may be limited by the lack of an appropriate loss function for the temporal domain. The explicit use of landmark trajectories could be helpful.

Chapter 5

Conclusion

Contents

The goal of our work was to investigate facial landmark detection under uncontrolled conditions. The first line of work focused on performance analysis of current methods. We highlighted a weakness in the commonly used evaluation protocols. They rely on datasets with limited annotations and do not allow for exhaustive performance analysis. It is especially complex to quantify the performance of approaches according to the challenges encountered. Our efforts were therefore directed towards this issue.

The second line of work focused on the development of a solution capable of coping with most of the difficulties encountered under uncontrolled conditions (e.g., pose, expression, illumination, occlusion, motion blur). Based on the results of our initial work, we considered video-based solutions. Unlike the popular still-image approaches which are unable to take advantage of temporal information, video-based approaches leverage temporal coherence to provide more robustness under uncontrolled conditions. We noted that the approaches proposed so far provide only partial temporal modeling and addressed this issue.

5.1 Summary of the Contributions

In this thesis, contributions for facial landmark detection are made in both the analysis of the performance of current approaches and in the modeling of temporal information.

5.1.1 Impact of Pose and Expression Variations

As part of the performance analysis, two benchmarks, one with image and one with video data, were conducted using methods compilation from the literature. They helped analyze the performance of a large number of methods and obtain useful insights about the current state of the problem. The datasets currently used in experiments remain limited and do not allow for a comprehensive analysis. Given the relevance of such an analysis, an experimental study was conducted using a recently released video dataset, SNaP-2DFe [2], with more exhaustive annotations. The impact of pose and expression variations on current facial landmark detection approaches has been quantified. Among the selected approaches are coordinate (DAN [111]) and heatmap (HG, SAN [23, 55]) deep regression models, including multi-task (TCDCN [280]) and dynamic solutions (SBR, FHR [56, 207]). Performance measurements for each pose, expression, and their different combination have been achieved. These measurements were also compared to the performances obtained on static neutral faces. In addition, an analysis at multiple scales, at the level of the shape and at the level of each landmark, was performed to identify the most critical regions of the face.

The overall performance in the presence of pose, expression or both is still worse than on static neutral face. Average ΔAUC / ΔFR for head poses and facial expressions are respectively -7.13 / +2.83 and -2.62 / -0.28. Eyebrows, facial edges, and the mouth are the regions that are the most affected. These observations are also valid for the best method despite its fair overall performance. Through this evaluation, the most challenging poses and expressions have been identified. With respect to SNaP-2DFe annotations, they include Pitch and Diag for pose (average ΔAUC / ΔFR of -8.31 / +0.22 and -15.32 / +7.56 respectively), and Disgust and Sadness for expressions (average ΔAUC / ΔFR of -7.52 / +0.02 and -3.89 / -0.51 respectively). They represent combination of complex motions which are very localized. The results illustrate the importance of suitable temporal modeling which can effectively capture the dynamics of the face. This complex dynamics involves both global motion, as in head movements, as well as local motion at the level of each components of the face. This distinction had not yet been considered in the literature; until now, temporal coherence has been exploited in global fashion only.

5.1.2 Spatio-Temporal Modeling

The objective then focused on improving the temporal modeling of landmark detection, by considering local motion in addition to global motion. By combining CNN with RNN, only limited temporal modeling on feature maps with high level of abstraction can be provided since CNN is not able to capture any motion. As a consequence, local motion can hardly be captured. An extensive review of the literature of spatio-temporal modeling highlighted the benefits of spatio-temporal convolution to address this problem. It was therefore this solution that has been chosen to improve the temporal connectivity of landmark detection and integrate local motion modeling.

To evaluate the contribution of local motion to facial landmark detection, several architectures have been designed based on the two main models in the literature: coordinate regression networks and heatmap regression networks. Experiments on two datasets confirm that local motion modeling improves results. Regarding challenging head poses such as Pitch and Diag on SNaP-2DFe, an increase of AUC@4 from 65.37 to 67.12 and from 64.58 to 65.54 respectively is achieved by adding early temporal connectivity to the proposed 2D coordinate baseline. A greater gain is observed for facial expressions. Regarding challenging facial expressions such as Disgust and Sadness, AUC@4 increased from 52.70 to 57.52 and from 53.49 to 58.31 respectively.

These experimentations were followed by the study of the complementarity between spatial and temporal information as well as local and global motion. The emphasis was placed on heatmap regression as it appears more suitable to handle spatio-temporal correlations. Through this study, the design of the initially proposed architectures has been revised and improved. In particular, it has been shown that decoupling spatial and temporal information processing has benefits, especially through the factorization of spatio-temporal convolutions. As a result, AUC@4 / FR@4 on 300VW improved from 53.46 / 4.37 to 54.90 / 4.22. By leveraging these complementarities more effectively, competitive performance with current state-of-the-art approaches have been achieved, despite the simplicity of the proposed models.

5.2 Future Work

Several perspectives have been identified to extend this work and go further in solving the problem of landmark detection. They concern the training strategy, model design, benchmarking, data, applications, and algorithm reliability.

5.2.1 Training Strategy

In Chapter 4, efforts have been focused on architecture design. The next step is the design of the loss function. In image-based work, some authors have shown the importance of taking into account geometric constraints [30, 273]. Assuming that constraints on trajectories can be just as important, an adversarial term could be added to the MSE loss by using a discriminator during training, which would distinguish between correct and incorrect trajectories. This would ensure consistent facial motion. Since the shape can be represented as a graph, the use of a graph neural network for the discriminator seems promising [254].

Among the major challenges are the occultations that could not be quantified in the benchmarks of Chapter 3 due to the absence of annotations. Some results are available for image-based approaches [235] and give a good illustration of the impact of occultations on facial landmark detection. Inspired by the work on deep generative models [110], one way to improve the robustness to occlusions could be to jointly train face deocclusion and landmark detection. Face deocclusion suffers from a lack of structural information which affects the appearance of reconstructed faces [266]. Landmarks are affected by occlusions but ultimately represent the structural information of the face. The joint training of both tasks could improve individual performance. Using a tem-

poral approach like the one proposed in this work, face deocclusion could also benefit from temporal information (i.e., landmarks trajectories) and therefore could have a better chance to correctly generate the occluded part.

5.2.2 Model Design

In Chapter 4, lightweight architectures have been used in order to better isolate the contribution of local motion to facial landmark detection. Experiments have shown the benefits of local motion modeling, especially for challenges such as expressions. These architectures could be enriched with some concepts, such as attention mechanism [267] or region of interest pooling [89], currently used in 2D models but extended to the temporal domain. A spatio-temporal attention mechanism could help to capture more relevant spatio-temporal features related to landmarks and thus further boost performance [131]. 3D region of interest pooling could also reduce computation complexity [250].

