

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Head Pose Estimation for Facial Biometric Recognition Systems

João Manuel Guedes Ferreira



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Ana Filipa Pinheiro Sequeira

Co-supervisor: Jaime dos Santos Cardoso

July 3, 2020

Head Pose Estimation for Facial Biometric Recognition Systems

João Manuel Guedes Ferreira

Mestrado Integrado em Engenharia Informática e Computação

July 3, 2020

Abstract

The human face is one of the most used biometric traits for identity recognition. The applications that use this biometric feature are vast and wide, from high-security systems to social-media interactions. With the growing use of the face recognition as an identification tool, several organizations combined their efforts in order to create a standard format that for codifying data describing the human faces effectively, allowing computer analysis of face images for automated face identification and authentication.

The task of verifying the identity of a person against a photograph is not easy, small changes in the image quality or the facial features can lead to a false positive or a false negative identification, making it necessary for investigation to be made in this area.

The performance of head pose estimation in the wild was evaluated in the biometric domain. The most promising state of the art approach was implemented and evaluated based on the compliance to the restrictions for machine readable travel documents.

The results obtained are very promising for a future implementation in the biometric domain, providing a tool that partial automates the head pose estimation task in this domain.

Keywords: Machine Readable Travel Documents, ISO Compliance, Biometric System, Face Analysis, Head Pose estimation, Machine Learning, Deep Learning

Resumo

A cara humana é uma das características biométricas mais usadas para reconhecimento de identidades. As aplicações que tomam partido desta característica biométrica são variadas, desde sistemas de alta segurança até interações em redes sociais. Com o uso de reconhecimento facial a aumentar em ferramentas de identificação, várias organizações uniram os seus esforços de forma em estabelecer normas para armazenamento e descrição de características faciais que permitem automatizar o reconhecimento e autenticação facial em computadores.

A tarefa de verificar a identidade de um sujeito através de uma fotografia não é fácil, pequenas modificações na qualidade da imagem ou características faciais levam à determinação de falsos positivos ou falsos negativos no processo de identificação, torna-se assim necessário investigar este domínio.

O desempenho de estimação de pose em ambientes não controlados foi avaliado no domínio biométrico. O método do estado de arte mais promissor foi implementado e avaliado contra as restrições dos documentos de identificação legíveis por máquina.

Os resultados obtidos são bastante promissores para a futura integração em sistemas biométricos, oferecendo uma ferramenta que automatiza parcialmente a estimação da pose para este domínio.

Acknowledgements

First, I would like to thank INESC-TEC and the CTM team for the fantastic opportunity to develop this thesis immersed in a multidisciplinary environment and full of motivated people. An outstanding thanks to my supervisor, Dr Ana Sequeira, for the immeasurable support and motivation that you gave me throughout this year and for all the flexibility shown, your ideas were fundamental for the accomplishment of this dissertation. An enormous thanks to Professor Jaime Cardoso, my co-supervisor, for all the insights and critical analysis of the work developed. Thank you both for being tireless when dealing with me and my pessimistic nature. Still, inside the group, I would like to thank, João Pinto, for all the patience to resolve my doubts and for all the extraordinary ideas and criticisms that greatly complemented this work.

To Simao and Bia to the immense support given during the period. For the motivation, you provided and above all else for the meals.

To my family, thank you for the unwavering support given during my academic journey, for all always believing in me. For all the selfless help you provided.

To all my friends, for the laughs, the memories and the experiences shared throughout the years.

To BEST Porto, thank for enabling an environment propitious for self-discovery and daring me to grow. For providing me with the best academic experience, one may desire. To everyone who worked alongside me and helped conquer unfamiliar territory.

Lastly, a special thanks to Cátia, just for being there, every time. For emotional support, clarity and unwavering love during these stressful times. You are my rock To everyone mentioned, this thesis would not be possible without your presence in my life. From the bottom of my heart, thank you!

João Ferreira

*“Until I began to learn to draw,
I was never much interested in looking at art.”*

Richard P. Feynman

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	2
1.3	Goals	2
1.4	Structure	2
2	Standards and Guidelines	5
2.1	Introduction	5
2.2	The ISO/IEC 19794–5 document	6
2.2.1	Face Tokenization	7
2.2.2	Frontal Pose	7
2.3	Benchmark	9
2.3.1	Requirements	9
2.3.2	BioLab-ICAO Database	9
2.3.3	Baseline Algorithms and Evaluation	10
3	Head Pose Estimation	15
3.1	Head Pose Representation	15
3.2	Head Pose Estimation Methods	15
3.2.1	Recent Approaches	21
3.2.2	Conclusion	22
3.3	Head Pose Datasets	22
3.3.1	BIWI	23
3.3.2	AFW	24
3.3.3	AFLW	24
3.3.4	300W-LP	24
3.3.5	AFLW2000-3D	25
3.3.6	Synthetic Datasets	25
3.4	Summary	26
4	Methodology	27
4.1	State of the Art method overview	27
4.2	Face Alignment	28
4.3	Regularized Image	30
4.3.1	Shape Alignment	30
4.4	Face Shape Heatmap	32
4.5	Artificial Neural Network used in the implementation	32
4.6	Euler Convention	33

4.7	Implementation Details	33
4.8	Summary	35
5	Experimental Settings	37
5.1	Datasets	37
5.1.1	Training Protocols	37
5.2	Evaluation Metrics	38
5.2.1	Mean Absolute Error	39
5.2.2	Compliance Classification	39
5.3	Evaluation Protocol	41
5.4	Summary	41
6	Experiments and Results	43
6.1	Baseline	43
6.2	Affine Image	44
6.3	Baseline with Heatmap	45
6.4	Affine with Heatmap	46
6.5	Examples Analysis	47
6.6	Compliance Assessment	48
6.6.1	Evaluation over different subsets	48
6.6.2	Evaluation over stricter thresholds	50
6.7	Summary	54
7	Conclusion	55
7.1	Limitations	55
7.2	Future Work	56
	References	57

List of Figures

2.1	Exemples of acceptable and unacceptable images for Machine Readable Travel Documents [18].	7
2.2	Geometric Requirements for a Token Type Image (from ISO 19794-5 [19]).	8
3.1	Representation of the human face with respect to the coordinate system and the Yaw, Pitch and Roll represented as rotation angles.	16
3.2	Appearance Template methods [33]	17
3.3	Detector Arrays Methods [33]	17
3.4	Nonlinear Regression Methods [33]	18
3.5	Manifold Methods [33]	19
3.6	Flexible Model Methods [33]	19
3.7	Geometric Methods [33]	20
3.8	Tracking Methods [33]	20
3.9	Hybrid Methods [33]	21
3.10	Example frames from the Biwi Kinect Head Pose Database. Both depth and RGB images are present in the dataset, annotated with head poses. [9]	23
3.11	Example frames from the 300W-LP dataset. [54]	25
4.1	Method proposed by <i>Xia et al.</i> [46]	28
4.2	Canonical face shape side by side with a face with overlaid landmarks	29
4.3	Example of the steps of the pipeline for intercepting the Face in an image with a example	29
4.4	Examples of affine transformation, the first row contains the original images and the second row the images after transformation (affine images).	31
4.5	Pipeline for generating the heatmap image	32
4.6	Structure of the adapted VGG16 [51] to head pose estimation problem	34
4.7	Pre Processing Pipeline	35
6.1	MAE of the baseline model over the restricted test set	44
6.2	Mean Absolute Error for the Affine model over the restricted test set	45
6.3	Mean Absolute Error for the Baseline with Heatmap model over the restricted test set	46
6.4	Mean Absolute Error for the Affine with Heatmap model over the restricted test set	47
6.5	Estimation of head pose in AFLW200-3D	49
6.6	Accuracy plotted over the test set with samples restricted in degrees	50
6.7	Accuracy plotted over different thresholds	52
6.8	ROC plot for Compliant Classification	52
6.9	Examples of compliance classification and the comparison between models	53

List of Tables

2.1	Tests to evaluate systems for the ISO/IEC 19794 Compliance [11]	11
2.2	Published Results of Facial Conformance Software, the first three are from the work published by [11] while the remaining are extracted from the FVC Ongoing Website	12
2.3	Equal error rate of each requirement and difficulty ordered in descent	14
3.1	Reported Mean Absolute Error (MAE) of the 3 angles(Yaw, Pitch, Roll) in degrees on the AFLW2000 and BIWI datasets.	21
3.2	Overview of Head Pose Datasets	23
5.1	The hyperparameters used for training every experiment	38
6.1	Results for baseline architecture evaluated on AFLW2000-3D segmented to 30° .	43
6.2	Results for affine image architecture evaluated on AFLW2000-3D segmented to 30°	44
6.3	Results for normal image with heatmap architecture evaluated on AFLW2000-3D segmented to 30°	45
6.4	Results for affine image with heatmap architecture evaluated on AFLW2000-3D segmented to 30°	47
6.5	MAEs for the models in degrees evaluated on the AFLW2000 segmented to 30° .	47

Abbreviations

ISO	International Standard Organization
NTWG	New Technologies Working Group
ICAO	International Civil Aviation Organization
MRTD	Machine Readable Travel Documents
YPR	Yaw Pitch Roll
HPE	Head Pose Estimation
CNN	Convolutional Neural Networks

Chapter 1

Introduction

1.1 Context

The human face is one of the most used biometric traits for human recognition. The applications that use this biometric feature are vast and wide, from high-security systems to social-media interactions. Comparing to other biometric traits like fingerprint, iris or voice, the face is one of the most flexible biometric identification methods. This biometric trait does not require user cooperation and can be used even when the subject is not aware that is being scanned. Another advantage of face based systems when compared to other methods is the limited collection time. Qualifying the human face for identification of large masses of humans that only stay in front of a sensor for a short period of time.

Facial recognition systems is based on analyzing certain features that are common to every individual's face and then using these features to differentiate each individual from the others. These features include the distance between the eyes, the position of the cheekbones or the chin, the jaw line, the width of nose, the shape of mouth, among others. The idea behind the use of face as a modality for biometric recognition is that every individual can be uniquely identified by combining the numerical quantities obtained from these anatomical characteristics into a single code - the biometric template.

These systems can be used to automatically identify or verify the identity of an individual from a digital image or a video frame. This is achieved by comparing the selected facial features from the image with a stored facial sample from a database.

Advances in recognition algorithms and approaches to facial recognition have greatly improved the matching accuracy over the last few years. Nevertheless, recognition errors remain significantly above zero, especially in systems where the image quality is substandard or where strict thresholds must be applied to reduce false positives.

This thesis explores the head pose estimation task in context to the biometric domain. The head pose estimation is one of the more challenging tasks in biometric analysis presenting high error in unconstrained conditions.

1.2 Motivation

The estimation of the 3D pose of the human head from sample images is a difficult problem that has numerous applications. The head pose estimation can be used as a pre-processing step for face recognition. Further applications include face analysis, driver monitor systems and human computer interaction. The performance of these systems are correlated with robustness of the head pose estimation techniques employed.

While humans quickly learn to estimate the orientation of the head very early in their life, computer vision systems have to overcome various challenges that puzzled researchers for decades. At present, the head pose estimation methods present excellent performance under very specific conditions, however when these conditions vary, performance is less than ideal in uncontrollable environments or commonly referred as in-the-wild. Approaches should demonstrate robustness to various factors like lighting, noise, distortion and occlusion. None of these challenges reduced momentum in the development of varied approaches to tackle the head pose estimation task. Some approaches focus on classifying pose in different discrete intervals while others aim to estimate the continuous angles for the head.

1.3 Goals

This dissertation investigates current state of the art approaches to head pose estimation. The implementation of state of the art approaches for head pose estimation is driven by following questions :

1. Are current state of the art methods based on in the wild datasets robust enough to be applied to the biometric domain ?
2. What is the expected error when pose angles are estimated ?
3. Can the system reliably classify faces according to the demands imposed to face images for machine readable travel documents ?

While its is possible to design a system that is dedicated to each of these goals, this dissertation aims to evaluate the head pose in the continuous and classification space.

1.4 Structure

The previous sections defined the context, motivation and goals of this dissertation. The structure of the remainder of this document is described as follows.

Chapter 2 presents the International Guidelines and operating characteristics for head pose estimation in the biometric domain.

Chapter 3 refers to the basic principles of head pose estimation, presents analysis the state of the art and selects the most performant algorithm suitable for this thesis goals.

Chapter 4 presents details regarding the methodology used for our problem and the preprocessing is described.

Chapter 5 presents the details regarding the experimental settings of our experiments, the training procedure is defined as well as the evaluation metrics.

In Chapter 6 the final results of the head pose estimation performance are presented. Finally, in Chapter 7 a summary of the thesis and results is given. The most relevant results are summarized in order to emphasize the strengths and weaknesses of the state of the art. Additionally, an outlook of the future work, which provides some avenues for possible improvement of the system is presented.

Chapter 2

Standards and Guidelines

2.1 Introduction

Starting in 1968 the International Civil Aviation Organization (ICAO), formed the ICAO's New Technologies Working Group (NTWG) with the purpose of "developing recommendations for a standardized passport book or card that would be machine readable, in the interest of accelerating the clearance of passengers through passport controls". The guidelines developed by this group were presented to the public in 1980, in Doc 9303, titled "A Passport with Machine Readable Capability". The publication of this document was the first effort towards a standard in machine readable travel documents.

Since then, ICAO worked to refine the concepts of machine readable documents, expand the use of such cards and increase the quality of the documents themselves with the subsequent release of updated editions.

In 2002, the NTWG announced the Berlin Resolution. The Berlin Resolution states that:

1. The face is chosen as the primary mechanism used as the globally interoperable biometric for machine assisted identity confirmation in Machine Readable Travel Documents.
2. The Member States can optionally use fingerprint and/or Iris recognition as additional biometric technologies in support of machine-assisted identity confirmation.

With the work developed previously by ICAO and in a cooperative effort with the International Standard Organization (ISO), a new initiative was created in order to work on specifications to strengthen the security and integrity of travel documents. In 2006, ISO published the standard ISO/IEC 19794 that specifies rules for encoding, recording, and transmitting the facial image information and defines scene constraints, photographic properties, and digital image attributes of facial images.

2.2 The ISO/IEC 19794–5 document

For decades, face images have been used to verify the identity of human beings. In recent years, these images are being replaced from traditional photographs to digital images.

The number of international airline travelers is increasing steadily due to the global economy and tourism. In addition, airport security is getting tighter every year. Therefore it is necessary to process identity documents like passports much quicker, more efficiently and possibly automatically while increasing security aspects. A special division of the ICAO for Machine Readable Travel Documents (MRTD) has therefore created an international standard (ISO/IEC 19794 [19]) for these documents in order to make them easier to read and process for machines.

The document ISO/IEC 19794 part 5 defines standard information format for facial images. Not only specifies data format but also provides guidelines regarding:

- **Scene constraints** define which requirements the person in the photo must fulfill:
 - A full-face frontal pose with a maximum deviation of ± 8 degrees for Yaw, Pitch and Roll angle from frontal pose. Where the Yaw angle is rotation about the vertical axis, Pitch the rotation along the horizontal side-to-side axis and Roll a rotation about the horizontal back to front axis.
 - A neutral expression with both eyes open and looking into the camera as well as a closed mouth is required.
 - The ICAO standard contains no limitations for the background. It only states that the background should allow a clear separation of the head. It should therefore be uniform with no visible shadows.
 - There should be no significant direction of the light visible in the picture. It is essential that no shadows occlude facial features.
 - Eye glasses are allowed as long as they are transparent and do not contain lighting artifacts like reflections. Also, the frames must not occlude the eyes.
- **Photographic constraints** define the requirements under which the picture must be taken. The image must be correctly exposed so that gradations in the texture are visible in all areas. The subject's eyes, nose, ears and chin must be in focus while a blurred background is allowed.
- **Digital image constraints** define how the image needs to be recorded and stored. Gray scale images must have at least a 7 bit intensity variation. Color images must allow a gray scale conversion that fulfills the first requirement.

ICAO completed the ISO standard publishing in Document 9303, part 3, volume 1, examples of acceptable or not acceptable photographs for travel document based on this standard (Figure 2.1). These guidelines are employed in the bureaucratic process of obtaining a new identification cards and passports. Most of the time, photograph of the subject the taken by a human operator with the help of a specialized tool.



Figure 2.1: Examples of acceptable and unacceptable images for Machine Readable Travel Documents [18].

