FLÁVIO HENRIQUE DE BITTENCOURT ZAVAN

NOSE POSE ESTIMATION IN THE WILD AND ITS APPLICATIONS ON NOSE TRACKING AND 3D FACE ALIGNMENT

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre. Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, Universidade
Federal do Paraná.
Orientadora: Profa. Dra. Olga R. P. Bellon

CURITIBA

2016

FLÁVIO HENRIQUE DE BITTENCOURT ZAVAN

NOSE POSE ESTIMATION IN THE WILD AND ITS APPLICATIONS ON NOSE TRACKING AND 3D FACE ALIGNMENT

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre. Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, Universidade
Federal do Paraná.
Orientadora: Profa. Dra. Olga R. P. Bellon

CURITIBA

2016

Z39n Zavan, Flávio Henrique de Bittencourt
Nose pose estimation in the wild and its applications on nose tracking and 3D face alignment / Flávio Henrique de Bittencourt Zavan. – Curitiba, 2016.
57 f. : il. color. ; 30 cm.

Dissertação - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática, 2016.

Orientador: Olga Regina Pereira Bellon .
Bibliografia: p. 53-57.

1. Processamento de imagens - Técnicas digitais. 2. Sistemas de reconhecimento de padrões. 3. Computação gráfica. I. Universidade Federal do Paraná. II.Bellon, Olga Regina Pereira. III. Título.

CDD: 621.367

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS EXATAS
Programa de Pós Graduação em INFORMÁTICA
Código CAPES: 40001016034P5

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **FLAVIO HENRIQUE DE BITTENCOURT ZAVAN**, intitulada: "**Nose Pose Estimation in the wild and its applications on nose tracking and 3D Face Alignment**", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua __APROVAÇÃO________________.

Curitiba, 31 de Agosto de 2016.

Olga Regina Pereira Bellon
Prof OLGA REGINA PEREIRA BELLON
Presidente da Banca Examinadora (UFPR)

Prof LUCIANO SILVA
Avaliador Interno (UFPR)

Iacopo Masi
Prof IACOPO MASI
Avaliador Externo (USC)

CONTENTS

LIST OF FIGURES iv

LIST OF TABLES vi

RESUMO vi

ABSTRACT vii

1 INTRODUCTION 1

2 NOSE POSE ESTIMATION 3

3 EXPERIMENTAL RESULTS 6

3.1 Multi-PIE 6

3.1.1 Multi-PIE Results 6

3.2 Pointing'04 12

3.2.1 Pointing'04 Experiments 12

3.3 McGillFaces 18

3.3.1 McGillFaces Results 18

3.3.2 Filtered McGillFaces Results 19

3.4 SFEW 25

3.4.1 SFEW Results 25

3.5 PaSC 28

3.5.1 PaSC Results 29

3.6 AFW 32

3.6.1 AFW Results 32

3.7 Nose Detection 40

4 APPLICATIONS 43

4.1 Face Quality Estimation For Face Tracking 43

4.2 3D Face Alignment in the Wild 47

5 FINAL REMARKS 51

BIBLIOGRAPHY 57

LIST OF FIGURES

2.1 Example subdivision of an image from 1 to 100 subregions 3

2.2 SVM-NosePose diagram 4

3.1 Image samples from the Multi-PIE dataset 7

3.2 Confusion matrices when estimating the yaw on the Multi-PIE dataset . . 10

3.3 Head pose estimation distribution on the Multi-PIE dataset 11

3.4 Image samples from the Pointing'04 dataset 12

3.5 Confusion matrix estimating the yaw on the Pointing'04 dataset 15

3.6 Confusion matrices when estimating the pitch on the Pointing'04 dataset . 16

3.7 Result distributions on the yaw and pitch axes 17

3.8 Image samples from the McGillFaces dataset 18

3.9 Confusion matrices estimating the yaw on the McGillFaces dataset 21

3.10 Results distribution on McGillFaces 22

3.11 Examples of inconsistent ground-truth annotations in the McGillFaces dataset 22

3.12 Confusion matrix and distribution on the Filtered McGillFaces dataset . . 24

3.13 Image samples from the SFEW dataset 25

3.14 Confusion matrix and distribution on the SFEW dataset 27

3.15 Image samples from the PaSC dataset 28

3.16 Different class distributions on the PaSC dataset 30

3.17 Results when estimating the yaw on the PaSC dataset 31

3.18 Image samples from the AFW dataset 32

3.19 Results on the AFW dataset, using 900 images for training 35

3.20 Results on the AFW dataset, using 7000 images for training 37

3.21 Results on the AFW dataset, using 268 augmented faces for training . . . 39

3.22 Intersection Coefficient curves on all datasets. The number in parenthesis is the area under the curve 41

4.1 Comparison of the intersection coefficient metric between the baseline and our approach. The area under the curve is displayed in parenthesis 45

4.2 Comparison of the precision metric between the baseline and our approach. The percentage of frames where the error was less than or equal to 20 pixels is displayed in parenthesis 46

4.3 Overview of the 3D face alignment method 47

4.4 a) Calibration image; b) Calibration landmarks; c) Landmark model viewed with the calibration pose 48

4.5 Example results of the model fitting stage, showing the face bounding-box in red, the detected nose region in blue and the estimated position of the landmarks in green. a) Near frontal good fit; b) Bad fit caused by bad head pitch estimation; c) and d) Half-profile good fit; e) Good fit in an image sourced from the MultiPIE dataset; f) Modest fit in one of the most challenging images in the dataset 49

LIST OF TABLES

2.1 Example results with different descriptors on a subset of the Multie-Pie dataset. Some are not available due to memory constraints when training . 4

3.1 Achieved accuracy with different numbers of subregions on the Multi-PIE dataset 7

3.2 Comparative results of the yaw estimation on Multi-PIE 8

3.3 Achieved accuracy with different numbers of subregions on the Pointing'04 dataset 13

3.4 Comparative results for estimating the yaw on Pointing'04 14

3.5 Comparative results when estimating the pitch on Pointing'04 14

3.6 Comparative results for simultaneous estimation of the pitch and yaw on Pointing'04 14

3.7 Achieved accuracy with different numbers of subregions on the McGillFaces dataset 19

3.8 Achieved accuracy with different numbers of subregions on the Filtered McGillFaces dataset 19

3.9 Original McGillFaces vs. Filtered McGillFaces both approaches. 20

3.10 Achieved accuracy with different numbers of subregions on the PaSC dataset 26

3.11 Comparative results on the SFEW dataset 26

3.12 Achieved accuracy with different numbers of subregions on the PaSC dataset 29

3.13 Results obtained on PaSC dataset 29

3.14 Achieved accuracy with different numbers of subregions on the AFW dataset 33

3.15 Comparative results when estimating the yaw on the AFW dataset 33

3.16 Percentage of images where the intersection coefficient of the detected nose region and the annotated ground-truth is at least 0.5 and the amount of false positives 40

3.17 Nose pose estimation performance when using the detected nose regions . . 42

RESUMO

Neste trabalho, estimamos a pose da cabeça em imagens 2D, tanto em ambientes controlados como não controlados, baseado apenas na região do nariz. Para este fim, propomos e comparamos uma metodologia livre de *landmarks*, baseado em *Support Vector Machines* (SVM-NosePose). O uso de apenas a região do nariz apresenta vantagens sobre o uso da face inteira; não apenas é menos provável a oclusão do nariz, mas ele também é visível e provado ser altamente discriminante em todas as poses de perfil a frontal. O SVM já foi utilizado para este tipo de tarefa em uma base pequena e controlada. Nosso SVM-NosePose adiciona novas idéias e experimentos à etapa da geração do vetor de características, tanto na extração destas, como na agregação dos dados. É comparado favoravelmente ao estado-da-arte, através de experimentos abrangentes cuidadosamente elaborados, utilizando seis bases de dados publicamente disponíveis, Pointing'04, Multi-PIE, McGillFaces, SFEW, AFW e PaSC, abrangendo diversos cenários possíveis na estimativa da pose da cabeça. A fim de realizar uma avalição completa e detalhada, apresentamos resultados tanto com as regiões anotadas do nariz quanto com a saída de um detector de narizes estado-da-arte. Adicionalmente, investigamos duas diferentes aplicações para nossa estimativa: a inclusão original de uma pontuação da pose da cabeça na estimativa da qualidade da face para a inicialização de um rastreador de narizes, alcançando maior precisão de rastreamento; e a execução de alinhamento 3D livre de landmarks em ambientes não controlados utilizando apenas a informação da região do nariz, permitindo que estimativas sejam geradas mesmo em cenários desafiadores.

Palavras-chave: pose da cabeça; rastreamento facial; alinhamento facial

ABSTRACT

We perform head pose estimation solely based on the nose region as input, extracted from 2D images in both constrained and unconstrained environments. To this end, we propose a landmark free methodology, based on Support Vector Machines (SVM-NosePose) and compare it against the state-of-the-art. Using the nose region has advantages over using the whole face; not only it is less likely to be occluded, it is also visible and proved to be highly discriminant in all poses from profile to frontal. SVM has been previously used for this task on a small, controlled dataset. Our SVM-NosePose adds new ideas and experiments on the feature vector generation stage, both in feature extraction and data aggregation. Our SVM-NosePose estimation favorably compares, through thoughtful and comprehensive experiments, against state-of-the-art approaches, using six publicly available datasets, Pointing'04, Multi-PIE, McGillFaces, SFEW, AFW and PaSC. To achieve a complete and detailed evaluation, we present results using both the nose region ground-truth and the output of a state-of-the-art nose detector. Additionally, two different applications for our approach are also investigated: the original inclusion of a head pose score for face quality estimation, for initializing a nose tracker, leading to higher tracking precision; and performing landmark-free 3D face alignment in the wild using only the information of the nose region, enabling coherent estimates to be generated even in challenging scenarios.

Keywords: head pose; face tracking; face alignment

CHAPTER 1

INTRODUCTION

The head pose estimation problem can be defined as determining at least one of the three parameters that configures the face relative to its three degrees of freedom, yaw, pitch and roll and the camera [1]. The growing interest in head pose estimation is mainly due to the advantages it brings to facial analysis tasks. Estimating the head pose can lead to higher accuracy rates in other computer vision problems, such as gaze estimation [2], face quality assessment [3], face recognition [4], facial landmark detection [5], automatic affect analysis in infants [6] and face frontalization [7].

Most of the previous works use 2D information from the whole face to perform head pose estimation [1]. Recently, due to the advent of real-time and low-cost 3D sensors, the focus of many researchers shifted towards estimating the head pose on facial depth images [8] [9]. However, one cannot rely on having depth information in unconstrained environments, where there is no control over the sensor that is being used to capture the images. According to Zhu and Ramanan [10], not only estimating extreme head poses (such as profile) is a difficult problem, but even face detection. Such poses are likely to be found in unconstrained environments and are not considered in many published works. In our work, the focus is kept on 2D RGB images, including those with extreme poses.

The use of manifold analysis for estimating 3D head pose in unconstrained environments is proposed by Peng *et al.* [11]. Pawelczyk and Kawulok [12] extract gradient information from the nose region and use SVM to classify it into a discrete set of angles. This approach was applied in a controlled environment dataset. For estimating the pose in uncontrolled environments, Demirkus *et al.* [13] proposed using a set of facial features to estimate a probability density function over the pose on each frame and aggregating the results using temporal information.

In this work we show that the nose region can be successfully used for head pose estimation on constrained and unconstrained environments. The use of the nose has already been proven efficient for biometrics [14] [15] and head pose estimation [12]. The nose has many properties that make it a good candidate to be used for estimating the head pose. Unlike the eyes and ears, it is visible even in profile faces; unlike the mouth, it cannot be easily deformed by speech and expressions; it is also less likely to be partly occluded by accessories and facial traits, such as sunglasses and beards, when compared to using the whole face.

We developed a method, named SVM-NosePose, for estimating the head pose based on the nose region. It uses Support Vector Machines (SVM) trained with the output of the LGIP filter [16] on the nose region, it is landmark-free, does not take advantage of temporal information, treats pose estimation as a classification problem and estimates the angles based on a predefined set of discrete poses that depends on the dataset used for training. To achieve completeness in our NosePose methodology, a state-of-the-art detection method [17] is combined with SVM-NosePose for finding the nose region.

In addition, we propose two applications for our pose estimation method. An existing face quality estimation method [18] is enhanced by a head pose score and is applied to initializing a nose tracker. The nose pose is also utilized for performing landmark-free 3D face alignment in the wild, such that consistent estimations are generated even in extreme poses and expressions.

This paper is organized as follows: Chapter 2 describes our approach in detail, Chapter 3 presents the datasets used for evaluating our method and the experimental results, Chapter 4 presents our applications and Chapter 5 suggests future work and includes final remarks.

CHAPTER 2

NOSE POSE ESTIMATION

Our SVM-NosePose strategy for feature extraction relies on experimenting with different descriptors and number of subregions of the input image. Pawelczyk and Kawulok [12] use the raw gradient values as the feature vector. We found that binary pattern descriptors histograms can be applied to achieve higher head pose classification accuracy. To find the best combination, three histogram-based descriptors and ten different numbers of subregions were tested. Results obtained using LBP [19], LGIP [16] and LGP [20] descriptors histograms with 1, 4, 9, 16, 25, 36, 49, 64, 81 and 100 subregions (Figure 2.1) were compared.

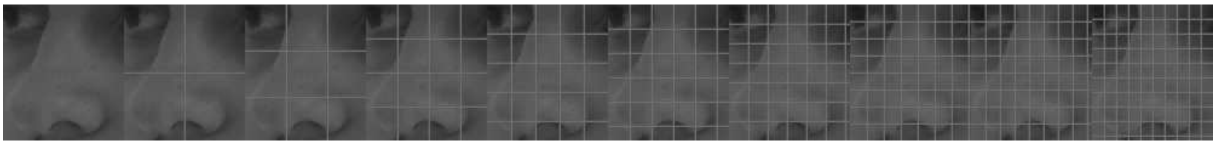


Figure 2.1: Example subdivision of an image from 1 to 100 subregions

Our tests indicate that LGIP almost always achieves higher recall rates than LBP and LGP for all number of subregions (Table 2.1), LGP comes second, but uses twice as much memory, slowing down classification and causing some tests to fail. LGIP robustness to local intensity variation proved to be able to properly describe the nose region for head pose estimation purposes in environments with variable lighting conditions, such as in the Multi-PIE [21] and PaSC [22] datasets. However, it was also noticed that the optimal number of subregions varies for each dataset, therefore, for each case, all possible number of subregions are tried and the one that gives the best results was chosen.

Extracting the histogram of the subregions of the ROI instead of the whole region

Table 2.1: Example results with different descriptors on a subset of the Multie-Pie dataset. Some are not available due to memory constraints when training

Subregions	**LBP**	**LGIP**	**LGP**
1	61.92%	73.27%	59.07%
4	76.52%	91.04%	82.53%
9	90.14%	91.75%	87.46%
16	87.67%	91.03%	89.84%
25	90.11%	90.16%	90.09%
36	83.46%	85.22%	87.98%
49	85.43%	94.01%	91.51%
64	90.84%	93.08%	91.14%
81	91.84%	94.13%	N.A.
100	92.03%	93.83%	N.A.