With the recent availability of multi-face datasets (i.e., images showing two or more faces) such as the AISA dataset [242], the evolution of the models proposed in this work towards the support of multiple faces is a possible direction of research. Existing image-based work generally integrate a bridge between face detection and landmark detection using a face proposal network [50], which helps remove the sensitivity issues related to the bounding box (see Section 2.1.2). A similar strategy could be used using a tube proposal network [72]. Lately, datasets such as MENPO [271] or 300VW [270] offer 3D landmark annotations. The integration of 3D information into the models of this work also seems interesting to match the 3D nature of the real world. 3D information can help to preserve the semantic meaning of landmarks, by handling invisible landmarks during extreme variations, and further improve performance [23, 287].

The memory footprint and execution time of DNNs also need to be considered. Spatio-temporal models, due to the integration of an additional dimension, require more effort in design optimization. It is worth investigating compression techniques (e.g., binarization) on spatio-temporal models and to compare the speed-accuracy trade-offs with 2D models. Some authors also suggest that motion modeling is not crucial across all layers of the model, but without any general consensus: is motion related to low-level features, high-level features, both, or is it problem-dependent [246]? A mixture of conv2D, conv3D, and convRNN has been manually explored [276], resulting in a slow sub-optimal design process.

Overall, architecture engineering can be difficult and time-consuming process as experienced in Section 4.5. A way to streamline this process that seems promising is neural

architecture search and evolution, i.e., automatically exploring models and finding the one that optimally captures features for the problem [170, 212]. This could be applied to landmark detection and could lead to the discovery of new architecture structures. At the same time, improving the interpretability of models through more appropriate visualization techniques could also contribute greatly to the design of models [155].

5.2.3 Benchmarking

The problem of facial landmark detection is currently being studied as an isolated problem when in fact it is a key component for many applications such as facial expression recognition. Although performance has improved considerably according to the common metrics such as RMSE, this does not guarantee that an approach with a lower RMSE, a higher AUC, or a lower FR would give better performances when used for a subsequent task. The community is working with, perhaps, a vision of the problem that is not broad enough, without knowing whether the current performance of facial landmark detection is good enough for these tasks [49]. In Chapter 3, the difficulty to quantify the performance of methods according to the different challenges encountered in uncontrolled conditions has been highlighted. Then a benchmark was conducted to address this problem. This work could be extended to measure the impact of landmark detection on expression recognition. By leveraging all the data provided by SNAP-2DFe (i.e., static and helmet camera and expression annotations), an in-depth benchmark could be achieved to identify which approach works best for expression recognition, which landmarks are the most critical, and evaluate whether the instability of static methods is a problem for dynamic analyses. Besides, answering these interrogations could lead to new metrics as well as provide some feedback to extend it to other applications. It is also desirable to find an appropriate way to integrate temporal and 3D approaches into this benchmark. Beyond SNaP-2DFe and expression recognition, other datasets could be used to evaluate facial landmark detection in more contexts. As an example, VGGFace2 appears relevant for face recognition [26].

5.2.4 Data

In Chapter 3, it was pointed out that sufficiently annotated data is needed to conduct a comprehensive analysis and thus be able to quantify the impact of the various difficulties encountered under uncontrolled conditions. SNaP-2DFe is a dataset that was used to achieve a benchmark focused on pose and expression variations. These results provide interesting feedback on the performance of current approaches. SNaP-2DFe could be extended to other major challenges, including occlusions. It is possible to artificially add static and/or dynamic occlusions to current data. This would be an opportunity to

extend the benchmark of Section 3.4. The impact of dynamic occlusions is an interesting topic regarding temporal approaches.

Currently the datasets for video-based facial landmark detection are limited. Although these datasets seem to contain much more data than still image datasets, this is only an illusion, given the redundancy present in videos. This results in datasets with much less variety. To limit these weaknesses, a first solution, presented in Section 4.2.2, consists in initializing video-based models based on pre-trained image-based models. However, the temporal dimension may not be correctly initialized. A second solution is to increase the amount of data through data augmentation (Section 3.2.2). Still, this is currently very limited since it only implies the use of spatial augmentation or resampling. There is a lack of relevant motion augmentation techniques. It seems useful to take an interest in these issues, which remains crucial considering the popularity of deep learning and could greatly contribute to video analysis. A potential lead would be to perform video augmentation on static datasets using video prediction from still images [129]. This would allow to benefit from the advantages of these datasets while considerably increase the amount of video data. An alternative could be to take advantage of photorealistic video games as already done for other tasks such as human pose estimation [60]. The annotation process can be considerably simplified with such a solution. Although the resulting data remains artificial, fine-tuning on real data can provide performance equivalent or even superior to a training with real data alone [214]. As opposed to these solutions which are intended to provide additional data, another alternative might be to develop a few-shot regression algorithm, i.e., able to learn from a limited amount of training data [68]. These approaches seem to be under-studied for regression problems and have not yet been considered for facial landmark detection.

5.2.5 Application

The visible spectrum is not the only light range studied in computer vision. Applications in the visible range can also make sense in the thermal infrared range (e.g., for surveillance). However, in the latter, it is difficult if not impossible to rely on the appearance and shape, which are missing or severely altered, especially within the face. The use of motion seems relevant to address facial landmark detection in thermal infrared images. Therefore, it seems interesting to extend the present work to this type of data. Given the lack of annotated data, another idea would be to perform domain change, from thermal infrared to visible spectrum, using GANs. It has already been experimented for facial recognition and can be used to exploit models trained on images from the visible present in larger quantities [277].

Still focusing on the face, the detection of landmarks can not only be performed on human faces but also on those of animals. Applications are also numerous (e.g., pain detection, surveillance). However, facial landmark detection on animals is much more difficult since the variability in terms of shape and appearance is more important, even within the same species. Some authors have already investigated the performance of algorithms initially designed for the human face [21] or have proposed solutions for transferring knowledge gained from human faces [176]. It would be interesting to evaluate the contribution of motion on animal faces, as it may help handle their high degree of variability.