2.2.1 Face Tokenization

A very important definition in the ICAO standard is the Token Face Image Type. It sets the geometric constraints about where the face must be located in a tokenized image. The only facial features used for this definition are the position of both eyes of the person in the image.

The tokenization procedure makes processing of the face by a computer very easy because many facial features can be found at almost the same position and scale in any tokenized image. Also, it provides a standardized image for which pose estimation can be performed.

2.2.2 Frontal Pose

According to the ICAO definition, the facial image must be a frontal face which means that the pitch and yaw angle of the face must be within a 8 degree range of the perfect frontal pose. Unfortunately the ICAO specification lacks one essential definition: the frontal pose itself. While there should be no head rotations in this pose, there is no written definition about what zero degree angles mean. Therefore the following convention is used in this dissertation:

- The yaw angle is zero, when the face shows a maximum symmetry along the vertical nose-mouth line.
- The pitch angle is zero when the eye corners are on the same height as the upper part of the ear. The easiest way to visualize this is to think of a person wearing glasses. Then the temple of the glasses must be horizontal.
- The roll angle is zero when the connecting line between the the two eye centers is horizontal.

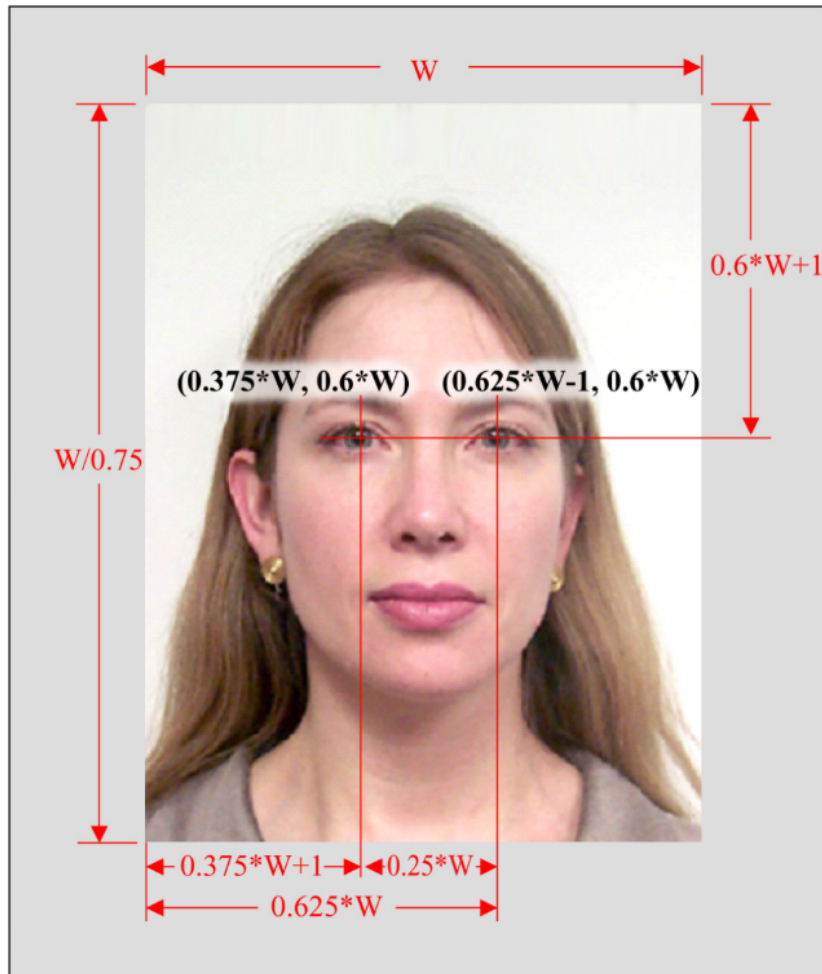


Figure 2.2: Geometric Requirements for a Token Type Image (from ISO 19794-5 [19]).

2.3 Benchmark

Ferrara et al. published a series of articles [11][10][12][27] on creating a framework that allows the evaluation of systems to the conformance of facial images to the International Standards. Directly from the articles we can extract 3 main contributions to this field:

1. Objective Requirements.
2. Dataset in compliance to ISO/ICAO standards.
3. Baseline Algorithms and evaluation.

2.3.1 Requirements

The authors states the guidelines presented by ICAO and ISO as

"quite generic guidelines and several examples of acceptable/unacceptable face images, a clear and unambiguous description of all the requirements is still missing"

With both documents not defining objective requirements, it is especially hard to create a system that can evaluate systems to the adherence of the international guidelines. The necessity of specific requirements was on the challenges for *Ferrara et al.* towards creating the benchmark. A list of objective requirements was directly derived from the generic guidelines provided in the standard documentation. The list of requirements contains the requirement specification and respective section with further information about the requirement. These requirements are presented in Table 6.5.

2.3.2 BioLab-ICAO Database

Another major contribution by the team was the creation and distribution of a database annotated with ground truth data for each of the requirements defined in Table 6.5.

Unfortunately this dataset is no longer available to the community nor neither training or testing.

The dataset created by *Ferrara et al.* was a collection of images from publicly available datasets with annotations for the compliance evaluation. The database consisted of 5588 images from 601 subjects, gathered from the following sources.

- 1741 images from the AR database [28].
- 1935 images from the Face Recognition Grand Challenge [36].
- 291 images from the PUT dataset [20].
- 804 synthetic expanded images from compliant images in the AR database by applying ink-marked/creased, pixelation and wash-out effects.

- 817 newly acquired images.

The resulting dataset and annotations have 310 fully compliant images to the standards and 5278 not compliant to one or multiple requirements. Despite of the unavailability of the datasets for developing the algorithms, the testing platform is still online and allows the submission of software to evaluate the conformance to the ICAO guidelines.

2.3.3 Baseline Algorithms and Evaluation

Ferrara et al. [11] present a baseline algorithm for each of the defined requirement and performed a comparison with the SDKs commercially available at the time of publication.

Since the benchmark was made public sometime after the article publication, there were a small number of biometric software providers that evaluated their software on the proposed platform. The results of the baseline algorithms implemented by *Ferrara et al.*, the SDKs evaluated and further assessments published in the FVC Ongoing website are present in Table 2.2. Despite the performance results being publicly available, none of the companies disclosed information regarding the algorithms used in each requirement besides the baseline algorithms evaluated in [11].

N°	Description of the test	Section
Feature extraction accuracy tests		
1	Eye Location Accuracy	
2	Face Location Accuracy (other points)	
Geometric tests (Full Frontal Image Format)		
3	Eye Distance (min 90 pixels)	8.4.1
4	Relative Vertical Position ($0.5B \leq BB \leq 0.7B$)	8.3.3
5	Relative Horizontal Position (no tolerances)	8.3.2
6	Head Image Width Ratio ($0.5A \leq CC \leq 0.71A$)	8.3.4
7	Head Image Height Ratio ($0.7B \leq DD \leq 0.8B$)	8.3.5
Feature extraction accuracy tests		
8	Blurring	7.3.3
9	Looking Away	7.2.3
10	Ink Marked/Creased	A3.2.3
11	Unnatural Skin Tone	7.3.4
12	Too Dark/Light	7.3.2
13	Washed Out	7.4.2.1
14	Pixelation	A3.2.3
15	Hair Across Eyes	A3.2.3
16	Eyes Closed	7.2.3
17	Varied Background	A2.4
18	Roll/Pitch/Yaw Greater than 8	7.2.2
19	Flash Reflection on Skin	7.2.10
20	Red Eyes	7.3.4
21	Shadows Behind Head	A3.2.3
22	Shadows Across Face	7.2.7
23	Dark Tinted Lenses	7.2.11
24	Flash Reflection on Lenses	7.2.11
25	Frames too Heavy	A4.3
26	Frame Covering Eyes	7.2.3
27	Hat/Cap	A3.2.3
28	Veil over Face	A3.2.3
29	Mouth Open	7.2.3
30	Presence of Other Faces or Toys too Close to Face	A3.2.3

Table 2.1: Tests to evaluate systems for the ISO/IEC 19794 Compliance [11]

Req.	SDK1		SDK2		BioLab		biometrika (2014)		id3 Technologies (2016)		VSOFT (2017)	
	EER	Rej.	EER	Rej.	EER	Rej.	EER	Rej.	EER	Rej.	EER	Rej.
Blurred	26.00%	8.90%	48.10%	0.60%	5.20%	0.00%	30.50%	36.00%	1.70%	0.20%	1.60%	3.30%
Looking Away	27.50%	7.10%	-	-	20.60%	0.00%	24.20%	3.10%	15.30%	15.80%	13.30%	3.30%
Ink Marked/Creased	-	-	-	-	3.40%	1.20%	3.60%	1.40%	-	-	4.80%	0.50%
Unnatural Skin Tone	18.70%	4.80%	50.00%	0.80%	4.00%	0.20%	5.10%	1.70%	2.10%	0.20%	1.90%	0%
Too Dark/Light	-	-	-	0%	4.20%	0.00%	4.60%	0.20%	2.90%	0%	3.10%	0.20%
Washed Out	-	-	3.10%	0%	9.60%	0.00%	9.20%	0%	0.20%	0%	0%	0%
Pixelation	-	-	40.80%	0.00%	1.30%	0.00%	32.40%	0.60%	0.20%	0.40%	1.30%	0%
Hair Across Eyes	50.00%	81.90%	-	-	12.80%	0.00%	12.40%	4.60%	-	-	13.00%	6.30%
Eyes Closed	2.90%	3.10%	-	-	4.60%	0.00%	6.70%	7.10%	0.20%	1.00%	4.60%	4.00%
Varied Background	7.50%	3.30%	17.90%	1.40%	5.20%	0.00%	3.70%	7.90%	-	-	5.20%	0.40%
Roll/Pitch/Yaw >8°	-	-	26.00%	2.90%	12.70%	0.20%	12.60%	3.80%	9.10%	6.90%	10.70%	1.20%
Flash Refl. on skin	5.00%	2.70%	50.00%	7.50%	0.60%	0.00%	1.20%	0.40%	1.70%	0.60%	1.40%	2.50%
Red Eyes	5.20%	4.50%	34.20%	0.00%	7.40%	0.00%	10.30%	8.40%	1.00%	2.00%	1.70%	0%
Shadows Behind Head	-	-	-	-	2.30%	0.20%	2.40%	7.90%	-	-	5.40%	8.40%
Shadows Across Face	36.40%	8.10%	-	-	13.10%	0.40%	15.90%	19.80%	10.50%	1.20%	9.90%	0.60%
Dark Tinted Lenses	-	-	-	-	1.90%	0.20%	2.10%	1.20%	2.70%	20.40%	1.80%	1.20%
Flash Refl. on Lenses	-	-	-	-	2.10%	0.00%	2.30%	0%	-	-	2.70%	0.80%
Frames Too Heavy	-	-	-	-	5.80%	0.00%	3.30%	8.40%	1.40%	15.80%	2.10%	12.60%
Frame Covering Eyes	50.00%	62.30%	-	-	6.30%	0.00%	4.00%	31.90%	6.60%	2.30%	10.70%	13.80%
Hat/Cap	-	-	-	-	14.00%	0.00%	16.50%	21.60%	6.80%	0.80%	9.80%	0.40%
Veil Over Face	-	-	-	-	2.50%	0.00%	3.70%	0%	-	-	1.40%	4.80%
Mouth Open	3.30%	52.00%	-	-	6.20%	0.00%	5.00%	2.70%	0.60%	0.40%	3.80%	0%
Objects Close to Face	-	-	-	-	21.60%	0.00%	15.40%	14.20%	-	-	1.20%	2.70%

Table 2.2: Published Results of Facial Conformance Software, the first three are from the work published by [11] while the remaining are extracted from the FVC Ongoing Website

The authors also argue that the requirements evaluated are separated in 3 groups based on assessment difficulty.

- $ERR < 3\%$ are relative easy to verify
- $3\% \leq ERR \leq 7\%$ are of medium difficult
- $ERR > 7\%$ are difficult to evaluate

With the software evaluations present in Table 2.2, it is highlighted the software provider that presents the lowest equal error rate (ERR) for each requirement. This observation is summarized in Table 2.3, where the requirements and their lowest error are ordered according to the assessment difficulty defined [11].

Despite the major advances in computer vision and deep learning techniques, the results reported on the platform do not seem to reflect the advances in the field. And the progress towards a robust system that can reliably assess the compliance of face images is not ideal.

In the Table 2.2, we can clearly see that the errors reported in some requirements did not improve, especially requirements of difficult category. The requirement "Hair across eyes" only saw a reduction of 0.40% in accuracy, the "Shadows Across Face" and "Roll/Pitch/Yaw" constraints saw a 4% in error reduction, in over 10 years.

The lack of progress may be correlated with the fact that there is no available information regarding the methods each of the companies used in their evaluations, the database used to perform the benchmarks is no longer distributed.

Requirement	ERR	Category
Looking Away	13.30%	Difficult
Hair Across Eyes	12.40%	Difficult
Shadows Across Face	9.90%	Difficult
Roll/Pitch/Yaw >8°	9.10%	Difficult
Hat/Cap	6.80%	Medium
Frame Covering Eyes	4.00%	Medium
Varied Background	3.70%	Medium
Ink Marked/Creased	3.40%	Medium
Too Dark/Light	2.90%	Easy
Shadows Behind Head	2.30%	Easy
Flash Refl. on Lenses	2.10%	Easy
Unnatural Skin Tone	1.90%	Easy
Dark Tinted Lenses	1.80%	Easy
Blured	1.60%	Easy
Frames Too Heavy	1.40%	Easy
Veil Over FAcE	1.40%	Easy
Objects Close to Face	1.20%	Easy
13Red Eyes	1.00%	Easy
Flash Refl. on skin	0.60%	Easy
Mouth Open	0.60%	Easy
Pixelation	0.20%	Easy
Eyes Closed	0.20%	Easy
Washed Out	0%	Easy

Table 2.3: Equal error rate of each requirement and difficulty ordered in descent

Chapter 3

Head Pose Estimation

This Chapter provides relevant background information on head pose estimation concepts, methods and existing databases.

3.1 Head Pose Representation

The position of the human face in the 3-dimensional space is traditionally defined by Euler Angles. Euler Angles describe the orientation of a rigid object with respect to a fixed coordinate system. To represent head pose, is used the Tait-Bryan notation, the notation of Euler angles commonly used in the aerospace industry. The advantage of the Tait-Bryan notation is the fact that zero degree in elevation represents the horizontal position of the head pointing forward. Figure 3.1 shows the head represented as a rigid object with the yaw, pitch, roll as rotation angles.

The research by *Ferrario et al.* [13] shows the range of head motion among average health young adults. The mean ranges derived from the study are the following :

- Yaw angle: -79.8° to $+75.3^\circ$
- Pitch angle: -60.4° to $+69.3^\circ$
- Roll angle: -40.9° to $+36.3^\circ$

Since the yaw angle represents the axis which is most flexible in terms of range of motion, it is usually the axis used to compare various approaches to head pose estimation.

3.2 Head Pose Estimation Methods

The survey performed by *Murphy et al.* [33] defined head pose estimation methods in 8 categories (Appearance Template Methods, Detector Array Methods, Non Linear Regression, Manifold Methods, Flexible Models, Geometric Models, Tracking and Hybrid), these methods are briefly described in this section, a more extensive description and research can be found in the original survey [33].

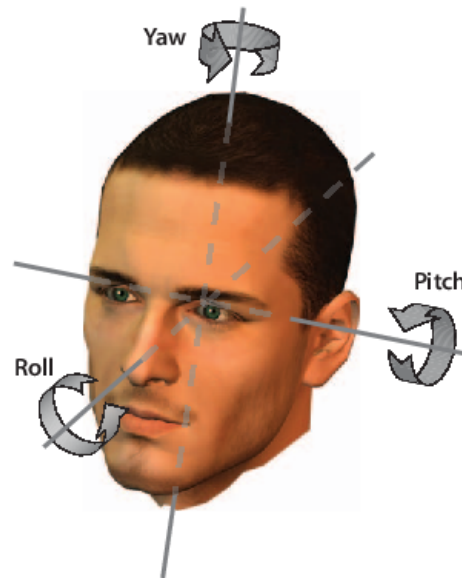


Figure 3.1: Representation of the human face with respect to the coordinate system and the Yaw, Pitch and Roll represented as rotation angles.

