Analysis and Improvement of Facial Landmark Detection

Philipp Kopp

Semester Project Report
February 2019

Dr. Derek Bradley
Dr. Thabo Beeler
Prof. Dr. Markus Gross

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Disney Research

cgl
computer graphics laboratory

Abstract

Facial landmark points capture rigid and non-rigid deformation of faces in a very compact description and are therefore valuable for many different face analysis tasks. For face recognition or different categorisation tasks such as gender, age, ethnicity or expressions, a rough pose normalisation is needed in order to apply other algorithms. For 3D face tracking or reconstruction facial landmarks are often used as initialisation for more sophisticated approaches. As input, landmark detectors use the image itself and a face box provided by a previous detector. In the field these consecutive detection steps are often looked at separately. One focus of this work is to evaluate the entire landmarking pipeline including face detectors. We evaluate how precision, robustness and temporal stability of landmark detectors change using face boxes provided by different face detectors. Temporal stability of landmarks is an issue becoming increasingly important with more applications shifting from single images to videos. There are only a few datasets with ground truth landmarks for videos. We annotate a sequence from Disney research particularly for this type of evaluation and propose a method to stabilise existing video data sets. Landmarking methods are tested exhaustively in several experiments and compared regarding their performance.

Contents

0

Acronyms

FB	Face box
LM	Landmark
IOD	Inter Ocular Distance
CNN	Convolutional Neural Network
HOG	Histogram of Oriented Gradients
GT	Ground truth
AAM	Active Appearance Models
DR	Disney Research
iBUG	Intelligent Behaviour Understanding Group, Imperial College London
OpenCV	Open Source Computer Vision Library
Dlib	A toolkit C++ Library
FAN	Face Alignment Network
SBR	Supervision-by-Registration
NME	Normalised Mean Error
LSE	Landmark Stability Error
CED	Cumulative Error Distribution

1

Introduction

Faces capture a lot of information used in face analysis, 3D face reconstruction and face recognition. These methods are used in various applications such as VR and AR and biometrics. Nevertheless faces can change appearance in various ways such as pose and expression. It is therefore essential to align faces and reduce complexity for following analysis. Facial landmarks can capture pose and some expression information in a sparse representation.

Facial landmarking algorithms are prevalent in most face analysis pipelines. It is also one of the first steps in such a pipe line, thus even small errors can have large effects on consecutive steps. Accuracy and temporal stability are therefore of high importance.

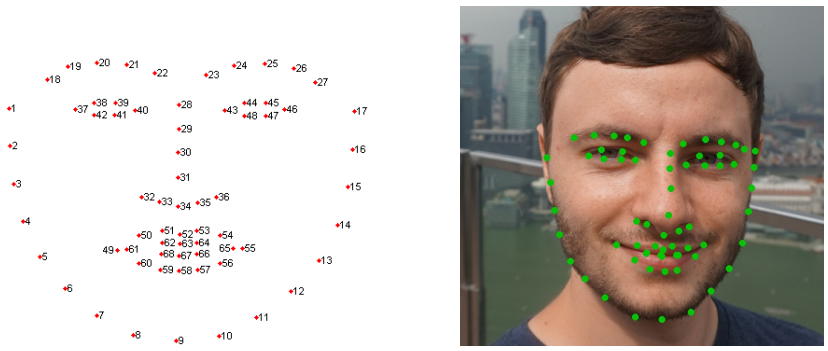

Figure 1.1.: *The 68 landmark locations defined by [GMC+10].* (left)*: schematic* (right)*: exemplar image with landmarks.*

Most landmarking methods need a face box to initialise with or crop the image to. In literature the robustness of landmarking methods to inaccurate face boxes or face boxes with different definitions is not evaluated yet.

With the rise of cameras in mobile phones and digital single lens reflex cameras being able to

shoot videos, the importance of landmarking on videos has become more important. For videos not only accuracy is important but also temporal stability.

Challenges beyond temporal stability in facial landmarking are large poses, occlusions, strong expressions, and rare lighting situations. Figure 1.2 shows some of these difficult cases. Strong light from one side can cause shadows or even totally unlit regions. Large poses are mainly a problem of missing training data. Most images of humans are taken from a frontal view. Strong expressions are also challenging for algorithms as the deformations of face regions can be extreme. Occlusions are another difficulty for landmarking methods. Glasses and hands are the most common occlusions.

According to Wu et al. [WJ18] and Feng et al. [FK18] landmarking methods can be roughly categorised into three classes. Holistic methods, constrained local model methods, and the regression-based methods. These methods were introduced historically in this order and will be briefly explained in the following.

Holistic methods incorporate the face shape into the model. The most famous algorithm of this type are Active Appearance Models (AAM) introduced by Cootes et al. [CET01]. AAMs build a model of how faces look from different angles and can therefor estimate the landmark positions after model fitting. Drawback of this method is the limited variability of the model and the consequential low performance on in-the-wild imagery.

Constrained Local Model methods consist of two parts. The landmark specific appearance variations are learned by a local model. These models usually output a probability map of the specific landmark they look for. The overall landmarking result is then computed by a second model capturing common face shapes and patterns between the landmarks. The final landmark positions are chosen in a way that maximises both the appearance probabilities of each of the local landmark specific models as well as the overall shape model. These types of models were first proposed by Cristinacce et al. [CC06] and have been extended in various ways. Among them [BRM13] and more recently by using CNNs as local experts [ZBM16].

The third group of approaches are regression based methods. Regression based methods are the most recent group of methods and estimate the landmark coordinates based on the face appearance in the image without building an explicit face shape model. They can be divided into approaches that directly regress the landmark positions from the input image, or cascaded approaches that need initialisation and then regress an update of the landmark positions based on the image. One of the first cascaded methods introduced was Cascaded Pose Regression (CPR) by Dollar et al. [DWP10], followed by Robust Cascaded Pose Regression (RCPR) [BAPD13] and [KS14]. In recent years, deep convolutional networks have gained popularity and improved robustness regarding the challenges shown in Figure 1.2. CNNs can be categorised as regression based models, as in most cases no explicit shape model is used. There is a variety of different architectures, while the hourglass structure as in [YLZ17] shows promising results. There are two major types of encoding of the landmark positions and loss functions associated with that. One option is to use heat maps in the dimensions of the input image. Each landmark is encoded in a separate heat map. A second option is to directly regress the landmark coordinates with networks that have fixed input size.

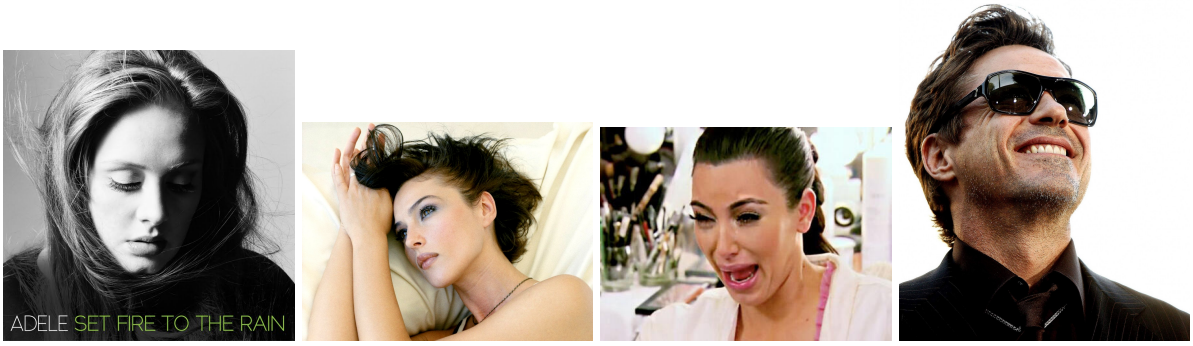

Figure 1.2.: *Examples of difficulties in landmarking: Strong lighting, pose, expressions and occlusions are the most common challenges. (Images from IBUG dataset [STZP13])*

We present three contributions to the field:

1. An overview of latest methods and implementations of landmarking.
2. Several experiments on landmark accuracy using face detectors for initialisation.
3. Analysis of the temporal stability of different landmarking methods.

The remaining report is structured as follows. We first present an overview of current face detection and landmarking methods in Chapter 2. Then we present the evaluation methods with which we want to compare the detectors in Chapter 3. In Chapter 4 we evaluate the combinations of face detectors and landmarking, followed by the Conclusion in Chapter 5.

2

Detectors

This chapter presents implementations of face and landmark detectors. This list does not claim to be a complete list of publicly available detectors. Nevertheless, it covers the big, well known libraries and the recently introduced methods. A library that is a focus of this evaluation and is also heavily used in industry is the dlib c++ library [Kin]. It comes with a boost software license and is therefore also usable free of charge in commercial applications.

Most landmark detectors need a bounding box around the face to either initialise the algorithm or to crop the image. Thus before applying a landmark detector in practice a face detector is needed to localise faces in the image. In research these two problems are looked at separately. When measuring performance of landmark detectors a ground truth face box, computed based on the ground truth landmarks, is used as initialisation. As we focus on practical performance of different landmark detectors we want to test them in combination with a face detector. This way we can measure performance in actual testing conditions. The following gives an overview of open source implementations of different face detection algorithms.

2.1. Face Detectors

Dlib hog face detector
The classic dlib face detector uses Histogram of Oriented Gradients (HOG) features [DT05] with a linear classifier. With an image pyramid and the sliding window scheme it covers the entire image while being computationally cheap. King introduces the Max-Margin Object Detection (MMOD) as cost function [Kin15]. The first version of this method was published in dlib in early 2014.

Dlib CNN face detector
The rise of deep learning has also introduced new methods of detecting faces. Dlib comes with its own CNN deep face detector. As loss function a variant of the MMOD objective is used. The implementation and the model were published in dlib in late 2016. Both dlib detectors output discrete square bounding boxes.

Multi-task Cascaded Neural Network (MTCNN)
In object detection one wants to detect the object as precisely as possible. Thus the resulting bounding box should cover all parts of the object while being tightly aligned to it. This method introduced by Zhang et al. [ZZLQ16] uses a cascade of 3 neural networks. The first network proposes areas where faces might be, applying a fully convolutional network to different image scales. The second network is called refine network that rejects face box candidates. The third network (output net) rejects further face box candidates but also regresses the final bounding box. As this last network aligns to the face area it can also output 5 landmarks. Between each of the networks non-maximum suppression is applied to filter similar face boxes.

Face detection with Faster R-CNN
The Faster R-CNN is a famous object detection method introduced by [overall method]. Jing et al. [JLM17] adapted the method to face detection. The method consists of two steps. First a fully convolutional network, called Regional Proposal Network (RPN), classifies each pixel as part of a face or not. This heat map is then fed into the second network, which refines the proposals and delivers the final detection score.

Face detection with deep neural networks in openCV
The face detector in the new dnn module of openCV uses the single shot detector architecture (SSD). The SSD method was introduced by Liu et al. [LAE+16] and builds on top of the VGG-16 architecture. This method was among the first to directly output bounding boxes of objects without any significant post processing.

Summary
As the main aspect of this work is the landmark detection, we focus on two methods, the HOG and CNN detector implemented in dlib. As third option the MTCNN implementation by David Sandberg would have been chosen, because of it's focus on face alignment and the additional landmarks it outputs.

2.2. Landmark Detectors

Following the face detection in a face analysis pipeline the landmark detector is applied. The following list gives an overview of different landmarking methods.

Dlib implementation of regression trees
The method of regression trees for landmark detection was introduced by Kazemi et al. [KS14]. The algorithm is a cascade of linear regressors where each regressor returns an update of the

current estimate of landmark positions. The initial position estimates are computed based on a given face box. This method requires comparably little training data and can be trained in several hours on a regular pc. During testing time it is very fast in the order of a millisecond.

Face Alignment Network (FAN)
The face alignment network introduced by Bulat et al. [BT17] consists of four stacked hour glass networks. The underlying hour glass structure is described in more detail in [NYD16]. These networks output a heatmap on which a mean squared error loss is applied comparing to a ground truth heat map. A python implementation of FAN is provided by the authors. Also different pretrained networks are available.

Supervision-by-Registration (SBR)
This method by Dong et al. [DYW+18] from facebook research uses weakly supervised methods to be able to use more training data. In particular a normal detection loss, comparing the predictions with ground truth labels in a supervised manner, is extended by a registration loss. This registration loss can be computed based on the Lucas Kanade [BM04] tracking on a video in an unsupervised manner. Therefore unlabeled videos can be used to extend the training set. The authors published their training and inference code.

Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks
Published by Feng et al. [FKA+18] this method focuses on the encoding of landmark positions in neural networks and the associated loss function. The authors propose to use a loss function penalising small errors more than in other approaches. This method regresses the landmark positions directly, instead of using heat maps as other approaches presented in this list. Open source code for applying models was published during the time of this work.

OpenFace
OpenFace is an open source project focusing on head pose estimation. It implements the methods of Baltrusaitis et al. [BRM13] and Zadeh et al. [ZBM16], although the later is an extension of the first. The implementation uses CNNs on local face patches as landmark specific detection experts. Each of these experts propose landmark positions and output their reliability. Using an iterative method called regularized landmark mean shift (RLMS) the final landmark positions are calculated considering the spacial dependencies of facial landmarks.

Look at Boundary: A Boundary-Aware Face Alignment Algorithm
In this work Wu et al. [WQY+18] propose to use facial boundaries as an intermediate computation step. They argue that most facial landmarks are defined to lie on facial boundaries and that providing a boundary heatmap improves performance of a landmarks regressor. The authors provide an open source implementation.

Cascade Multi-view Hourglass Model for Robust 3D Face Alignment
As many other algorithms Deng et al. use hourglass networks [DZCZ18]. There are two networks, where the first network predicts 2D positions used to estimate the rigid face pose. After eliminating the pose a second network predicts the final 3D landmark positions. The authors published there code as well as some models.

Robust Facial Landmark Detection via a Fully-Convolutional Local-Global Context Network

The novelty of this approach is that it does not depend on face detection. Merget et al. [MRR18] apply a CNN in a fully convolutional manner over the entire image. To regress the exact faces and landmark positions a PCA model is fed with the heat map output of the CNN. All models and code are provided publicly by the authors.