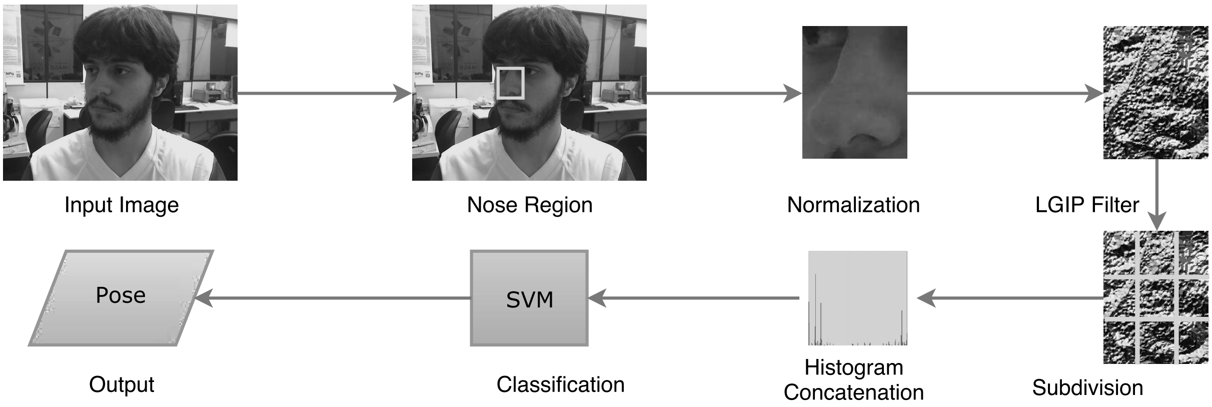


Figure 2.2: SVM-NosePose diagram

enables some of the spatial information to be kept, while allowing some variations to occur. This also allows the final size of the feature vector to remains constant independent of the image size, this way, more data can be extracted from higher resolution images. Our SVM-NosePose method is outlined in Algorithm 1 and shown as a diagram in Figure 2.2.

To properly apply our method for assessing the head pose in unconstrained environments, the nose must first be detected. To this end, we propose the use of a state-of-the-art object detection approach, Faster R-CNN [23].

Algorithm 1 SVM-NosePose Classification Algorithm

```
function EstimatePose($img$)
    Detect the nose
    Crop $img$
    Normalize $img$

    $lgip \leftarrow$ LGIP filter on $img$

    Initialize an empty feature vector $fv$

    for all subregion $sr$ do
        $h \leftarrow sr$'s histogram
        Concatenate $fv$ and $h$
    end for

    $c \leftarrow$ SVM classification of $fv$
    return $c$
end function
```

CHAPTER 3

EXPERIMENTAL RESULTS

Our SVM-NosePose was tested on six datasets using ground-truth nose region annotations to allow for better evaluating the performance of the pose estimation. We present the results individually in this section, including the optimal number of subregions, confusion matrices, grand-truth and result class distribution, a comparison against other published methods when possible and a brief discussion of the achieved results using both strict and weak (off-by-one errors are considered hits) evaluation protocols [12] when applicable. All experiments were performed on Arch Linux running on an Intel Xeon E5-2640 with 64GB of RAM.

3.1 Multi-PIE

The CMU Multi-PIE [21] is a controlled environment dataset composed of 755370 high resolution images of 337 subjects taken in four distinct sessions (Figure 3.1). Each scene was captured using 15 cameras in 19 different illumination conditions. Each camera represents a different head yaw angle, ranging from $-90°$ to $90°$ in steps of $15°$. The remaining two cameras are mounted near the ceiling, simulating a surveillance camera setting. Different sessions have different facial expressions and amount to a total of six distinct possibilities: neutral, smiling, surprised, squinting, disgusted and screaming.

3.1.1 Multi-PIE Results

The ground-truth noses for the Multi-PIE dataset were generated semi-automatically, a 54x60 region was cropped around the annotated tip of the nose [24] on all images taken with all 13 cameras around the head, except for exactly two images, which had broken annotations. Random visual inspection of a few thousand images showed that this heuristic provided adequate precision.

Figure 3.1: Image samples from the Multi-PIE dataset

Two main subsets were stipulated, one for testing and one for training, in such a way that all images from a given subject belong to the same subset. Due to the highly increasing training and testing complexity when using more images with the SVM approach, both subsets had to be further simplified to allow for experiments to be performed. Exactly 7000 random images were used from the training subset to train the SVM and 10000 random images from the testing subset were used for quickly evaluating the performance. However, once all parameters were set, a more thorough test was performed with the whole testing subset (355900 images). The ideal number of subregions was estimated using the smaller testing subset (Table 3.1), 81 regions yield the best results.

Table 3.1: Achieved accuracy with different numbers of subregions on the Multi-PIE dataset

Subregions	**1**	**4**	**9**	**16**	**25**	**36**	**49**	**64**	**81**	**100**
Accuracy	73.27%	91.04%	91.75%	91.03%	90.16%	85.22%	94.01%	93.08%	**94.13%**	93.83%

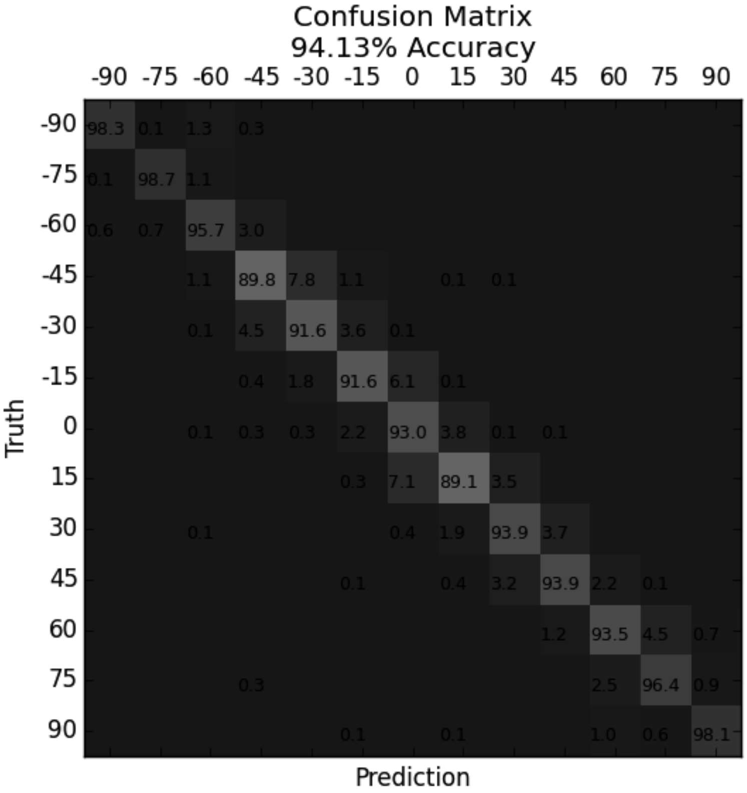
Confusion matrices for the tests on the reduced subsets using SVM-NosePose are presented in Figure 3.3(a) and 3.3(b). Results are similar when using the whole dataset, as seen in Figure 3.3(c) and 3.3(d), meaning that the training performed on 7000 images generalizes well for the whole dataset.

The comparative results between our approaches and [10] and [13] are shown in Table 3.2. Both our approaches achieve comparable or better results than the compared methods even when tested in much larger subsets. Peng *et al.* [11] reported the mean absolute error (MAE) and standard deviation (SD) when estimating the pose on 3600 images from the Multi-PIE dataset, the values are 4.62°and 3.89°, respectively. Our SVM-NosePose achieves 1.04°MAE and 4.85°SD for 10000 images and 1.06°MAE and 5.07°SD for 355900.

We also present the distributions of our results, for visual evaluation. Figure 3.3(e) and Figure 3.3(f) show the results for SVM-NosePose using 10000 and 355900 images, respectively.

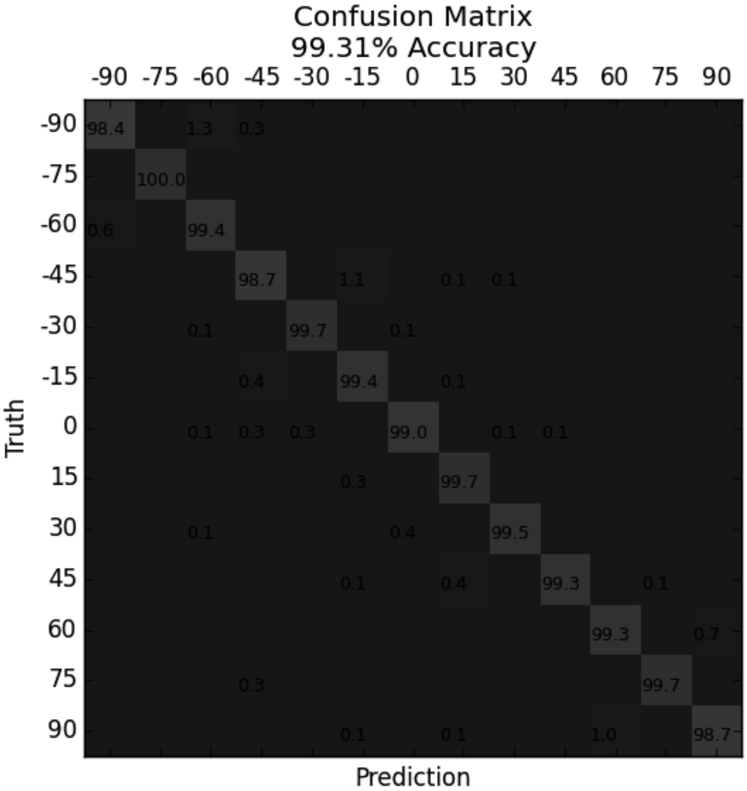
Table 3.2: Comparative results of the yaw estimation on Multi-PIE

	Strict	Weak
SVM-NosePose (10000 images)	94.13%	99.31%
SVM-NosePose (355900 images)	94.11%	99.29%
[10] (900 images)	91.40%	99.99%
[13] (5200 images)	94.46%	–



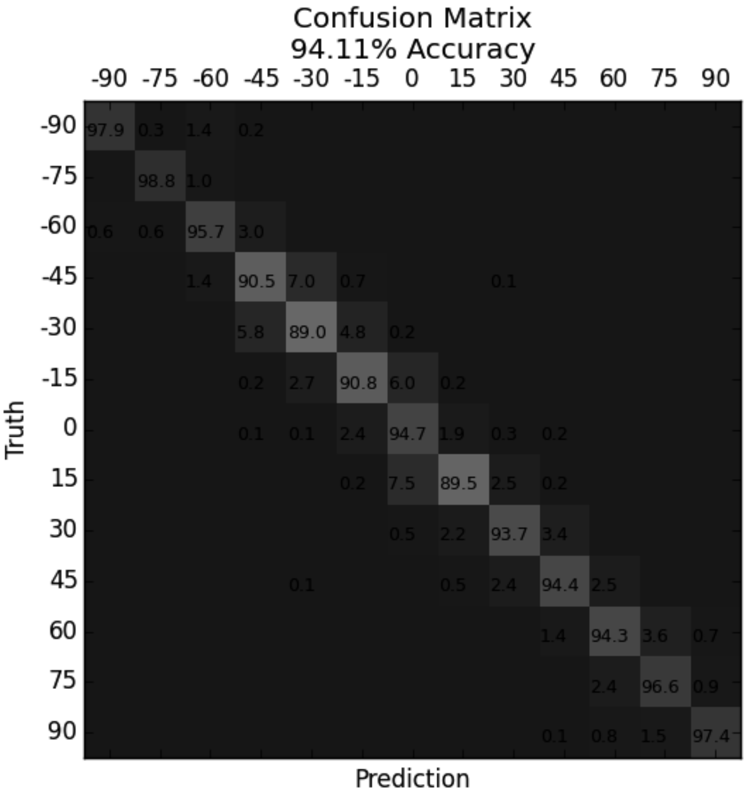
	-90	-75	-60	-45	-30	-15	0	15	30	45	60	75	90
-90	98.3	0.1	1.3	0.3									
-75	0.1	98.7	1.1										
-60	0.6	0.7	95.7	3.0									
-45			1.1	89.8	7.8	1.1		0.1	0.1				
-30			0.1	4.5	91.6	3.6	0.1						
-15				0.4	1.8	91.6	6.1	0.1					
0			0.1	0.3	0.3	2.2	93.0	3.8	0.1	0.1			
15						0.3	7.1	89.1	3.5				
30			0.1				0.4	1.9	93.9	3.7			
45						0.1		0.4	3.2	93.9	2.2	0.1	
60										1.2	93.5	4.5	0.7
75				0.3							2.5	96.4	0.9
90						0.1		0.1			1.0	0.6	98.1

(a) Strict evaluation 10000 images SVM-NosePose

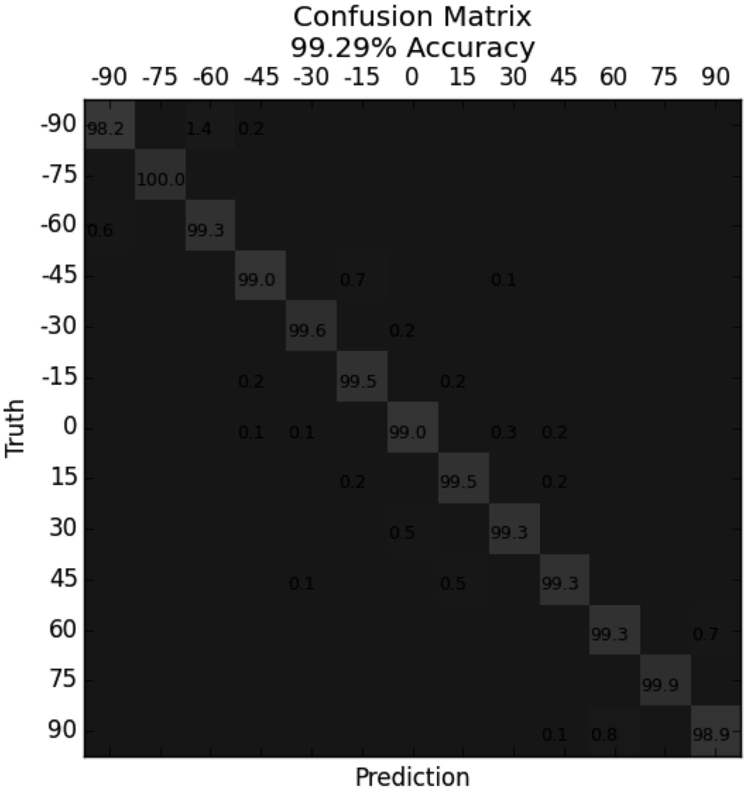


	-90	-75	-60	-45	-30	-15	0	15	30	45	60	75	90
-90	98.4		1.3	0.3									
-75		100.0											
-60	0.6		99.4										
-45				98.7		1.1		0.1	0.1				
-30			0.1		99.7		0.1						
-15				0.4		99.4		0.1					
0			0.1	0.3	0.3		99.0		0.1	0.1			
15						0.3		99.7					
30			0.1				0.4		99.5				
45						0.1		0.4		99.3		0.1	
60											99.3		0.7
75				0.3								99.7	
90						0.1		0.1			1.0		98.7