Generally, methods can be categorized in two main approaches, methods that are based on landmarks, and methods that do not need landmarks annotations to estimate pose. For the context of this thesis, methods that use landmark information are preferable, since this information is required to perform other assessments tasks, for example check if the mouth is open, if the eyes are closed or to measure the distance between the eyes.

3.2.0.1 Appearance Template Methods

The appearance template methods compare a new view of a head to a set of exemplar templates annotated with poses. A new image is overlaid and evaluated with the comparison metrics with the templates to find the most similar template. The overlay can be processed under different evaluation metrics to find the best match. Beymer approach is to calculate mean squared error on sliding windows [4]. The feature finder detects the two eyes and at least one nose feature. To reduce the search space, detected features in the window are hierarchically searched in course-to-fine strategy over the tree shaped head template pool. At any level a branch is pruned if the template correlation values are not above a threshold.

The advantages of template appearance methods is the relative simple implementation, the template pool is flexible to be expanded or reduced in regard to the needs of the domain, do not require facial feature points or training examples.

The accuracy of the head pose task is inherently attached to the accuracy of the matching to the templates. The main drawback of this approach is the assumption that similar images also have similar poses. The head pose is estimated based on the success of this match as the templates are annotated with discrete pose angles.

To decrease this problem, the input images are often filtered with Laplacian-of-Gaussians [14] or Gabor-wavelets [41][45].

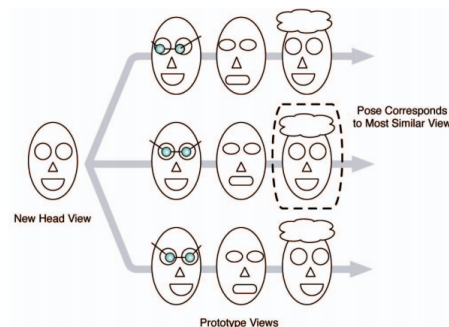


Figure 3.2: Appearance Template methods [33]

3.2.0.2 Detector Array Methods

The array methods is the classification approach to the problem using machine learning techniques. A series of head pose detectors are trained to different discrete poses. Figure 3.3 presents this approach.

Detector Arrays are similar to Appearance Template Methods. The later compares the input image to a poll of template faces while the former the input image is evaluated on a neural network trained a large dataset with supervised learning [17].

The biggest advantage over template methods is that learning algorithms tend to be more robust against appearance variance and put more emphasis on pose related features. Nevertheless it can be difficult to train two classifiers for similar poses.

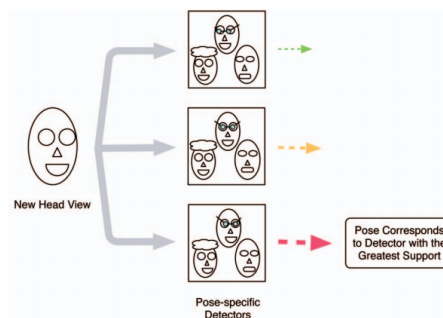


Figure 3.3: Detector Arrays Methods [33]

3.2.0.3 Nonlinear Regression Methods

The Nonlinear regression method, as pictured in Figure 3.4, estimates pose from a image or facial feature set by learning a non linear functional mapping between the domains. In this approach a model can be trained on a set of labeled data to predict the discrete or continuous head pose for new data.

Different techniques can be used. In the earlier years multilayer perceptrons and support vector regressors were the most common. To the best of our knowledge the first approaches to this problem using neural networks were proposed by Shiele and Waibel [40] in 1995 to predict the yaw angle of the head orientation. Rainer Stiefelhagen [42] also use a neural network with one hidden layer to learn the yaw and pitch angle.

The other approach, support vector regressors was successfully used in combination with localized gradient orientation histograms [31] and dimensional reduction methods such as principal component analysis [26].

The main disadvantage of these methods is that they are prone to error from a faulty head localization. To suppress this error, convolutional neural networks were used due to their high tolerance shift and distortion variance. One of the first methods using convolutional neural networks was proposed by *Osadchy et al.* [35]

These methods have innumerable advantages. The methods are very fast, work well both in near field and far field imagery, only require labeled face images for training and according to the research are most accurate.

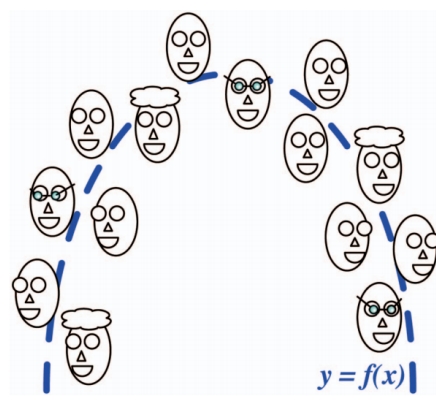


Figure 3.4: Nonlinear Regression Methods [33]

3.2.0.4 Manifold embedding Methods

These methods assume that even though an image consists of hundreds of dimensions spanned by the pixels of the image, only a few dimensions define the pose [29]. Thus they map the image to a low-dimensional manifold that is defined by the continuous pose. A good algorithm can do this mapping without the influence of variations in faces. The biggest problem with manifold learning is that it is unsupervised and thus may learn the wrong features. Several approaches have been developed that overcome this problem by making the learning supervised and they show promising pose estimation results [2].

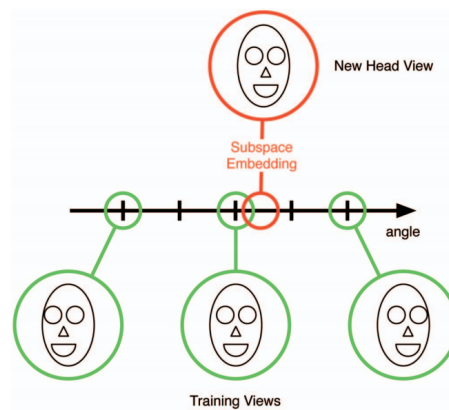


Figure 3.5: Manifold Methods [33]

3.2.0.5 Flexible models

Other approaches treat the head pose estimation as a classification, regression or geometrical problem. Flexible models take a different approach. These methods iteratively fit a non-rigid model of the human head to the facial features of the subject, as presented in Figure 3.6. For example, graphs of facial features can be deformed until they fit a face [45]. The successful training of these approaches require datasets with the pose annotations as well as landmark annotations, such training set allows the comparison at local feature level instead of global appearance level. The main advantage is that a precise localization of head features is not required initially since these algorithms are able to adapt to the optimal positions. Yet all required facial features must be visible and detectable.

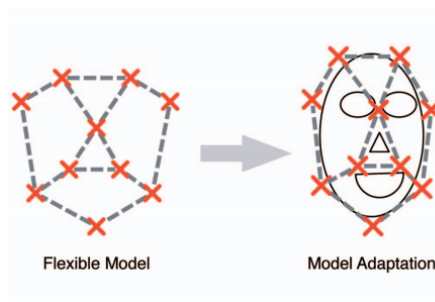


Figure 3.6: Flexible Model Methods [33]

3.2.0.6 Geometric Methods

Methods based on this approach use facial landmarks, geometric measures and models of the face to estimate pose. This approach is based on the assumption that humans determine head orientation based on bi-lateral symmetry properties of the human head [44]. Another approach to geometric models is in the baseline algorithm described in *Ferrara et al.* [11] research.

The advantage of geometric models is that they are fast and simple. But they require accurate localization of facial features like eye corners which may be occluded by glasses, or mouth corners which may or may not exist depending on facial expressions.

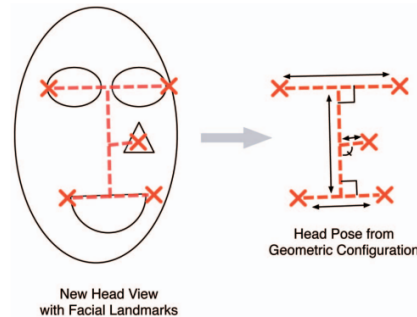


Figure 3.7: Geometric Methods [33]

3.2.0.7 Tracking Methods

Tracking Methods, as the name suggests, tracks the movement of the head between the sequence frame of a video to estimate pose over time (Figure 3.8). The survey also shows, at the time of publication, that tracking methods are significantly more accurate than other methods [32][34]. But these methods have a major limitation by design, since the algorithm requires the initialization of a known head pose.

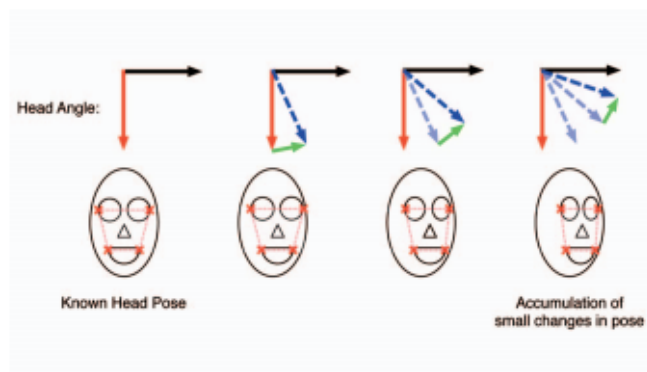


Figure 3.8: Tracking Methods [33]

3.2.0.8 Hybrid Methods

These methods represent approaches that used any combination of the previous methods presented in Figure 3.9 for representation. Hybrids approaches foster the achievement of better results through the combination of several approaches. Despite the fact that this strategy allows to overcome limitations of the methods individually. The process of selecting and implementing multiple methods must be done with care as the final algorithm could be very complex.

Method Name	Year	MAE on the AFLW2000	MAE on the BIWI
KEPLER [25]	2017	-	13.9°
3DDFA [54]	2017	7.39°	-
HOPENET [38]	2018	6.16°	5.18°
FSA-NET [48]	2019	5.07°	4.00°
Landmark assisted CNN [46]	2019	1.46°	-

Table 3.1: Reported Mean Absolute Error (MAE) of the 3 angles(Yaw, Pitch, Roll) in degrees on the AFLW2000 and BIWI datasets.

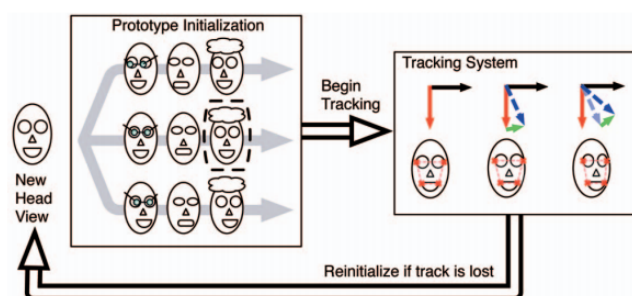


Figure 3.9: Hybrid Methods [33]

3.2.1 Recent Approaches

The previous taxonomy was presented in the survey published in 2008 by *Murphy et al.* [33]. Since the publication, the state of the art methods have converged to approaches that leverage the use of the neural networks mainly due to the increase in computational power available and the high accuracy of these methods. This section refers to the most successful methods in the recent years. An overview of these methods is presented in Table 3.1.

There are several recent approaches based on different methodologies. These works use mainly two datasets, the BIWI [9] and AFLW2000 [54].

HopeNET [38] is based on ResNet50 architecture combined with three mean squared error and cross entropy losses, one for each degree of freedom. This method reports a Mean Absolute Error (MAE) of 6.16 and 5.18 on the AFLW2000 and BIWI dataset respectively.

FSA-Net [48] is a method that uses direct regression without landmarks on a single image. This method proposes to learn feature mapping and feature aggregation to estimate pose. The MAE of this method is 5.07 on the AFLW2000 and 4.00 on the BIWI dataset.

Kumar et al. [25] present a novelty architecture called Heatmap-CNN, which can learn local and global structure dependencies. This method is used to predict the facial landmarks as the main task and extrapolate the head pose as a bi product of this method. This method was only evaluated on the BIWI dataset and reports a MAE of 7.39.

Another approach was proposed by *Xiangyu Zhu et al.* [54] where they fit a non-rigid 3D model of the human to the input image, they report good results on the AFLW2000 and BIWI datasets. A better description of this method is presented in [54]. The reported results for this method is a MAE of 7.39 in the AFLW2000 dataset.

A promising approach presented by *Xia et al.* [46] achieved the best reported results on the AFLW2000 and BIWI challenging datasets. The authors achieve this by firstly use standard methods for face alignment to extract the landmarks; with the landmarks they create a heatmap of the landmarks. Once the heatmap is obtained, a stacking of the image and heatmap is done and set as the input of the CNN. The introduction of the heatmap makes the CNN focus on the areas around the landmarks and as such reduce the interference of background information present in the image. This method present a Mean Absolute Error of 1.46 on the AFLW2000 benchmark.

For the context of this dissertation this method presents very promising results as the consistent accuracy of head pose estimation is necessary to assess if a image is compliant to the standards guidelines. While the average error of the other methods stay around 5 degrees, this last method decreased the error in 4 degrees making possible the applicability of such model to our target threshold.

3.2.2 Conclusion

Based on the study of a thorough evaluation of research in this field [33] it was concluded that there is there is a variety of different approaches that are possible to solve the problem efficiently. Each approach presents their strengths and their weaknesses. The next step is to select the approach that can meet the goals of the context of this dissertation. One of the criteria defined was to choose a method that estimates the continuous angle of head pose on the three degrees of freedom in facial images. Another criterion is to select a method that uses landmark, as the conformance system will require these keypoints to assess other requirements. With this context in mind the method proposed by *Jiahao Xia et al.* [46] seems the most appropriate choice, as the method utilizes landmarks and provides a MAE of less than 2° .

3.3 Head Pose Datasets

Head pose estimation datasets have changed from a few years ago, due to governmental regulations and technological advances. In the earlier years the datasets used were created in laboratories around the world under very specific conditions such as the Point'04 dataset [15].

The systems that use these datasets should be robust enough for real-life applicability, as such the design of new datasets moved from laboratories to use real-life images for training and testing. Most of the datasets described in this section are of such nature and called in-the-wild datasets.

The datasets in this section also have another property worthy mention. Some datasets are synthetic or synthetic-expanded. The Synthetic dataset used a average 3D model of the face to create multiple 2D views of the face to use for training while synthetic expanded datasets use real

Dataset	Yaw	Pitch	Roll	Angles values	N ^o of Images
AFW [53]	$\pm 105^\circ$	$\pm 45^\circ$	$\pm 105^\circ$	Discrete in 15° steps	468
BIWI [9]	$\pm 60^\circ$	$\pm 75^\circ$	-	Continuous	15 678
AFLW [22]	$\pm 170^\circ$	$\pm 90^\circ$	$\pm 170^\circ$	Continuous	25 993
300W-LP [54]	$\pm 90^\circ$	$\pm 40^\circ$	$\pm 45^\circ$	Continuous	122 450
Synhead [16]	-	-	-	Continuous	51 096
Synthetic Dataset [43]	$\pm 75^\circ$	$\pm 50^\circ$	$\pm 20^\circ$	Continuous	310 000

Table 3.2: Overview of Head Pose Datasets

life images and perform transformations to the source dataset to create a new dataset with a bigger sample pool.

A description of the most commonly used datasets is present in this section and an overview of the datasets analyzed is present in Table 3.2.

3.3.1 BIWI

The BIWI [9] dataset contains 15,678 images of 20 subjects (16 male, 4 female and 4 of the 20 subjects were using glasses) recorded from 1 meter away with a Kinect Camera. The dataset provides RGB images, depth information, and head pose estimation that ranges from $\pm 75^\circ$ on yaw and $\pm 60^\circ$ for the pitch angle. The dataset can be downloaded for research purposes¹. The ground truth annotations were created by the facial motion capture software Faceshift, unavailable since 2015 as Disney bought the company². The software captures and describes a person facial movement, head pose and eye gaze. The information captured can then be translated to virtual characters to be used in movies, animations or games.