Summary

There is an astonishing variety of different landmarking methods. There are major trends identifiable such as CNNs in general and hour glass networks in particular. As one example of this group of methods FAN was chosen to evaluate against the dlib implementation of regression trees. Two further networks were compared. The SBR method by facebook research, using an unsupervised training procedure, and the wingloss network. The dlib regression trees is the oldest among the compared methods and the only method not based on deep learning. The four other algorithms were introduced more recently and are examples of recent developments in the field of facial landmarking.

3

Evaluation Methodology

This chapter describes the methodology used to compare the different combinations of face and landmark detectors. Section 3.1 deals with different landmark sets, advantages and disadvantages of these and the sets used for the following experiments. As a public sequence for evaluation a 300VW video was chosen and ground truth landmarks improved. The steps undertaken are presented in Section 3.2. The third section of this chapter 3.3 describes how a precise ground truth base line was created on a Disney Research internal video sequence. Finally section 3.4 proposes a method of measuring temporal stability of landmarking methods.

3.1. Set of Landmarks

There is no common set of landmarks used across databases. This is also due to the different applications landmarks can be used for. In general one can differ between 2D and 3D landmarks. Most landmarking methods, especially in the past predicted 2D landmarks. This was mainly due to the fact, that there were no data sets that contained 3D landmarks. The second main landmark set criterion is the number of landmarks. There are definitions with 5 landmarks covering the eyes, nose and mouth only [ZZLQ16]. And other sets with up to 194 Landmarks, such as the HELEN data set [LBL$^+$12].

Most definitions include key points of the face such as the eye corners. These landmarks are the inner face landmarks. For many applications it is also important to record the overall face size. Landmarks on the face silhouette can be one way of capturing the face size more precisely. Figure 3.2 shows sets of 68 2D landmarks visualised on a mean head from different poses. Green landmarks are 51 inner landmarks. The blue landmarks are 17 outer landmarks. For frontal pose images the outer landmarks are located at the silhouette of the face. The lower parts follow the chine line and end up close to the ears. Nevertheless there are two different definitions of how the outer landmarks should be placed for non-frontal views, visualised in

the *(middle)* and *(right)* of the figure. One option is defining the landmarks as to lei on the 2D silhouette. For non-frontal views the outer landmarks remain at the contour of the face. Following this definition the landmarks have no fixed physical position on the face. As the yaw angle of the face changes the landmarks of the back facing side wander closer to the middle of the face. Following an alternative definition the landmarks for non-frontal views follow the physical chine line for both the front facing and back facing face side. This leads to physically fixed correspondences between landmark and location on the head.

Inner face landmarks are reasonably well defined up to a certain image resolution. Nevertheless in high resolution imagery there can be several pixels difference between defining the eye corner as the boarder of the eye ball or on the eye lid. Figure 3.1 shows that landmarks clicked by humans are not consistent. There is significant uncertainty among the annotations.

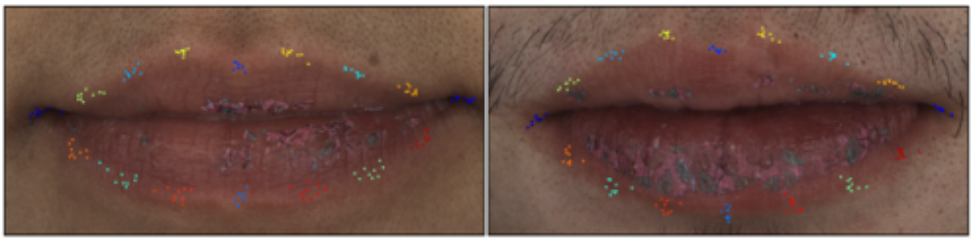

Figure 3.1.: *Figure from [DYW+18]. Landmark positions clicked by nine annotators on two images. The annotations are imprecise and not well defined.*

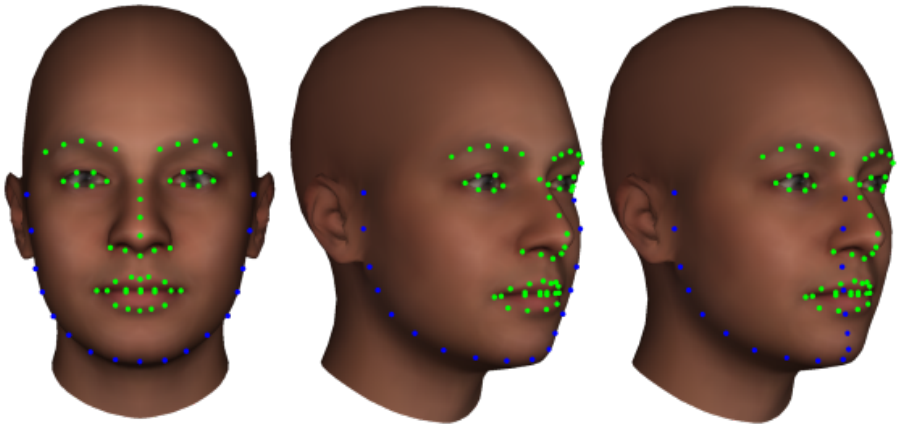

Figure 3.2.: *Sets of 68 2D landmarks visualised on mean head from different poses. Green landmarks are 51 inner landmarks. The blue landmarks are 17 outer landmarks. For frontal views the landmark definitions are straight forward. For larger poses there are different options.* (left)*: Frontal view.* (middle)*: Non-frontal view with the outer landmarks on the silhouette of the face.* (right)*: Non-frontal view with the outer landmarks following the physical chin line for both the front facing and back facing face side.*

To be able to compare different landmarking methods in Chapter 4 without need to train on the same set, the 51 inner 2D landmarks were chosen. These landmarks have the same definitions across different databases and are part of most landmark sets.

3.2. Stability Improvement of 300VW Ground Truth Labels

In order to measure temporal stability of landmarks we need ground truth landmarks on a video sequence. Face datasets of videos that come with landmarks are CK [KTC00] and CK+ [LCK+10]. These focus on expression analysis and the landmarks are labelled fully automatically by an AAM. The only larger dataset of videos with ground truth landmarks up to date is 300VW [SZC+15]. The landmarks are labelled in a semi automatic approach. First the single frames were labelled by the Project-out cascaded regression method [Tzi15]. Then using Gauss-newton deformable part models [TP14] a person specific model is learned for each video and used to adjust the landmarks. Finally human annotators checked and corrected the landmarks. The final landmarks are robust to pose and different lighting, nevertheless the landmarks appear very jittery in the video, thus not temporally stable. The videos of 300VW show in the wild situations captured by hand-held cameras or difficult lighting situations. Some show extreme expressions. Most videos in 300VW are of low resolution and the faces have an inter-ocular-distance of 70-130 pixels. Sequence 007 is an exception to that with an average IOD of 285 pixels, which was also the main reason to choose this sequence as test sequence.

The temporal stability of the ground truth landmarks is non the less not satisfactory. The landmarks appear very jittery on the video. A possible way of improving temporal stability is using optical flow to remove minor jumps from one frame to another. One could not just track a ground truth landmark over an entire video, as the landmark would drift and most likely loose it's original semantic location. Nevertheless as we have the location of the landmark labelled by the iBUG group, we could crop a region around that position and then apply optical flow from the first crop in the sequence to the other crops in the sequence. Therefore no drift is possible and optical flow is always used to compare to the first frame. The chosen optical flow method was introduced by Brox et al. [BBPW04]. Using an in house implementation at DR the optical flow of the crops was calculated. The resulting optical flow vector at the position of the landmark is then applied to shift the landmark to a new location. This method was only used to adjust the landmark on the nose tip.

3.3. Automatic Labelling of Disney Research Sequence

To overcome the issue of temporally unstable landmarks and low resolution faces an own ground truth sequence for evaluation was created. As manually labelling all frames is prone to errors and not a scalable approach the Anyma system by Wu et al. [WBGB16] was chosen to fully automatise the labelling. This reconstruction method uses local deformation models constrained by a global anatomical model. For the given sequence several views were available which leads to a very precise reconstruction. For evaluation purposes we only use the frontal most camera in the setting. Additionally, optical flow is used to constrain the reconstruction across frames which leads to temporally stable results. Given the reconstructed face meshes one can project the vertex coordinates to image space using camera parameters found during calibration. As the meshes are in dense correspondence the vertex IDs corresponding to different landmarks stay the same and have to be clicked only once. As described in Section 3.1 the landmarks evaluated

and thus also the 51 inner face landmarks are annotated using this method.

Figure 3.3 shows exemplar frames of the both sequences used for evaluating the detectors.

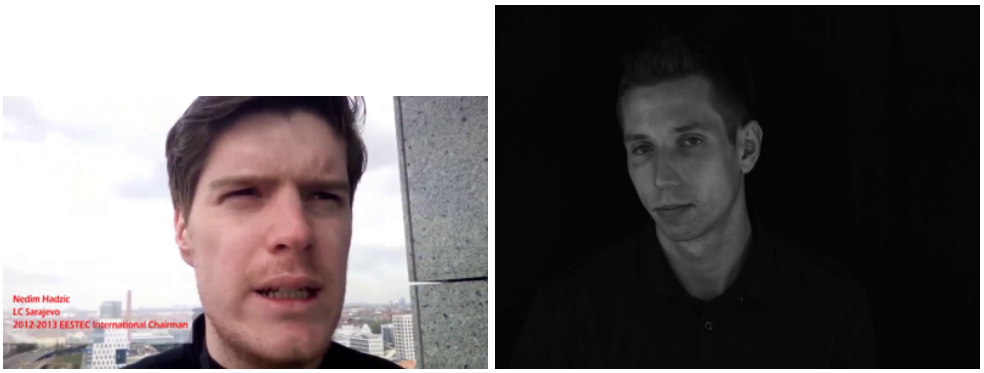

Figure 3.3.: *Frames of the evaluation sequences.* (left)*: Frame of video 007 of the 300VW data set [SZC+15].* (right)*: Frame of the Disney Research (DR) cam03 sequence used.*

3.4. Measurement of Temporal Stability

The quality of landmark detections can be measured by different methods. The most common error metric is to take the normalised mean error (NME). One can normalise the error between the detection and the ground truth either by the inter ocular distance or based on the face box size. Using the IOD as normaliser has the characteristic of normalising large yaw poses by a smaller factor as frontal poses. Thus errors on pose images are higher than on frontal pose images. Nevertheless this method is more common among the community and also chosen as metric for this work. Finally, the NME metric used for images was:

$$\text{NME} = \frac{1}{N} \sum_{k}^{N} \frac{\|p_k - g_k\|_2}{\|g_l - g_r\|_2} \tag{3.1}$$

where p_k denotes the prediction of landmark k and g_k the associated ground truth label. g_l and g_r are the ground truth positions of the outer eye landmarks left and right. The NME is measured on each image of a face separately. Given a test set one can then plot a curve of the cumulative distribution to compare different detectors.

Figure 3.4 shows the disadvantage of this approach. In videos the accuracy of the landmarks measured by the NME metric does not penalise jittery detectors. The figure shows two consecutive frames of a test sequence, the ground truth landmark (blue) three detectors (red, green, yellow). The three detectors have equal accuracy on both frames. Nevertheless the yellow detector is much more stable as the others.

As metric to measure this temporal stability we propose the landmark stability error:

$$\text{LSE}(k,i) = \frac{\|\Delta p_{k,i} - \Delta g_{k,i}\|_2}{\|g_{l,i} - g_{r,i}\|_2} = \frac{\|(p_{k,i} - p_{k,i-1}) - (g_{k,i} - g_{k,i-1})\|_2}{\|g_{l,i} - g_{r,i}\|_2} \tag{3.2}$$

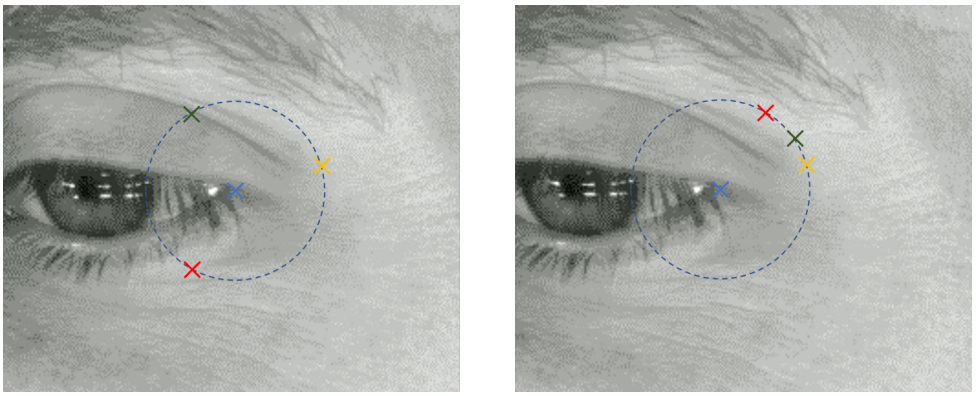

Figure 3.4.: *Temporal stability and accuracy. Two consecutive frames, the ground truth landmark in blue and three detectors in green, red, yellow.*

For a given landmark ID k and a frame number i the error measures the difference of derivatives between ground truth and detector from one frame to another. As with the NME the error is normalised by the inter ocular distance. This is necessary to not only be able to compare results across video sequences, but also to weigh frames in a video equally. One can then get the mean LSE of a video, of length T, and all landmarks, N as:

$$\text{MLSE} = \frac{1}{NT} \sum_{i}^{T} \sum_{k}^{N} \text{LSE}(k, i) \tag{3.3}$$

The LSE is also stable in regards of different landmark definitions. As not the distance between detection and ground truth is measured, but the difference between the deltas from one frame to another, slightly different landmark locations are not considered as error.

4

Experiments

In the following the experiments conducted will be presented. In Section 4.1 we present each of the detectors in more detail and look at their accuracy with different face detectors used. Section 4.2 compares the methods using dlib CNN face boxes. The experiments in which the temporal stability is measured are presented in Section 3.4.