5.2.6 Reliability

When using facial analysis in real-world applications, it is relevant to think about how malicious people will try to compromise algorithms. Regarding facial analysis, glasses have been specially designed to impersonate another person in the eyes of a DNN-based recognition system [192]. In various contexts (e.g., surveillance, security, autonomous agents) this kind of attack can lead to dangerous situations. Beyond spoofing attacks [70], which are generally independent of the type of algorithm used, some are instead targeted at specific approaches. Deep learning has become more popular over the years and it is all the more important to focus on these security aspects since it has been shown that it is easy to deceive a neural network [1]. A small imperciptible perturbation added to the input [74], which can be applied only to a single pixel [201], can result in errors, even with a well-trained model. These carefully crafted modifications are called adversarial perturbation and make DNNs vulnerable. Adversarial attacks were initially performed in the digital domain but have been successfully generalized to the physical domain, i.e., perturbations are physically added to the objects themselves, making them even more critical [59, 115, 20]. Although these attacks have shown generic properties called transferability and are therefore likely to succeed on different models and problems, there doesn't seem to be any work on facial landmark detection. Landmarks are rather directly used for geometry-based attack [42]. Adversarial attacks are widely studied in the context of classification but only a few work has been conducted on regression problems [153]. It might be interesting to evaluate adversarial robustness of the models of this work, which are representative of the literature, to the different possible attacks and to develop defense solutions if needed. Such adversarial attacks could interfere with any facial analysis system, landmark detection being commonly one of their major building block. An approach could be to give little confidence or attention to unseen regions.

Appendix A

Additional Experiments

Contents

This appendix contains additional independent experiments related to Chapter 4. The protocol used may vary from the one described in Section 4.4.1, especially the amount of training data, which can be more limited due to the nature of the experiment, e.g., larger input size, sub-sampling. The results obtained influenced our choices when designing the architectures from Section 4.3 and 4.5.

A.1 Influence of the Size of the Temporal Window

One of the parameters supposed to be critical for our spatio-temporal models from 4 is the size of the temporal window. Three values have been evaluated, i.e., 3, 6, and 9 (see Table A.1). The worst performances are obtained with a temporality of 6 images. Similar performances are observed with values of 3 and 9. We were unable to evaluate any larger value due to the limited amount of training data. As already mentioned in Section 4.4.1, each video of the training set is split into several non-overlapping image sequences of size T. When $T > 9$, the number of training data is considerably reduced, which leads to overfitting problems. Based on these results, a short-term temporal dependency might be sufficient for facial landmark detection. Long-term motion may provide irrelevant information. Different sizes of convolution kernel, i.e., 7, 5, 3, have also been investigated. As in [211], a constant size of 3 over all layers provides the best results.

Table A.1: Influence of the size of the temporal window on the three categories of the 300VW test set. AUC and FR at a threshold of 8% are reported. Short term temporal dependency might be sufficient for facial landmark detection. A constant kernel size of 3 over all convolution layers provides the best results.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
FCN 3D (a) w/ $T = 3$	**74.99**	2.27	73.44	0.26	**71.85**	**4.27**
FCN 3D (a) w/ $T = 6$	74.48	2.50	73.42	0.28	71.21	4.45
FCN 3D (a) w/ $T = 9$	74.96	**2.06**	**74.02**	**0.25**	71.65	4.30
FCN 3D (a) w/ $T = 9$ + kernel variation	72.21	4.17	70.36	0.61	70.08	4.64

A.2 Encoder-Decoder Design

There is currently a lack of insights about the design and parameterization of heatmap regression architectures for facial landmark detection. Therefore, we conducted several

experiments to inform our design decisions. We compared the most commonly used encoding-decoding strategies in the literature:

- *conv* with stride (> 1) and *conv* with max pooling for encoding;
- transposed *conv* and oversampling with *conv* for decoding.

No significant differences can be observed. Among the other experiments that have been conducted, we investigated the importance of multiple 1x1 *conv* output layers and residual connections between the encoder and the decoder, which are two characteristics found in the HG [23]. We also investigated the use of dropout regularization to prevent overfitting. The results are reported in Table A.2.

Table A.2: 2D designs evaluation on the three categories of the 300VW test set. AUC and FR at a threshold of 8% are reported. Regarding the different encoding and decoding strategies, no significant differences can be observed. One output layer appears to be sufficient. The addition of dropout and residual connections alters performance.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
2DFCN w/ stride + deconv	75.69	2.39	**74.64**	0.23	71.91	3.91
2DFCN w/ stride + upsamp	**76.16**	1.54	73.87	0.19	71.17	3.43
2DFCN w/ pool + upsamp	75.89	**1.44**	74.18	**0.15**	71.52	**2.86**
2DFCN w/ stride + upsamp + connections	74.88	2.48	71.39	0.31	70.11	4.31
2DFCN w/ stride + upsamp + dropout	75.69	1.57	72.77	0.22	70.80	3.41
2DFCN w/ stride + upsamp w/o 1x1	76.12	2.06	**74.64**	0.30	**72.06**	3.94

One output layer appears to be sufficient in our case. The removal of the first 1x1 *conv* layer did not affect the performance. However, the addition of dropout on the *conv* layers or residual connections between the encoder and the decoder can significantly alters performance. We already discussed in Section 4.3 that the combination of dropout with batch normalization can lead to a decrease in performance [128]. Regarding residual connections, they may be unnecessary for shallow networks such as the ones proposed in this work.

We also investigated the use of max and average pooling for global temporal pooling between the 3D encoder and the 2D decoder. Both seem relevant to process motion information. Max pooling shows the best performance but with a relatively marginal gain (see Table A.3). Important motion clues may be missed by average pooling.

Table A.3: 3D designs evaluation on the three categories of the 300VW test set. AUC and FR at a threshold of 8% are reported. Max pooling shows the best performance but with a relatively marginal gain.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
3DFCN-a w/ max pool	**75.79**	**2.23**	74.80	**0.26**	**72.34**	**4.31**
3DFCN-a w/ average pool	75.27	2.69	**75.24**	0.27	72.26	4.39

A.3 Data Augmentation

Since we encountered overfitting problems when training our models from Chapter 4, we investigated video-based data augmentation. Temporal sub-sampling of 2x was performed on all video sequences of 300VW training set. The resulting data has been used, in addition to the original data, to train our heatmap regression model with early temporal connectivity (3DFCN-a) from Section 4.5.1. However, as shown in Table A.4, a decrease in performance is observed compared to the model trained without such an augmentation. The motion introduced may not be natural enough and irrelevant to the test set.

Table A.4: Influence of 2x sub-sampling data augmentation on the three categories of the 300VW test set. AUC and FR at a threshold of 8% are reported. A decrease in performance is observed on all categories when the sub-sampled data are used.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
3DFCN-a w/o sub-sampling	**74.26**	**2.73**	**73.22**	**0.31**	**70.96**	4.51
3DFCN-a w/ sub-sampling	73.47	3.06	72.70	0.33	70.72	**4.50**

Bibliography

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6:14410–14430, 2018.