(b) Weak evaluation 10000 images SVM-NosePose

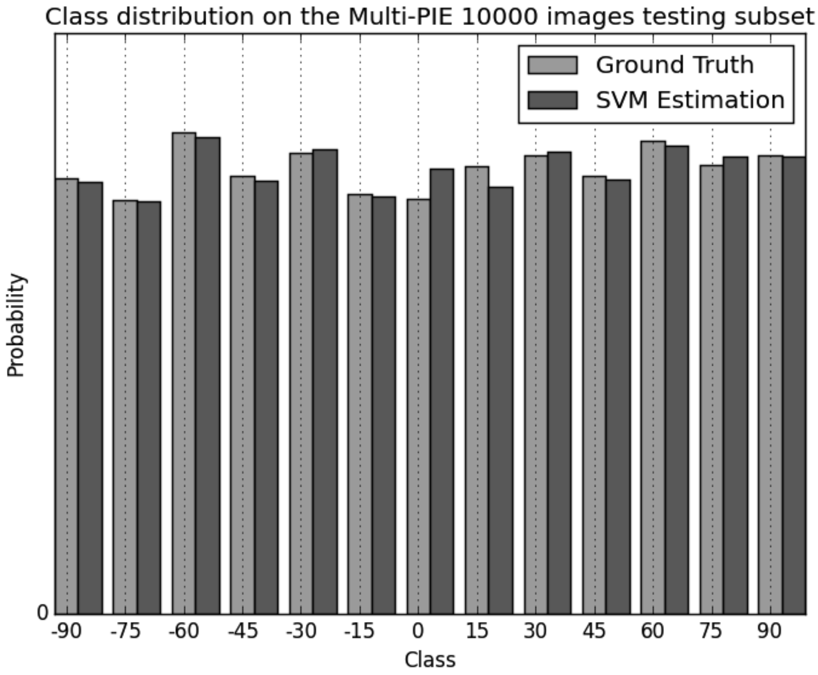


(c) Strict evaluation 355900 images SVM-NosePose

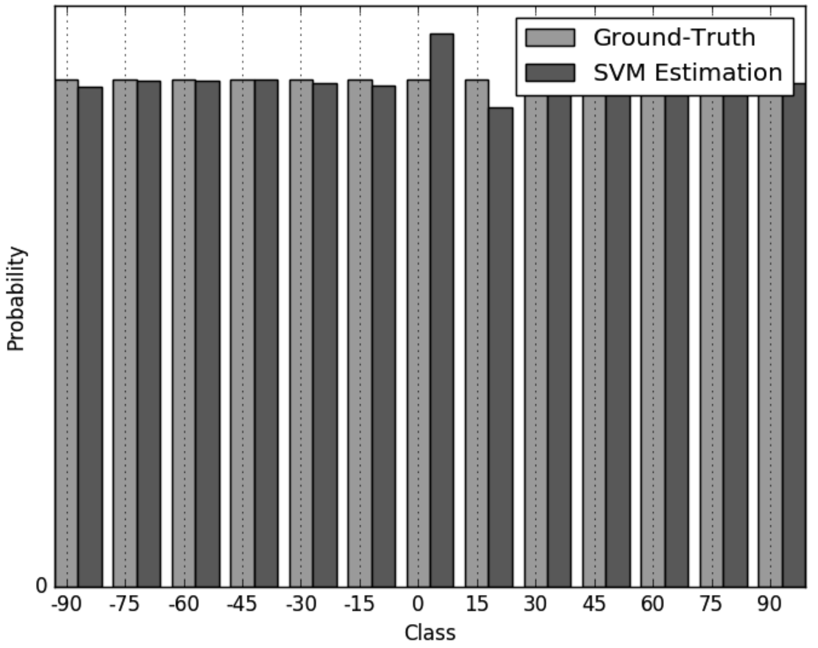


(d) Weak evaluation 355900 images SVM-NosePose

Figure 3.2: Confusion matrices when estimating the yaw on the Multi-PIE dataset



(e) Distribution for 10000 images



(f) Distribution for 355900 images

Figure 3.3: Head pose estimation distribution on the Multi-PIE dataset

3.2 Pointing'04

The Pointing'04 dataset [25] includes 15 subjects and two different series of 93 images of each. Each image represents a different combination of yaw and pitch in the range from $-90°$ to $90°$ in both axes. Totaling 2790 images (Figure 3.4) , the dataset is mostly composed of white Europeans (73% of the subjects). Only a small portion of the dataset (27%) is composed of other ethnicities. No expressions or lighting variations are present.

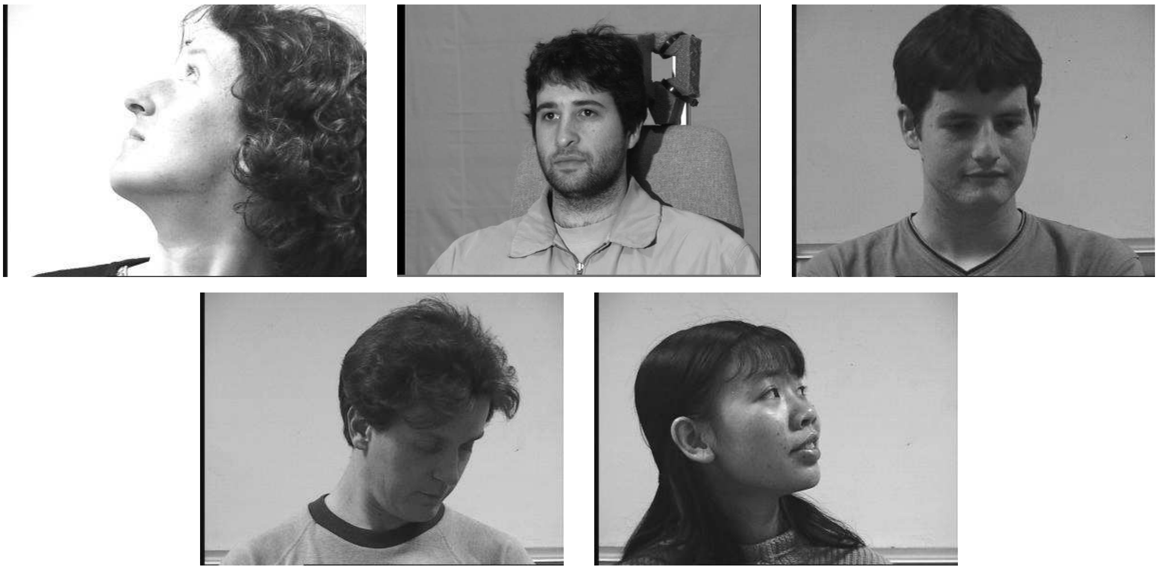


Figure 3.4: Image samples from the Pointing'04 dataset

3.2.1 Pointing'04 Experiments

Experiments performed on the Pointing'04 dataset were conducted similar to Pawelczyk and Kawulok [12]. Ten subjects were chosen for the training subset and 5 for testing, amounting to 1842 training images and 836 test images. The pose annotations are already part of the dataset, even though they are not precise [12], and Pawlczyk and Kawulok [12] shared the nose region annotations via personal communication to allow for a fairer comparison.

Tests were performed estimating the head yaw, pitch and both yaw and pitch at the same time, the optimal number of subregions are presented in Table 3.3. Confusion matrices were generated for both strict and weak evaluation protocols when estimating the yaw (Figure 3.5) and the pitch (Figure 3.6) separately. When estimating the pose on both axes at the same time, due to the large number of classes, the resulting confusion matrices are unreadable.

Table 3.3: Achieved accuracy with different numbers of subregions on the Pointing'04 dataset

Subregions	**1**	**4**	**9**	**16**	**25**	**36**	**49**	**64**	**81**	**100**
Yaw Acc.	48.21%	54.90%	58.97 %	57.18%	49.64%	48.56%	58.97%	**61.36%**	61.00%	59.09%
Pitch Acc.	41.27%	57.89%	56.22%	51.67%	46.41%	39.47%	56.94%	**58.49%**	57.18%	55.02%
Both Acc.	25.12%	35.53%	38.28%	**40.55%**	37.68%	37.92%	36.36%	36.96%	38.40%	38.52%

When comparing against Pawelczyk and Kawulok's [12] method, we implemented the approach and estimated the SVM parameters, but were unable to achieve the same accuracy as reported. Kawulok shared the original SVM parameters that were used, but we were still unable to achieve similar accuracy. We present all our results and compare them against all of Pawelczyk and Kawulok's [12] method's possibilities in Tables 3.4, 3.5, 3.6 when estimating the yaw, pitch and both, respectively.

The ground-truth distribution and the distribution we obtained when estimating the yaw with both our methods are shown in Figure 3.7(a). A similar graph, but for the pitch estimation is shown in Figure 3.7(b).

Out method outperforms the compared approach in all aspects. When the weak evaluation protocol is considered, good performance is reached even for estimating the pitch and yaw simultaneously when each class only has a few samples in the training set.

Table 3.4: Comparative results for estimating the yaw on Pointing'04

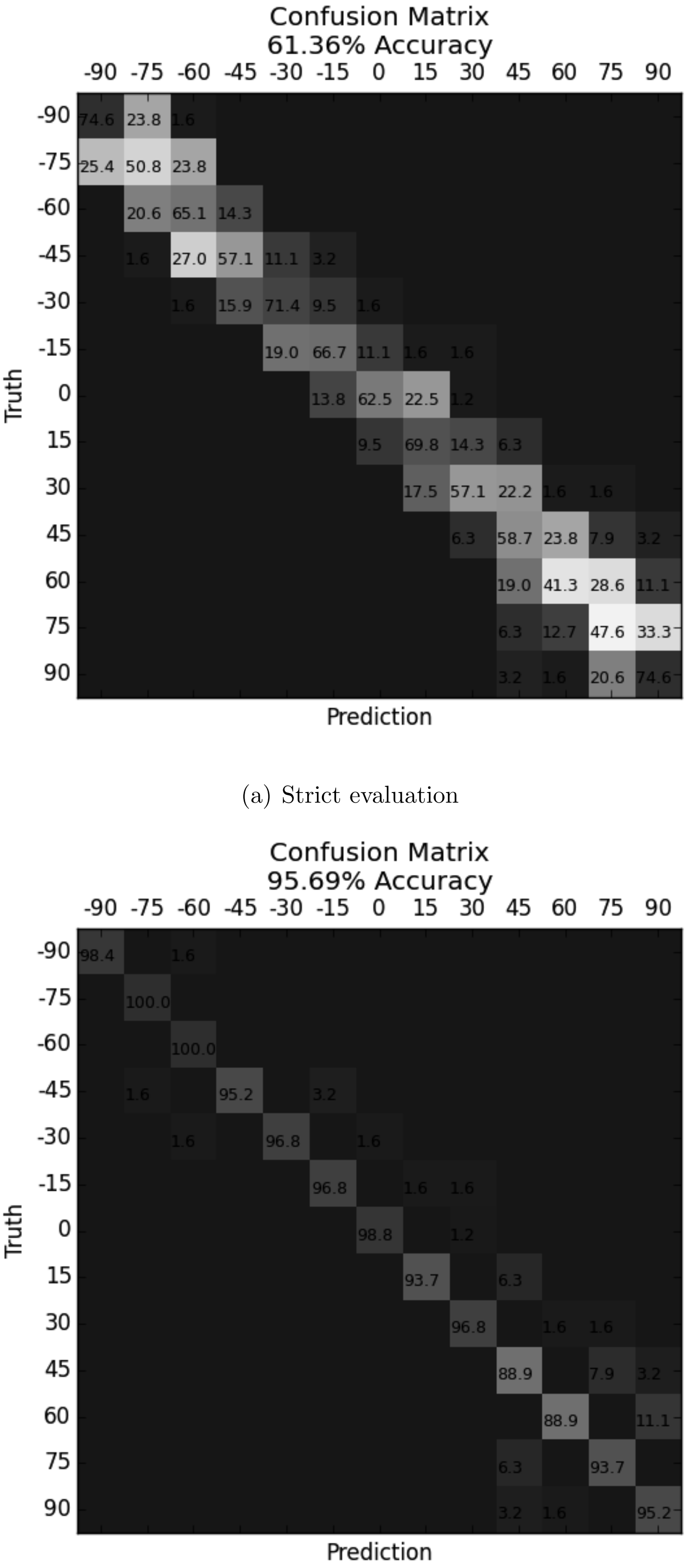
	Strict	Weak
[12] (reported)	56.99%	93.41%
[12] (original parameters)	18.06%	45.45%
[12] (estimated parameters)	28.59%	69.38%
SVM-NosePose	61.36%	95.69%

Table 3.5: Comparative results when estimating the pitch on Pointing'04

	Strict	Weak
[12] (reported)	47.91%	77.80%
[12] (original parameters)	13.88%	37.44%
[12] (estimated parameters)	28.23%	59.57%
SVM-NosePose	58.49%	94.50%

Table 3.6: Comparative results for simultaneous estimation of the pitch and yaw on Pointing'04

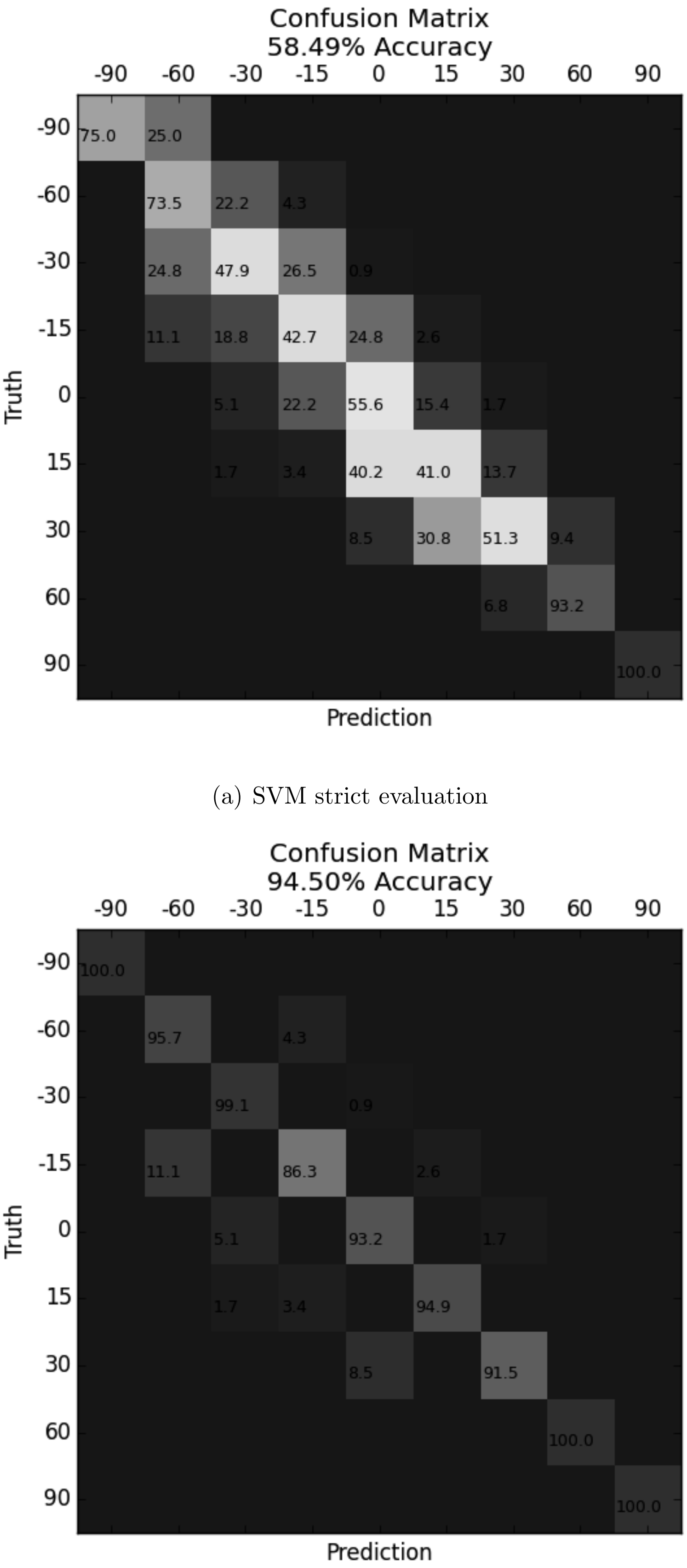
	Strict	Weak
[12] (reported)	27.41%	73.46%
[12] (original parameters)	3.47%	27.27%
[12] (estimated parameters)	14.11%	66.27%
SVM-NosePose	40.55%	91.87%