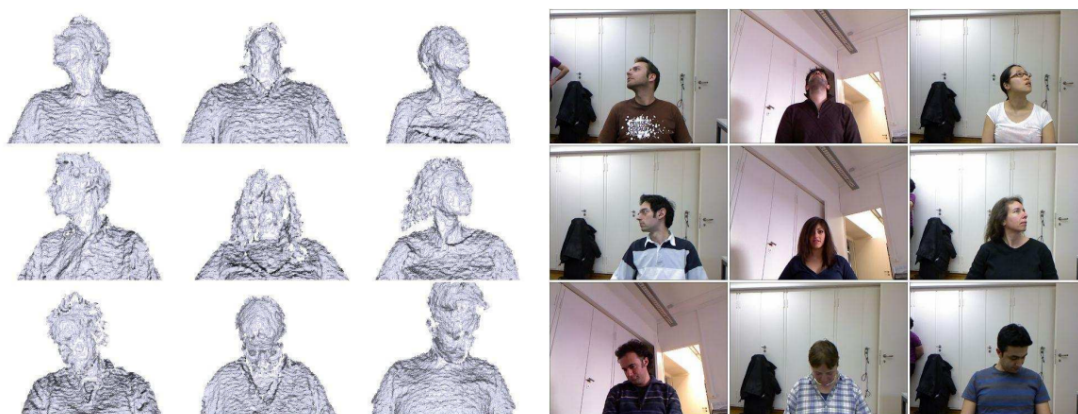


Figure 3.10: Example frames from the Biwi Kinect Head Pose Database. Both depth and RGB images are present in the dataset, annotated with head poses. [9]

¹https://data.vision.ee.ethz.ch/cvl/gfanelli/head_pose/head_forest.html

²<https://www.alphr.com/apple/1002098/apple-buys-star-wars-motion-capture-firm-faceshift>

3.3.2 AFW

To the best of our knowledge, this dataset is already deprecated or impossible to obtain. In spite of this, it is worth a mention as it was one of the first examples of an in-the-wild dataset.

AFW means Annotated Faces in-the-wild, in [53] was introduced as a novelty dataset for head pose estimation. Subsequently this dataset was used by some authors, as such the description of this dataset is widely available.

The dataset provides 468 face images and annotated with five landmarks (the center of eyes, tip of nose, the two corners and center of mouth). The head pose information ranges from $\pm 90^\circ$ degrees in yaw, $\pm 90^\circ$ degrees in Pitch and left, center, right view points in Roll. The angles represented in this dataset are discrete in 15° increments.

To capture real world scenarios, the images that compose this dataset were gathered from Flickr³, a public media hosting service.

3.3.3 AFLW

The AFLW [22] dataset is similar the AFW while providing more data. It focus on provide a wide range of different faces (poses, ethnicity, age, gender, expression, occlusion, etc). Akin to AFW, the images were also extracted from Flickr.

The resulting dataset contains 25.993 images annotated with up to 21 facial landmarks if visible, head pose information, face bounding box and ellipse. The landmarks were annotated manually.

Besides the landmarks, the rest of the annotated information was performed automatically using the POSIT algorithm [8]. POSIT works by fitting a mean 3D Model of the head to the 2D landmarks. Since the original purpose for this dataset is for face alignment, the head pose annotations present in the dataset is coarse in nature.

The subjects of this dataset have a distribution of 56% female faces and 44% male faces. The dataset can be downloaded⁴ for research purposes after approval by the creators.

3.3.4 300W-LP

The 300W-LP [54] dataset is derived from the 300W dataset [39]. The 300W dataset standardizes multiple source datasets (AFW [53], LFPW [3], HELEN [52], IBUG [39] and XMSVTS [30]) with 68 facial landmarks. *Xiangyu Zhu et al.* adapted the proposed face profiling [54] to generate 61,225 samples across large poses with the proposed face profiling method [54], which is expanded to 122450 samples using vertical flip on the images. The name given to this dataset is 300W Across Large Poses (300W-LP). The specific values for the head pose information is not reported.

The nature of this dataset is synthetically expanded. The dataset is obtained by performing transformation of the source images to create new images with different poses.

³<https://www.flickr.com/>

⁴<https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/>

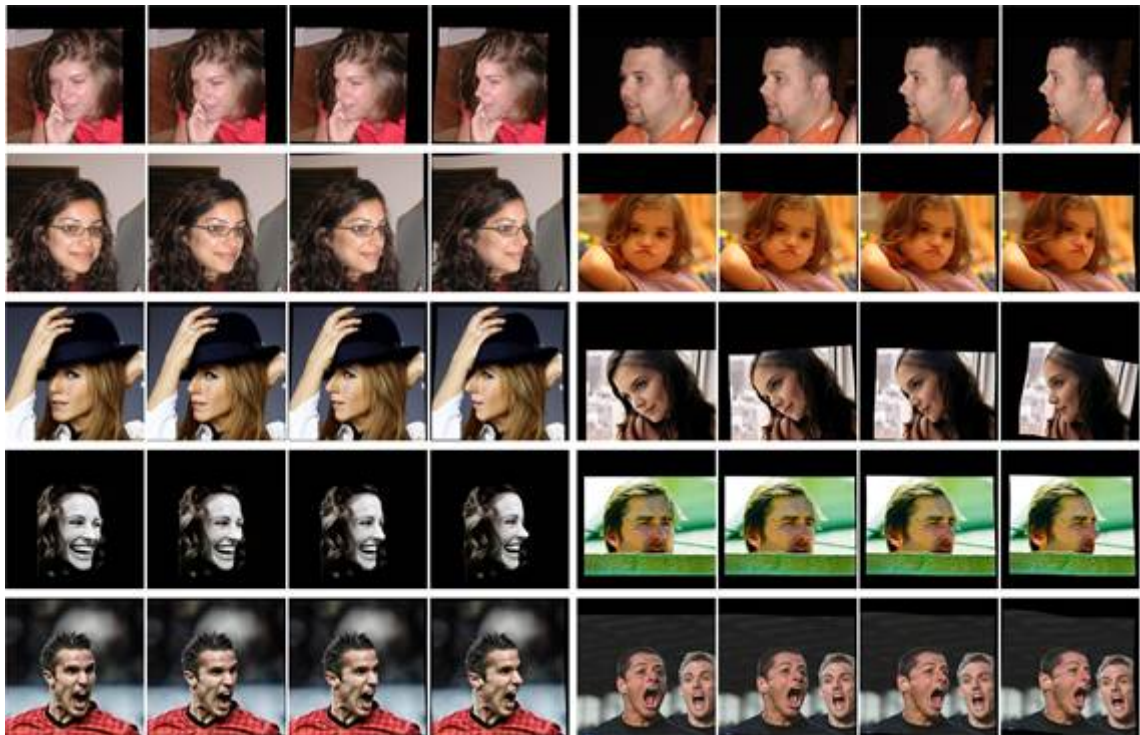


Figure 3.11: Example frames from the 300W-LP dataset. [54]

3.3.5 AFLW2000-3D

This dataset appeared in the work published by *Xiangyu Zhu et al.* [54]. This dataset represents the first 2000 images of the AFLW dataset [22] that was reannotated with 68 landmarks with a 2D and 3D representation. Since the publication the dataset has been commonly used as standard benchmark both for landmark detection and head pose estimation and is particularly challenging because it contains images in very large poses.

3.3.6 Synthetic Datasets

Synthetic datasets [16] [24] [43] [49] are generated from a 3D Model of the Human Head. These type of datasets are created on the assumptions that head pose datasets are not sufficient for deep learning because of the limited pose variation, number of subject samples and partially incomplete annotations [43].

These datasets usually are not available to download but the pipeline to render such datasets are described in the articles. The pipeline can render innumerable samples over the desired range of rotation and can take into account gender, expression, occlusion, etc.

The number of synthetic images used is not reported in some cases [24], while others do report it, the sample size varies. Synhead [16] reports using 51 096 different head pose images, *Wang et al.* [43] trains their model on over 310 000 images.

The advantages of these datasets is the ability to provide discrete or continuous pose annotations, a large number of samples, and extremely accurate annotations.

3.4 Summary

In this chapter, the state of the art was reviewed and the most common datasets presented. Of the presented methods, *Xia et al.* [46] approach was the only one that achieved the performance level required to the domain of our problem. To evaluate this approach two datasets were selected. One dataset for training and another for validation. These dataset need to fulfill a certain requirements to evaluate the head pose estimation task. The list that follows gives an overview of the requirements used to select the dataset.

- A large number of samples with a wide variety of age, ethnicity and gender is required to achieve person independent training.
- A fine grained head pose angle is beneficial.
- The actual pose should be measured and not be left to the subjective decision of humans. The different subjects with the same pose should have the same angles for yaw, pitch and roll.
- The background should be non-uniform and the different types of occlusions need to present so the models can learn their effects.
- The database should consist of real human faces in order to train a system that is applicable to real-world scenarios

Based on the requirements presented and the information in the previous sections. The only datasets that present the necessary characteristics are : [300W-LP](#), [AFLW2000-3D](#).

Chapter 4

Methodology

This chapter aims describe the methodology of the algorithm evaluated in this work to address the Head Pose Estimation Problem. Despite the many applications of head pose estimation, it is important to keep in mind what motivated this dissertation. The application of the head pose estimation for the face image assessment in the biometric domain was the starting point of this work. It is only natural to employ the method that provides the highest competing performance among the different methods published. The high level concepts introduced by *Xia et al.* [46] are the foundation for this dissertation. The method proposes several increments towards solving the head pose estimation task. The regularization of the input image, the integration of a heatmap to help focus the Convolutional Neural Network (CNN) and Euler angles convention change.

For the remaining of this chapter presents a description of our approach to replicate the method proposed [46].

4.1 State of the Art method overview

Figure 4.1 illustrates the method proposed in [46]. Head pose angles are the final result of the several steps after the image is fed as input into the pipeline. First, the facial landmarks are obtained a using state of the art Face Alignment method. After obtaining the landmarks, an area around the face is cropped. Section 4.2 presents the details regarding the face alignment methodology.

Section 4.3 presents the methodology for generating the regularized face image from the previous step. This process produces the affine transformation required to transform the input image into the regularized Image. The regularization procedure improves learning by narrowing the variation of head poses by rotating the images along the Z axis.

The heatmap generator improves the performance of the model by helping the cnn focus on the facial features around the landmarks, eliminating background variations. The heatmap of the facial landmarks is generated from the landmarks obtained after the affine transformation of the original landmarks, the details of the generation pipeline are presented in Section 4.4.

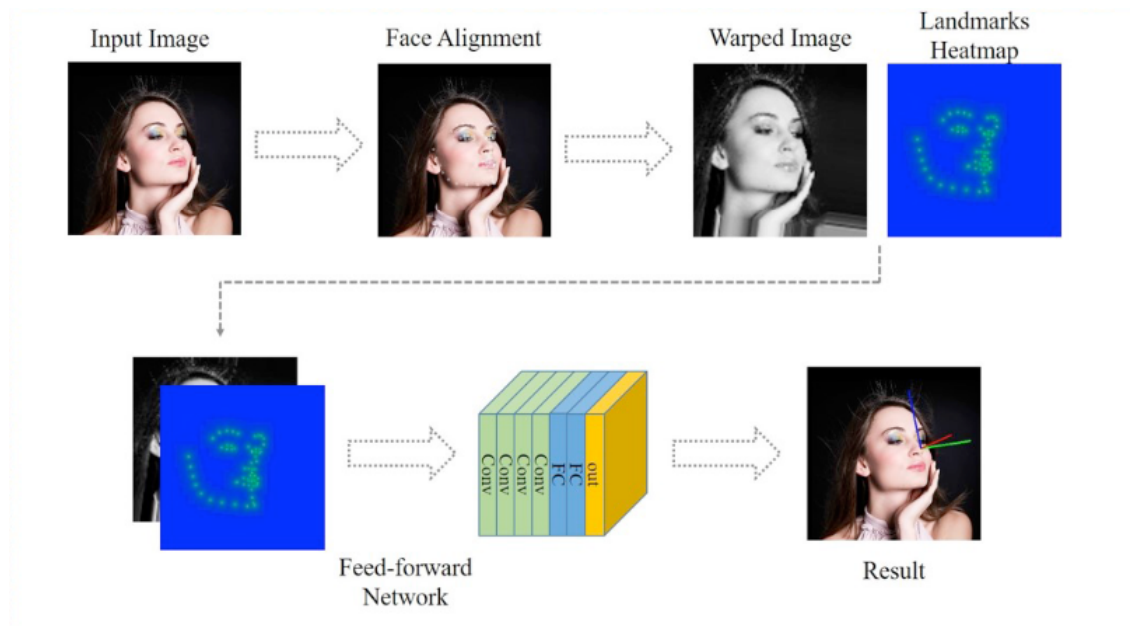


Figure 4.1: Method proposed by *Xia et al.* [46]

The Affine image and the Heatmap are stacked and utilized as the input of the network to estimate the head pose angles. The network details are presented in Section 4.5.

The method proposed additionally changes the convention of the Euler Angles. Section 4.6 details the conversion of Euler Angles convention.

The next sections present the approach to the several steps described in this section.

4.2 Face Alignment

Given that the method proposed by *Xia et al.* [46] requires landmarks to compute the affine image and the heatmap. Similarly to *Xia et al.*, our approach utilizes a state of the art face alignment method. FAN [5] and SFD [50] are used for landmark detection. The method used for landmark extraction provides face shapes S with 68 points, represented with the coordinates (x, y) of the point in the 2D plane. Figure 4.2 presents an illustration of the face with the landmarks on the right side. While the left is the representation of the Canonical Shape.

The detected landmark are subjected to additional computations to modify the original face to the input size of the network.

With the set of landmarks, our first step is to extract the maximum and minimum value for landmark along the X-axis and Y-axis. These values allow us to define the bounding boxes for the faces present in the image, consequently the bounding boxes permits us to determine the image that has the most significant footprint on the image by calculating the area of each bounding box. This step is necessary as the image can contain multiple faces, while only providing ground truth annotations for a single face.

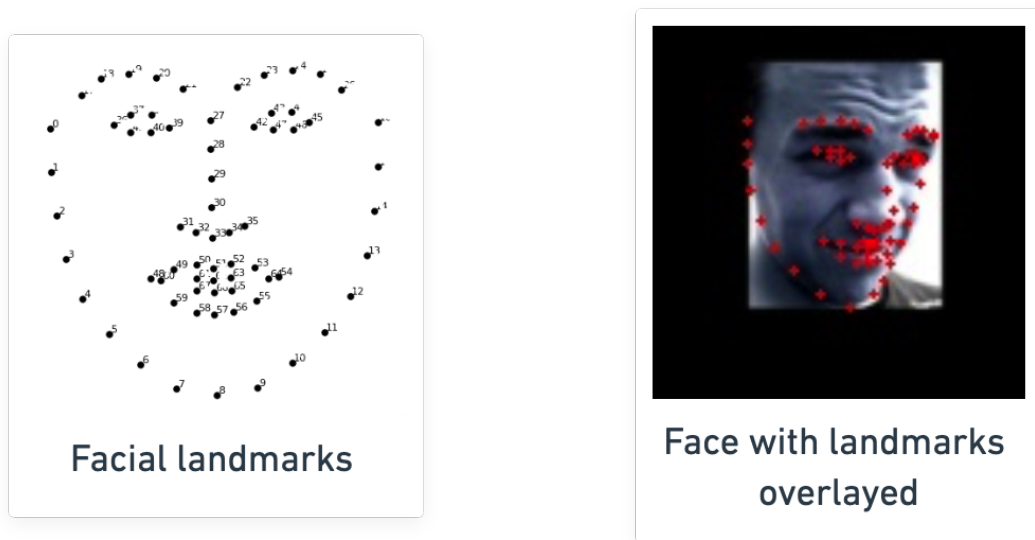


Figure 4.2: Canonical face shape side by side with a face with overlaid landmarks

The face on the original image only represents a small area in the image. With the landmarks of the most prominent face defined, our next is to crop the face.

The bounding box created from the landmarks does not contain the whole face of the subject. As such, we need to expand the area of the landmarks as to contain the whole face. When expanding, a factor of 0.5 for the height and width of the bounding box was employed. The factor expands the image by half on each side of the bounding box, intercepting an area where the face represents 25% of the total area.