4.1. Accuracy of Detectors across Face Box Initialisations

We evaluate each of the detectors more closely on both test sequences. Note that for the low resolution but more frontal sequence 300VW 007 we plot the NMR between 0% and 6%. Errors on the high resolution and more difficult poses and expression sequence DR cam03 are plotted between 2% and 8%. Failure cases are defined as above the plot range, thus 6% and 8% respectively. In both cases the entire sequence data available was used although for DR cam03 after the 200^{th} frame only every 10^{th} frame was provided.

4.1.1. Dlib Regression Trees [KS14]

The regression tree method implemented in dlib is trainable in a few hours on a standard pc. All results presented in the following are therefore trained by the author. Initial tests showed that the dlib implementation yields similar performance as originally stated in [KS14]. These initial comparison tests were conducted on the LFPW [BJKK13] subset of the 300W challenge [SAT+16] data set.

Figure 4.1 shows the cumulative error distribution on the 300VW 007 sequence. The best curve denoted as "dlib on GT trained on GT FB" is by training and running the landmark detector on

a face box regressed based on the ground truth landmarks. The other combinations of face box during training and testing lead to similar performance, except running the detector trained on HOG face boxes and during testing on CNN face boxes. This leads to around 5% failure cases.

On the more challenging DR cam03 sequence, results are similar (Figure 4.2). The "dlib on CNN trained on HOG FB" is dramatically worse than the rest and fails on ca. 35% of all test images. Figure 4.3 shows examples of these failure cases. A qualitative investigation showed that the CNN face boxes are often higher than the HOG face boxes which leads to poor alignment of the face landmarks. Interestingly using the ground truth faces boxes does not deliver the best results. This might be because of slightly different landmark definitions. As the detector is trained on face boxes that are perfectly aligned with the ground truth landmarks, slightly different landmark locations can lead to inaccuracies.

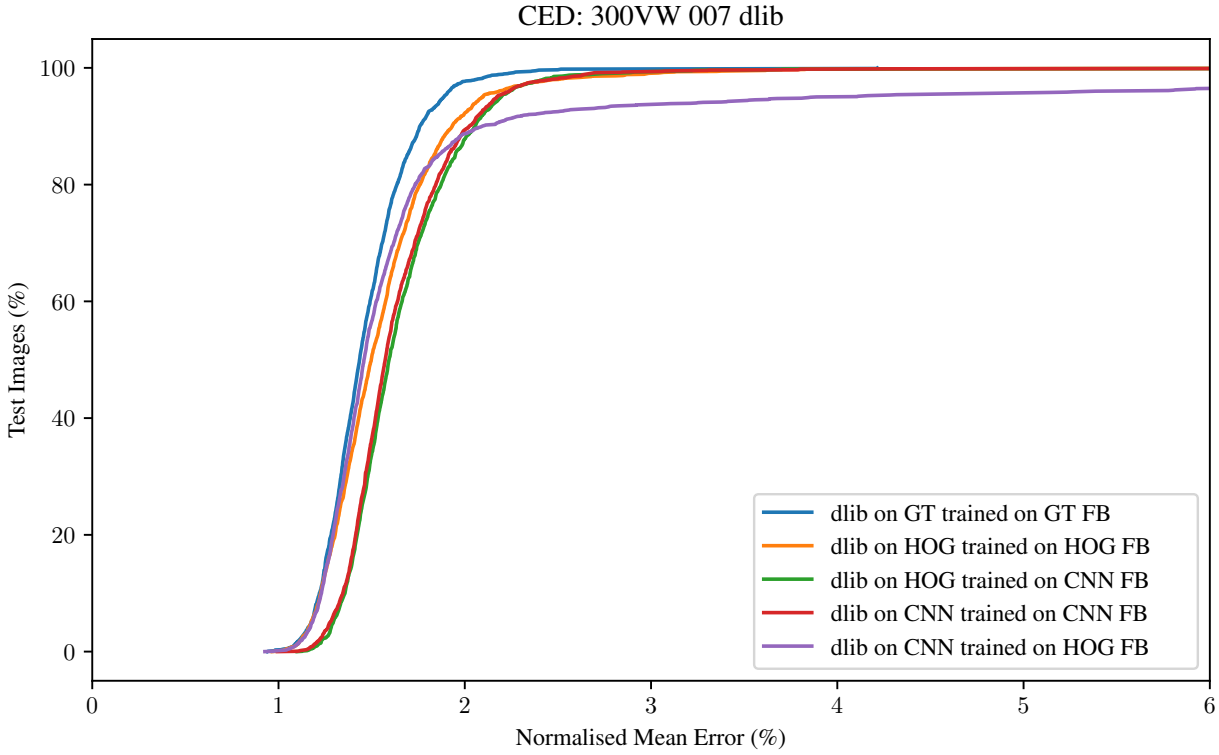

Figure 4.1.: *Comparison of the dlib landmark detector using different face boxes on the 300VW 007 sequence.*

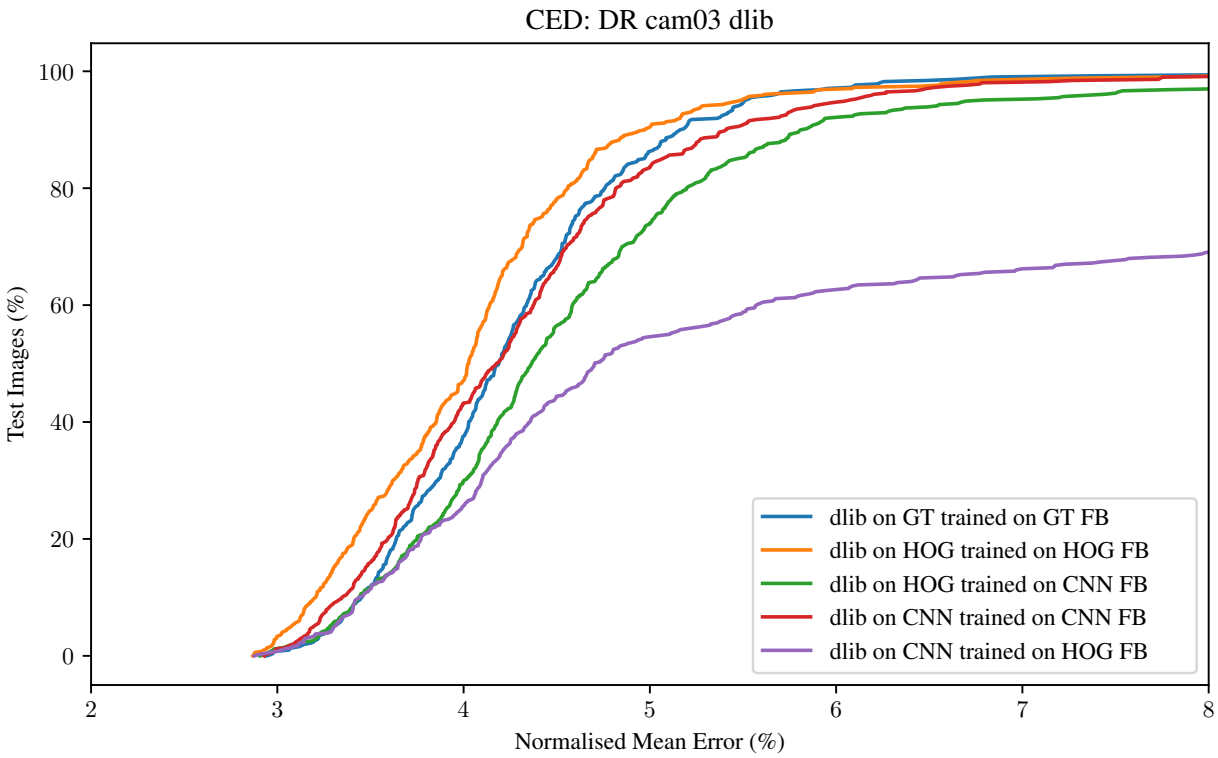

Figure 4.2.: *Comparison of the dlib landmark detector using different face boxes on the DR cam03 sequence.*

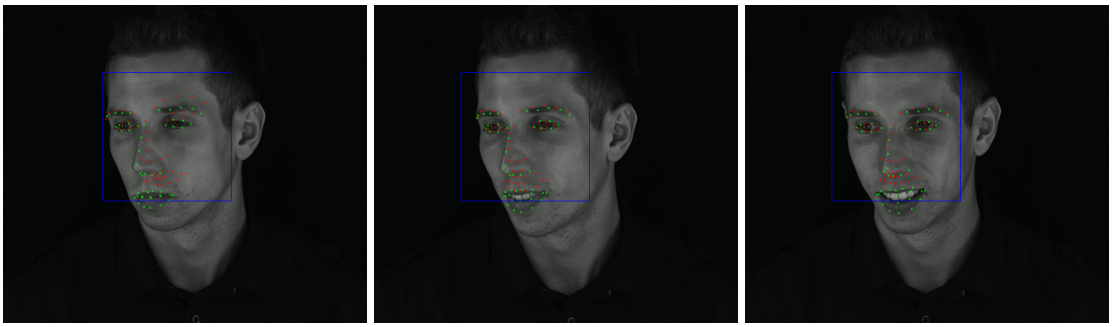

Figure 4.3.: *Failure cases of dlib detector using CNN face boxes but trained on HOG face boxes. Green crosses are ground truth landmarks. Red crosses are predicted landmarks.*

4.1.2. Face Alignment Network (FAN) [BT17]

There are two different FAN models we evaluate, the FAN2D and FAN3D, both provided by the FAN authors. For the FAN3D we do not use the depth coordinate but project orthographically and only use the x and y coordinates. The FAN networks were trained on 300-W-LP, a synthetically enlarged version of 300W [ZLL+16]. The FAN2D was additionally fine-tuned for a few epochs on the original 300W trainingset. For training face boxes computed by the ground truth landmarks were used with 10% noise for FAN2D and 20% noise for FAN3D. The inference on a standard laptop PC using CPU takes around 5 sec for the FAN2D and 10 sec for FAN3D. Comparing performance on 300VW 007 (Figure 4.4) one can see that FAN2D is more precise but less robust than FAN3D. Nevertheless this can not be observed on the Disney Research sequence shown in Figure 4.5. Regarding face box initialisation there is no significant difference between HOG or CNN face boxes on either sequences, except with FAN3D on 300VW 007 where the CNN initialisation performs better.

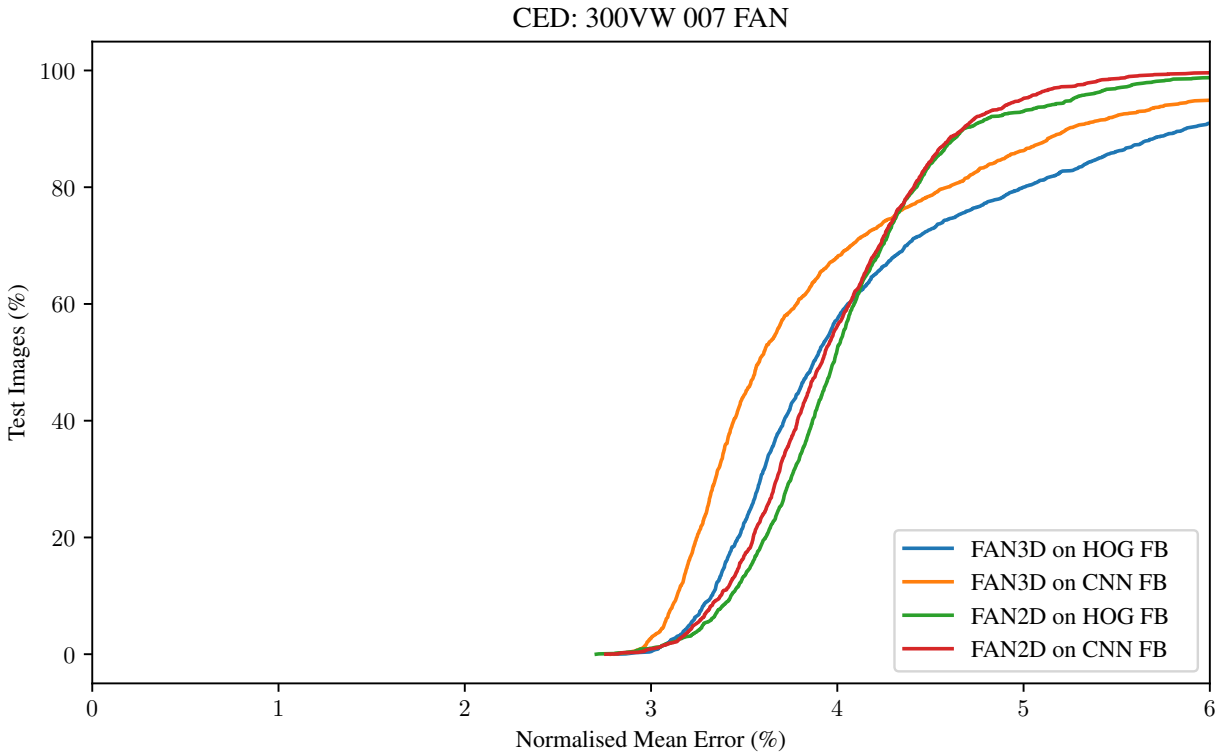

Figure 4.4.: *Comparison of the FAN 2D and 3D landmark detectors using different face boxes on the 300VW 007 sequence.*