(a) Strict evaluation

(b) Weak evaluation

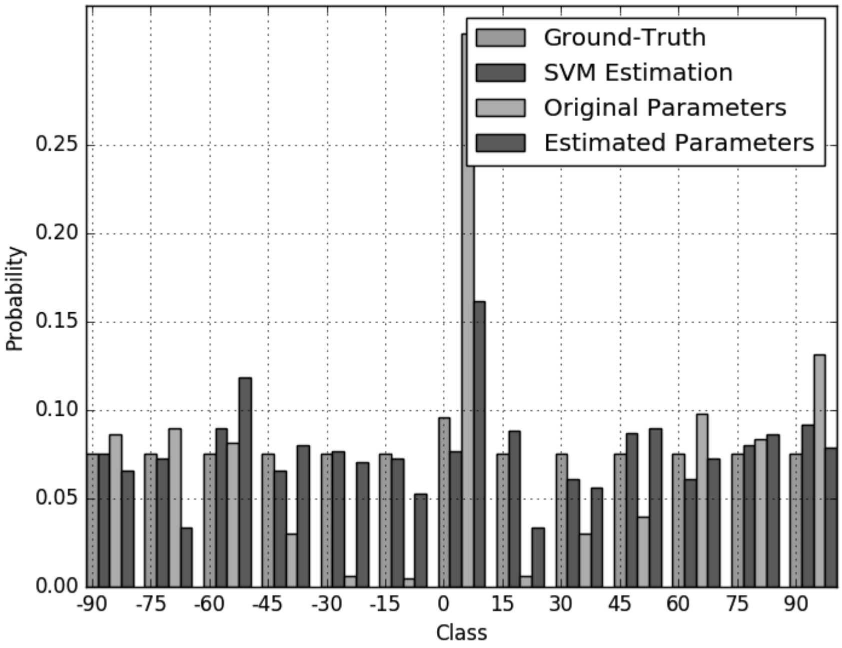
Figure 3.5: Confusion matrix estimating the yaw on the Pointing'04 dataset



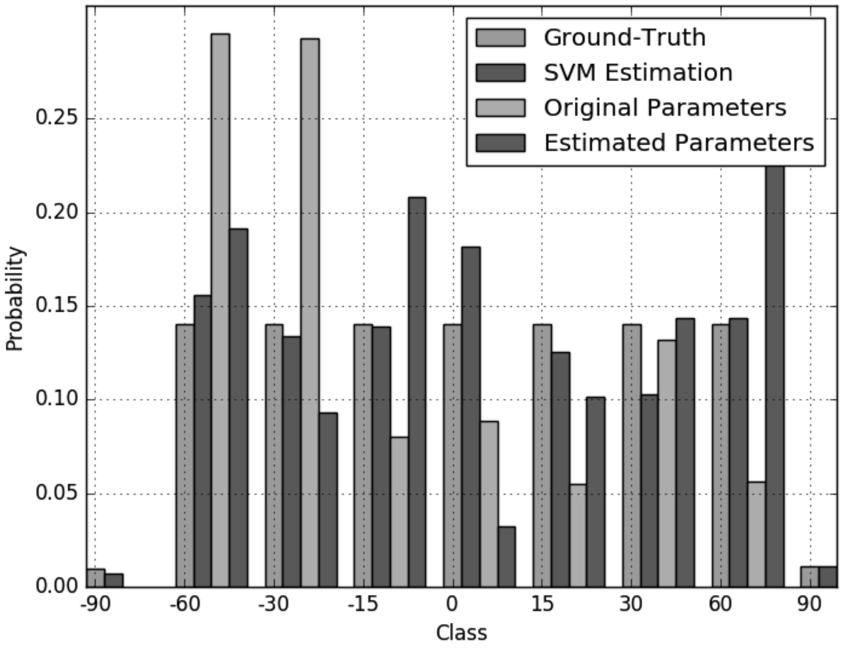
(a) SVM strict evaluation

(b) SVM weak evaluation

Figure 3.6: Confusion matrices when estimating the pitch on the Pointing'04 dataset



(a) Result distribution on the yaw axis



(b) Result distribution on the pitch axis

Figure 3.7: Result distributions on the yaw and pitch axes

3.3 McGillFaces

The McGillFaces database [26] consists of 18000 unconstrained frames extracted from video sequences of 60 unique subjects and the corresponding labels (face mask, gender and head yaw). However, only 10500 frames are available publicly and only 6665 frames have the head pose annotation. During recording, the subjects were placed in different illumination and background conditions and were allowed free movement and object interaction. This resulted in a variety of arbitrary face scales, expressions, viewpoints and occlusions. Figure 3.8 shows sample images taken from the McGillFaces dataset.

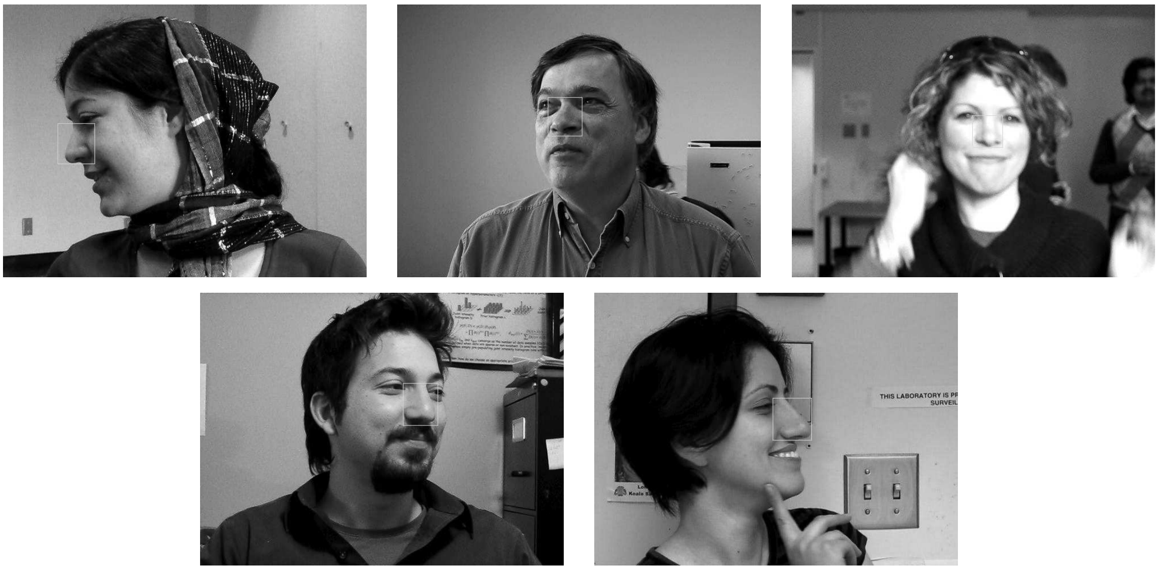


Figure 3.8: Image samples from the McGillFaces dataset

3.3.1 McGillFaces Results

To perform our experiments, we manually annotated the nose region in all provided images with the head yaw annotation and used 6665 total images for our tests, 3208 for training and 3457 for testing, without overlapping subjects. We present the achieved accuracy for all possible number of subregions in Table 3.7.

Table 3.7: Achieved accuracy with different numbers of subregions on the McGillFaces dataset

Subregions	**1**	**4**	**9**	**16**	**25**	**36**	**49**	**64**	**81**	**100**
Accuracy	53.08%	55.34%	53.34%	55.71%	53.69%	58.06%	**59.24%**	58.66%	58.57%	56.78%

The results obtained with our approach are shown in Figure 3.9 in the form of confusion matrices for both strict and weak evaluation protocols. We provide the distribution of our estimation and the ground-truth labels in Figure 3.10. Table 3.9 shows our results compared to Demirkus *et al.*, who reported, via personal communication, to have used all 18000 images for training and testing. Because of this, the comparison is, unfortunately, biased as our method would have benefited from using almost three times as many images.

3.3.2 Filtered McGillFaces Results

We investigated and evaluated the reliability of the provided pose annotations, since they were annotated semi-automatically [26]. Each image in the dataset with a label was evaluated by at least two different people, one by one in random order and was tagged either good or inconsistent. This visual analysis of the provided ground-truth annotation showed that approximately one fifth of the images were assigned inconsistent labels (Figure 3.11).

Because of this, we also evaluate our algorithm in a filtered version of the McGillFaces dataset, containing only the images tagged as good. It contains 5329 total images, 2475 for training and 2854 for testing. The increase in accuracy is evident, our results in this subset are presented in Figure 3.12 as confusion matrices and distributions, Table 3.8 presents the optimal number of subregions.

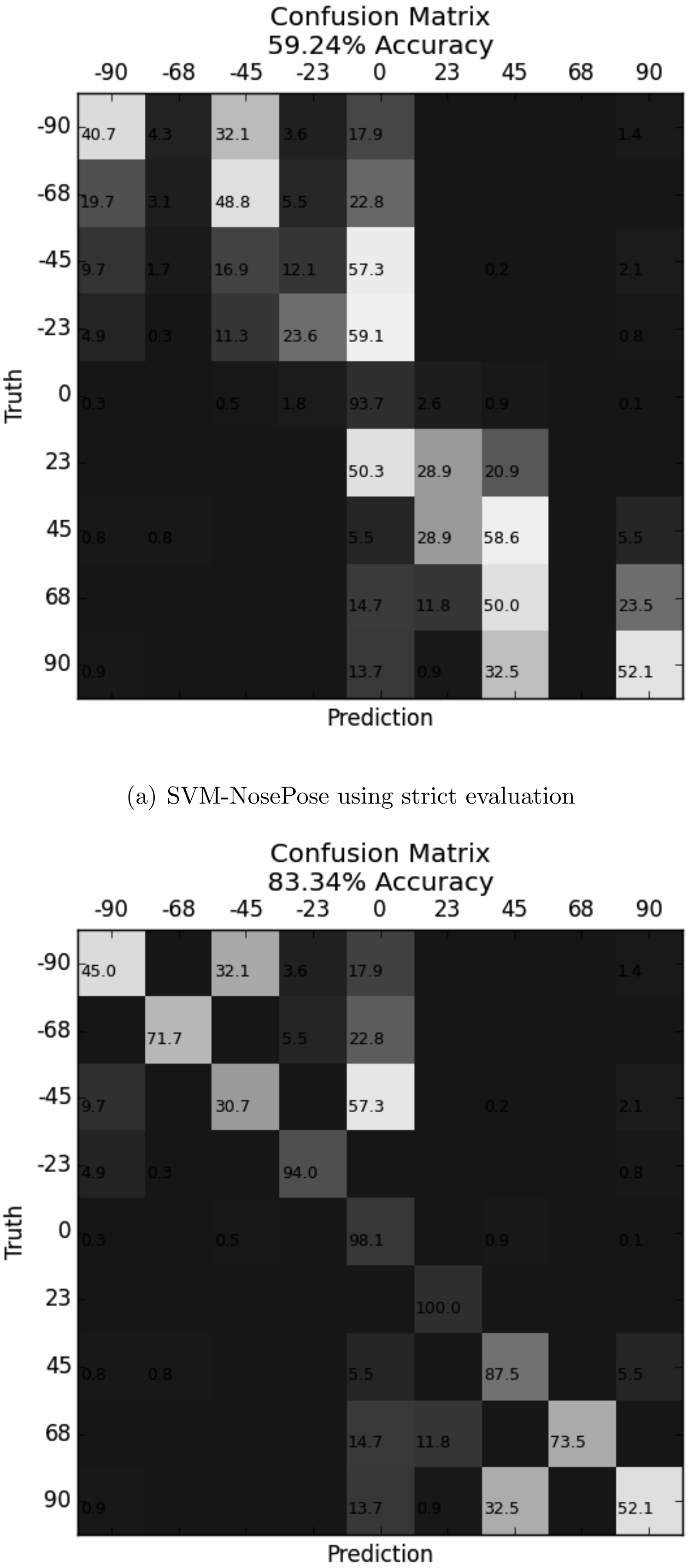
Table 3.8: Achieved accuracy with different numbers of subregions on the Filtered McGillFaces dataset

Subregions	**1**	**4**	**9**	**16**	**25**	**36**	**49**	**64**	**81**	**100**
Accuracy	57.95%	62.37%	66.54%	67.55%	61.88%	69.38%	**70.71%**	69.80%	69.48%	68.04%

We also present all results in the McGillFaces dataset in Table 3.9. Our accuracy is lower than Demirkus *et al.* [13], however we used a smaller training subset, due to the whole dataset not being available, and make no use of temporal information. Higher accuracy is achieved in the filtered dataset, indicating that some labels are inconsistent.

Table 3.9: Original McGillFaces vs. Filtered McGillFaces both approaches.