[2] Benjamin Allaert, José Mennesson, Ioan Marius Bilasco, and Chabane Djeraba. Impact of the face registration techniques on facial expressions recognition. Signal Processing: Image Communication, 61:44–53, 2018.

[3] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: a comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164, 2018.

[4] Epameinondas Antonakos, Joan Alabort-i Medina, and Stefanos Zafeiriou. Active pictorial structures. In Computer Vision and Pattern Recognition, pages 5435–5444. IEEE, 2015.

[5] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In Computer Vision and Pattern Recognition, pages 3444–3451. IEEE, 2013.

[6] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In Computer Vision and Pattern Recognition, pages 1859–1866. IEEE, 2014.

[7] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In International Workshop on Human Behavior Understanding, pages 29–39. Springer, 2011.

[8] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. arXiv preprint arXiv:1511.06432, 2015.

[9] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In International Conference on Computer Vision Workshops, pages 354–361. IEEE, 2013.

[10] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In European Conference on Computer Vision, pages 404–417. Springer, 2006.

[11] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. Transactions on Pattern Analysis and Machine Intelligence, 35(12):2930–2940, 2013.

[12] Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. Transactions on Neural Networks, 5(2):157–166, 1994.

[13] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Generative or discriminative? getting the best of both worlds. Bayesian Statistics, 8(3):3–24, 2007.

[14] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In International Conference on Computer Vision, pages 3980–3989. IEEE, 2017.

[15] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In SIGGRAPH, volume 99, pages 187–194, 1999.

[16] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In Computer Vision and Pattern Recognition, pages 2544–2550. IEEE, 2010.

[17] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning, 3(1):1–122, 2011.

[18] Matteo Bregonzio, Shaogang Gong, and Tao Xiang. Recognising action as clouds of space-time interest points. 2009.

[19] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[20] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. arXiv preprint arXiv:1712.09665, 2017.

[21] Adrian Bulat and Georgios Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. 2016.

[22] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In International Conference on Computer Vision. IEEE, 2017.

[23] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In International Journal of Computer Vision, volume 1, page 8, 2017.

[24] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In Computer Vision and Pattern Recognition, pages 109–117. IEEE, 2018.

[25] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In International Conference on Computer Vision, pages 1513–1520. IEEE, 2013.

[26] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In Automatic Face and Gesture Recognition. IEEE, 2018.

[27] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. International Journal of Computer Vision, 107(2):177–190, 2014.

[28] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[29] Xi Chen, Erjin Zhou, Yuchen Mo, Jiancheng Liu, and Zhimin Cao. Delving deep into coarse-to-fine framework for facial landmark localization. In Computer Vision and Pattern Recognition Workshops, pages 142–149. IEEE, 2017.

[30] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial learning of structure-aware fully convolutional networks for landmark localization. arXiv preprint arXiv:1711.00253, 2017.

[31] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014.

[32] Grigorios G Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, and Stefanos Zafeiriou. A comprehensive performance evaluation of deformable face tracking "in-the-wild". International Journal of Computer Vision, 126(2-4):198–232, 2018.

[33] Grigorios G Chrysos, Epameinondas Antonakos, and Stefanos Zafeiriou. Ipst: Incremental pictorial structures for model-free tracking of deformable objects. Transactions on Image Processing, 27(7):3529–3540, 2018.

[34] Grigorios G Chrysos and Stefanos Zafeiriou. Pd 2 t: Person-specific detection, deformable tracking. Transactions on Pattern Analysis and Machine Intelligence, 40(11):2555–2568, 2018.

[35] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). 2016.

[36] Tim F Cootes, Mircea C Ionita, Claudia Lindner, and Patrick Sauer. Robust and accurate shape model fitting using random forest regression voting. In European Conference on Computer Vision, pages 278–291. Springer, 2012.

[37] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. Transactions on Pattern Analysis and Machine Intelligence, 23(6):681–685, 2001.

[38] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. Computer Vision and Image Understanding, 61(1):38–59, 1995.

[39] Timothy F Cootes, K Walker, and Christopher J Taylor. View-based active appearance models. In Automatic Face and Gesture Recognition, pages 227–232. IEEE, 2000.

[40] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In British Machine Vision Conference, volume 1, page 3, 2006.

[41] David Cristinacce and Timothy F Cootes. Boosted regression active shape models. In British Machine Vision Conference, volume 2, pages 880–889, 2007.

[42] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. Fast geometrically-perturbed adversarial faces. In Winter Conference on Applications of Computer Vision, pages 1979–1988. IEEE, 2019.

[43] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In International Conference on Computer Vision, pages 764–773. IEEE, Oct 2017.

[44] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, volume 1, pages 886–893. IEEE, 2005.

[45] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In European Conference on Computer Vision, pages 428–441. Springer, 2006.

[46] Matthias Dantone, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Real-time facial feature detection using conditional regression forests. In Computer Vision and Pattern Recognition, pages 2578–2585. IEEE, 2012.

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, pages 248–255. IEEE, 2009.

[48] Jiankang Deng, Qingshan Liu, Jing Yang, and Dacheng Tao. M3 csr: Multi-view, multi-scale and multi-component cascade shape regression. Image and Vision Computing, 47:19–26, 2016.

[49] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. International Journal of Computer Vision, pages 1–26, 2018.

[50] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. arXiv preprint arXiv:1708.06023, 2017.

[51] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3D convnets: New architecture and transfer learning for video classification. arXiv preprint arXiv:1711.08200, 2017.

[52] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In International Workshop on

Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pages 65–72. IEEE, 2005.

[53] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In Computer Vision and Pattern Recognition, pages 1078–1085. IEEE, 2010.

[54] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Computer Vision and Pattern Recognition, pages 2625–2634. IEEE, 2015.

[55] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In Computer Vision and Pattern Recognition, pages 379–388. IEEE, 2018.

[56] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In Computer Vision and Pattern Recognition, pages 360–368. IEEE, 2018.

[57] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. California Mental Health Research Digest, 8(4):151–158, 1970.

[58] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture, pages 27–46, 1997.

[59] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In Computer Vision and Pattern Recognition, pages 1625–1634. IEEE, 2018.

[60] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In European Conference on Computer Vision. Springer, 2018.

[61] Haoqiang Fan and Erjin Zhou. Approaching human level facial landmark localization by deep learning. Image and Vision Computing, 47:27–35, 2016.

[62] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In Neural Information Processing Systems, pages 3468–3476, 2016.

[63] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In Computer Vision and Pattern Recognition, pages 1933–1941. IEEE, 2016.