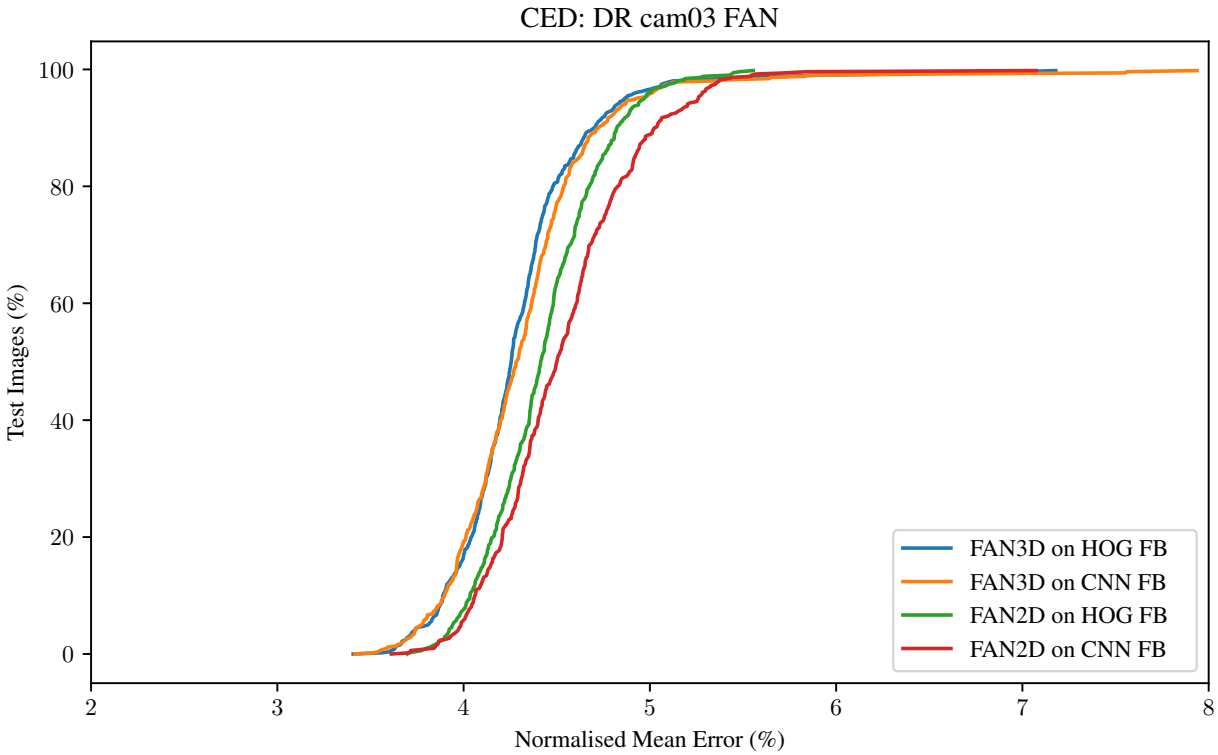

Figure 4.5.: *Comparison of the FAN 2D and 3D landmark detectors using different face boxes on the DR cam03 sequence.*

4.1.3. Supervision by Registration (SBR) [DYW+18]

The authors of the supervision by registration detector do not publish a pretrained model. The model used for training was provided by Industrial Light and Magic (ILM). As training set 300W and AFLW [KWRB11] were used. It remains unclear what face boxes were used for training as well as if and how the registration was applied without having videos in the training set. Figures 4.6 and 4.7 plot the CED curves. On the DR sequence the CNN face box initialisation leads to slightly better performance.

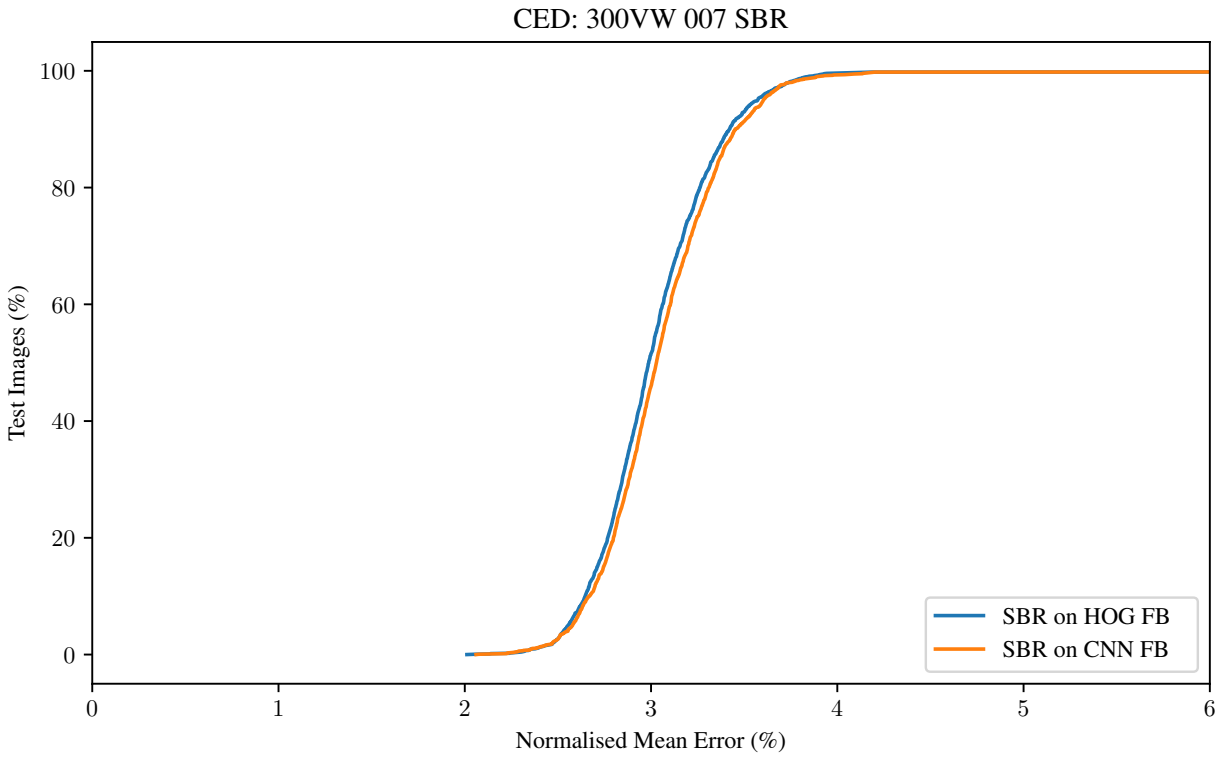

Figure 4.6.: *Comparison of the SBR landmark detector using different face boxes on the 300VW 007 sequence.*

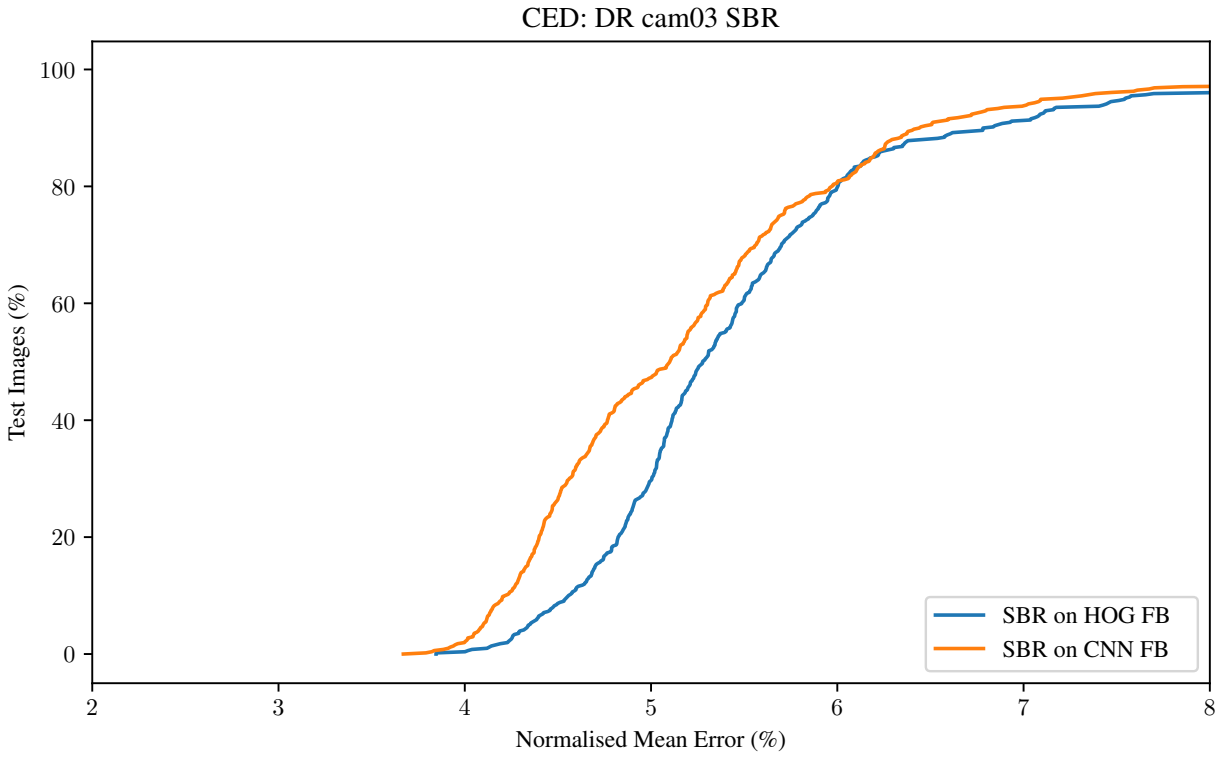

Figure 4.7.: *Comparison of the SBR landmark detector using different face boxes on the DR cam03 sequence.*

4.1.4. Wingloss [FKA+18]

The wingloss network was specifically trained for evaluation in this work by Dr. Feng. It is based on the ResNet50 architecture [HZRS16]. Two networks with different training sets were provided. Wingloss1 which was trained on 300W only and wingloss2 was trained on 300W and 300-W-LP. In all cases the dlib CNN face detector was used to crop the images to network input size. This can also be seen in both Figure 4.8 and 4.9 as the performance when using HOG face boxes is lower. On the 300VW 007 sequence the performance of wingloss1 is significantly higher. This might be due to the landmark definitions provided in the training sets. 300W and 300VW are both from the Intelligent Behaviour Understanding Group (iBUG) and therefore share definitions. 300-W-LP is synthetically generated and not by iBUG. These differences lead to decreased performance of wingloss2 on 300VW 007. On the DR cam03 sequence the landmarks are generated as described in Sector 3.3 and both wingloss networks perform equally well.

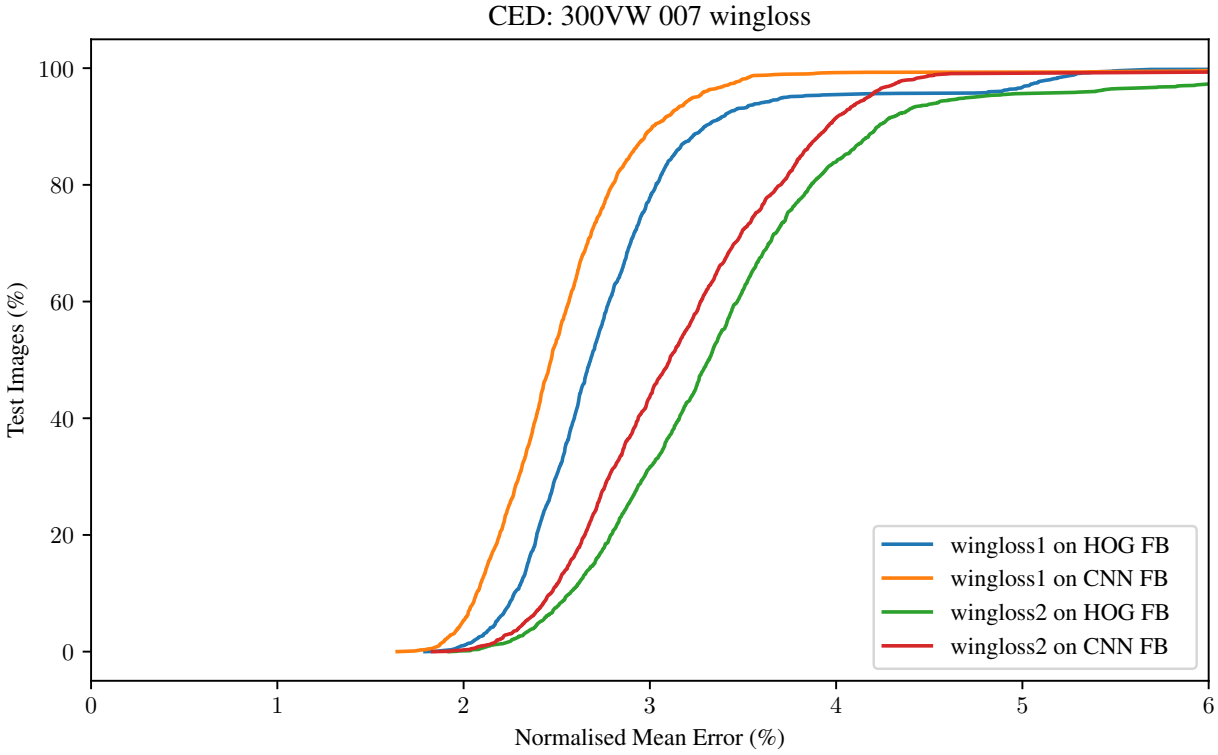

Figure 4.8.: *Comparison of the wingloss landmark detector using different face boxes on the 300VW 007 sequence.*

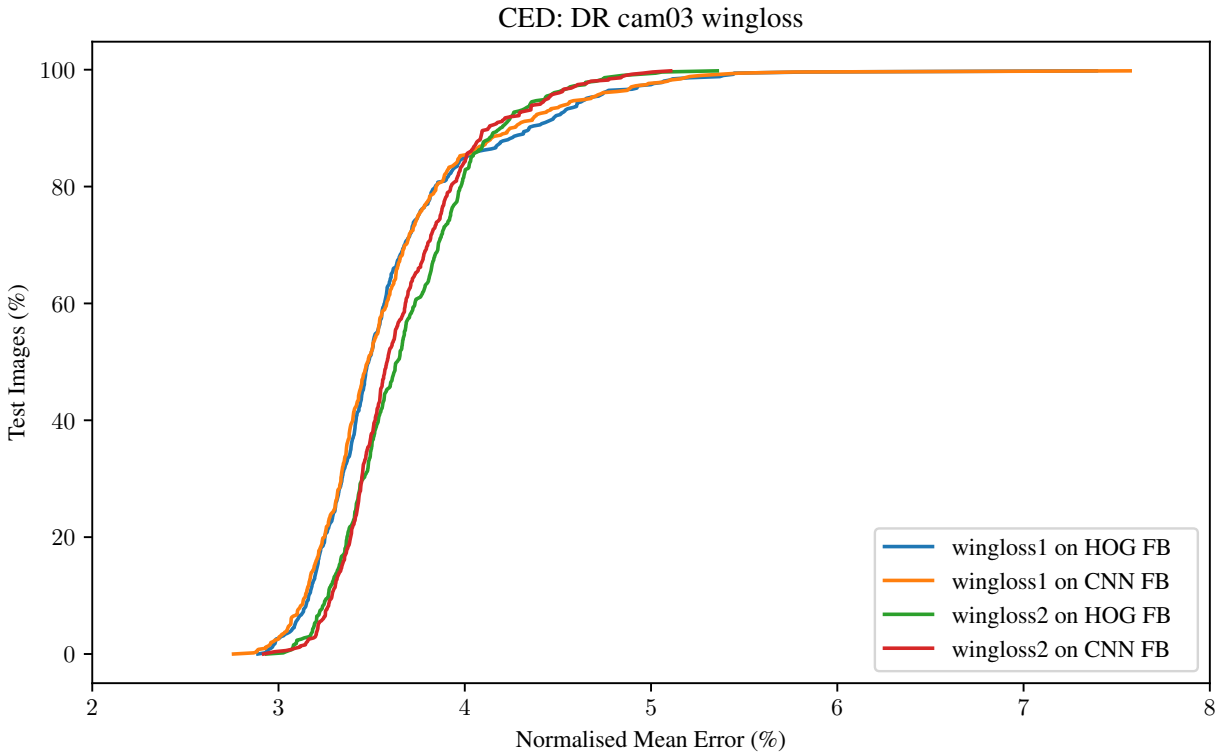

Figure 4.9.: *Comparison of the wingloss landmark detector using different face boxes on the DR cam03 sequence.*

4.2. Accuracy Comparison of Detectors with CNN Face Boxes

The CNN face detector is more robust against pose and lighting conditions than the older HOG detector. Thus, when using a face detector for initialisation of a landmark detector one would prefer the CNN face detector as it detects more faces overall and the entire pipeline would fail less often. Therefore we will compare the landmark detectors using the CNN face boxes.