The pipeline used in this pre-processing step as illustrated by [Figure 4.3](#)

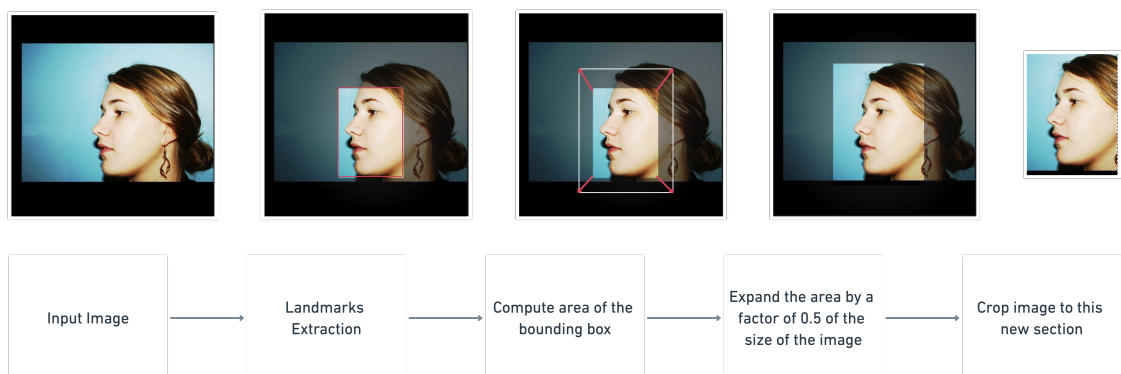


Figure 4.3: Example of the steps of the pipeline for intercepting the Face in an image with an example

4.3 Regularized Image

The objective of generating the regularized image is to transform the input face into the canonical representation. Figure 4.4 presents several examples of images and their regularized counterparts.

The procedure to obtain the regularized image comprises several actions. First, utilizing facial landmarks and canonical landmarks, we need to define the affine transformation that aligns the two shapes. Once the transformation is defined, the same parameters to create the affinity map that transforms the face image into canonical representation.

4.3.1 Shape Alignment

The alignment of two shapes is a recurrent task in the face alignment domain. Frequently problems of this domain require to transform a canonical representation of the face to the detected face in an image. Our approach is heavily inspired by Cascade Shape Regression [6] [37] [47] and Active Shape Models [7]. The two shapes are aligned via an affine transformation. More specifically a similarity transform, is a subset of affine transformation that uses scale, rotation and translation. Similarity Transforms can be written as :

$$T = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \quad (4.1)$$

where a, b, c, d, e, f are the parameters of the similarity transform.

To determine these affine parameters, we used the following process:

1. Determine the translation vector between the two face shapes

The translation vector is written by $\vec{v} = \begin{bmatrix} x \\ y \end{bmatrix}$. The vector \vec{v} represents the translation between the input face landmarks $S_1(w, z)$ to the canonical landmarks $S_0(x, y)$. x and y , are calculated by subtracting the centroid of the canonical shape S_0 to the centroid of the input shape S_1 . The centroid of a shape is the mean value along the x and y axis.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \bar{x} - \bar{w} \\ \bar{y} - \bar{z} \end{bmatrix} \quad (4.2)$$

2. Determine the scale and rotation

To find out the scale and rotation to fit S_1 into S_0 , our approach follows the methodology proposed by Tim Cootes in [7]. The scale and rotation is obtained by minimizing $|sAS_1 - S_0|$, where s represents the scaling factor and A an orthogonal rotation matrix. Following the set of steps and equations provided by Tim Cootes [7]:

- (a) Centering S_1 and S_0 on the origin by subtracting the mean values of the coordinate point.
- (b) Determine the values that will be used for the orthogonal matrix

$$A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \quad (4.3)$$

$$a = (S_1 \cdot S_0) / |S_1|^2 \quad (4.4)$$

$$b = \left(\sum_{i=1}^n (S_1(x_i)S_0(y_i) - S_1(y_i)S_0(x_i)) \right) / |S_1|^2 \quad (4.5)$$

From a and b , the scale is $s^2 = a^2 + b^2$ and the rotation is defined by $\theta = \tan^{-1}(-b/a)$

3. Create a Similarity Matrix that aligns the face shape to canonical shape.

With the values obtained from previous equations, now our approach has everything necessary to create the similarity transform.

The matrix from equation 4.1 can be written as follows:

$$T = \begin{bmatrix} a & b & x' \\ -b & a & y' \\ 0 & 0 & 1 \end{bmatrix} \quad (4.6)$$

4. Apply the transform and generate the affine images

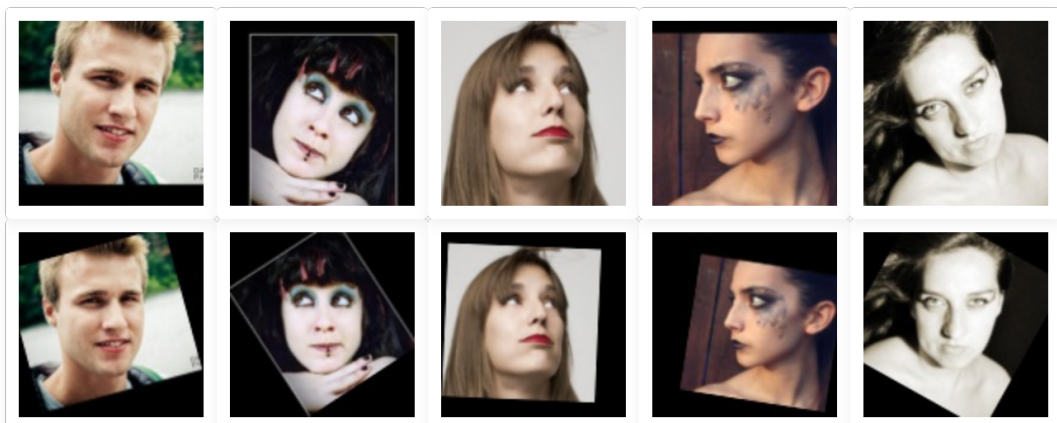


Figure 4.4: Examples of affine transformation, the first row contains the original images and the second row the images after transformation (affine images).

4.4 Face Shape Heatmap



Figure 4.5: Pipeline for generating the heatmap image

The landmarks heatmap produce an image where the highest intensity values locations are near the landmark, and the intensity decreases as the distance to the landmark increases. This technique borrowed concepts from Face Alignment task and was introduced by *Kowalski et al.* [23]. The introduction of the heatmap helps the network focus on the areas surrounding the landmarks, increasing performance by reducing background changes. The pipeline illustrated in Figure 4.5 presents our approach for the heatmap generation and is detailed below.

To create the landmarks heatmap. First, its necessity to generate the heatmap for every landmark using equation 4.7.

$$H(x,y) = \frac{1}{1 + \min_{s_i \in T(S)} \|(x,y) - S_i\|} \quad (4.7)$$

$H(x,y)$ represents the intensity at point (x,y) of the heatmap and S_i is the i -th landmark after an affine transformation. In our implementation, the heatmap intensity is computed in a circle with a radius of 8 for each landmark.

After generating a heatmap for each landmark, these local heatmaps are concatenated to a single image, forming the heatmap of the face. When stacking the local heatmaps, the intensity of the final image is the maximum intensity on every local heatmap for that point.

4.5 Artificial Neural Network used in the implementation

The network used by *Xia et al.* [46] was inspired by VGG16 as proposed in [51]. VGG is a commonly deep learning network. Deep learning is an umbrella term commonly used to define complex neural networks, is a network that comprises of several hidden layers between the input layer and the output layer, complex neurons and architectures. The inner mechanism's of deep learning networks are more complex than simpler networks and tend to be more abstract. However, the authors adapted the original VGG16 network to fit better the HPE problem. HPE is generally simpler than object detection or image classification, as such the authors decreased the

input size of the network from 224 x 224 as reported in [51] to 112x112. Additionally, the inputs channels were modified from three channels corresponding to RGB images to two channels using for the grayscale input image and the heatmap generated. Figure 4.6 describes the network used for the problem.

The network consists of four blocks with two convolutions layers in each block. A max-pooling layer at the end of each block is employed to create a down-sample of the input representation reducing its dimensionality. The max-pooling layer uses a filter of 2x2 and a stride of 2.

Every convolution block takes advantage of batch normalization and utilizes Rectified Linear Unit for activation.

A dropout layer is employed before the first fully connected layers for regularization to prevent overfitting. The Euler angles in radians are in the range $[-\pi, \pi]$ and the output layer uses tanh function for activation, which output values in the range $[-1, 1]$. To normalize the results to Euler angles, the output layer is multiplied by π .

In total there are 8 hidden layers in the network.

4.6 Euler Convention

Existing datasets for head pose estimation, usually utilize the rotation axes for Euler Angles as Z-Y-X. The Z-Y-X ordering means that the final orientation of the head is obtained first by rotating the Z-axis, followed by the Y-axis and X-axis sequentially. In other words, after the rotation is performed along the Z-axis, the Euler angles for the Y and X axes change accordingly. However, after the Euler convention change from ZYX to XYZ, the rotation along the Z-axis is the last operation in the rotation matrix. Thus the rotations of the Y-axis and X-axis are independent from the rotation along the Z-axis. The implementation of the component that changes the Euler convention was made with the SciPy library¹. The choice to use this library was due to the fact that this library provides an interface that can instantiate and convert rotations quite easily. Additionally, the library provides another advantage. The Scipy library uses Quaternions for spatial rotations². Quaternions are a better approach to define the 3D position of the object in space because when converting Euler conventions, the original angles are recover integrally.

4.7 Implementation Details

Our implementation deviates from *Xia et al.* [46] approach. While their approach was end to end, in other words, the pre-processing of the sample images was done at runtime, our approach generates the affine images and heatmaps in a phase separate from training.

¹<https://www.scipy.org/>

²<https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.transform.Rotation.html#scipy.spatial.transform.Rotation>

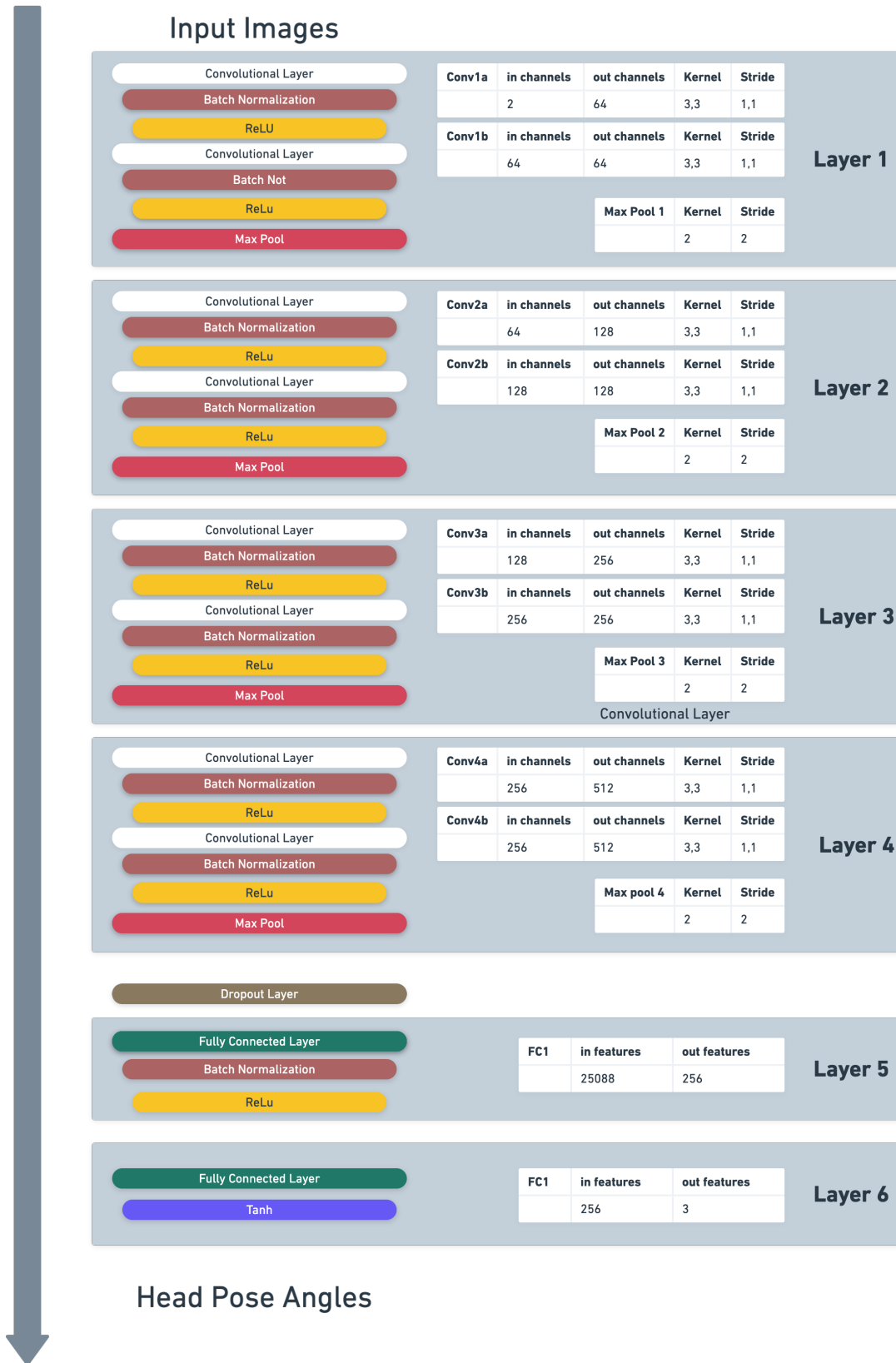


Figure 4.6: Structure of the adapted VGG16 [51] to head pose estimation problem

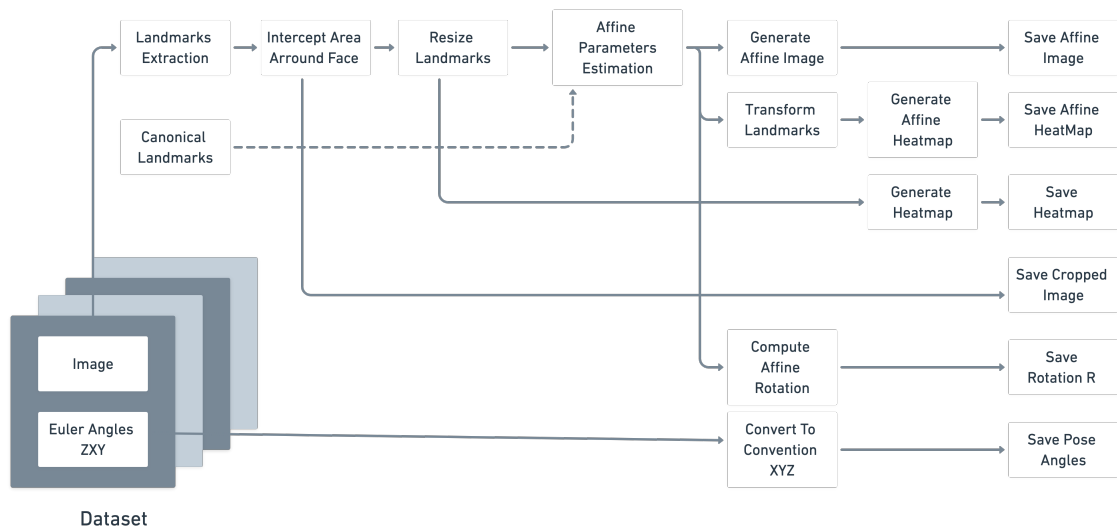


Figure 4.7: Pre Processing Pipeline

The datasets used for training and evaluation are not ready to be for their intended purposes. Additional steps are required to modify the original images to the input of the network. The pre-processing is separated into two phases. The first phase determines the canonical shape of the training set, while the second phase creates the set of images handled for training and evaluation. The obligation to separate pre-processing in different phases is associated with the fact that computation of the canonical shape needs is performed on the all dataset. Only then the canonical shape is obtained, and thus all the variables are gathered to perform the face normalisation step. Figure 4.7 presents the Pre-Processing pipeline.

Additionally, another step of pre-processing is applied. Our approach intends to study the evaluation of the model across different pose restrictions. This step generates text files that contain the images used for training and evaluation where head pose angles are restricted to various intervals. These intervals were created by dividing the range of 0 to 30 degrees in increments of 5 degrees.

At the end of the pre-processing, all the required images and range restriction were generated to proceed with training and evaluating the models.

4.8 Summary

In this chapter was provided a thoroughly explanation regarding the inner workings of the method proposed by Xia *et al.* [46]. The method consists of three components that help improve the accuracy. The affine image, the landmarks heatmap and reordering of the Euler convention. Each of these techniques help the model learn better by increase the focus on more predominant areas or reducing the complexity of the problem.

While Xia *et al.* [46] approach employed the techniques at runtime, our approach is separated in two phases, pre-processing phase and training phase. The pre-processing generates all the

required information for the method ahead of training. While training strictly focus on training the model with the information pre-processed.