	Strict	Weak
SVM (Original McGill, 6665 images)	59.24%	83.34%
SVM (Filtered McGill, 5329 images)	70.71%	92.68%
[13] (18000 images)	79.02%	–



	-90	-68	-45	-23	0	23	45	68	90
-90	40.7	4.3	32.1	3.6	17.9				1.4
-68	19.7	3.1	48.8	5.5	22.8				
-45	9.7	1.7	16.9	12.1	57.3		0.2		2.1
-23	4.9	0.3	11.3	23.6	59.1				0.8
0	0.3		0.5	1.8	93.7	2.6	0.9		0.1
23					50.3	28.9	20.9		
45	0.8	0.8			5.5	28.9	58.6		5.5
68					14.7	11.8	50.0		23.5
90	0.9				13.7	0.9	32.5		52.1

(a) SVM-NosePose using strict evaluation

	-90	-68	-45	-23	0	23	45	68	90
-90	45.0		32.1	3.6	17.9				1.4
-68		71.7		5.5	22.8				
-45	9.7		30.7		57.3		0.2		2.1
-23	4.9	0.3		94.0					0.8
0	0.3		0.5		98.1		0.9		0.1
23						100.0			
45	0.8	0.8			5.5		87.5		5.5
68					14.7	11.8		73.5	
90	0.9				13.7	0.9	32.5		52.1

(b) SVM-NosePose using weak evaluation

Figure 3.9: Confusion matrices estimating the yaw on the McGillFaces dataset

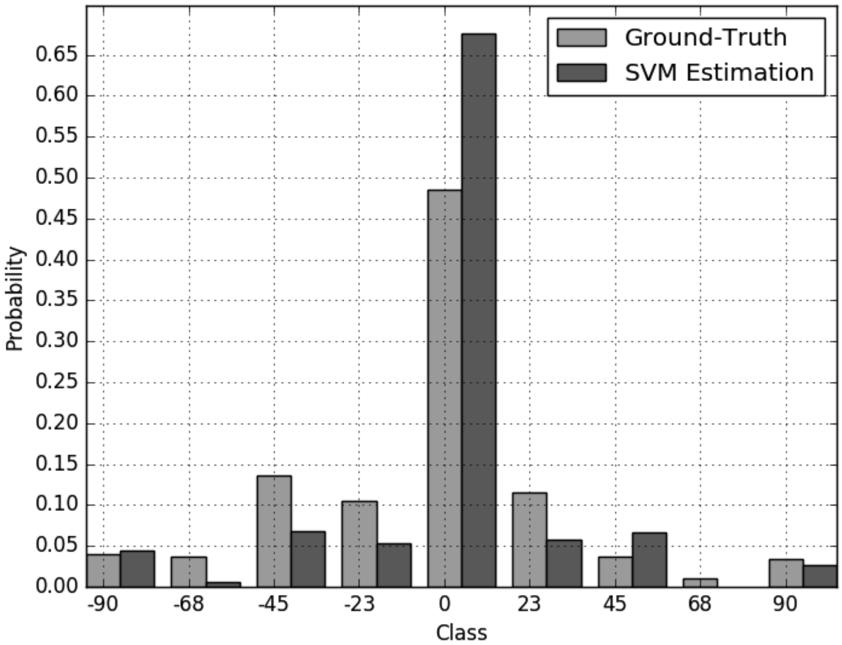


Figure 3.10: Results distribution on McGillFaces

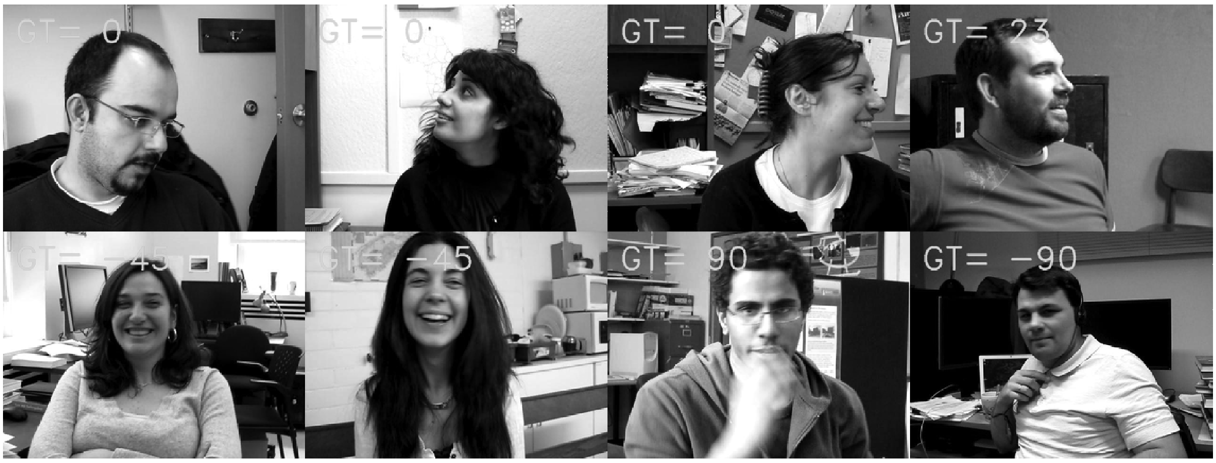
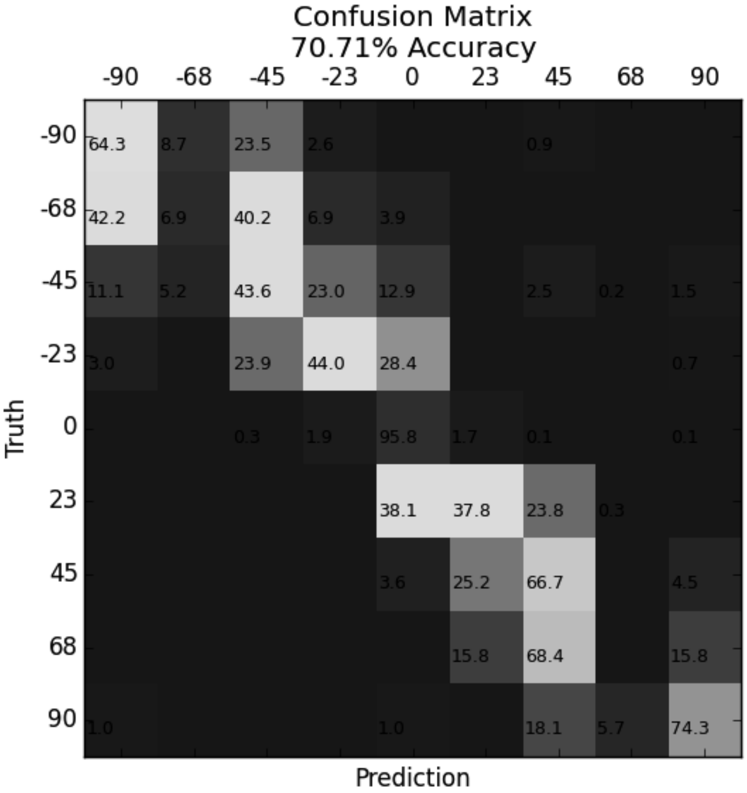
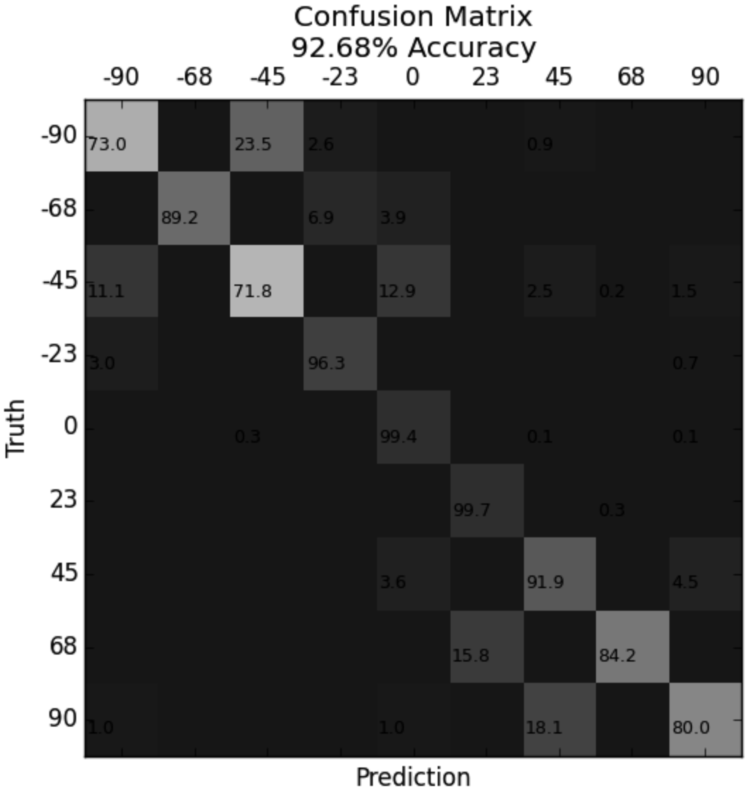


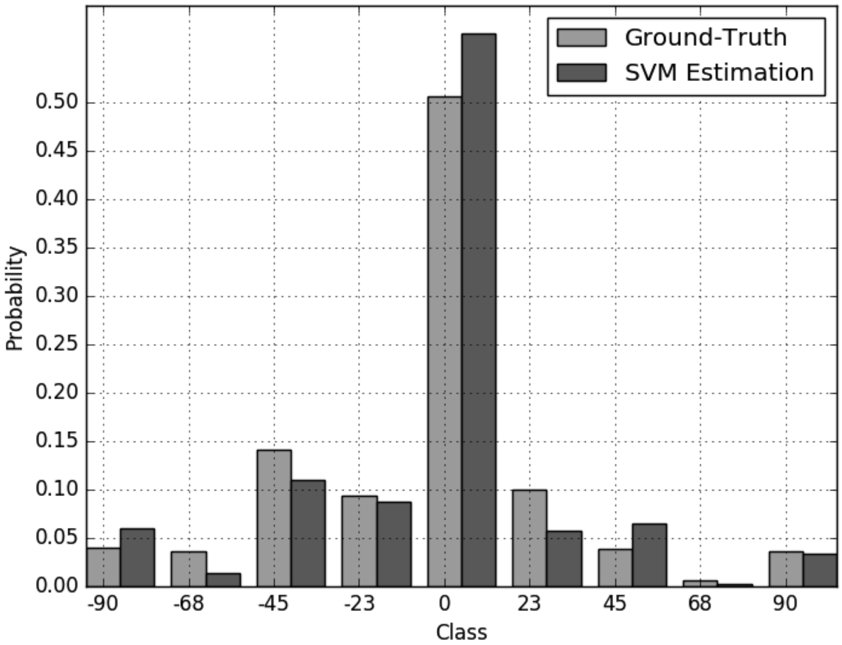
Figure 3.11: Examples of inconsistent ground-truth annotations in the McGillFaces dataset



(a) SVM using strict evaluation



(b) SVM using weak evaluation



(c) Results Distribution

Figure 3.12: Confusion matrix and distribution on the Filtered McGillFaces dataset

3.4 SFEW

SFEW (Static Facial Expressions in the Wild) [27] is a dataset dedicated for expression recognition evaluation, it contains approximately 1700 images divided into three subsets, training, validation and test. All images are automatically extracted movie frames based on the detected expression and contain challenging variations in illumination, pose, expression and scale. Sample images from the dataset can be seen in Figure 3.13.

Figure 3.13: Image samples from the SFEW dataset

3.4.1 SFEW Results

The SFEW dataset is divided into three subsets, training, validation and testing. As the SFEW dataset only comes with annotated expressions, we annotated both the head pose and nose region in all images, totaling 957 training images, 436 validation images and 372 testing images. We trained our method using both the training and validation subsets and evaluated the performance on the testing subset. Figure 3.14 contains our resulting confusion matrix and distributions, Table 3.10 presents the best number of subregions.

The lack of profile head pose images in the training subset causes our estimation to

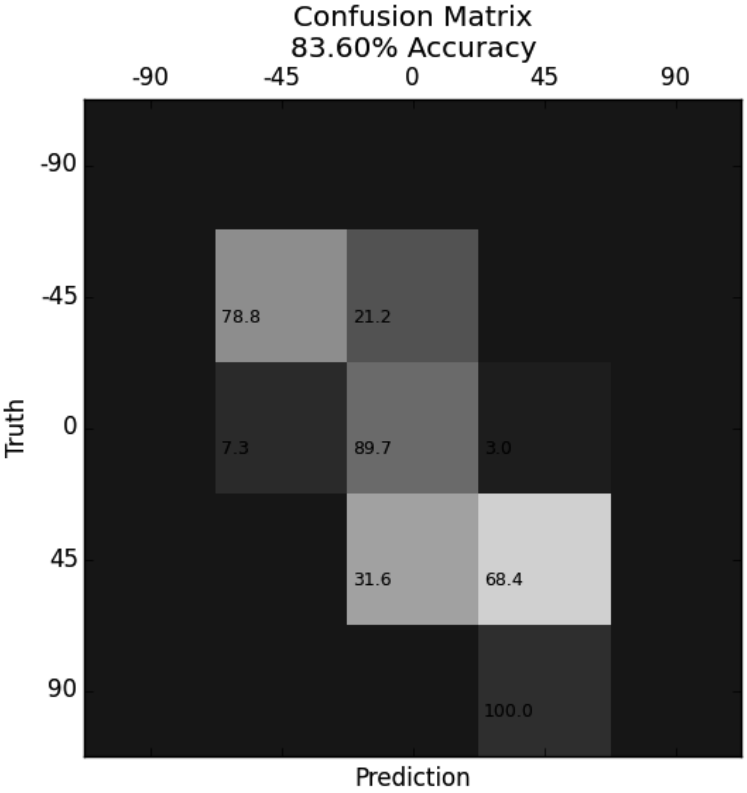
Table 3.10: Achieved accuracy with different numbers of subregions on the PaSC dataset

Subregions	**1**	**4**	**9**	**16**	**25**	**36**	**49**	**64**	**81**	**100**
Accuracy	65.86%	72.31%	77.42%	75.27%	76.61%	81.45%	**83.60%**	79.57%	81.99%	76.34%

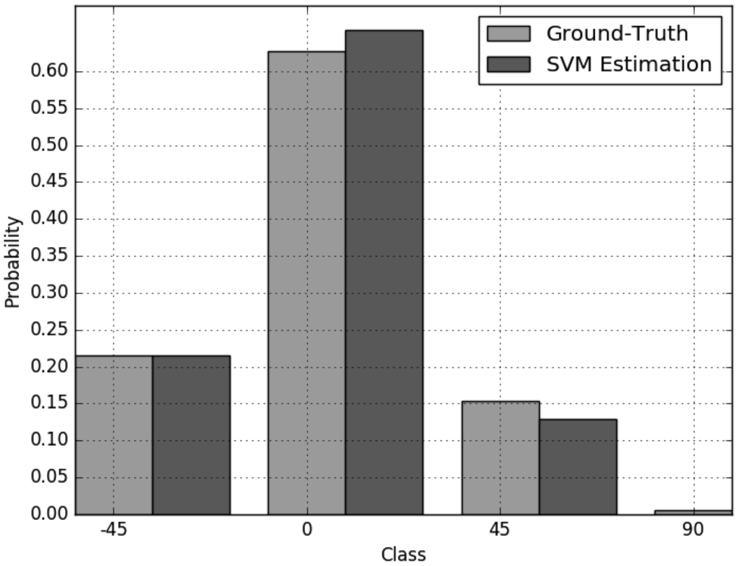
favor frontal and half-frontal poses even when presented with profile images. Our comparative results (displayed in Table 3.11) show that our algorithm can achieve satisfactory performance in extreme cases with facial expressions, poor illumination and other adversities present in the SFEW dataset.

Table 3.11: Comparative results on the SFEW dataset

	Accuracy
SVM-NosePose	83.60%



(a) Confusion Matrix



(b) Distribution

Figure 3.14: Confusion matrix and distribution on the SFEW dataset

3.5 PaSC

PaSC (Point-and-Shoot Challenge) [22] is an in-the-wild dataset with both videos and still frames subsets for face recognition. In this paper, we focus on the stills subset. It contains 9376 challenging images of 293 subjects (Figure 3.15) of different ethnical backgrounds in different environments, illumination conditions, poses and sensors.