[64] Daniel J Felleman and DC Essen Van. Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex, 1(1):1–47, 1991.

[65] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. International Journal of Computer Vision, 61(1):55–79, 2005.

[66] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In Computer Vision and Pattern Recognition. IEEE, 2018.

[67] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In Computer Vision and Pattern Recognition, pages 3681–3690. IEEE, 2017.

[68] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning, pages 1126–1135, 2017.

[69] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981.

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

[71] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In Computer Vision and Pattern Recognition, pages 2385–2392. IEEE, 2014.

[72] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In Computer Vision and Pattern Recognition, pages 350–359. IEEE, 2018.

[73] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Neural Information Processing Systems, pages 2672–2680, 2014.

[74] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

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

[76] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. Transactions on Neural Networks and Learning Systems, 28(10):2222–2232, 2017.

[77] Ralph Gross, Iain Matthews, and Simon Baker. Generic vs. person specific active appearance models. Image and Vision Computing, 23(12):1080–1093, 2005.

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

[79] J. Gu, X. Yang, S. D. Mello, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In Computer Vision and Pattern Recognition, pages 1531–1540. IEEE, July 2017.

[80] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. 2018.

[81] Minghao Guo, Jiwen Lu, and Jie Zhou. Dual-agent deep reinforcement learning for deformable face tracking. In European Conference on Computer Vision, pages 768–783. Springer, 2018.

[82] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885, 2016.

[83] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Computer Vision and Pattern Recognition, pages 6546–6555. IEEE, 2018.

[84] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In Alvey Vision Conference, volume 15, pages 10–5244, 1988.

[85] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In Computer Vision and Pattern Recognition, pages 4295–4304. IEEE, 2015.

[86] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Computer Vision and Pattern Recognition, pages 770–778. IEEE, 2016.

[87] Zhenliang He, Meina Kan, Jie Zhang, Xilin Chen, and Shiguang Shan. A fully end-to-end cascaded cnn for facial landmark detection. In Automatic Face and Gesture Recognition, pages 200–207. IEEE, 2017.

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

[89] Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, and Yihong Gong. Face alignment recurrent network. Pattern Recognition, 74:448–458, 2018.

[90] Gee-Sern Hsu, Kai-Hsiang Chang, and Shih-Chieh Huang. Regressive tree structured model for facial landmark localization. In International Conference on Computer Vision, pages 3855–3861. IEEE, 2015.

[91] Yuchi Huang, Qingshan Liu, and Dimitris Metaxas. A component based deformable model for generalized face alignment. In International Conference on Computer Vision, pages 1–8. IEEE, 2007.

[92] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160(1):106–154, 1962.

[93] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360, 2016.

[94] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In Asian Conference on Computer Vision, pages 302–315. Springer, 2014.

[95] Mihir Jain, Herve Jegou, and Patrick Bouthemy. Better exploiting motion for better action recognition. In Computer Vision and Pattern Recognition, pages 2555–2562. IEEE, 2013.

[96] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[97] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. Transactions on Pattern Analysis and Machine Intelligence, 35(1):221–231, 2013.

[98] Xin Jin and Xiaoyang Tan. Face alignment by robust discriminative hough voting. Pattern Recognition, 60:318–333, 2016.

[99] Xin Jin and Xiaoyang Tan. Face alignment in-the-wild: A survey. Computer Vision and Image Understanding, 162:1–22, 2017.

[100] Amin Jourabloo and Xiaoming Liu. Pose-invariant face alignment via cnn-based dense 3d model fitting. International Journal of Computer Vision, 124(2):187–203, 2017.

[101] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Pose-invariant face alignment with a single cnn. In International Conference on Computer Vision, pages 3200–3209. IEEE, 2017.

[102] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In International Conference on Machine Learning, pages 2342–2350, 2015.

[103] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078, 2015.

[104] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Computer Vision and Pattern Recognition, pages 1725–1732. IEEE, 2014.

[105] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In Computer Vision and Pattern Recognition, pages 1867–1874. IEEE, 2014.

[106] Muhammad Haris Khan, John McDonagh, and Georgios Tzimiropoulos. Synergy between face alignment and tracking via discriminative global consensus optimization. In International Conference on Computer Vision, pages 3811–3819. IEEE, 2017.

[107] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[108] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In British Machine Vision Conference, pages 275–1. British Machine Vision Association, 2008.

[109] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In International Conference on Computer Vision Workshops, pages 2144–2151. IEEE, 2011.

[110] Jean Kossaifi, Linh Tran, Yannis Panagakis, and Maja Pantic. Gagan: Geometry-aware generative adversarial networks. In Computer Vision and Pattern Recognition, pages 878–887. IEEE, 2018.

[111] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In Computer Vision and Pattern Recognition Workshops, volume 3, page 6. IEEE, 2017.

[112] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Neural Information Processing Systems, pages 1097–1105, 2012.

[113] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In International Conference on Computer Vision, pages 2556–2563. IEEE, 2011.

[114] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In Computer Vision and Pattern Recognition, pages 430–439. IEEE, 2018.

[115] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[116] Hanjiang Lai, Shengtao Xiao, Yan Pan, Zhen Cui, Jiashi Feng, Chunyan Xu, Jian Yin, and Shuicheng Yan. Deep recurrent regression for facial landmark detection. Transactions on Circuits and Systems for Video Technology, 28(5):1144–1157, 2018.

[117] Ivan Laptev. On space-time interest points. International Journal of Computer Vision, 64(2-3):107–123, 2005.

[118] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008.

[119] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. Transactions on Neural Networks, 8(1):98–113, 1997.

[120] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. 2011.

[121] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In European Conference on Computer Vision, pages 679–692. Springer, 2012.

[122] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436, 2015.

[123] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In Neural Information Processing Systems, pages 396–404, 1990.

[124] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[125] Donghoon Lee, Hyunsin Park, and Chang D Yoo. Face alignment using cascade gaussian process regression trees. In Computer Vision and Pattern Recognition, pages 4204–4212. IEEE, 2015.

[126] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Unsupervised learning of hierarchical representations with convolutional deep belief networks. Communications of the ACM, 54(10):95–103, 2011.

[127] Christy Yuan Li, Tadas Baltrušaitis, and Louis-Philippe Morency. Constrained ensemble initialization for facial landmark tracking in video. In Automatic Face and Gesture Recognition, pages 697–704. IEEE, 2017.

[128] Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. Understanding the disharmony between dropout and batch normalization by variance shift. arXiv preprint arXiv:1801.05134, 2018.

[129] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In European Conference on Computer Vision, pages 600–615. Springer, 2018.