Table 4.1 summarises the detectors compared in this chapter with their training sets. 300-W-LP is a synthetically augmented version of 300W and thus these sets are based on the same images. The training set of 300W consists of 3837 images. In 300-W-LP these were augmented to a data set of 61,225 images. AFLW comes with 25,993 images gathered from Flickr, but with only up to 21 landmarks.

The dlib detector we use for these experiments was trained on CNN face boxes. The wingloss networks were also trained on these face boxes. The other detectors are either not optimised for dlib CNN face boxes or the used face boxes are unknown.

First we show the performance of the detectors on both sequences on the first 200 frames each in Figures 4.10 and 4.11. The order of the performance of the detectors stays the same for most of the frames. Thus the detectors do not show strong differences of performance in varying situations of the sequence.

In the following we show the cumulative error distributions of the detectors taking all available

	300W [SAT+16]	300-W-LP [ZLL+16]	AFLW [KWRB11]
dlib [KS14]	x		
FAN3D [BT17]		x	
FAN2D [BT17]	x	x	
SBR [DYW+18]	x		x
wingloss1 [FKA+18]	x		
wingloss2 [FKA+18]	x	x	

Table 4.1.: *List of detectors evaluated and the training sets of these.*

frames of the sequence into account. Comparing the detectors on 300W 007 in Figure 4.12 one can see that the dlib regression trees implementation leads to best performance. It is also by far the fastest method. The wingloss1 network is second to best, using the same training set. Wingloss2, that was additionally trained on 300-W-LP has lower performance. The cumulative distributions of wingloss2 and SBR cross at around 40%. It highly depends on the application which of these detectors should be considered better. The FAN networks come last.

In Figure 4.13 the detectors are compared on the more high resolution and challenging DR cam03 sequence. For the wingloss and the FAN networks the difference between the training sets does not cause a big performance change, as it did on the 300VW 007 sequence. The wingloss networks are best, followed by the FAN networks with about 1% higher error rate. Dlib has high accuracy on some images similar to the wingloss networks but is also worse than FAN on about 20% of the frames. SBR is last and can not reach the performance of the other detectors.

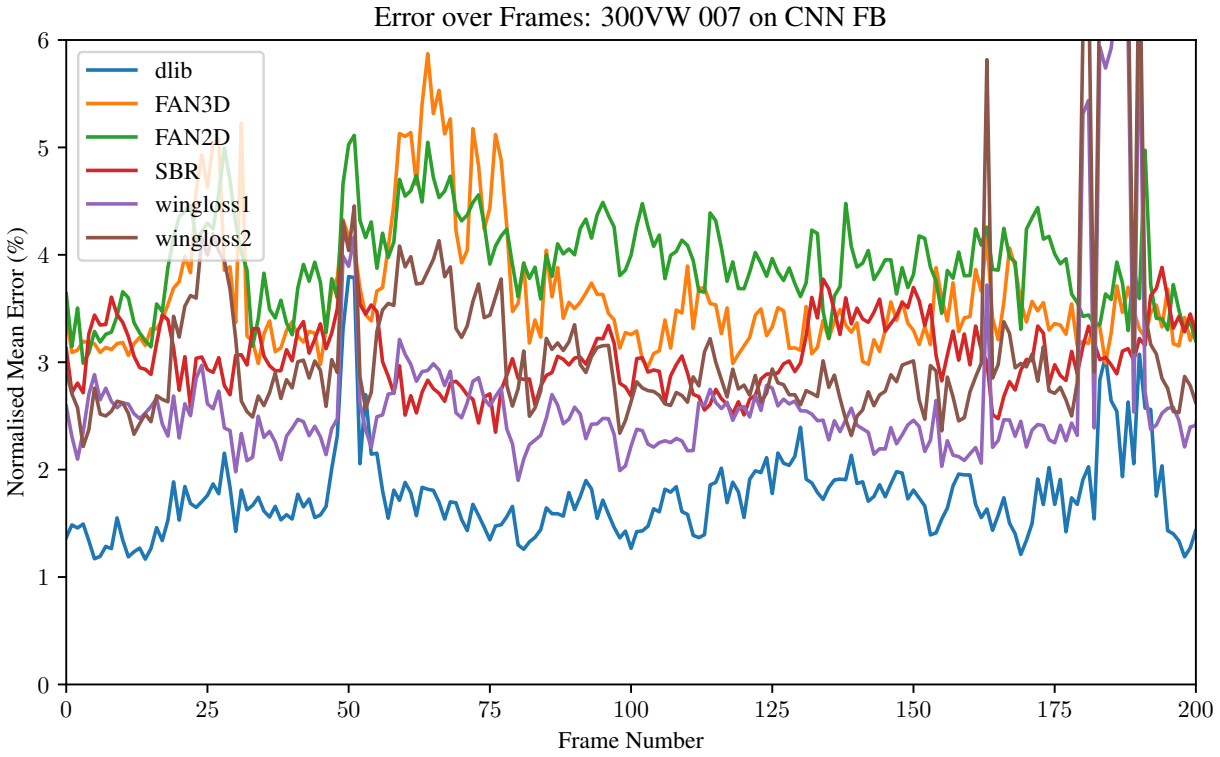

Figure 4.10.: *Frame-wise comparison of all detectors using CNN face boxes on the 300VW 007 sequence.*

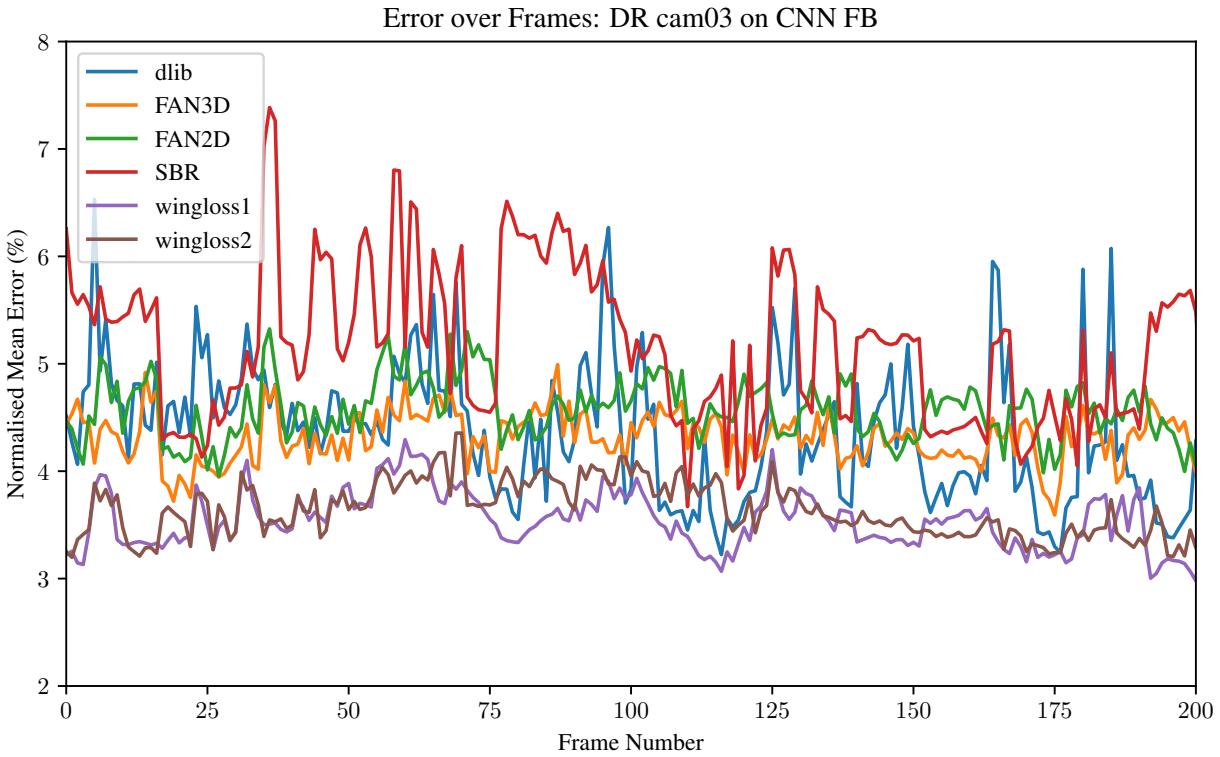

Figure 4.11.: *Frame-wise comparison of all detectors using CNN face boxes on the DR cam03 sequence.*

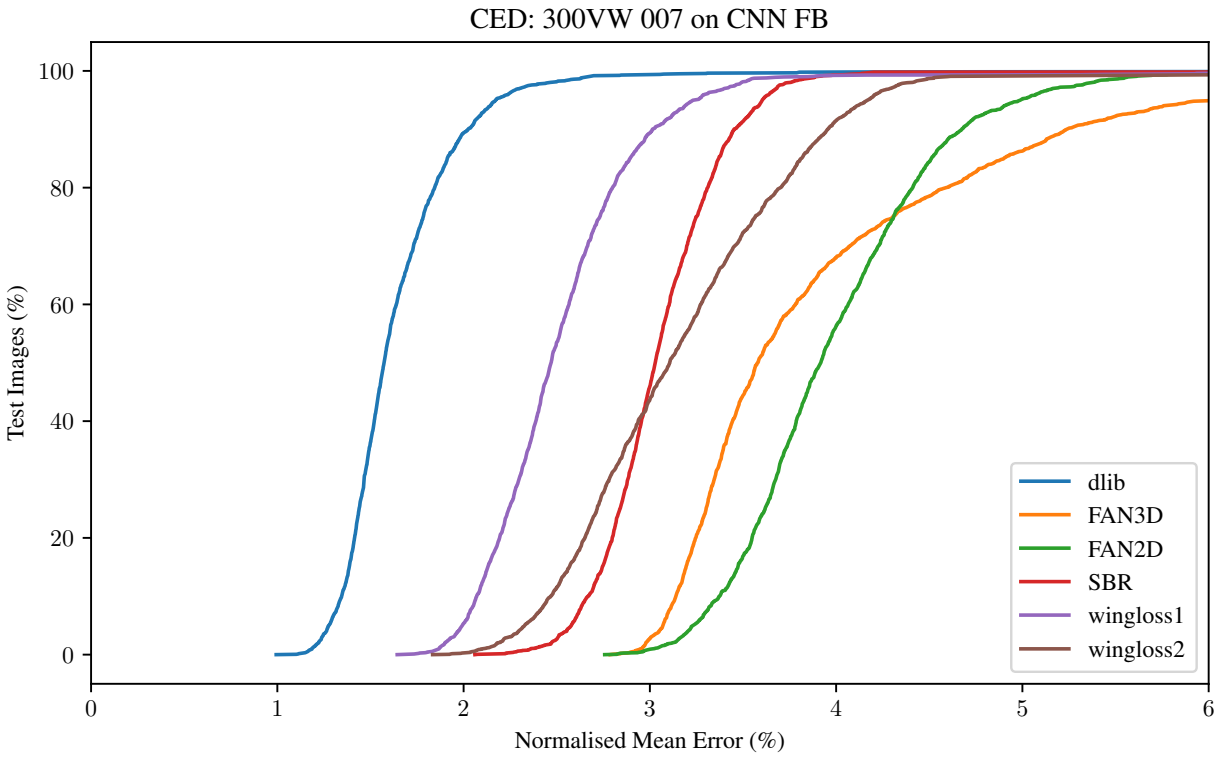

Figure 4.12.: *Comparison of all detectors using CNN face boxes on the 300VW 007 sequence.*