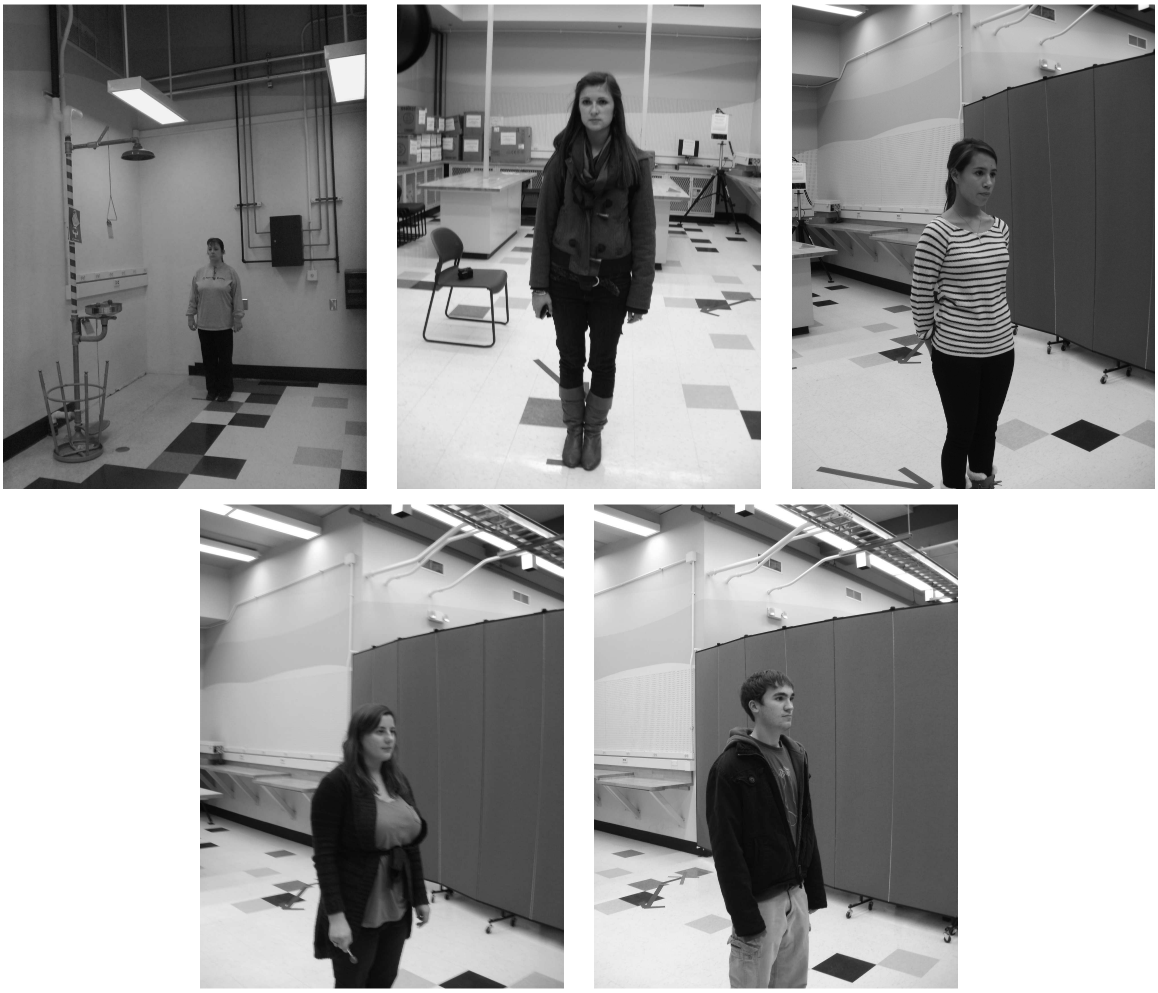


Figure 3.15: Image samples from the PaSC dataset

3.5.1 PaSC Results

The PaSC dataset is pre-divided into training and testing subsets optimized for evaluating face recognition. This subdivision proved to be poor for evaluating head pose estimation, as the distribution of the poses in the subsets vary greatly (Figure 3.16(a)). We redivided the images in a way that this difference would be less noticeable (Figure 3.16(b)) while guaranteeing that no subject is present in both subsets.

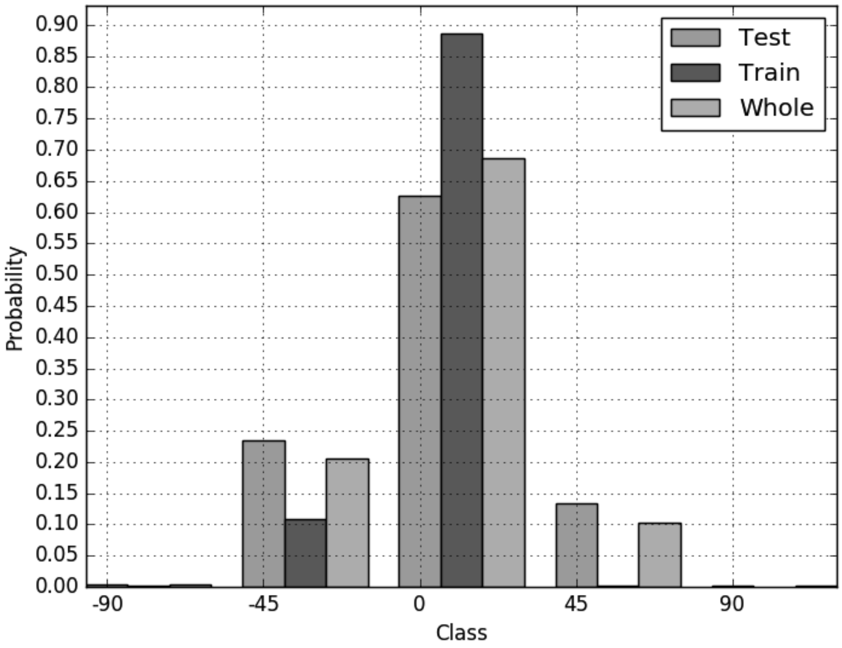
Table 3.12: Achieved accuracy with different numbers of subregions on the PaSC dataset

Subregions	**1**	**4**	**9**	**16**	**25**	**36**	**49**	**64**	**81**	**100**
Accuracy	75.65%	81.05%	82.80%	75.51%	74.63%	77.32%	86.13%	86.42%	**86.91%**	85.41%

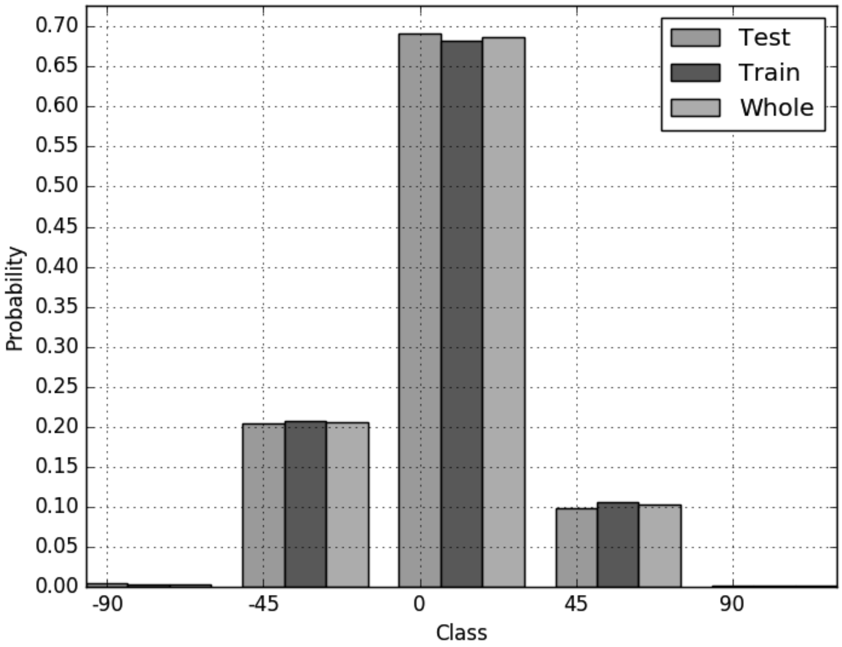
For this experiment, we used only the images on which we were able to manually annotate the nose region and the head pose (into 5 classes $[-90, 45, 0, 45, 90]$), resulting in 5784 training and 6243 testing still images. The best number of subregions is presented in Table 3.12, Figure 3.17 shows the confusion matrices and label distribution. Table 3.13 shows our results for both approaches. As a consequence of the head pose not being very well distributed in the dataset, both algorithms tend to be biased towards estimating the pose as frontal.

Table 3.13: Results obtained on PaSC dataset

	Strict
SVM-NosePose	86.91%

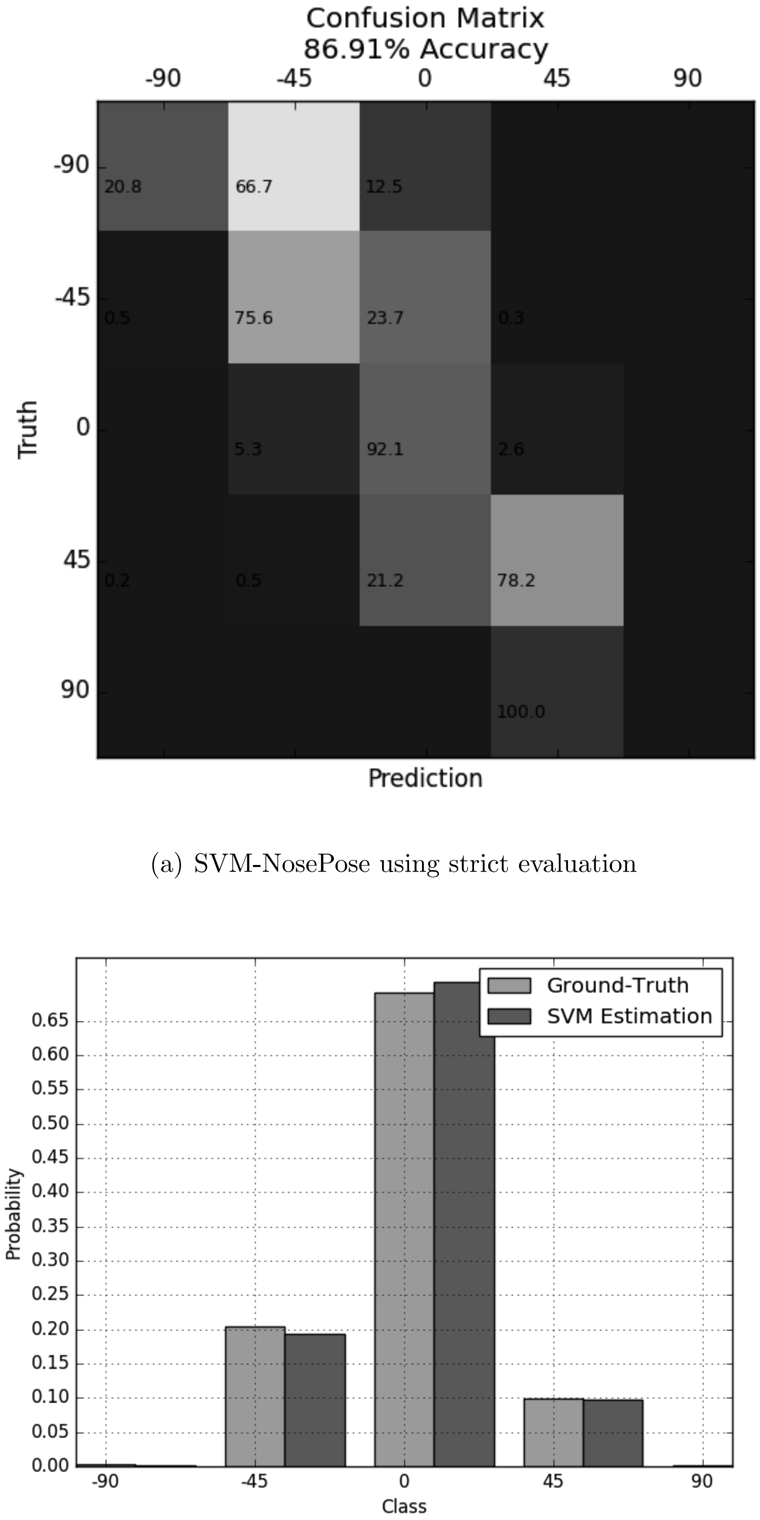


(a) Original PaSC subset distribution for face recognition



(b) Proposed PaSC subset distribution for head pose estimation

Figure 3.16: Different class distributions on the PaSC dataset



(a) SVM-NosePose using strict evaluation

(b) Distribution of our results

Figure 3.17: Results when estimating the yaw on the PaSC dataset

3.6 AFW

The annotated faces in-the-wild (AFW) [10] dataset containing 468 faces with landmark and pose annotations. Large variations in the background, pose, expression and subject appearance are present (Figure 3.18), as the images were extracted from Flickr and are all from real world in-the-wild scenarios.

Figure 3.18: Image samples from the AFW dataset

3.6.1 AFW Results

The AFW dataset contains no training subset. Just like Zhu and Ramanan [10], we trained using the provided 900 images subset from the Multi-PIE dataset, the same training data we used when experimenting on the dataset and split and augmented the subjects 14-fold, using 300 of them for training to evaluate our performance on AFW, yielding three distinct results. The annotations include, sometimes, multiple subjects in a single image with a precision of 15 degrees on the yaw pose. We manually annotated all the corresponding nose regions. Our results are shown as confusion matrices and distributions in Figures

3.19, 3.20 and 3.21. The optimal number of subregions is shown in Table 3.14. We compare our results against Zhu and Ramanan [10] in Table 3.15.