[130] Yongqiang Li, Shangfei Wang, Yongping Zhao, and Qiang Ji. Simultaneous facial feature tracking and facial expression recognition. Transactions on Image Processing, 22(7):2559–2573, 2013.

[131] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding, 166:41–50, 2018.

[132] Michael L Littman. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521(7553):445, 2015.

[133] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Two-stream transformer networks for video-based face alignment. Transactions on Pattern Analysis and Machine Intelligence, 40(11):2546–2554, 2018.

[134] Hao Liu, Jiwen Lu, Minghao Guo, Suping Wu, and Jie Zhou. Learning reasoning-decision networks for robust face alignment. Transactions on Pattern Analysis and Machine Intelligence, 2018.

[135] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Computer Vision and Pattern Recognition, pages 3431–3440. IEEE, 2015.

[136] David G Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004.

[137] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[138] Simon Lucey, Yang Wang, Mark Cox, Sridha Sridharan, and Jeffery F Cohn. Efficient constrained local model fitting for non-rigid face alignment. Image and Vision Computing, 27(12):1804–1813, 2009.

[139] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In Computer Vision and Pattern Recognition. IEEE, 2017.

[140] Milner AD MA. Separate visual pathways for perception and action. Trends in Neurosciences, 15(1):2025Grave, 1992.

[141] Elman Mansimov, Nitish Srivastava, and Ruslan Salakhutdinov. Initialization strategies of spatio-temporal convolutional neural networks. arXiv preprint arXiv:1503.07274, 2015.

[142] Brais Martinez, Michel F Valstar, Xavier Binefa, and Maja Pantic. Local evidence aggregation for regression-based facial point detection. Transactions on Pattern Analysis and Machine Intelligence, 35(5):1149–1163, 2013.

[143] Pedro Martins, Rui Caseiro, and Jorge Batista. Non-parametric bayesian constrained local models. In Computer Vision and Pattern Recognition, pages 1797–1804. IEEE, 2014.

[144] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In International Conference on Computer Vision Workshops, pages 514–521. IEEE, 2009.

[145] Iain Matthews and Simon Baker. Active appearance models revisited. International Journal of Computer Vision, 60(2):135–164, 2004.

[146] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In Computer Vision and Pattern Recognition, pages 781–790. IEEE, 2018.

[147] Ross Messing, Chris Pal, and Henry Kautz. Activity recognition using the velocity histories of tracked keypoints. In International Conference on Computer Vision, pages 104–111. IEEE, 2009.

[148] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In Computer Vision and Pattern Recognition, pages 5040–5049. IEEE, 2018.

[149] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In Computer Vision and Pattern Recognition, pages 4207–4215. IEEE, 2016.

[150] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In International Conference on Machine Learning, pages 807–814, 2010.

[151] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. In SIGGRAPH, 2018.

[152] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In European Conference on Computer Vision, pages 483–499. Springer, 2016.

[153] Andre T Nguyen and Edward Raff. Adversarial attacks, regression, and numerical stability regularization. In AAAI Conference on Artificial Intelligence Workshops, 2019.

[154] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. Distill, 2016.

[155] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. Distill, 2017. https://distill.pub/2017/feature-visualization.

[156] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In Computer Vision and Pattern Recognition, pages 4594–4602. IEEE, 2016.

[157] George Papandreou and Petros Maragos. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008.

[158] Dennis Park, C Lawrence Zitnick, Deva Ramanan, and Piotr Dollár. Exploring weak stabilization for motion feature extraction. In Computer Vision and Pattern Recognition, pages 2882–2889. IEEE, 2013.

[159] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026, 2013.

[160] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In International Conference on Machine Learning, pages 1310–1318, 2013.

[161] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572, 1901.

[162] Xi Peng, Rogerio S Feris, Xiaoyu Wang, and Dimitris N Metaxas. A recurrent encoder-decoder network for sequential face alignment. In European Conference on Computer Vision, pages 38–56. Springer, 2016.

[163] Xi Peng, Rogerio S Feris, Xiaoyu Wang, and Dimitris N Metaxas. Red-net: A recurrent encoder–decoder network for video-based face alignment. International Journal of Computer Vision, pages 1–17, 2018.

[164] Xi Peng, Qiong Hu, Junzhou Huang, and Dimitris N Metaxas. Track facial points in unconstrained videos. arXiv preprint arXiv:1609.02825, 2016.

[165] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In Computer Vision and Pattern Recognition, pages 2226–2234. IEEE, 2018.

[166] Xi Peng, Shaoting Zhang, Yu Yang, and Dimitris N Metaxas. Piefa: Personalized incremental and ensemble face alignment. In International Conference on Computer Vision, pages 3880–3888. IEEE, 2015.

[167] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding, 150:109–125, 2016.

[168] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In European Conference on Computer Vision, pages 581–595. Springer, 2014.

[169] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In European Conference on Computer Vision, pages 143–156. Springer, 2010.

[170] AJ Piergiovanni, Anelia Angelova, Alexander Toshev, and Michael S Ryoo. Evolving space-time neural architectures for videos. arXiv preprint arXiv:1811.10636, 2018.

[171] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large-Margin Classifiers, 10(3):61–74, 1999.

[172] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, 89(4):344–350, 2001.

[173] Utsav Prabhu, Keshav Seshadri, and Marios Savvides. Automatic facial landmark tracking in video sequences using kalman filter assisted active shape models. In European Conference on Computer Vision, pages 86–99. Springer, 2010.

[174] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In International Conference on Computer Vision, pages 5534–5542. IEEE, 2017.

[175] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In Automatic Face and Gesture Recognition, pages 17–24. IEEE, 2017.

[176] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In Computer Vision and Pattern Recognition, pages 6894–6903. IEEE, 2017.

[177] Bhargava Reddy, Ye-Hoon Kim, Sojung Yun, Chanwon Seo, and Junik Jang. Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. In Computer Vision and Pattern Recognition Workshops, pages 121–128, 2017.

[178] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In Computer Vision and Pattern Recognition, pages 1685–1692. IEEE, 2014.

[179] Herbert Robbins and Sutton Monro. A stochastic approximation method. The Annals of Mathematical Statistics, pages 400–407, 1951.

[180] F Rosenblatt. The perception: a probabilistic model for information storage and organization in the brain, neurocomputing: foundations of research, 1988.

[181] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[182] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. Cognitive Modeling, 5(3):1, 1988.

[183] James A Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161, 1980.

[184] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. Image and Vision Computing, 47:3–18, 2016.

[185] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In International Conference on Computer Vision Workshops, pages 397–403. IEEE, 2013.