Chapter 5

Experimental Settings

This chapter aims to describe experiments performed to address the problem of this dissertation. All evaluations are performed in a similar setup in order to assure the results of different models can be compared objectively. A description of the Training Protocols and Evaluation Protocols are presented in this chapter. Additionally, a description of the methods and datasets used to evaluate the experiments are given.

5.1 Datasets

The datasets used for Training and Evaluation are defined in chapter 3.3. In order to train the models that estimate the head pose, training data that represent the wide variety of poses, interpersonal differences and environment conditions is required. Therefore the training and evaluations are performed on the 300W-LP and AFLW2000-3D datasets, respectively.

The 300W-LP dataset is split into a training set used for training the different models, and a validation set used for validating the model performance throughout training. While both datasets are from the same database, they contain images from different individuals and are therefore independent from each other.

Additionally, we evaluate the final result of the training procedure in the AFLW2000-3D. This evaluation guarantees that the models do not only perform well on the validation set but show good generalisation performance across datasets

5.1.1 Training Protocols

This section describes all the aspects of the training process. The definition of the training protocol defines the consistent process for the training loop.

The inputs of the method are a combination of the pre-processed images, the labelled head pose angles, and in cases where the affine image (when applicable) was employed, the rotation component of the affine matrix is also provided. Throughout this section is described the selection of hyper-parameters. The loss function, optimiser and learning rate used during the training procedure are identical to the parameters used by *Xia et al.* [46]

The loss function is defined as follows:

$$L = \frac{1}{N_b} \sum_{i=1}^{N_b} \sqrt{\frac{\|P_i - \hat{P}_i\|_2^2}{3}} \quad (5.1)$$

where P is a vector with the ground truth pose, and \hat{P} is the predicted pose, and N_b is the number of samples in training.

The optimiser employed to update the weight parameters is referred to as ADAM [21]. The decaying beta factors and learning rate employed during training are default values recommend by Xia *et al.* [46] in their experiments.

In our implementation, the learning rate scheduler chosen is the step learning rate available in Pytorch. The parameters used to instantiate the scheduler are defined as follows: step size of 2 and gamma of 0.96. The step size of the scheduler defines the rate at which the learning rate decays. Whereas gamma defines the multiplying factor for learning rate decay. During training, the learning rate decays every two epochs.

Additionally, the training protocol defines the batch size and the split validation factor. The batch size defines the number of samples that will be propagated through the network. The split validation factor represents the percentage of dataset samples that are used for validation of the model every epoch. The values for these parameters are 256 and 0.1 for batch size and split validation factor, respectively.

Table 5.1 presents a summary of the training protocol and hyperparameters for all the experiments

Training Configuration		
Number of Epochs		35
Batch Size		256
	ADAM	
Optimizer	Beta 1	0.9
	Beta 2	0.999
	Learning Rate	0.001
	StepLR	
Learning Rate Scheduler	step size	2
	gamma	0.96

Table 5.1: The hyperparameters used for training every experiment

5.2 Evaluation Metrics

To provide a high degree of comparability between the developed methods and the state of the art competing methods, a method of measurement is needed. The aim is to obtain significant

measures for both the head pose estimation as well as the classification scenario.

5.2.1 Mean Absolute Error

A typical measuring tool for the accuracy of head pose estimation is the mean absolute error (MAE) of continuous values. The head pose angles and the mean values are represented in degrees.

The equation 5.2 represents the mean absolute error along each one of the head pose axis, where n is the number of testing samples, \hat{P}_i represents the predicted pose and P_i the ground truth pose. This equation represents the three MAEs possible values regarding roll, pitch and yaw: the subscript s denotes head pose estimation for either Yaw, Pitch or Roll. While equation 5.3 represents the MAE of the three types of angles, therefore it takes in account the difference between the predicted values and the ground-truth values for all pairs of roll, pitch and yaw angles.

$$MAE^s = \frac{1}{n} \sum_{i=1}^n \left| \hat{P}_{i,s} - P_{i,s} \right|, \text{ for } s \in \{\text{yaw}, \text{pitch}, \text{roll}\} \quad (5.2)$$

$$MAE_{avg} = \frac{MAE^{yaw} + MAE^{pitch} + MAE^{roll}}{3} \quad (5.3)$$

5.2.2 Compliance Classification

In addition to the mean absolute error estimation, we were also interested in evaluating the compliance of each image with the requirements of face image quality for MRTD. Therefore, it was necessary to define a way to perform the evaluation of "compliance classification" or "frontal pose classification". This classification is performed comparing the prediction of the head pose angles with the defined value $T_{compliant}$. This value $T_{compliant}$ in the context of MRD is defined by ICAO [18] and its recommended value is 8° :

$$\hat{C}_{Yaw}(n) = \begin{cases} True & \left| \hat{P}_{Yaw}(n) \right| < T_{compliant} \\ False & otherwise \end{cases} \quad (5.4)$$

$$\hat{C}_{Pitch}(n) = \begin{cases} True & \left| \hat{P}_{Pitch}(n) \right| < T_{compliant} \\ False & otherwise \end{cases} \quad (5.5)$$

$$\hat{C}_{Roll}(n) = \begin{cases} True & \left| P_{Roll}(n) \right| < T_{compliant} \\ False & otherwise \end{cases} \quad (5.6)$$

$\hat{C}_{Yaw}, \hat{C}_{Pitch}, \hat{C}_{Roll}$ are boolean variables that are true when the predicted pose of the n -nth image from the test set is classified as compliant. The ground truth data, $C_{Yaw}, C_{Pitch}, C_{Roll}$ is classified as compliant if the pose is below threshold $T_{compliant}$ along every axis :

$$C_{Yaw}(n) = \begin{cases} True & \left| P_{Yaw}(n) \right| < T_{compliant} \\ False & otherwise \end{cases} \quad (5.7)$$

$$C_{Pitch}(n) = \begin{cases} True & |P_{Pitch}(n)| < T_{compliant} \\ False & otherwise \end{cases} \quad (5.8)$$

$$C_{Roll}(n) = \begin{cases} True & |P_{Roll}(n)| < T_{compliant} \\ False & otherwise \end{cases} \quad (5.9)$$

For both ground truth data and validation data the value chosen for Compliance Treshold is 8° , as defined in Chapter 2. With the labels for each pose angle, it is now required to assess the compliance of the head pose as a whole. The equations that follows determines the compliance labels for the head pose:

$$C_{y,p,r}(i) = C_{yaw}(i) \wedge C_{pitch}(i) \wedge C_{roll}(i) \quad (5.10)$$

In Equation 5.10, $C_{y,p,r}$ denotes the ground truth compliance classification for Yaw, Pitch and Roll. The resulting label will be true if every angle of the ground truth is compliant, otherwise false if one or more angles are not compliant.

$$\widehat{C}_{y,p,r}(i) = \widehat{C}_{yaw}(i) \wedge \widehat{C}_{pitch}(i) \wedge \widehat{C}_{roll}(i) \quad (5.11)$$

In Equation 5.11, $\widehat{C}_{y,p,r}$ denotes the compliance classification for the predictive Yaw, Pitch and Roll. The resulting label will be true if every predicted angle is compliant, otherwise false if one or more angles are not compliant.

The number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are then calculated :

$$TP = \sum_{i=1}^n (C_{y,p,r}(i) \wedge \widehat{C}_{y,p,r}(i)) \quad (5.12)$$

$$FP = \sum_{i=1}^n (\neg C_{y,p,r}(i) \wedge \widehat{C}_{y,p,r}(i)) \quad (5.13)$$

$$TN = \sum_{i=1}^n (\neg C_{y,p,r}(i) \wedge \neg \widehat{C}_{y,p,r}(i)) \quad (5.14)$$

$$FN = \sum_{i=1}^n (C_{y,p,r}(i) \wedge \neg \widehat{C}_{y,p,r}(i)) \quad (5.15)$$

With the confusion matrix complete, the accuracy can be determined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5.16)$$

With a Binay decision (compliant vs non-compliant), a Receiver Operator Characteristics curve can be plotted that shows the effect of the threshold $T_{compliant}$ on the true positive rate (TPR)

and false positive rate (FPR). These values are determined as follows :

$$TPR = \frac{TP}{TP + FN} \quad (5.17)$$

$$FPR = \frac{FP}{FP + TN} \quad (5.18)$$

For predictions the value of $T_{compliant}$ is varied from 0° to 10° in one degree increments. While for the ground truth data the $T_{compliant}$ stays at 8° . Additionally, the accuracy is also calculated along different threshold variations.

5.3 Evaluation Protocol

Our objective with the definition of an evaluation protocol is to define objective metrics and qualitative results that can provide insights into the performance of the model. The complete evaluation of head pose estimation is comprised of several steps, which are summarized in the list below :

- The different variations of the models are trained with according to the trained protocol defined in 5.1.1.
- After training, several evaluations are performed on the another database:
 - For all the images of the database, the Yaw, Pitch and Roll angles are estimated.
 - Then the MAE is calculated for individual angles and the for the group
 - Then the confusion matrix are completed and the accuracy plot which shows the compliance performance of the model.

This standard evaluation allows a unified comparison of all the models for the head pose estimation investigated in this dissertation.

The models are evaluated on the whole test set as well as over different subsets of the test set. The restrictions of the test set is done by restricting the images evaluated to several intervals from 0° to 30° . The division of subsets allows the evaluation of the models on different ranges of angles.

5.4 Summary

In this chapter, the training procedure and evaluation procedure for our models were defined. The definition of such procedures create a unified approach for both training and evaluation. The results of each model can be compared with each other.

The metrics defined report the regression and classification performance. The mean absolute error provides a clear interpretation of the results between ground truth and predictions. For this reason, MAE is the most common evaluation metric for head pose estimation.

The conversion of continuous values to discrete categories has inherently problematic. When measuring the accuracy of the discrete categories, the process of binning the values to the categories can distort the results. For example, when presented with these two values -7.35 and 8.00 . After the conversion to discrete categories, the values will have different categories associated and thus the predicted value is classified as incorrect, when in this case the predicted value are really close to the expected value. However this conversion is necessary to evaluate the compliance for Machine Readable Travel Documents.

Chapter 6

Experiments and Results

Since our approach is based on the proposed architecture of *Xia et al.* [46]. Before any study is performed regarding the compliance assessment of the model, it is necessary to establish a baseline. The baseline refers to the model training ignoring the heatmap and the affine transformation. Training with the face images ground truth based on the Z-Y-X convention.

Since our approach was trained on a subset of 300W-LP (restricted to 30 degrees and fewer epochs). This restriction was enforced due two main reasons. First, our models need to be accurate on low range poses while errors on large poses should not affect the compliance accuracy of the system. Second the processing power of the infrastructure to train for long times was not feasible for the scope of this dissertation. Our models are not strictly comparable to the results reported by *Xia et al.* [46] because of the training procedure employed on our models. However the models should display the same effect with the introduction of the heatmap and affine image on the results.

6.1 Baseline

In Table 6.1, the results of our baseline model are presented. The values reported establish the baseline our models will be compared to when employing the affine image and the heatmap. Additionally, Figure 6.1 reports the evolution of the error when tested with different set of samples restricted to different intervals of angles.

Table 6.1: Results for baseline architecture evaluated on AFLW2000-3D segmented to 30°

Method	MAE^{yaw}	MAE^{pitch}	MAE^{roll}	MAE_{avg}°
<i>Xia et al.</i> [46]	3.99	7.32	6.50	5.94
Baseline	1.74	2.58	4.10	2.80

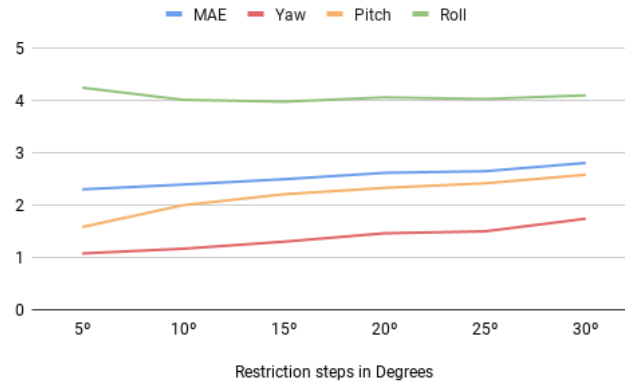


Figure 6.1: MAE of the baseline model over the restricted test set

On the test set the baseline model reports the error of 1.74° for MAE^{yaw} , 2.58° for MAE^{pitch} and 4.10° for MAE^{roll} . These are consistent with the hypothesis that low pose samples are easier to detect than samples with medium or large poses.

The evolution of the errors across different pose intervals is reported in figure 6.1. This plot demonstrates that the curves of the errors are very stable across the whole test set. The curves for the Yaw and the Pitch increase around 1° while the Roll error stabilizes around 4° . Overall the baseline model performs reasonably well.

6.2 Affine Image

Table 6.2: Results for affine image architecture evaluated on AFLW2000-3D segmented to 30°

Method	MAE^{yaw}	MAE^{pitch}	MAE^{roll}	MAE_{avg}
<i>Xia et al.</i> [46] Baseline	3.99	7.32	6.50	5.94
<i>Xia et al.</i> [46] Affine	2.51	4.81	3.19	3.50
Baseline	1.74	2.58	4.10	2.80
Affine	1.71	4.31	4.85	3.63

In this section, the analysis of training the model with the affine images is detailed. Table 6.2 reports the results obtained when compared to the baseline. When comparing the baseline model to the affine model, it is noticeable that the affine model is performing worse than the baseline when evaluated on the test set.

This fact is not in agreement with the values reported by *Xia et al.* [46]. The affine model should have performed better, especially estimating the roll angle, since the input images are rotated along the Z axis to minimize the variations of face images.

Upon further analysis, the results reported in Figure 6.2, clearly shows that the affine images does help reduce the errors, especially at lower degrees. However the inclination of the errors

curve is bigger than the inclination presented in Fig 6.1. The mean absolute error increases as the pose angles of the images increase.

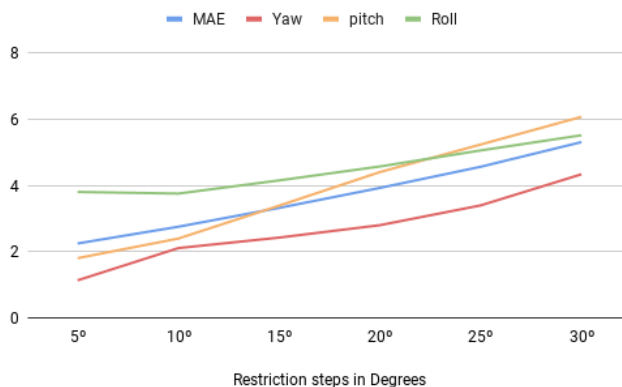


Figure 6.2: Mean Absolute Error for the Affine model over the restricted test set

The fact that the affine image does perform worse in this experiment, does not necessarily mean that the employment of affine image for head pose estimation is inherently bad. There are several possible explanations for this to occur. The affine model produces errors that are above the baseline. There are some possible explanation as to why this happens. Referring to Figure 4.4, the rotation applied to face image does suffer from reduction in scale and the transformation performed incurs loss in information. A loss in information decreases the learning capability of the model.

Another factor that may explain the errors reported from our approach, its threshold defined to stop early the training. If the validation loss does not decrease for 5 epochs it is sensible to stop training, as to not waste computing resources. However training the models with more face samples and for more epochs, may enable the model to extract features that are more prominent for head pose estimation, thus approximating our implementation to [46].

6.3 Baseline with Heatmap

Table 6.3: Results for normal image with heatmap architecture evaluated on AFLW2000-3D segmented to 30°

Method	MAE^{yaw}	MAE^{pitch}	MAE^{roll}	MAE_{avg}
<i>Xia et al.</i> [46] Baseline	3.99	7.32	6.50	5.94
<i>Xia et al.</i> [46] Heatmap	1.95	3.91	3.99	3.05
Baseline	1.74	2.58	4.10	2.80
Baseline with Heatmap	2.22	2.37	3.97	2.86

The next experiment performed was to evaluate the model trained with the baseline samples and the heatmap stacked on the input for the network. The results pertaining to this experiment are present in Table 6.3. The addition of the heatmap decreases the error on both MAE^{pitch} and MAE^{roll} , while increasing the Yaw error. However the MAE increases a small amount, mainly due to the increase in the MAE^{yaw} error. The reduction of error when compared to [46] is not as drastic as the results gathered, but prove that employing the heatmap does help in the head pose estimation task.