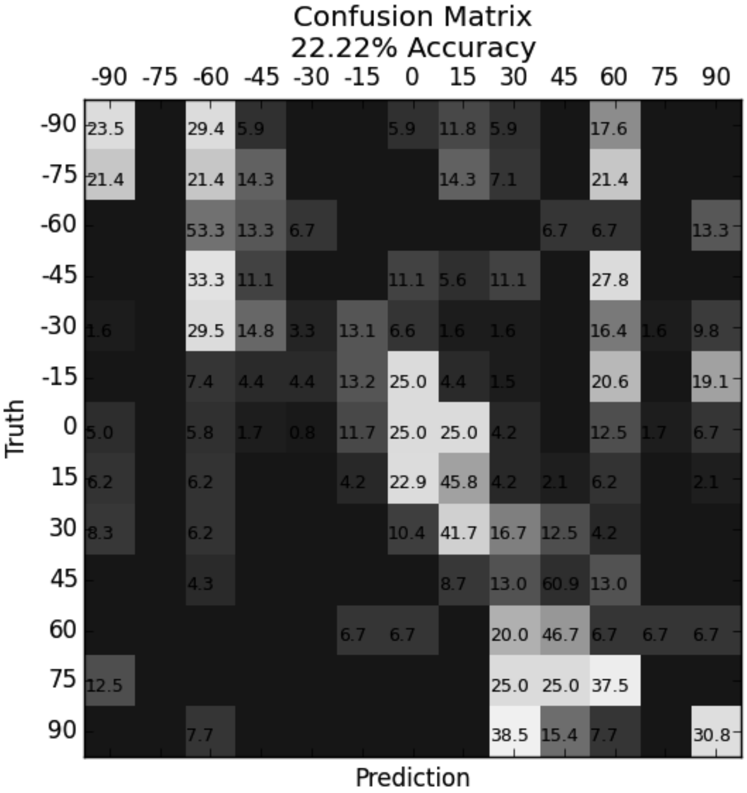
Table 3.14: Achieved accuracy with different numbers of subregions on the AFW dataset

Subregions	**1**	**4**	**9**	**16**	**25**	**36**	**49**	**64**	**81**	**100**
900	12.18%	16.03%	23.08%	23.50%	22.01%	**22.22%**	22.01%	20.30%	21.15%	20.51%
7000	10.90%	14.11%	20.09%	21.37%	22.86%	23.50%	**25.00%**	25.00%	23.93%	25.00%
Augmented	28.21%	37.07%	36.25%	36.00%	35.46%	35.25%	44.71%	44.46%	**44.82%**	44.21%

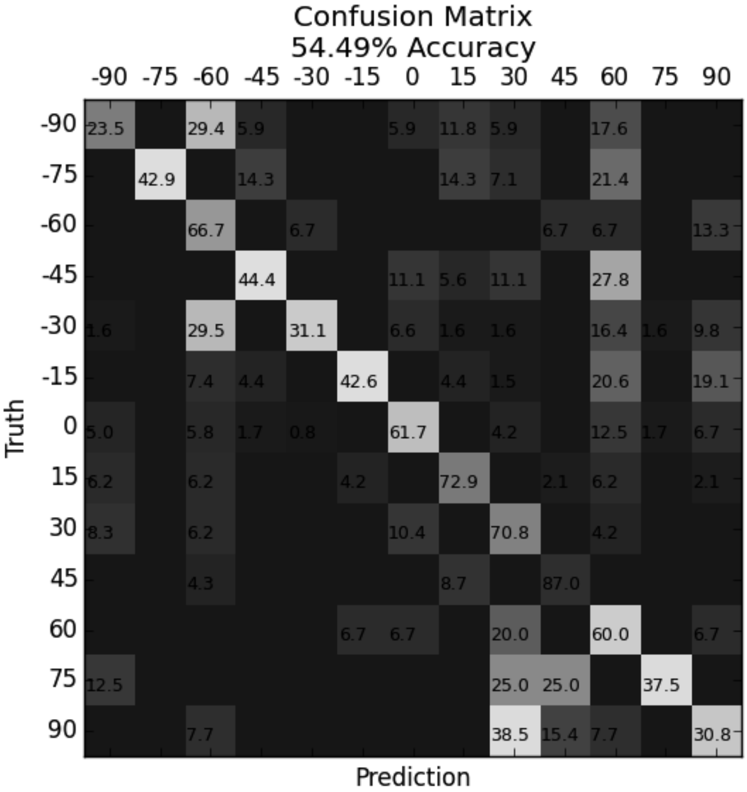
Table 3.15: Comparative results when estimating the yaw on the AFW dataset

	Strict	Weak
SVM-NosePose (900 images)	22.22%	54.49%
[10] (900 images)	–	81.00%
SVM-NosePose (7000 images)	25.00%	70.09%
Augmented ($268 * 14$ faces)	44.71%	81.26%

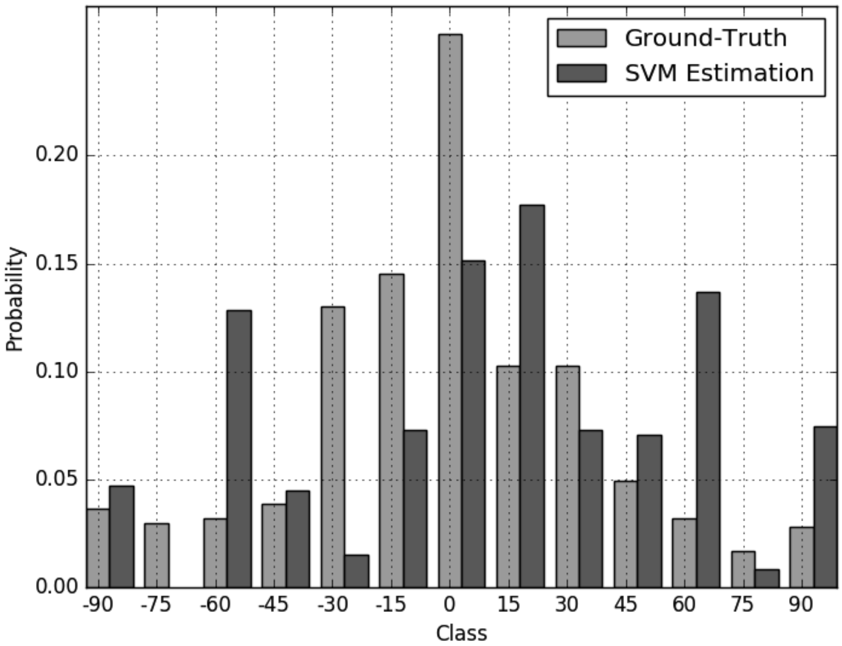
The increase in accuracy is clear when more images are used for training. Our algorithm has no problem estimating the pose for multiple subjects in a single image. However, the achieved mediocre accuracy is expected when training using a controlled environment dataset and estimating the pose on an in-the-wild dataset, when splitting the dataset for training, the results are consistent with the previous ones.



(a) SVM-NosePose using strict evaluation

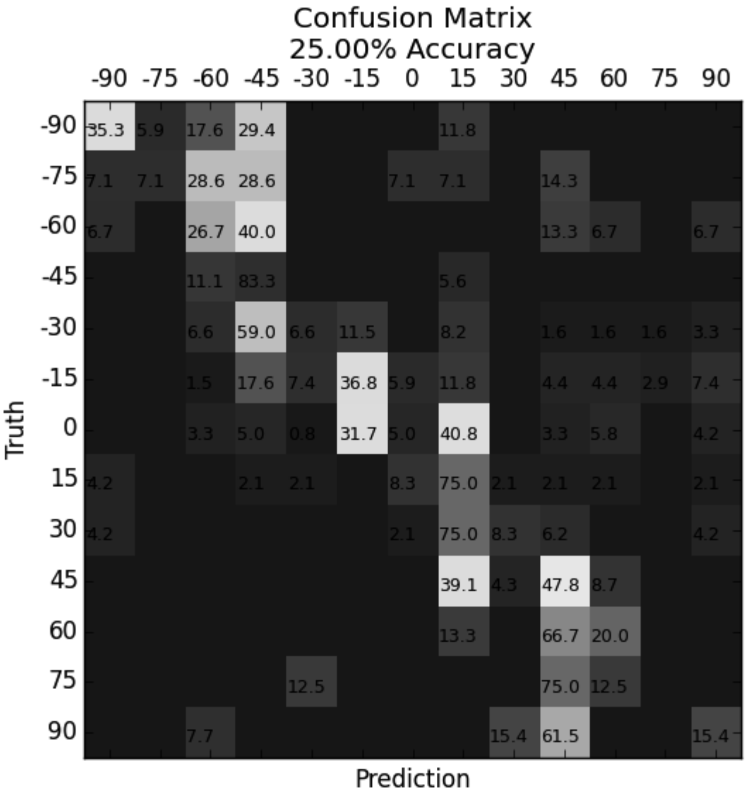


(b) SVM-NosePose using weak evaluation

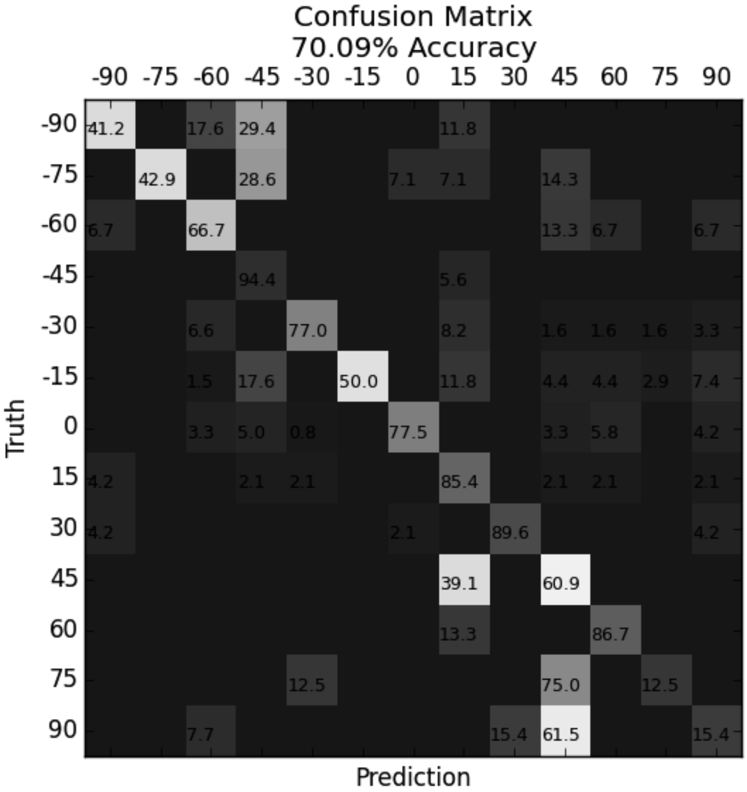


(c) Distribution of the results

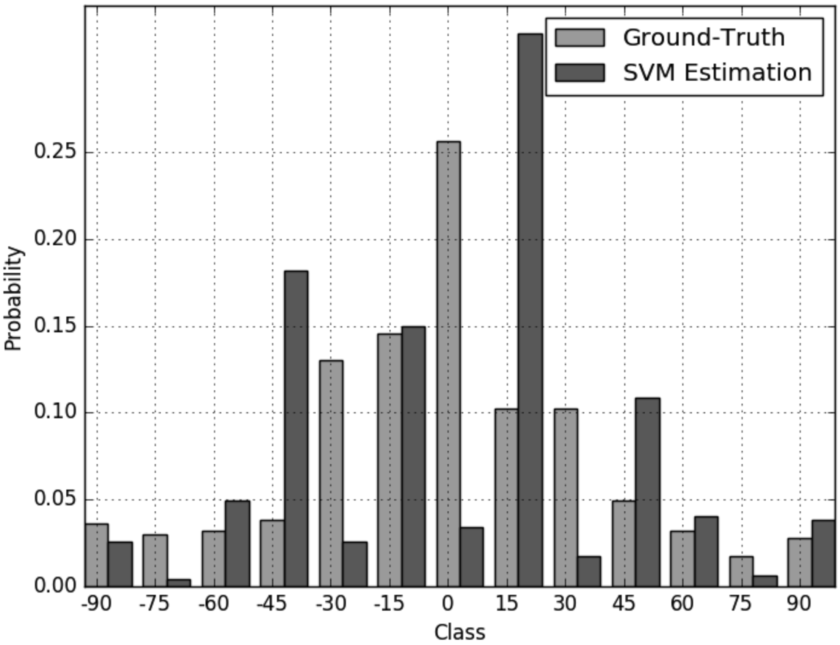
Figure 3.19: Results on the AFW dataset, using 900 images for training



(a) SVM-NosePose using strict evaluation

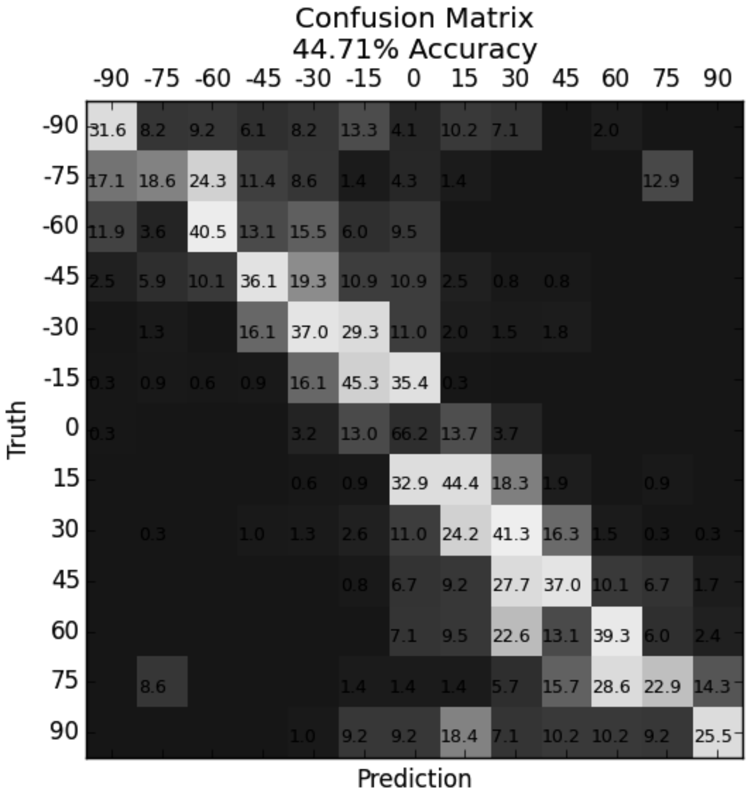


(b) SVM-NosePose using weak evaluation

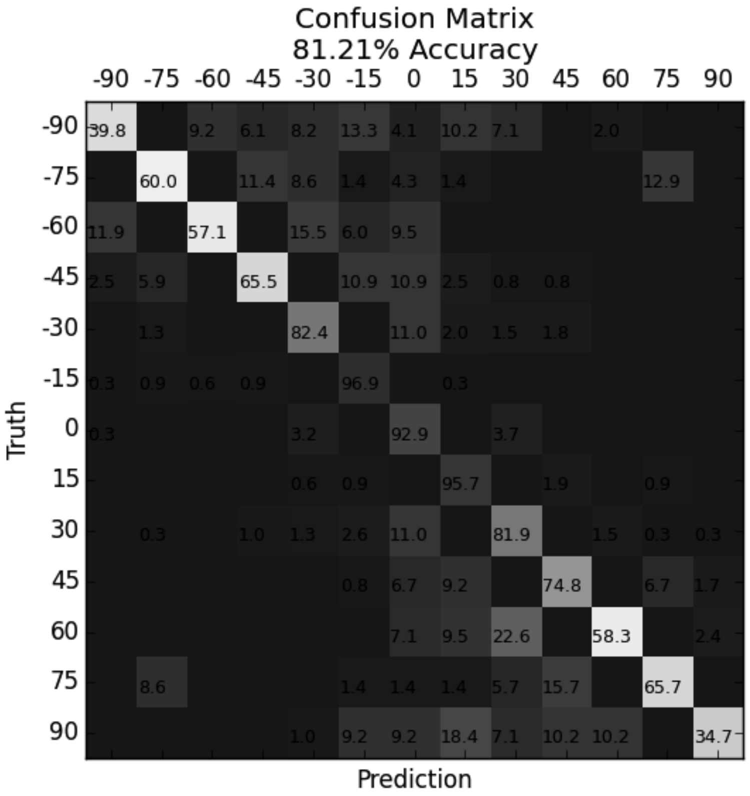


(c) Distribution of the results

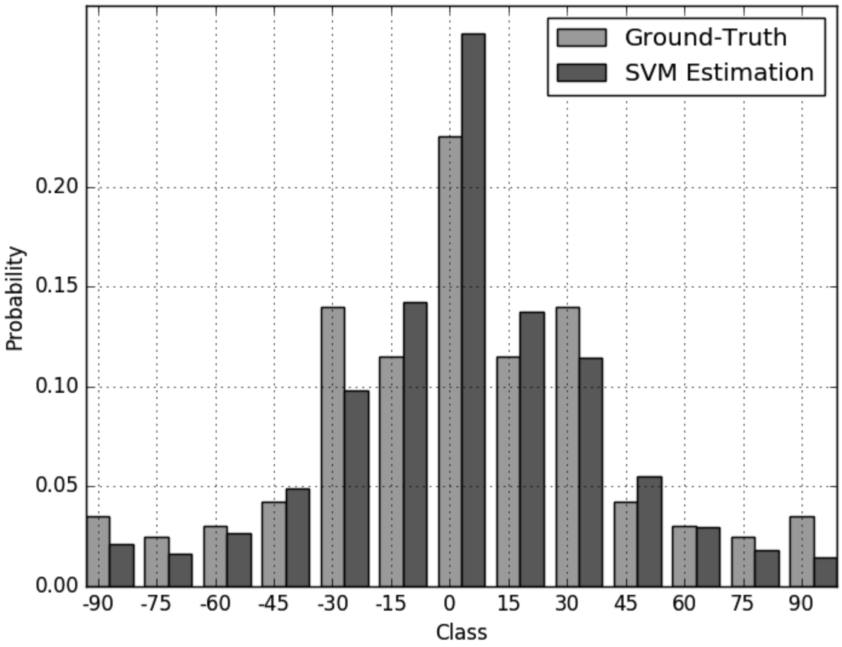
Figure 3.20: Results on the AFW dataset, using 7000 images for training



(a) SVM-NosePose using strict evaluation



(b) SVM-NosePose using weak evaluation



(c) Distribution of the results

Figure 3.21: Results on the AFW dataset, using 268 augmented faces for training

3.7 Nose Detection

The Faster R-CNN [23] detection method was evaluated on all datasets for finding the nose in the whole image, to assess the suitability of combining our head pose estimation method with a state-of-the-art detection step. All default Faster R-CNN parameters and models were used for training.