[186] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, and Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. In European Conference on Computer Vision, pages 645–661. Springer, 2016.

[187] Jason Saragih and Roland Goecke. A nonlinear discriminative approach to aam fitting. In International Conference on Computer Vision, pages 1–8. IEEE, 2007.

[188] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision, 91(2):200–215, 2011.

[189] Konrad Schindler and Luc Van Gool. Action snippets: How many frames does human action recognition require? 2008.

[190] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In ACM International Conference on Multimedia, pages 357–360. ACM, 2007.

[191] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In European Conference on Computer Vision, pages 705–720. Springer, 2018.

[192] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In ACM SIGSAC Conference on Computer and Communications Security, pages 1528–1540. ACM, 2016.

[193] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In International Conference on Computer Vision Workshops, pages 50–58. IEEE, 2015.

[194] Jongju Shin and Daijin Kim. Robust face alignment and tracking by combining local search and global fitting. Image and Vision Computing, 51:69–83, 2016.

[195] Bernard W Silverman. Density estimation for statistics and data analysis. Routledge, 2018.

[196] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Neural Information Processing Systems, pages 568–576, 2014.

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

[198] Brandon M Smith, Jonathan Brandt, Zhe Lin, and Li Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In Computer Vision and Pattern Recognition, pages 1741–1748. IEEE, 2014.

[199] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.

[200] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In International Conference on Machine Learning, pages 843–852, 2015.

[201] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. Transactions on Evolutionary Computation, 2019.

[202] Ju Sun, Yadong Mu, Shuicheng Yan, and Loong-Fah Cheong. Activity recognition using dense long-duration trajectories. In International Conference on Multimedia and Expo, pages 322–327. IEEE, 2010.

[203] Ju Sun, Xiao Wu, Shuicheng Yan, Loong-Fah Cheong, Tat-Seng Chua, and Jintao Li. Hierarchical spatio-temporal context modeling for action recognition. In Computer Vision and Pattern Recognition, pages 2004–2011. IEEE, 2009.

[204] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In International Conference on Computer Vision, pages 4597–4605. IEEE, 2015.

[205] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In Computer Vision and Pattern Recognition, pages 3476–3483. IEEE, 2013.

[206] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Computer Vision and Pattern Recognition, pages 1–9. IEEE, 2015.

[207] Ying* Tai, Yicong* Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In AAAI Conference on Artificial Intelligence, 2019.

[208] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In European Conference on Computer Vision, pages 140–153. Springer, 2010.

[209] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In International Conference on Computer Vision, pages 4481–4490. IEEE, 2017.

[210] Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji. Robust facial feature tracking under varying face pose and facial expression. Pattern Recognition, 40(11):3195–3208, 2007.

[211] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In International Conference on Computer Vision, pages 4489–4497. IEEE, 2015.

[212] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038, 2017.

[213] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In Computer Vision and Pattern Recognition, pages 6450–6459. IEEE, 2018.

[214] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In Computer Vision and Pattern Recognition Workshops, pages 969–977. IEEE, 2018.

[215] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In Computer Vision and Pattern Recognition, pages 4177–4187. IEEE, 2016.

[216] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In Computer Vision and Pattern Recognition, pages 3659–3667. IEEE, 2015.

[217] Georgios Tzimiropoulos, Joan Alabort-i Medina, Stefanos Zafeiriou, and Maja Pantic. Generic active appearance models revisited. In Asian Conference on Computer Vision, pages 650–663. Springer, 2012.

[218] Georgios Tzimiropoulos and Maja Pantic. Optimization problems for fast aam fitting in-the-wild. In International Conference on Computer Vision, pages 593–600. IEEE, 2013.

[219] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In Computer Vision and Pattern Recognition, pages 1851–1858. IEEE, 2014.

[220] Hirofumi Uemura, Seiji Ishikawa, and Krystian Mikolajczyk. Feature tracking and motion compensation for action recognition. In British Machine Vision Conference, pages 1–10, 2008.

[221] M Uricar and V Franc. Real-time facial landmark tracking by tree-based deformable part model based detector. In International Conference on Computer Vision Workshops. IEEE, 2015.

[222] Michal Uřičář, Vojtěch Franc, and Václav Hlaváč. Detector of facial landmarks learned by the structured output svm. Computer Vision Theory and Applications, 12:547–556, 2012.

[223] Michel Valstar, Brais Martinez, Xavier Binefa, and Maja Pantic. Facial point detection using boosted regression and graph models. In Computer Vision and Pattern Recognition, pages 2729–2736. IEEE, 2010.

[224] Heng Wang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin. Action recognition by dense trajectories. In Computer Vision and Pattern Recognition, pages 3169–3176. IEEE, 2011.

[225] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103(1):60–79, 2013.

[226] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In International Conference on Computer Vision, pages 3551–3558. IEEE, 2013.

[227] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In British Machine Vision Conference, pages 124–1. BMVA Press, 2009.

[228] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In Computer Vision and Pattern Recognition, pages 4305–4314. IEEE, 2015.

[229] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision, pages 20–36. Springer, 2016.

[230] Wei Wang, Sergey Tulyakov, and Nicu Sebe. Recurrent convolutional shape regression. Transactions on Pattern Analysis and Machine Intelligence, (1):1–1, 2018.

[231] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions˜ transformations. In Computer Vision and Pattern Recognition, pages 2658–2667. IEEE, 2016.

[232] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In European Conference on Computer Vision, pages 650–663. Springer, 2008.

[233] Shu-Fai Wong and Roberto Cipolla. Extracting spatiotemporal interest points using global information. In International Conference on Computer Vision, pages 1–8. IEEE, 2007.

[234] Shandong Wu, Omar Oreifej, and Mubarak Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In International Conference on Computer Vision, pages 1419–1426. IEEE, 2011.

[235] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In Computer Vision and Pattern Recognition, pages 2129–2138. IEEE, 2018.

[236] Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In Computer Vision and Pattern Recognition Workshops, pages 150–159. IEEE, 2017.

[237] Yue Wu and Qiang Ji. Discriminative deep face shape model for facial point detection. International Journal of Computer Vision, 113(1):37–53, 2015.

[238] Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion. In International Conference on Computer Vision, pages 3658–3666. IEEE, 2015.

[239] Yue Wu and Qiang Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In Computer Vision and Pattern Recognition, pages 3400–3408. IEEE, 2016.

[240] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. International Journal of Computer Vision, pages 1–28, 2018.

[241] Yue Wu, Ziheng Wang, and Qiang Ji. A hierarchical probabilistic model for facial feature detection. In Computer Vision and Pattern Recognition, pages 1781–1788. IEEE, 2014.