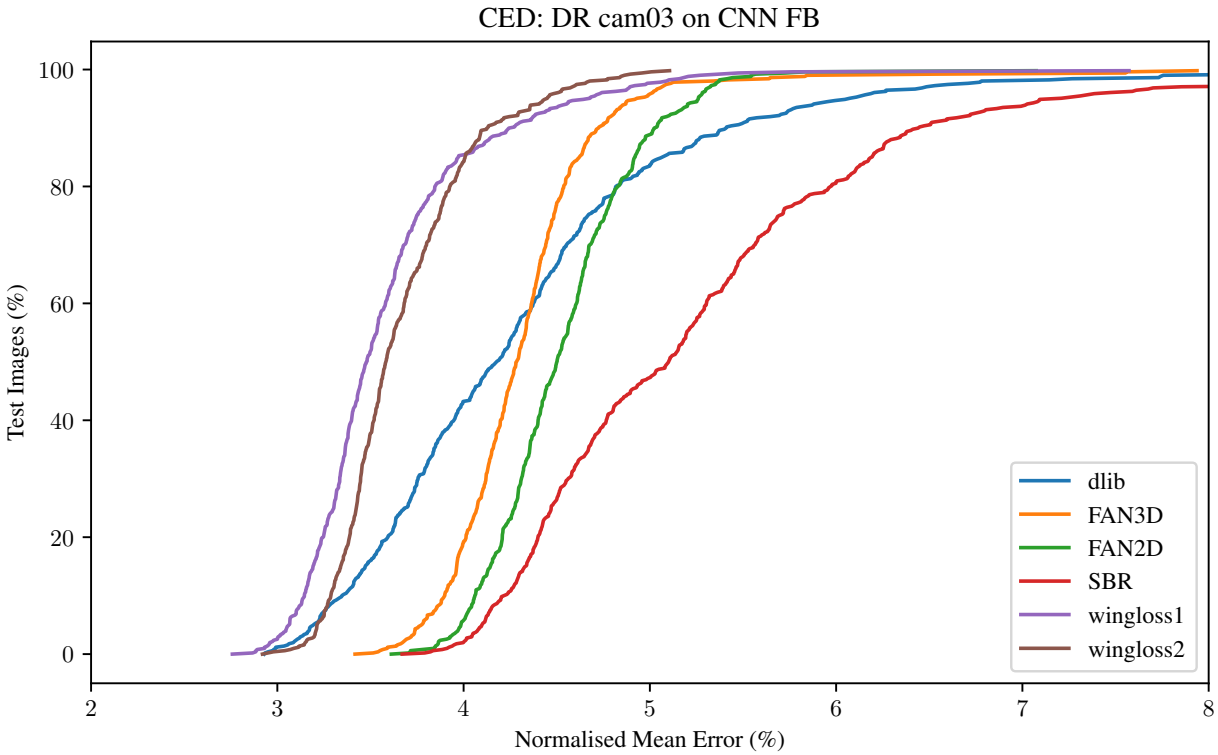

Figure 4.13.: *Comparison of all detectors using CNN face boxes on the DR cam03 sequence.*

4.3. Evaluation of Temporal Stability

In this Section we present the results achieved by the novel evaluation metric introduced in Section 3.4. We plot the L2 norm of the difference between ground truth and detection movement from one frame to another. At first we focus on a single landmark, in particular the nose tip landmark in Subsection 4.3.1. Then in Subsection 4.3.2 we look at the influence of the face detector for landmarks stability using all landmarks.

4.3.1. Stability of the Nose Landmark and Optical Flow Fix

In Figure 4.14 we compare the methods on the 300VW 007 sequence. The FAN networks show error rates 2-3 times higher than the other methods on average. The accuracy presented in Figure 4.12 of the FAN networks was also lowest but not by this margin. Thus it appears that FAN is not only not accurate on the lower resolution sequence but also highly unstable.

As described in Section 3.2 optical flow was used to improve the stability of the ground truth nose tip landmark. In Figure 4.15 the error against this new ground truth is plotted. The overall appearance is very similar suggesting that the optical flow fix did not change much. Nevertheless, these small adjustments can make a big difference when measuring smaller errors. The average SBR error for example changed from 0.662 to 0.566, which is a decrease of 14%. Table 4.2 shows the landmark tracking error of all detectors with and without the optical flow stabilisation of the ground truth landmarks. The new ground truth landmark leads to a lower error for all detectors indicating that the new landmarks are actually more stable than the old ones.

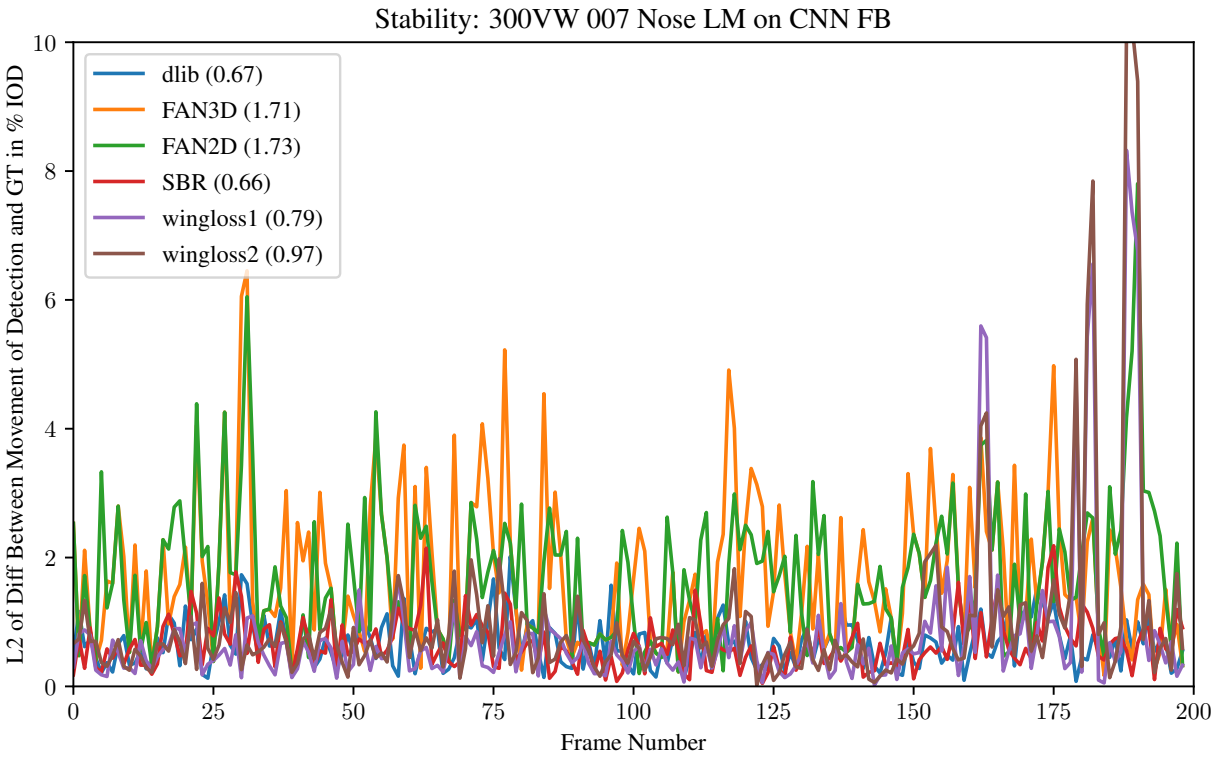

Figure 4.14.: *Comparison the tracking error from one frame to another of all detectors using CNN face boxes on the 300VW 007 sequence. The numbers in brackets are the average errors.*

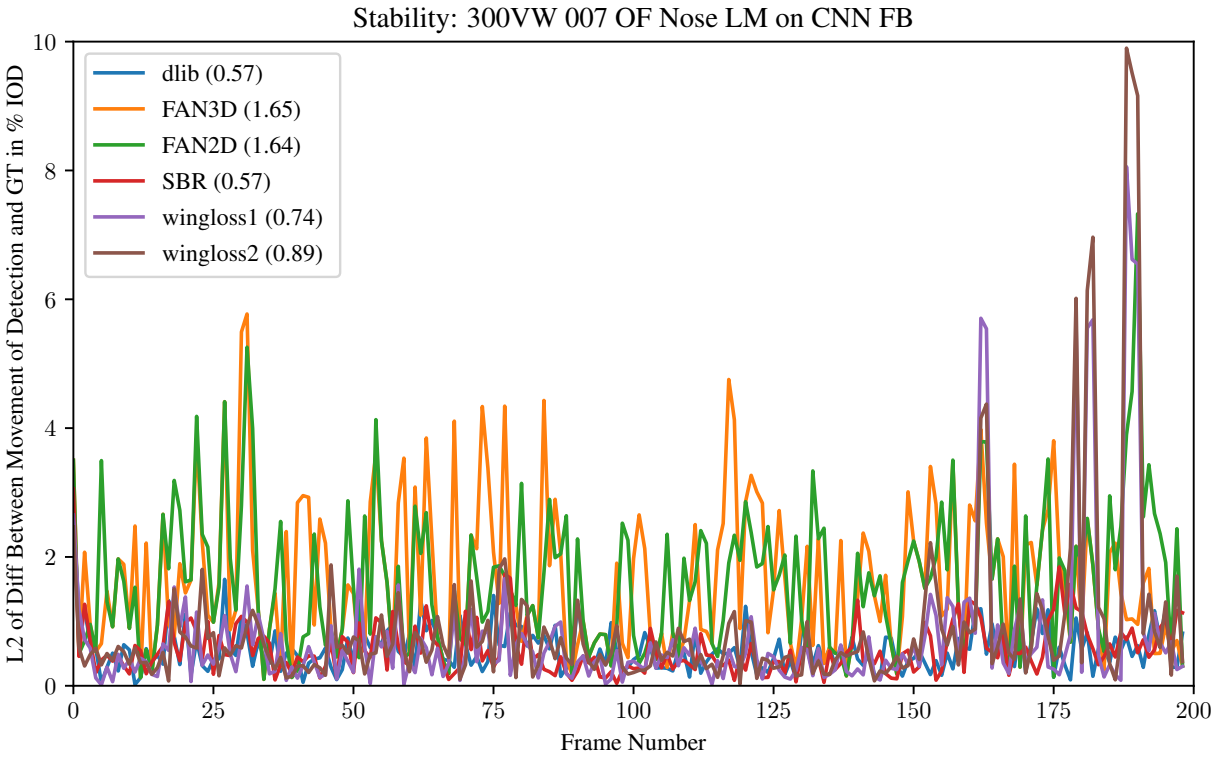

Figure 4.15.: *Comparison the tracking error from one frame to another of all detectors using CNN face boxes on the 300VW 007 sequence. The ground truth landmark position was stabilised before using optical flow. The numbers in brackets are the average errors.*

	dlib	FAN3D	FAN2D	SBR	wingloss1	wingloss2
without OF	0.67	1.71	1.73	0.66	0.79	0.97
with OF	0.57	1.65	1.64	0.57	0.74	0.89

Table 4.2.: *Comparison of stability error rates (LSE) with and without the optical flow fix on the nose tip landmark.*

Figure 4.16 shows the errors for the first 200 frames of the DR sequence. In general one can see that the errors are very unsteady. There are jumps from one frame to another. The results are consistent with Figure 4.13, with SBR and dlib having high error rates. The FAN networks have slightly lower error rates and the wingloss method performs best.

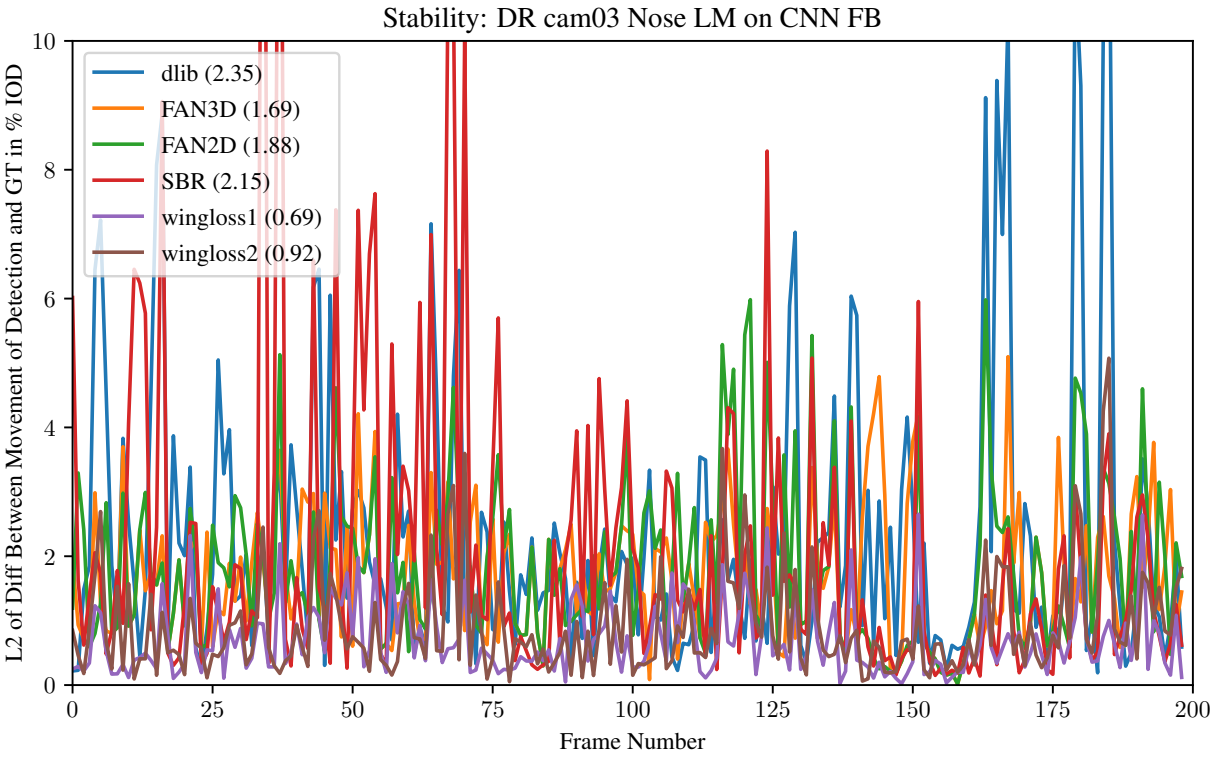

Figure 4.16.: *Comparison the tracking error from one frame to another of all detectors using CNN face boxes on the DR cam03 sequence. The numbers in brackets are the average errors.*

4.3.2. Stability of All Landmarks

For reasons of completeness we plot the average error over all landmarks on the 300VW 007 video in Figure 4.17. As initialisation CNN face boxes were used because of higher detection rates. Nevertheless because of the unstable ground truth landmarks the evaluation shall focus on the DR cam03 sequence.