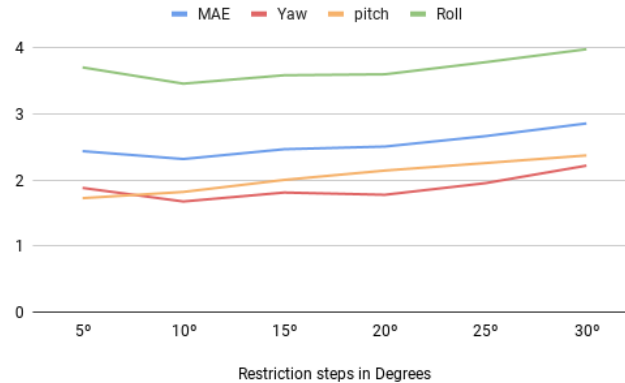


Figure 6.3: Mean Absolute Error for the Baseline with Heatmap model over the restricted test set

When comparing the chart in Figure 6.3 with Figure 6.1, the curvature of the errors of both charts seem identical and stable.

6.4 Affine with Heatmap

The model evaluated in this section represent the model where the input is the affine image and heatmap. The goal was to reach the same results the reported in [46]. However this did not occur, as the reader can see in Table 6.4.

When employing the affine and heatmap, the errors along the different axis were expected to be significantly lower, the results produced show a decrease in error on the MAE^{pitch} and MAE^{roll} but a higher error on MAE^{yaw} , when compared to our baseline.

Figure 6.4 shows that the errors across the test set, start significantly lower along every rotation axis for poses close to 0 but increase as the poses move to higher poses.

Table 6.4: Results for affine image with heatmap architecture evaluated on AFLW2000-3D segmented to 30°

Method	MAE^{yaw}	MAE^{pitch}	MAE^{roll}	MAE_{avg}
<i>Xia et al. [46]</i> Baseline	3.99	7.32	6.50	5.94
<i>Xia et al. [46]</i> Affine Heatmap	0.63	2.05	1.10	1.46
Baseline	1.74	2.58	4.10	2.80
Affine with Heatmap	2.37	2.60	3.59	2.85

Table 6.5: MAEs for the models in degrees evaluated on the AFLW2000 segmented to 30°

Method	MAE^{yaw}	MAE^{pitch}	MAE^{roll}	MAE_{avg}
Baseline	1.74	2.58	4.10	2.80
Affine	1.71	4.31	4.85	3.63
Heatmap	2.22	2.37	3.97	2.85
Affine Heatmap	2.38	2.60	3.59	2.86

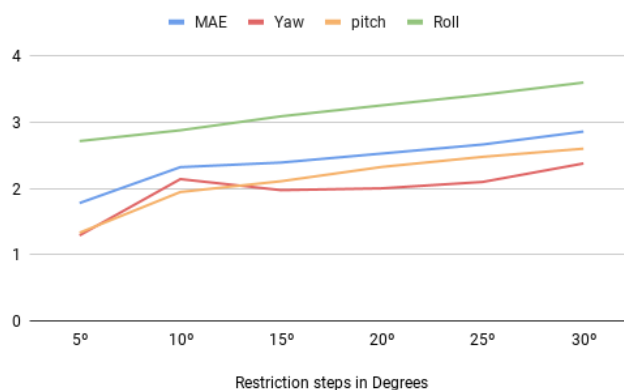


Figure 6.4: Mean Absolute Error for the Affine with Heatmap model over the restricted test set

6.5 Examples Analysis

This section presents several examples of inaccurate pose estimation from our models. For each model, the some examples that reported high estimation error were gathered from all the models. This high degree of error represents an anomaly and may provide insights into the limitations and strengths of our models.

In order for the images to be comparable, a plot representing the head pose in the 3d space was drawn over the samples. The X-axis is represented by the colour green and points to the right. The Y-axis is pointing down and is represented by the colour red. Finally, the Z-axis points towards the reader and uses the colour blue.

Figure 6.5 presents these results. On the first column the plot drawn shows the head pose ground truth. For the remaining columns the plot represent the head pose estimated by the corresponding model specified at the top of figure. These plots enable the comparison of the ground truth head pose to the estimated head poses of the different models.

On a general note, the images in Figure 6.5 present several known challenges for accurate head pose estimation. The presence of multiple subjects, occlusions, low-quality images and missing facial features affect the performance of some models.

The images also show that the variations of the model components do not always increase the accuracy of the head pose estimation task. For example, in the third row, the model with the affine image has a higher error than the model trained with the heatmap. However, the model with the affine image and heatmap performs slightly better than the models with affine. Despite this, the error of the affine with heatmap model is still very high.

On the other hand, in the last row, where face image is in favourable conditions for accurate head pose estimation, the affine model has a high degree of error. The model where the affine image and the heatmap were employed report a very drastic reduction in error when compared to the affine model.

In conclusion the different task simplification methods have advantages and disadvantages when predicting head pose angles in different situations and the combination of both methods work with each other to negate the consequences of each task simplification method. Due to the fact that the affine model does not reduce background interference.

6.6 Compliance Assessment

This section shows the results obtained from the evaluation of the models to face image compliance with the requirements of MRTD. The results and analysis presented in this section refer to differences in evaluation methodologies.

The first methodology of evaluation is to assess the accuracy of the model to different restricted subset of the test set while classifying the compliance of the images to the international standards.

On the second methodology, the models are evaluated on the whole test set while iterating over the compliance threshold for the predicted head poses. The threshold variation allows to evaluate the performance of the models on stricter thresholds.

6.6.1 Evaluation over different subsets

The models are evaluated by defining the threshold for compliance of the predicted head pose to 8 degrees defined by the international standards [19] and evaluate over the restricted subsets of the test set. The AFLW2000-3D test set is restricted to the range of $[0, 30]$ in 5 degrees step. With the method defined in Section 5.2.2, our models are evaluated based on the classification prediction of the image as compliant or not compliant. By classifying the regression of the model, we construct multiple confusion matrices for each step in the range and we plot the accuracy. Figure 6.6 shows the results obtained from this evaluation.

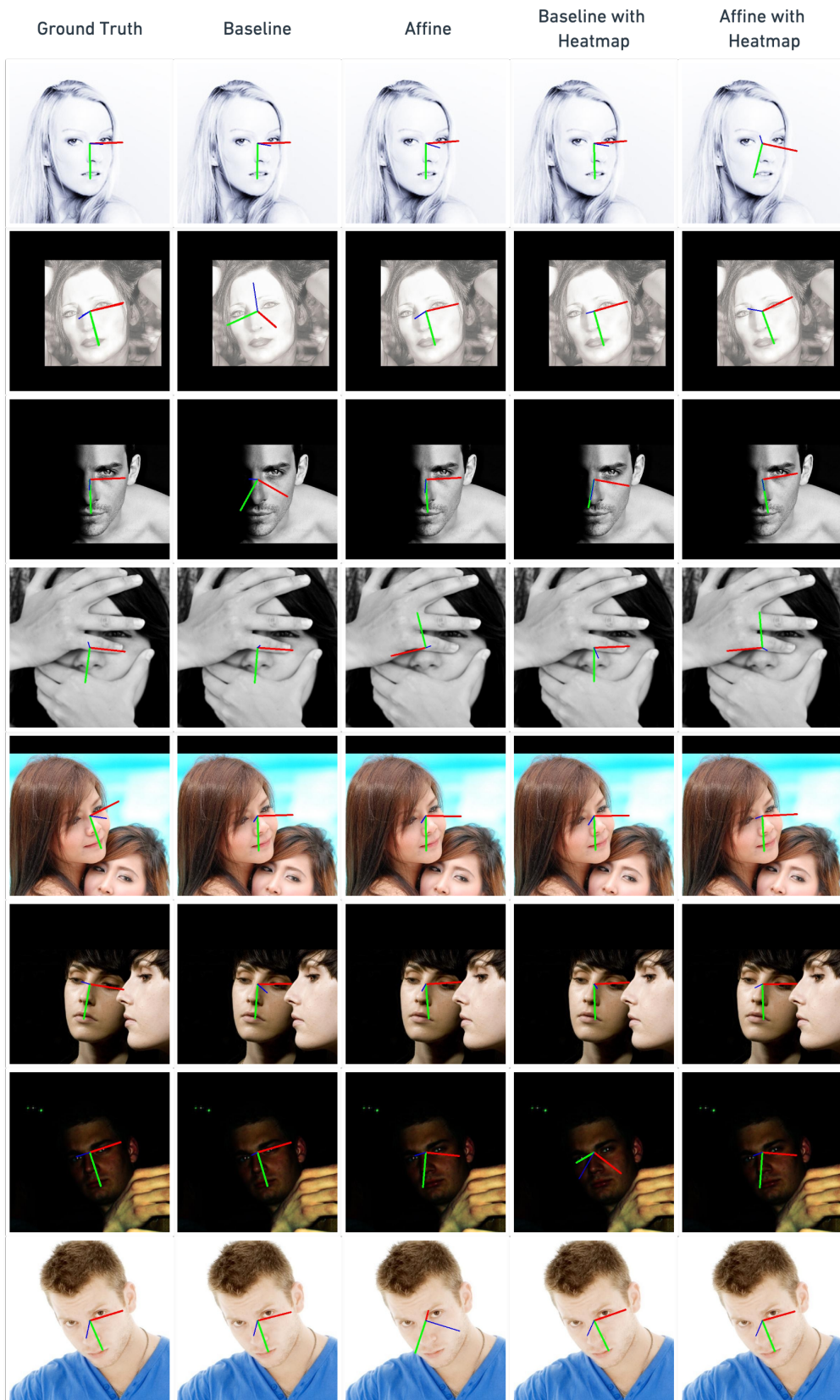


Figure 6.5: Estimation of head pose in AFLW200-3D

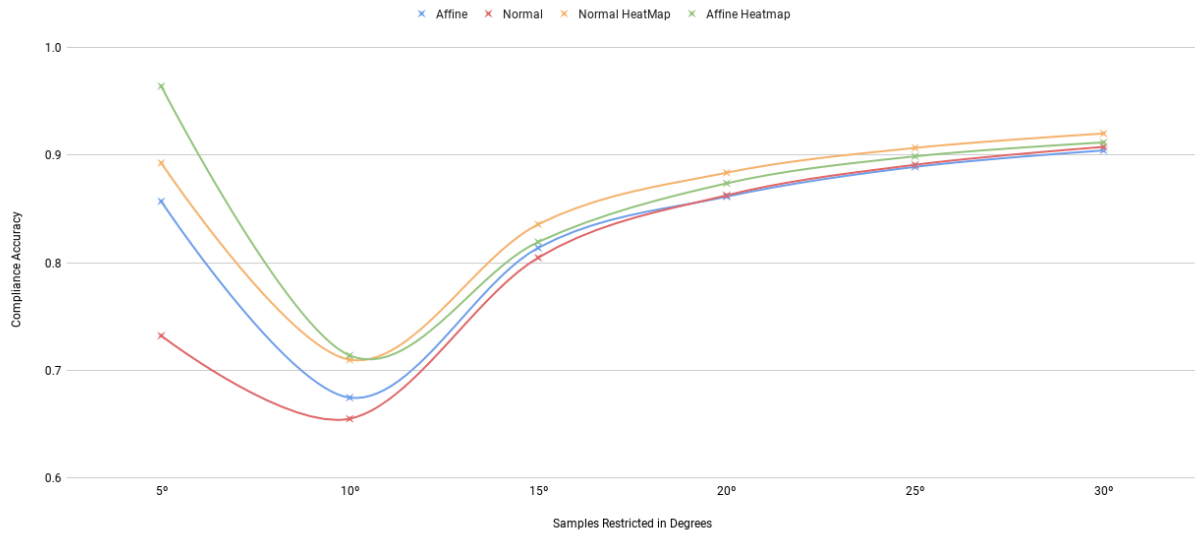


Figure 6.6: Accuracy plotted over the test set with samples restricted in degrees

On the lower bound of the chart, the baseline model performs the worst with an accuracy of 73% while the affine with heatmap performs the highest with 96% accuracy. Every model decreases their accuracy when labelling images near the critical point of 8°. After the surpassing the critical point, the accuracy increases and converges to above 90%.

Near the highest bound, evaluated on sets of samples that include medium poses, the model trained with baseline and heatmap performs better than other models by 1% with 92% accuracy.

It is possible to conclude the influence of the different approaches to the models, specially on the subset of 5°, where all the images are compliant. However due to the imbalanced nature of the subset regarding compliance classification, the accuracy of the models increase to upper bounds as the number of true negative grows as angles of the restricted test set increases.

6.6.2 Evaluation over stricter thresholds

Additionally another evaluation made was by using the whole test set but changing the threshold of compliance when estimating head pose. Referring to Section 5.2.2, the threshold for the ground truth stays the same while the threshold for the predicted head pose is varied from $[1^\circ, 10^\circ]$ in 1 degree steps. This incremental change in the prediction threshold allows the study of the optimal threshold for the compliance classification.

While Figure 6.7 represents the overall accuracy of the models for all the samples that were correctly classified. The Receiver Operating Characteristics (ROC) plots, presented in Figure 6.8, focus on more fine grained evaluation of the tests by comparing the true positive rate versus false positive rate.

Accuracy

The results presented in figure 6.7 are not surprising, the models perform consistently over the different thresholds. Starting from threshold value 1° and increasing until threshold of 8° , the plot shows that the accuracy increases until reaching the critical point of 8° . In turn, from thresholds 8° to 10° , the accuracy actually decreases. The loss in accuracy is natural, the number of images that are compliant to threshold 8 stays the same for every test. This, in combination with the fact that threshold value that the model uses for classification is above 8° means that the number of false positives is increasing. Still when observing the accuracy on the lower bounds, it is clear that the affine with heatmap model performs better than the rest.

The variation of the threshold provide consistent results when compared Figure 6.6. With the threshold defined at 8° , the model where the heatmap was employed performed better than the other models. And by reducing the reducing the threshold the model that performs better is the affine heatmap. However, it is worth mentioning that the results obtained for the affine heatmap may have a degree of error induced by the affine image as discussed in Section ??.

Receiver Operator Characteristics

The receiver operating characteristics is the plot of the true positive rate over the false positive rate over of the classification over different predictive compliance thresholds. These plots are often used to assess the classification performance because it is possible to assess the different trade offs between true positive rate and false positive rate in various threshold for classification. The overall accuracy of the classification is performed by observing the area under the ROC curve and allows for comparison of different models for the same domain, in our case for head pose estimation. For the test set, the model that performs the best by changing the threshold is the model that uses heatmap.

Examples Analysis

In this section, Figure 6.9 presents a few examples of compliance classification across several images. The same image was evaluated on the every model and the resulting head pose and the comparison to the ground truth label is reported.

From the various examples, it is possible to observe that some examples are consistently miss classified or correctly classified across the different models. However examples near the compliance critical point, the classification of the models are inconsistent, some models accurately predict the pose angles while others do not.

Across the different examples, the affine heatmap appears to be the model that is predicting the values more precisely than the other models.

Moreover the affine model on the third row has an absolute error of the yaw of 50° .

On the seventh row, it is noticeable the effect the heatmap and the affine have on the prediction. The baseline miss classifies the image as compliant while the other models do not.