[242] Yuxiang Wu, Zehua Cheng, Bin Huang, Yiming Chen, Kele Xu, and Weiyang Wang. Foxnet: A multi-face alignment method. arXiv preprint arXiv:1904.09758, 2019.

[243] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition, 2019.

[244] Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, and Ashraf Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In International Conference on Computer Vision, pages 1633–1642. IEEE, 2017.

[245] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In European Conference on Computer Vision, pages 57–72. Springer, 2016.

[246] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In European Conference on Computer Vision, pages 305–321. Springer, 2018.

[247] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Neural Information Processing Systems, pages 802–810, 2015.

[248] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In Computer Vision and Pattern Recognition, pages 532–539. IEEE, 2013.

[249] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In Computer Vision and Pattern Recognition, pages 2664–2673. IEEE, 2015.

[250] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In International Conference on Computer Vision, pages 5783–5792. IEEE, 2017.

[251] Xiang Xu and Ioannis A Kakadiaris. Joint head pose estimation and face alignment framework using global and local cnn features. In Automatic Face and Gesture Recognition, pages 642–649. IEEE, 2017.

[252] Junjie Yan, Zhen Lei, Dong Yi, and Stan Li. Learn to combine multiple hypotheses for accurate face alignment. In International Conference on Computer Vision Workshops, pages 392–396. IEEE, 2013.

[253] Shuicheng Yan, Xinwen Hou, Stan Z Li, Hongjiang Zhang, and Qiansheng Cheng. Face alignment using view-based direct appearance models. International Journal of Imaging Systems and Technology, 13(1):106–112, 2003.

[254] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI Conference on Artificial Intelligence, 2018.

[255] Heng Yang, Xuming He, Xuhui Jia, and Ioannis Patras. Robust face alignment under occlusion via regional predictive power estimation. Transactions on Image Processing, 24(8):2393–2403, 2015.

[256] Heng Yang and Ioannis Patras. Privileged information-based conditional regression forest for facial feature detection. In Automatic Face and Gesture Recognition, pages 1–6. IEEE, 2013.

[257] Heng Yang and Ioannis Patras. Sieving regression forest votes for facial feature detection in the wild. In International Conference on Computer Vision, pages 1936–1943. IEEE, 2013.

[258] Jing Yang, Jiankang Deng, Kaihua Zhang, and Qingshan Liu. Facial shape tracking via spatio-temporal cascade shape regression. In International Conference on Computer Vision Workshops, pages 41–49. IEEE, 2015.

[259] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In Computer Vision and Pattern Recognition Workshops, pages 79–87. IEEE, 2017.

[260] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In Computer Vision and Pattern Recognition. IEEE, 2016.

[261] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Multilayer and multimodal fusion of deep neural networks for video classification. In ACM International Conference on Multimedia, pages 978–987. ACM, 2016.

[262] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Making convolutional networks recurrent for visual sequence learning. In Computer Vision and Pattern Recognition, pages 6469–6478. IEEE, 2018.

[263] Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In International Conference on Computer Vision, pages 492–497. IEEE, 2009.

[264] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In International Conference on Computer Vision, pages 1944–1951. IEEE, 2013.

[265] Xiang Yu, Zhe Lin, Jonathan Brandt, and Dimitris N Metaxas. Consensus of regression for occlusion-robust facial feature localization. In European Conference on Computer Vision, pages 105–118. Springer, 2014.

[266] Xiaowei Yuan and In Kyu Park. Face de-occlusion using 3D morphable model and generative adversarial network. arXiv preprint arXiv:1904.06109, 2019.

[267] Lei Yue, Xin Miao, Pengbo Wang, Baochang Zhang, Xiantong Zhen, and Xianbin Cao. Attentional alignment networks. In British Machine Vision Conference, 2018.

[268] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In Computer Vision and Pattern Recognition, pages 4694–4702. IEEE, 2015.

[269] Kevan Yuen and Mohan M Trivedi. An occluded stacked hourglass approach to facial landmark localization and occlusion estimation. Transactions on Intelligent Vehicles, 2(4):321–331, 2017.

[270] Stefanos Zafeiriou, Grigorios G Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. The 3d menpo facial landmark tracking challenge. In International Conference on Computer Vision, pages 2503–2511, 2017.

[271] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In Computer Vision and Pattern Recognition Workshops, pages 2116–2125. IEEE, 2017.

[272] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: past, present and future. Computer Vision and Image Understanding, 138:1–24, 2015.

[273] Jiajian Zeng, Siyuan Liu, Xi Li, Debbah Abderrahmane Mahdi, Fei Wu, and Gang Wang. Deep context-sensitive facial landmark detection with tree-structured modeling. Transactions on Image Processing, 27(5):2096–2107, 2018.

[274] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In European Conference on Computer Vision, pages 1–16. Springer, 2014.

[275] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. Signal Processing Letters, 23(10):1499–1503, 2016.

[276] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In International Conference on Computer Vision, pages 3120–3128. IEEE, 2017.

[277] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In International Conference on Biometrics, pages 174–181. IEEE, 2018.

[278] Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In Computer Vision and Pattern Recognition, pages 718–726. IEEE, 2017.

[279] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In European Conference on Computer Vision, pages 94–108. Springer, 2014.

[280] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. Transactions on Pattern Analysis and Machine Intelligence, 38(5):918–930, 2016.

[281] Xiaowei Zhao, Tae-Kyun Kim, and Wenhan Luo. Unified face analysis by iterative multi-output random forests. In Computer Vision and Pattern Recognition, pages 1765–1772. IEEE, 2014.

[282] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In International Conference on Computer Vision Workshops, pages 386–391. IEEE, 2013.

[283] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In Computer Vision and Pattern Recognition, pages 4747–4756. IEEE, 2017.

[284] Hongyuan Zhu, Romain Vial, and Shijian Lu. Tornado: A spatio-temporal convolutional regression network for video action proposal. In International Conference on Computer Vision, pages 5813–5821. IEEE, 2017.

[285] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In Computer Vision and Pattern Recognition, pages 4998–5006. IEEE, 2015.

[286] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In Computer Vision and Pattern Recognition, pages 2879–2886. IEEE, 2012.

[287] Xiangyu Zhu, Zhen Lei, Stan Z Li, et al. Face alignment in full pose range: A 3D total solution. Transactions on Pattern Analysis and Machine Intelligence, 2017.

[288] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In Computer Vision and Pattern Recognition, pages 146–155. IEEE, 2016.

[289] Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. Transactions on Affective Computing, 9(4):578–584, 2017.