The following Figures 4.18 and 4.19 show the average tracking error over all landmarks using a CNN face box and HOG face box, respectively. Comparing the overall magnitude of errors

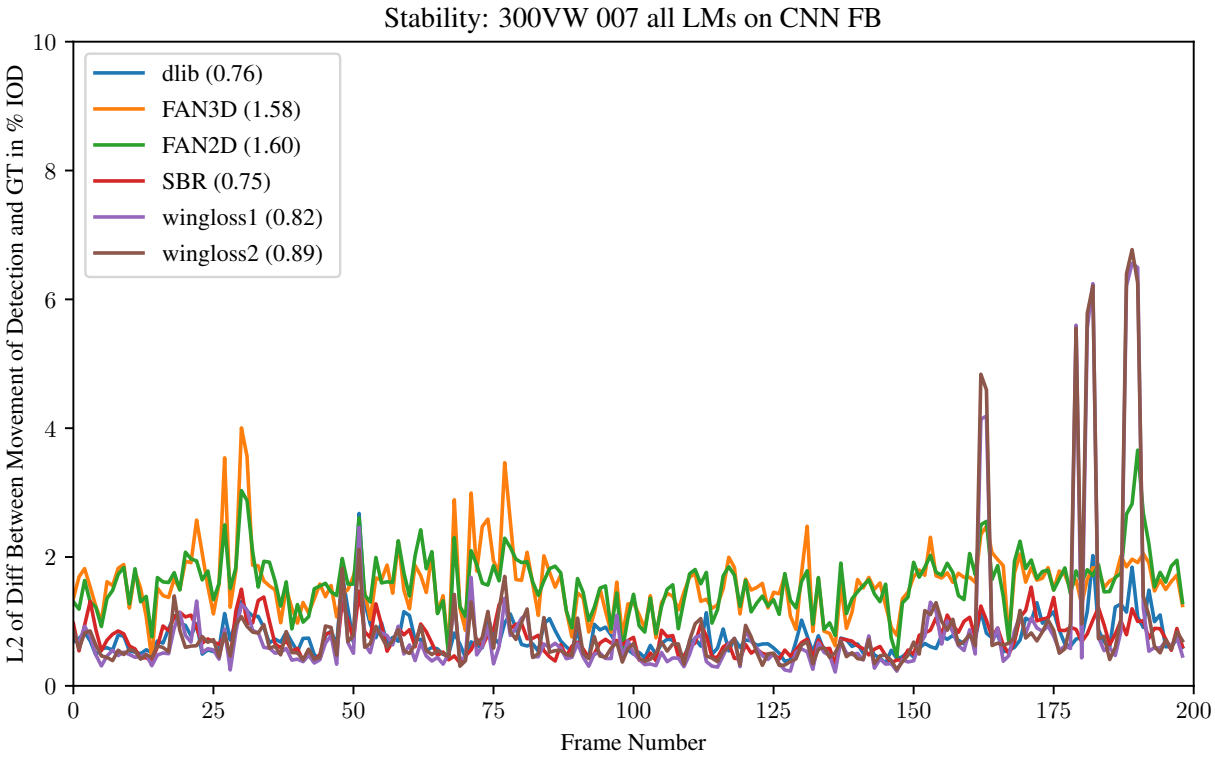

Figure 4.17.: *Comparison the tracking error from one frame to another of all detectors using CNN face boxes on the 300VW 007 sequence on all landmarks. The numbers in brackets are the average errors.*

with the landmark error in the previous subsection one can see that the error is lower when averaging. One can conclude that the jittering is independent across landmarks and jumps do not appear at the same time across all landmarks. The most unstable detector are the FAN networks, followed by the dlib regression trees and SBR. Wingloss appears to be most stable. These results match the conclusion drawn from Figure 4.14 and 4.15. Also in Figure 4.19 when using HOG face boxes instead of CNN face boxes, the FAN networks produce the most unstable landmarking. The dlib detector used in this experiment was also trained on the same training set but using HOG face box initialisations. In direct comparison the HOG face boxes lead to more stable results. One can also see some correlation between the detectors which can be a result of imprecise or unstable face boxes. When using the CNN face boxes (Figure 4.18) it seems the frames around number 60 and number 125 lead to unstable results. This cannot be observed when using HOG face boxes (Figure 4.19) where these parts are stable, but around frame number 100 bigger peaks appear. This difference can only be caused by the face detectors. Table 4.3 directly compares the average stabilities using different face boxes. For all landmark detectors the HOG face boxes lead to more stable results and using face boxes created from the ground truth landmarks are most stable, as shown in Figure 4.20. Even as we do not know the exact mapping between ground truth landmarks and optimal face box for a landmark detector we can use an arbitrary mapping as for stability measurements only the deltas between frames matter.

In Figure 4.21 the changes of the face box in sum of left, right, top and bottom pixels is plotted. One can see that it matches the curves in Figures 4.18 and 4.19. The CNN face boxes are

unstable around frames 60 and 125, whereas the HOG face boxes are jittery around frame 100. The face boxes generated from the ground truth landmarks are not discrete as the dlib detectors and the changes show a relatively smooth curve. The average changes per frame are denoted in brackets and show the same order of stability as the resulting landmarks. One can conclude that unstable face boxes lead to unstable landmarks.

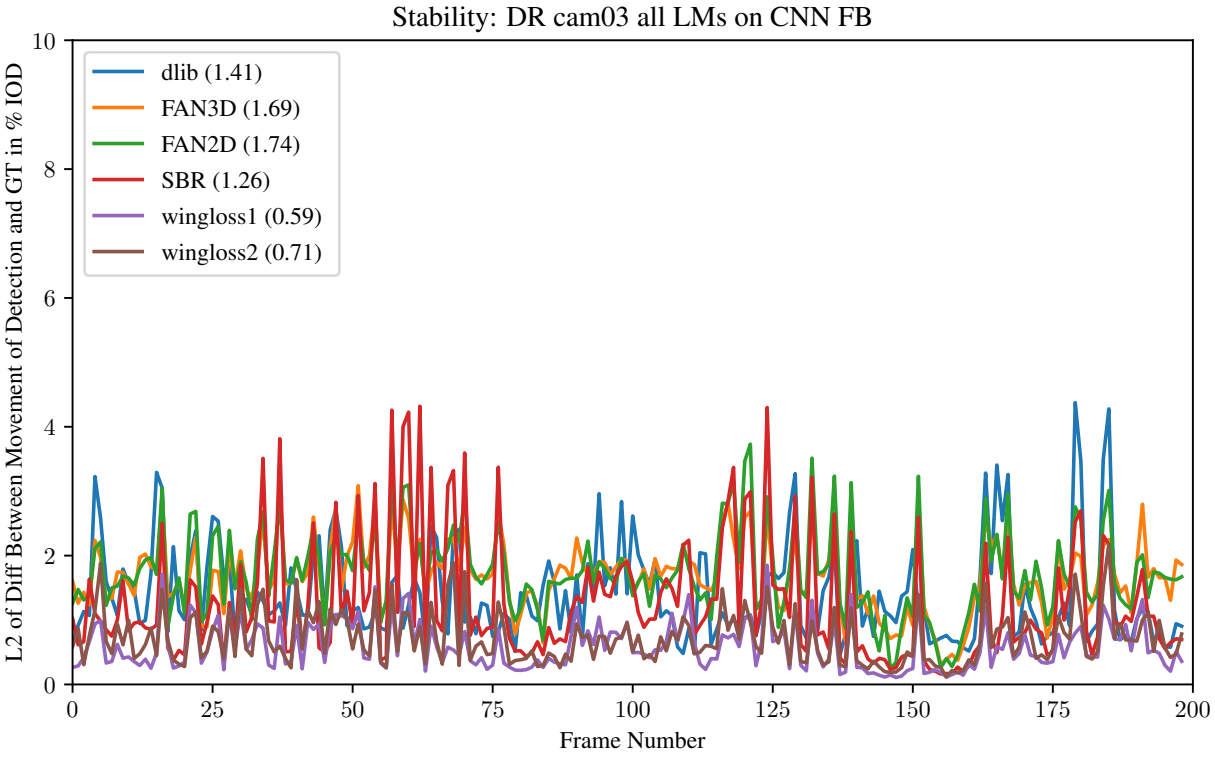

Figure 4.18.: *Comparison the tracking error from one frame to another of all detectors using CNN face boxes on the DR cam03 sequence on all landmarks. The numbers in brackets are the average errors.*

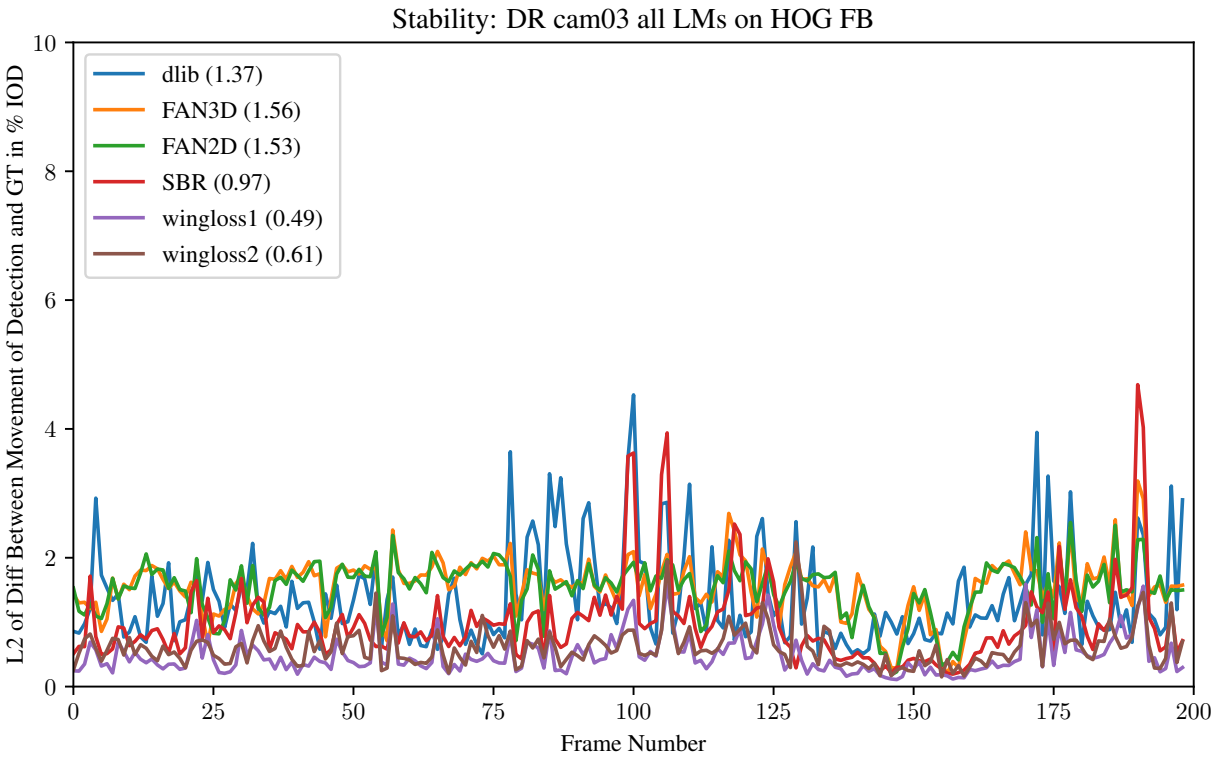

Figure 4.19.: *Comparison the tracking error from one frame to another of all detectors using HOG face boxes on the DR cam03 sequence on all landmarks. The numbers in brackets are the average errors.*

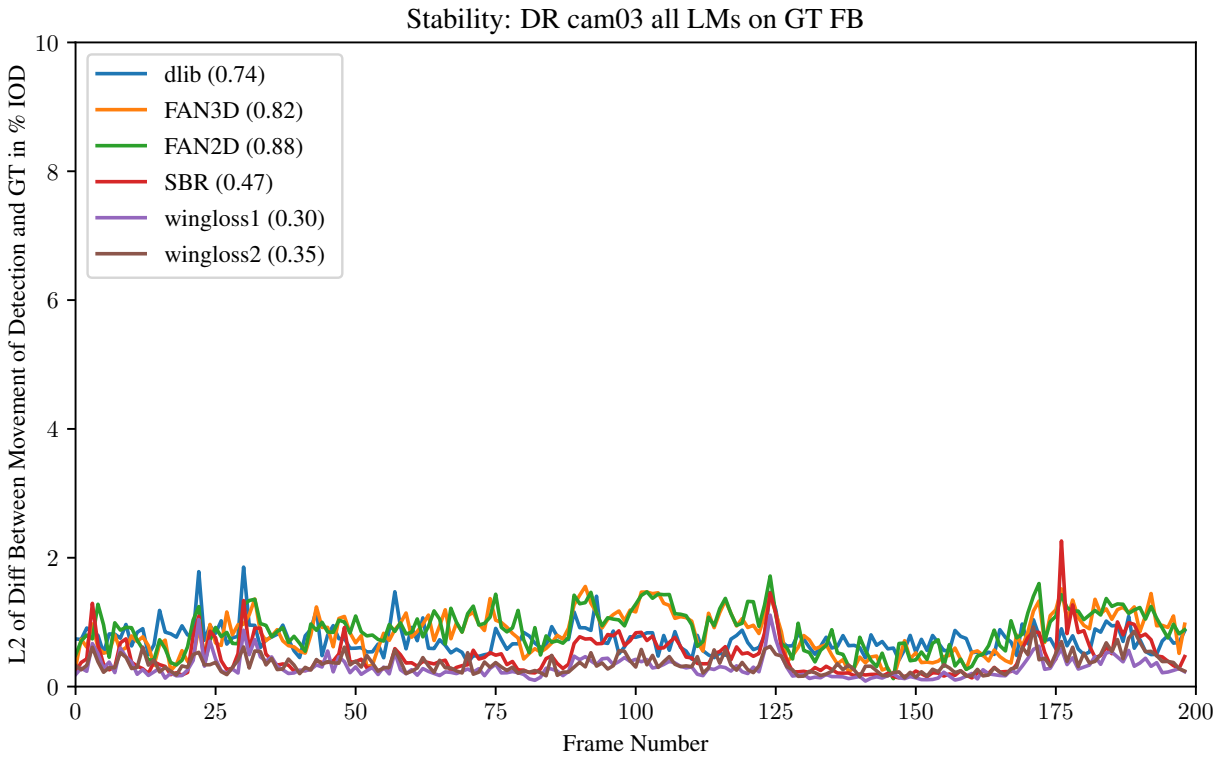

Figure 4.20.: *Comparison the tracking error from one frame to another of all detectors using face boxes generated from ground truth landmarks on the DR cam03 sequence on all landmarks. The numbers in brackets are the average errors.*

	dlib	FAN3D	FAN2D	SBR	wingloss1	wingloss2
CNN FB	1.41	1.69	1.74	1.26	0.59	0.71
HOG FB	1.37	1.56	1.53	0.97	0.49	0.61
GT FB	0.74	0.82	0.88	0.47	0.30	0.35