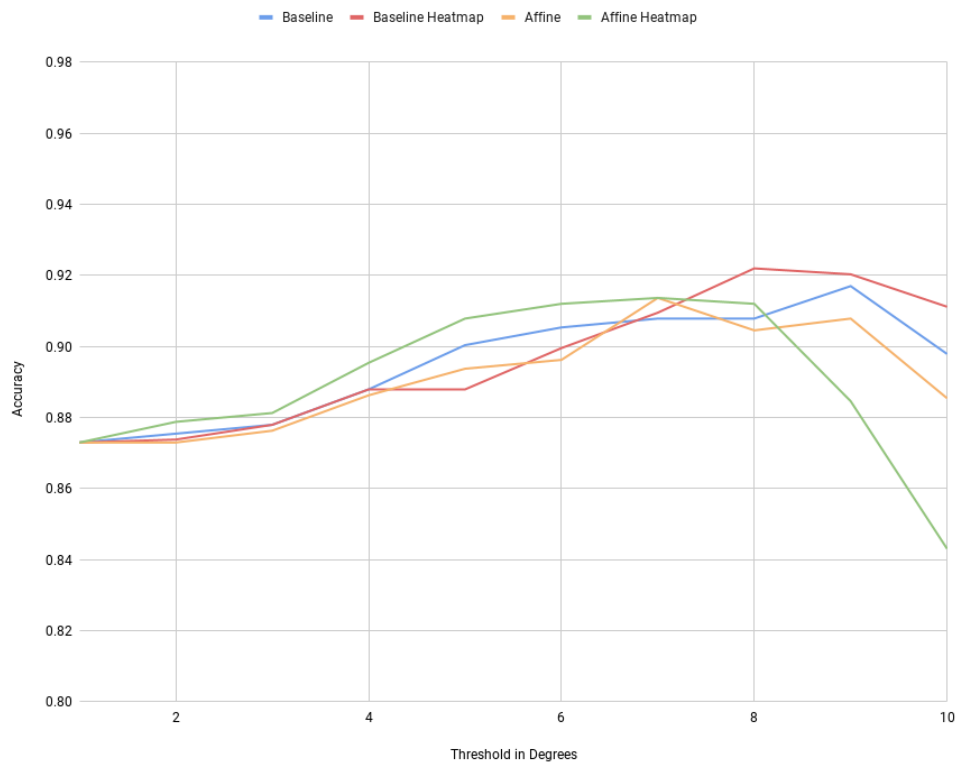


Figure 6.7: Accuracy plotted over different thresholds

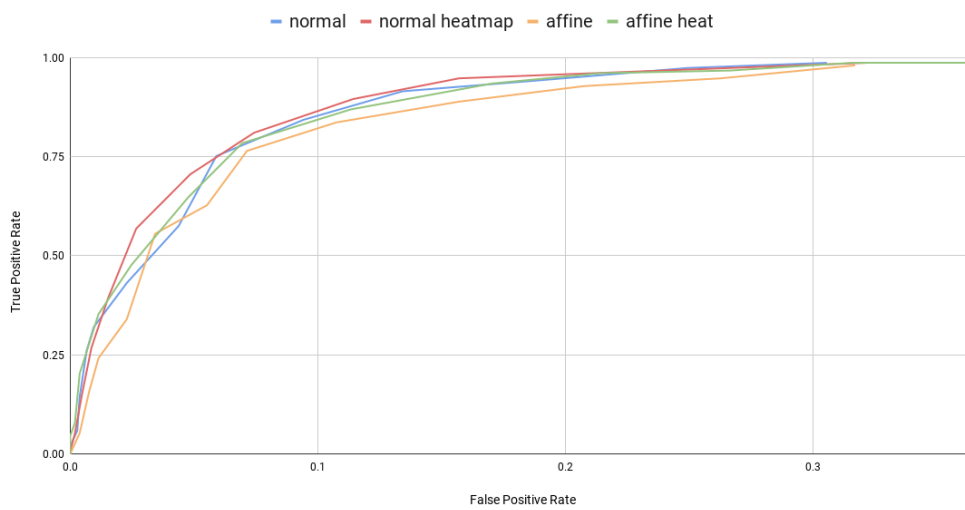


Figure 6.8: ROC plot for Compliant Classification

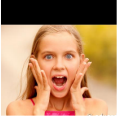
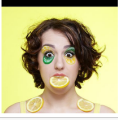

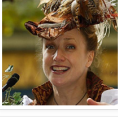

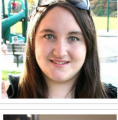

	Ground Truth	Baseline	Baseline with Heatmap	Affine	Affine with Heatmap	
Label		True Positive	True Positive	True Positive	True Positive	
Head Pose	Ground Truth Yaw -1°	Predicted Yaw -5°	Predicted Yaw -6°	Predicted Yaw -4°	Predicted Yaw -2°	
	Ground Truth Pitch -4° Ground Truth Roll -1°	Predicted Pitch -4° Predicted Roll -1°	Predicted Pitch -6° Predicted Roll -1°	Predicted Pitch -6° Predicted Roll 0°	Predicted Pitch -4° Predicted Roll -1°	
Label		True Positive	True Positive	True Positive	True Positive	
Head Pose	Ground Truth Yaw 0°	Predicted Yaw 2°	Predicted Yaw -1°	Predicted Yaw -2°	Predicted Yaw 1°	
	Ground Truth Pitch -2° Ground Truth Roll 1°	Predicted Pitch -3° Predicted Roll 0°	Predicted Pitch -3° Predicted Roll -1°	Predicted Pitch -4° Predicted Roll 0°	Predicted Pitch -2° Predicted Roll 0°	
Label		True Negative	True Negative	True Negative	True Negative	
Head Pose	Ground Truth Yaw 9°	Predicted Yaw 12°	Predicted Yaw 13°	Predicted Yaw 59°	Predicted Yaw 13°	
	Ground Truth Pitch -11° Ground Truth Roll -14°	Predicted Pitch -9° Predicted Roll -10°	Predicted Pitch -6° Predicted Roll -14°	Predicted Pitch 18° Predicted Roll 11°	Predicted Pitch -9° Predicted Roll -14°	
Label		True Negative	True Negative	True Negative	True Negative	
Head Pose	Ground Truth Yaw 11°	Predicted Yaw 14°	Predicted Yaw 12°	Predicted Yaw 15°	Predicted Yaw 13°	
	Ground Truth Pitch 7° Ground Truth Roll 1°	Predicted Pitch 0° Predicted Roll 3°	Predicted Pitch -2° Predicted Roll 1°	Predicted Pitch -6° Predicted Roll 0°	Predicted Pitch 6° Predicted Roll 1°	
Label		False Positive	False Positive	False Positive	False Positive	
Head Pose	Ground Truth Yaw -3°	Predicted Yaw -1°	Predicted Yaw -5°	Predicted Yaw -6°	Predicted Yaw -2°	
	Ground Truth Pitch -10° Ground Truth Roll 0°	Predicted Pitch -6° Predicted Roll 0°	Predicted Pitch -5° Predicted Roll -1°	Predicted Pitch -5° Predicted Roll 0°	Predicted Pitch -7° Predicted Roll 0°	
Label		False Positive	True Negative	True Negative	True Negative	
Head Pose	Ground Truth Yaw 1°	Predicted Yaw -1°	Predicted Yaw -1°	Predicted Yaw -1°	Predicted Yaw 0°	
	Ground Truth Pitch -12° Ground Truth Roll 3°	Predicted Pitch -7° Predicted Roll 3°	Predicted Pitch -8° Predicted Roll 2°	Predicted Pitch -9° Predicted Roll 2°	Predicted Pitch -10° Predicted Roll 1°	
Label		False Positive	False Positive	True Positive	False Positive	
Head Pose	Ground Truth Yaw 6°	Predicted Yaw 8°	Predicted Yaw 8°	Predicted Yaw 7°	Predicted Yaw 9°	
	Ground Truth Pitch 0° Ground Truth Roll -2°	Predicted Pitch -2° Predicted Roll -1°	Predicted Pitch -2° Predicted Roll -2°	Predicted Pitch -4° Predicted Roll -1°	Predicted Pitch -1° Predicted Roll -1°	

Figure 6.9: Examples of compliance classification and the comparison between models

6.7 Summary

In this chapter several experiments were made on the models developed for this thesis. The experiment are divided in two parts. The first part presented an objective evaluation of the different models developed for this thesis as well as a relative comparison to the inspiration for our work. Although the results obtained showed some slight differences regarding the method implemented and other from when compared to the state of the art, it is possible to conclude the effectiveness of the different variations of the models in our area of research.

When evaluating the models according to the compliance classification necessary for machine readable travel documents and enforced by ICAO [1] several conclusions can be drawn. Only about 1 in 9 images is miss classified as correct. If the head pose estimation is required to have high sensitivity or high specificity the model that performs better by enforcing stricter thresholds is the model where the input is the heatmap and the non transformed face image. However, if the best accuracy is the criteria desired, then the best model for the head pose estimation task is the affine image with heatmap.

Chapter 7

Conclusion

In this thesis, the head pose estimation task was reviewed and several approaches to the problem were presented. After a thorough evaluation of head pose estimation approaches, a method which promises low error in the wild scenarios was chosen as the basis for this work. The main idea for this approach is to simplify the task of pose estimation by combining a regularized representation of the face image with a heatmap and regress the head pose from this mapping. The structure of this method was implemented as close as possible to the original work. Different variations of the models were trained to study the effect of task simplification components. The models are trained to regress on the head pose from a dataset with large background variations and uncontrolled conditions. The performance of the model was then evaluated on another dataset. The main goal of this thesis was to estimate the pose angle of arbitrary face image with the lowest possible error and to verify if the state of the art produces good results when classifying the frontal pose according to the international standards. On the evaluation dataset, the mean absolute error for the head pose was 2.80 degrees. Also the compliance classification achieves a performance of 90% success rate when a threshold of 8° , as required by the international standards is enforced.

The main contributions of this thesis is the study of the performance of the state of the art in head pose estimation for in the wild environments and the assessment of the applicability to the biometric domain.

7.1 Limitations

The models presented in this thesis have limitations when compared to other approaches.

- Facial Landmarks: The performance of this model is dependent of the performance of the process used to obtain the landmarks to the corresponding face image. If the landmarks detection method have low accuracy the model may not learn well enough the mapping between the images and the pose angles.
- Dataset : The dataset used for training is challenging and balanced along the different rotation axis. However the number of images near frontal pose is low.

7.2 Future Work

The head pose estimation models presented in this thesis is able to achieve good performance when assessing compliance in unconstrained environments, yet there is still room for improvements until such models can be deployed for automation in the biometric domain. The ideas that follow are some of the possible optimizations for the models presented in this thesis.

- **Better Training Data:** The dataset used is a limitation and a possibility for improvement in future work, as the number of images for training and testing that have near frontal pose is quite small. It is possible to expect an improvement when training the model with more images with small pose angles.
- **More Training:** Due to computational limitations, it was not possible to fully train the models, the whole dataset. Improvements in performance can be expected since the original work this dissertation presents a model that was trained for more epochs and with more data.
- **Combined Losses:** The loss function used in this dissertation treats the head pose problem as a whole. While several works divided the head pose angles and treat each angles a sub problem of head pose estimation.
- **Spatial Transformer:** The effect of the regularization module that produces the affine image could be substituted by a Spatial Transformer, a technique introduced by Google Deep Mind, a plugin component placed before the network that helps the model learn with invariance to translation, scale and rotation.
- **Fine Tuning:** The model is trained using in the wild dataset for generalization purposes. However in the biometric domain, the main focus is on images with low to medium poses. It is therefore reasonable to invest in fine tuning the model for images with lower poses and compliant images.

References

- [1] M.S.A. Abdel-mottaleb and M. H. Mahoor. Application notes - algorithms for assessing the quality of facial images. *IEEE Computational Intelligence Magazine*, 2:10–17, 2007.
- [2] Vineeth N. Balasubramanian, Sreekar Krishna, and Sethuraman Panchanathan. Person-independent head pose estimation using biased manifold embedding. *EURASIP Journal on Advances in Signal Processing*, 2008:1–15, 2007.
- [3] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.
- [4] David Beymer. Face recognition under varying pose. *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 756–761, 1994.
- [5] 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 Conference on Computer Vision*, 2017.
- [6] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107:177–190, 2012.
- [7] Tim F. Cootes. An introduction to active shape models *. 1992.
- [8] Daniel DeMenthon and Larry S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141, 1995.
- [9] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101:437–458, 2012.
- [10] Matteo Ferrara, Annalisa Franco, and Dario Maio. A multi-classifier approach to face image segmentation for travel documents. *Expert Syst. Appl.*, 39:8452–8466, 2012.
- [11] Matteo Ferrara, Annalisa Franco, Dario Maio, and Davide Maltoni. Face image conformance to iso/icao standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security*, 7:1204–1213, 2012.
- [12] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Evaluating systems assessing face-image compliance with icao/iso standards. In *BIOID*, 2008.
- [13] Virgilio Ferruccio Ferrario, Chiarella Sforza, Graziano Serrao, Gianpiero Grassi, and Erio Mossi. Active range of motion of the head and cervical spine: a three-dimensional investigation in healthy young adults. *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, 20 1:122–9, 2002.

- [14] Rafael C. González and Richard E. Woods. Digital image processing, 3rd edition. 2008.
- [15] Nicolas Gourier, Daniela Hall, and James L. Crowley. Estimating face orientation from robust detection of salient facial structures. 2004.
- [16] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1531–1540, 2017.
- [17] Jeffrey Huang, Xuhui Shao, and Harry Wechsler. Face pose discrimination using support vector machines (svm). *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)*, 1:154–156 vol.1, 1998.
- [18] Machine Readable Travel Documents. Standard, International Civil Aviation Organization, 2015.
- [19] Information technology – biometric data interchange formats – part 5: Face image data. Standard, International Organization for Standardization, Geneva, CH, 2011.
- [20] Andrzej J. Kasinski, Andrzej Florek, and Adam Schmidt. The put face database. 2008.
- [21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.
- [22] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151, 2011.
- [23] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcíński. Deep alignment network: A convolutional neural network for robust face alignment. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2034–2043, 2017.
- [24] Felix Kuhnke and Jörn Ostermann. Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In *ICCV 2019*, 2019.
- [25] Amit Kumar, Azadeh Alavi, and Rama Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, May 2017.
- [26] Yongmin Li, Shaogang Gong, and Heather M. Liddell. Support vector regression and classification based multi-view face detection and recognition. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 300–305, 2000.
- [27] Davide Maltoni, Annalisa Franco, Matteo Ferrara, Dario Maio, and Antonio Nardelli. Biolab-ica: A new benchmark to evaluate applications assessing face image compliance to iso/iec 19794-5 standard. *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 41–44, 2009.
- [28] Aleix M. Martínez. The ar face database. 1998.
- [29] Stephen J. Mckenna and Shaogang Gong. Real-time face pose estimation. *Real-Time Imaging*, 4:333–347, 1998.

- [30] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luetttin, and Gilbert Maître. Xm2vtsdb : The extended m2vts database. 1999.
- [31] Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. *2007 IEEE Intelligent Transportation Systems Conference*, pages 709–714, 2007.
- [32] Erik Murphy-Chutorian and M. Trivedi. Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking. *2008 IEEE Intelligent Vehicles Symposium*, pages 512–517, 2008.
- [33] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:607–626, 2009.
- [34] Shay Ohayon and Ehud Rivlin. Robust 3d head tracking using camera pose estimation. *18th International Conference on Pattern Recognition (ICPR'06)*, 1:1063–1066, 2006.
- [35] Margarita Osadchy, Yann LeCun, and Matthew L. Miller. Synergistic face detection and pose estimation with energy-based models. *J. Mach. Learn. Res.*, 8:1197–1215, 2004.
- [36] P. Jonathon Phillips, Patrick J. Flynn, W. Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William J. Worek. Overview of the face recognition grand challenge. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:947–954 vol. 1, 2005.
- [37] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.
- [38] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2155–215509, 2017.
- [39] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.
- [40] Bernt Schiele and Alex H. Waibel. Gaze tracking based on face-color. 1995.
- [41] Jamie Sherrah, Shaogang Gong, and Eng-Jon Ong. Face distributions in similarity space under varying head pose. *Image Vision Comput.*, 19:807–819, 2001.
- [42] Rainer Stiefelhagen. Estimating head pose with neural networks-results on the pointing 04 icpr workshop evaluation data. 2004.
- [43] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. A deep coarse-to-fine network for head pose estimation from synthetic data. *Pattern Recognition*, 94:196–206, 2019.
- [44] Hugh R. Wilson, Frances Wilkinson, Li ming Lin, and Maja Castillo. Perception of head orientation. *Vision Research*, 40:459–472, 2000.

- [45] Junwen Wu and Mohan Manubhai Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41:1138–1158, 2008.
- [46] Jiahao Xia, Libo Cao, Guanjuan Zhang, and Jiakai Liao. Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks. *IEEE Access*, 7:48470–48483, 2019.
- [47] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013.
- [48] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019.
- [49] Hui Yuan, Mengyu Li, Junhui Hou, and Jimin Xiao. A single image based head pose estimation method with spherical parameterization. *ArXiv*, abs/1907.09217, 2019.
- [50] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. s^3 fd: Single shot scale-invariant face detector. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 192–201, 2017.
- [51] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):1943–1955, 2015.
- [52] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 386–391, Dec 2013.
- [53] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.
- [54] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. *CoRR*, abs/1511.07212, 2015.