Table 4.3.: *Comparison of mean stability error rates (MLSE) over all landmarks with different face box initialisations.*

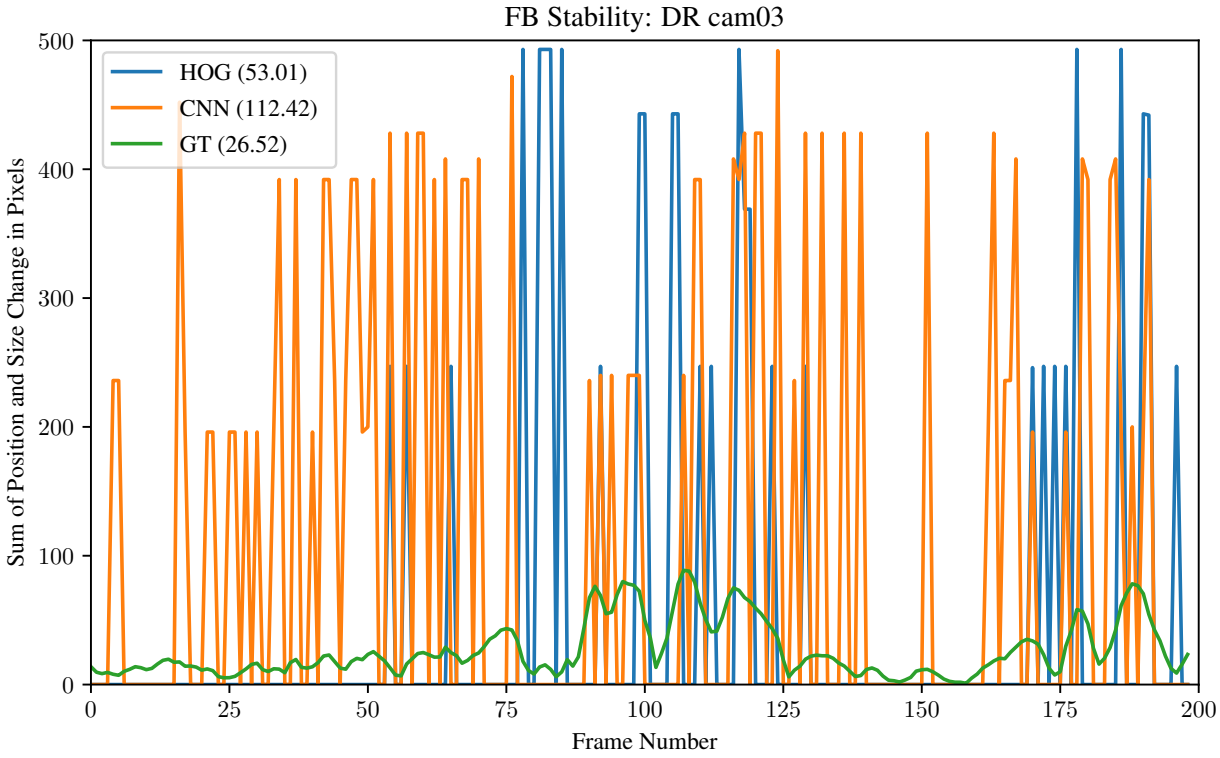

Figure 4.21.: *Comparison of the changes of face box size and position from one frame to the next. The numbers in brackets are the average changes.*

5

Conclusion

5.1. Summary of Results

This work deals with landmarking methods that are important steps in most face analysis pipelines. There have been tremendous advances in landmarking over the last 20 years. Nevertheless pose, lighting, expressions and occlusions remain challenges in this field. We give a brief overview of the history of methods used to improve landmarking and present a list of detectors with publicly available implementations. This gives a good overview of the most used methods in academia as well as industry today. As in the entire perception field deep learning based methods have become increasingly popular. The list includes networks proposed just recently. Most landmarking methods need to be given a face box that is either used to initialise the landmarking method or crop the image to the face region. In research the problems are often looked at separately. All experiments conducted in this work evaluate the entire pipeline of face detector and consecutive landmark detector. For some combinations of face and landmark detectors we see a big loss of performance if the landmarking was not trained on appropriate box initialisations. Even though the regression tree method from 2014 it performs reasonably well in terms of accuracy. On the low resolution sequence it is even the most accurate detector. The CNN based detectors show better performance on higher resolution input. A recently introduced method using a special loss function called wingloss achieves highest accuracy.

Another characteristic of landmark detectors that is not evaluated thoroughly yet by academia is landmark stability across consecutive frames in video sequences. We propose a metric that measures stability and conduct experiments on two testing sequences. As there is no big database with stable ground truth landmarks publicly available we propose a method of how to stabilise existing ground truth data, and show in experiments that stability errors decrease across all detectors. Additionally we label a sequence with ground truth landmark positions using Disney Research internal methods. The experiments confirm differences in stability experienced qualitatively and show high dependence of landmark and face box stability.

5.2. Future Work

A big limitation of this work is the amount of testing data that was available. The detectors and metrics were only applied to two sequences. The frames in a sequence are highly correlated. Pose and expressions change but the face itself stays the same with gender, age, ethnicity. Both sequences used for testing show middle aged men. Because of limited training data for the regression trees we also focused on more or less frontal sequences. A future work package of high priority would be to use more testing sequences with more variety in people and pose.

All detectors except the dlib regression trees were not trained by the author, but either publicly available or provided by others. This has the disadvantage that comparing the methods is not straight forward as differences in training set, augmentation and face box have to be taken into account. Training the methods oneself was beyond the scope of this work, but would help to gain further insights in the advantages and disadvantages of the methods.

More and more researchers in the field use 3D landmarks for their algorithms. These have the advantage that they stay in correspondence with physical positions on the face. One could compare these methods directly or using model based methods to reconstruct 3D positions from 2D landmarks.

To measure landmark stability the authors propose to take the L2 norm of the difference between movement of detection and ground truth landmarks. Instead one could penalise angle and length of this difference separately.

The experiments measuring stability show that the results depend on the face boxes used. One could use more sophisticated face detectors that not only roughly detect the face but also predict an aligned face box. This could improve stability or even accuracy.

To the best of our knowledge there is no study that evaluates if the size of the face box changes performance. Especially for methods that crop the image to the area around the face the size of the box might make a difference.

A

Links to Detectors

In the following there will be a list of resources given where to find implementations of the detectors presented in Chapter 2.

A.1. Face Detectors

Dlib hog face detector
Example python script: `https://github.com/davisking/dlib/blob/master/python_examples/face_detector.py`

Dlib CNN face detector
C++ implementation with network description: `https://github.com/davisking/dlib/blob/master/examples/dnn_mmod_face_detection_ex.cpp`
Blog post: `http://blog.dlib.net/2016/10/easily-create-high-quality-object.html`

Multi-task Cascaded Neural Network (MTCNN)
Implementation: `https://github.com/davidsandberg/facenet/tree/master/src/align`

Face detection with Faster R-CNN
Implementation: `https://github.com/playerkk/face-py-faster-rcnn`
Comes with model trained on WIDER dataset. Needs caffee.

Face detection with deep neural networks in openCV
Blogpost: `https://www.pyimagesearch.com/2018/02/26/face-detection-with-ope`

Source: https://github.com/opencv/opencv/tree/master/samples/dnn/face_detector

A.2. Landmark Detectors

Dlib implementation of regression trees
Blogpost: http://blog.dlib.net/2014/08/real-time-face-pose-estimation.html

Face Alignment Network (FAN)
Python version used for testing: https://github.com/1adrianb/face-alignment
Paper implementation using LUA (recommended for numerical evaluations): https://github.com/1adrianb/2D-and-3D-face-alignment

Supervision-by-Registration (SBR)
Implementation: https://github.com/facebookresearch/supervision-by-registration
Project website: https://research.fb.com/publications/supervision-by-registration
No pretrained model publicly available. Used one from ILM.

Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks
Just recently made public: https://github.com/FengZhenhua/Wing-Loss Used implementation and models given personally.

OpenFace
Implementation: https://github.com/TadasBaltrusaitis/OpenFace

Look at Boundary: A Boundary-Aware Face Alignment Algorithm
Project website: Implementation:https://github.com/TadasBaltrusaitis/OpenFace Implementation: https://github.com/wywu/LAB

Cascade Multi-view Hourglass Model for Robust 3D Face Alignment
Implementation: https://github.com/jiankangdeng/Face_Detection_Alignment
Models: https://drive.google.com/file/d/1DKTeRlJjyo_tD1EluDjYLhtKFPJ9vIVd/view

Robust Facial Landmark Detection via a Fully-Convolutional Local-Global Context Network
Code and models https://www.mmk.ei.tum.de/cvpr2018/

Bibliography

[BAPD13] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr DollÃąr. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[BBPW04] Thomas Brox, AndrÃl's Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004.

[BJKK13] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing Parts of Faces Using a Consensus of Exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2930–2940, December 2013.

[BM04] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004.

[BRM13] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild. pages 354–361, 2013.

[BT17] 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). *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, October 2017. arXiv: 1703.07332.

[CC06] David Cristinacce and Timothy F. Cootes. Feature detection and tracking with constrained local models. In *Bmvc*, volume 1, page 3. Citeseer, 2006.

[CET01] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.

[DWP10] Piotr DollÃąr, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1078–1085. IEEE, 2010.

[DYW+18] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–368, 2018.

[DZCZ18] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zaferiou. Cascade Multi-View Hourglass Model for Robust 3d Face Alignment. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 399–403, Xi'an, May 2018. IEEE.

[FK18] Zhen-Hua Feng and Josef Kittler. Advances in facial landmark detection. *Biometric Technology Today*, 2018(3):8–11, March 2018.

[FKA+18] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing Loss for Robust Facial Landmark Localisation With Convolutional Neural Networks. pages 2235–2245, 2018.

[GMC+10] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JLM17] H. Jiang and E. Learned-Miller. Face Detection with the Faster R-CNN. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 650–657, May 2017.

[Kin] Davis E King. Dlib-ml: A Machine Learning Toolkit. page 4.

[Kin15] Davis E. King. Max-Margin Object Detection. *arXiv:1502.00046 [cs]*, January 2015. arXiv: 1502.00046.

[KS14] Vahid Kazemi and Josephine Sullivan. One Millisecond Face Alignment with an Ensemble of Regression Trees. pages 1867–1874, 2014.

[KTC00] Takeo Kanade, Yingli Tian, and Jeffrey F. Cohn. Comprehensive database for facial expression analysis. In *fg*, page 46. IEEE, 2000.

[KWRB11] 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 *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[LAE+16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed,

Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[LBL+12] 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.

[LCK+10] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[MRR18] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust Facial Landmark Detection via a Fully-Convolutional Local-Global Context Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 781–790, 2018.

[NYD16] 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.

[SAT+16] 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.

[STZP13] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[SZC+15] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1003–1011, December 2015.

[TP14] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[Tzi15] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2015.

[WBGB16] Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. An anatomically-constrained local deformation model for monocular face capture. *ACM Transactions on Graphics*, 35(4):1–12, July 2016.

[WJ18] Yue Wu and Qiang Ji. Facial Landmark Detection: a Literature Survey. *International Journal of Computer Vision*, May 2018. arXiv: 1805.05563.

[WQY+18] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look

at Boundary: A Boundary-Aware Face Alignment Algorithm. *arXiv:1805.10483 [cs]*, May 2018. arXiv: 1805.10483.

[YLZ17] J. Yang, Q. Liu, and K. Zhang. Stacked Hourglass Network for Robust Facial Landmark Localisation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2025–2033, July 2017.

[ZBM16] Amir Zadeh, Tadas BaltruÅąaitis, and Louis-Philippe Morency. Convolutional Experts Constrained Local Model for Facial Landmark Detection. *arXiv:1611.08657 [cs]*, November 2016. arXiv: 1611.08657.

[ZLL+16] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[ZZLQ16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016. arXiv: 1604.02878.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Eigenständigkeitserklärung

Die unterzeichnete Eigenständigkeitserklärung ist Bestandteil jeder während des Studiums verfassten Semester-, Bachelor- und Master-Arbeit oder anderen Abschlussarbeit (auch der jeweils elektronischen Version).

Die Dozentinnen und Dozenten können auch für andere bei ihnen verfasste schriftliche Arbeiten eine Eigenständigkeitserklärung verlangen.

Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuer und Betreuerinnen der Arbeit.

Titel der Arbeit (in Druckschrift):

Analysis and Improvement of Facial Landmark Detection

Verfasst von (in Druckschrift):
Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.

Name(n):	**Vorname(n):**
Kopp	Philipp

Ich bestätige mit meiner Unterschrift:
- Ich habe keine im Merkblatt „Zitier-Knigge“ beschriebene Form des Plagiats begangen.
- Ich habe alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu dokumentiert.
- Ich habe keine Daten manipuliert.
- Ich habe alle Personen erwähnt, welche die Arbeit wesentlich unterstützt haben.

Ich nehme zur Kenntnis, dass die Arbeit mit elektronischen Hilfsmitteln auf Plagiate überprüft werden kann.

Ort, Datum	**Unterschrift(en)**
28.02.19	P. Kopp

Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich. Durch die Unterschriften bürgen sie gemeinsam für den gesamten Inhalt dieser schriftlichen Arbeit.