Algorithm 2 Intersection Coefficient

```
function iCoefficient($pred, gt$)
    $intersection \leftarrow getIntersection(pred, gt)$
    $iArea \leftarrow intersection.width * intersection.height$
    $pArea \leftarrow pred.width * pred.height$
    $gArea \leftarrow gt.width * gt.height$
    return $min(iArea/pArea, iArea/gArea)$
end function
```

The intersection coefficient (Algorithm 2), proposed by Hoover *et al.* [28], was used as metric, the rates presented in Table 3.16 are the number of images where the coefficient is at least 0.5 and include all detections. The results are also presented as curves in Figure 3.22 after being filtered according to the detection score internally calculated by Faster R-CNN, limiting one region per subject.

Table 3.16: Percentage of images where the intersection coefficient of the detected nose region and the annotated ground-truth is at least 0.5 and the amount of false positives

Dataset	**Accuracy**	**False Positives**
Multi-PIE	99.66%	15.52%
Pointing'04	99.89%	29.38%
McGillFaces	97.21%	22.56%
SFEW	90.86%	14.21%
PaSC	82.89%	92.51%
PaSC (face only)	73.66%	47.97%
AFW	51.19%	79.18%

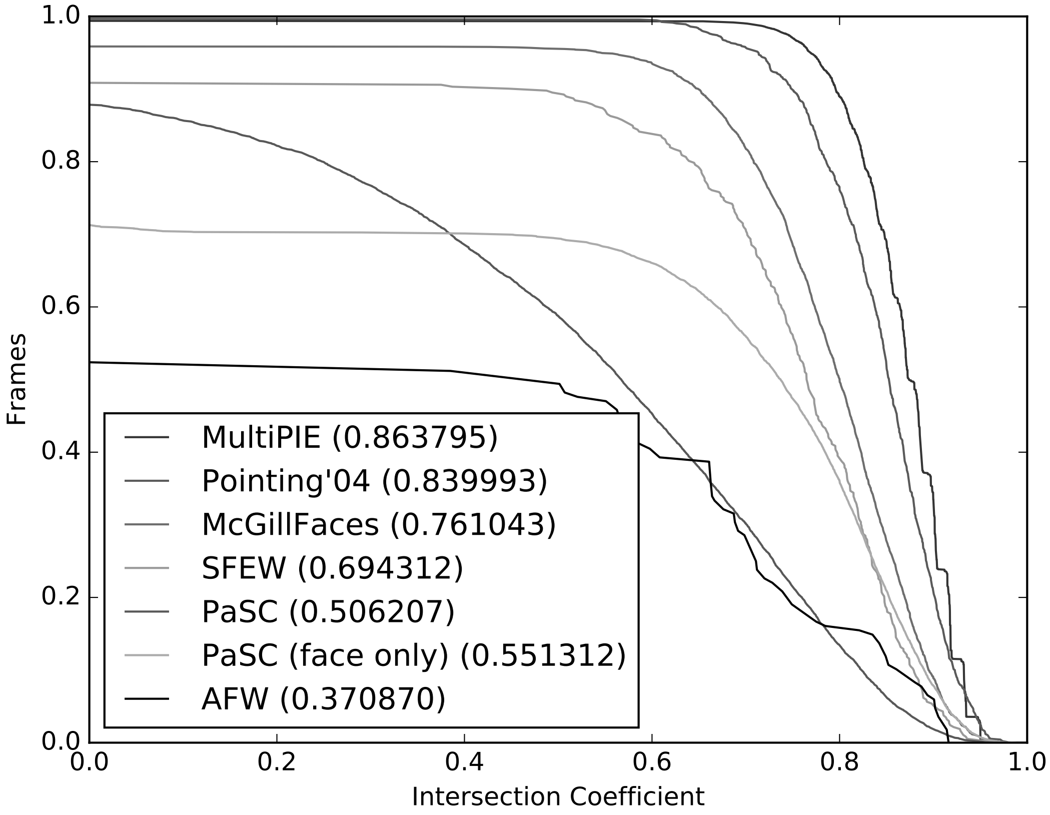


Figure 3.22: Intersection Coefficient curves on all datasets. The number in parenthesis is the area under the curve

While the detection performance on the Multi-PIE, Pointing'04, McGillFaces and SFEW is very high, Faster R-CNN failed to produce useful estimations on both PaSC and AFW. We believe this is due to different reasons: the very limited number of images in the AFW dataset, the low quality of many faces in the PaSC dataset (some are just a blur) and the large size of the images in the same dataset, which is incompatible with the default training parameters. To better understand the obtained results, we also retrained on the PaSC dataset using only the cropped face regions and achieved much lower false positive rates, particularly after filtering the detections.

We also present the pose estimation accuracy on all datasets when the filtered detected regions are used for extracting the features (Table 3.17). Performance is similar on most datasets, with the exception of those where the detection did not yield good results.

Table 3.17: Nose pose estimation performance when using the detected nose regions

Dataset	**Ground-Truth**		**Detections**	
	Strict	Weak	Strict	Weak
Multi-PIE	94.13%	99.31%	76.67%	97.13%
Pointing'04 (yaw)	58.49%	94.50%	45.53%	83.96%
Filtered McGillFaces	70.71%	92.68%	55.08%	82.98%
SFEW	83.60%	–	66.67%	–
PaSC (faces only)	86.91%	–	75.65%	–
AFW	25.00%	70.09%	7.14%	31.55%

CHAPTER 4

APPLICATIONS

We present two different applications for our head pose estimation method: Face quality estimation for nose tracking and 3D face alignment in the wild. Tests were performed to assess the real possibility and performance of integrating head pose estimation to these ends.

4.1 Face Quality Estimation For Face Tracking

We propose an initialization step for in the wild nose tracking, such that the initial frame is chosen based on face quality and head pose. We perform experiments using a state-of-the-art generic visual tracking method [17] and compare the accuracy when initializing the tracker with the first frame (baseline) and with the selected frame.

Given all frames in a video, the quality score for each is calculated according to the following steps: 1. Detect the face using Faster R-CNN [23]; 2. Detect the nose inside the face; 3. Obtain face quality asessment score using Abaza *et al.*'s method [18]; 4. Estimate the head pose yaw; 5. Calculate a pose score favoring half-profile and frontal faces; 6. Multiply the face quality and pose scores. Nam and Han's [17] tracker is initialized using the frame with the highest quality score and tracking is performed both forwards and backwards in time.

The proposed pose score can be adapted to the problem being solved, our empirical experiments showed that initializing the tracker with the nose region on frontal and half-frontal head poses yield better results compared to profile noses, due to the difference in the included background. By multiplying it by the geometrical average of the face illumination, brightness, focus, sharpness and contrast scores calculated by Abaza *et al.*'s method [18], our algorithm is able to select the best quality nose for the problem.

Experiments were performed on the 300 Videos in the Wild (300VW) [29] dataset,

Algorithm 3 The precision metric. The predicted region is represented as pred and the ground-truth region as gt

```
function PRECISION($pred, gt$)
    return $l2norm(center(pred), center(gt))$
end function
```

where all 64 videos in the testing subset where used to evaluate our results. We present the achieved increase in performance when our initial frame selection approach is applied and compare it against the baseline using two different metrics: intersection coefficient and precision (Algorithms 2 and 3).

Using our initialization, the tracker is able to estimate the size of of the bounding boxes with greater accuracy (Figure 4.1), 79.65% of the estimated regions are have an intersection coefficient of at least 0.5, compared to only 75.36% achieved by the baseline. Our frame selection did not influence the precision of the nose tracker (Figure 4.2).

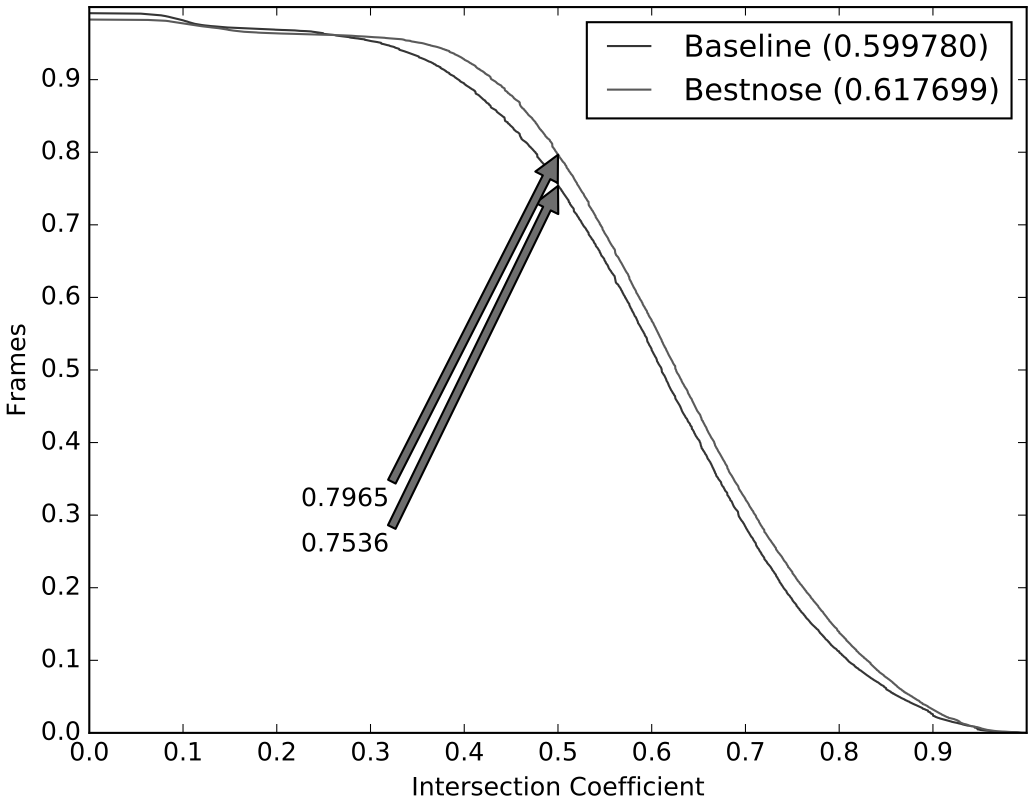


Figure 4.1: Comparison of the intersection coefficient metric between the baseline and our approach. The area under the curve is displayed in parenthesis

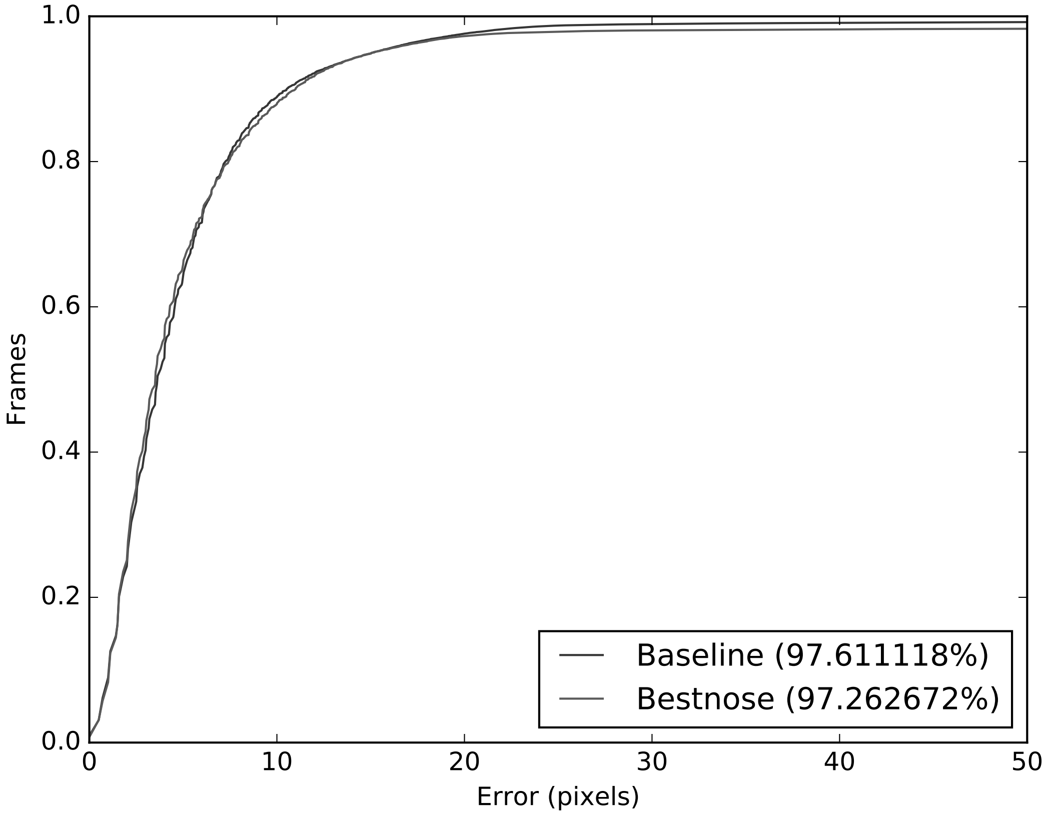


Figure 4.2: Comparison of the precision metric between the baseline and our approach. The percentage of frames where the error was less than or equal to 20 pixels is displayed in parenthesis

4.2 3D Face Alignment in the Wild

The 3D face alignment in the wild problem is defined by estimating the position of the facial landmarks in the 3D space given only a 2D image of the face. We propose a landmark-free, in the sense that it does not use any individual face characteristics, methodology using only the nose region. It is trained and evaluated on the available subset of the 3D Face Alignment in the Wild (3DFAW) challenge The face points are estimated by scaling, rotating and translating a generic face landmark model according to the detected nose region, estimated head pose and an optional face region (Figure 4.3), such that no specific facial trait or expression information is taken into account.

The 3DFAW challenge presents a set of images and annotations for evaluating the performance of in the wild 3D sparse face alignment methods. Part of the data is from the MultiPIE dataset [21] or from images and videos collected on the internet, having its depth information been recovered through a dense 3D from 2D videos alignment method [30]. The rest of the data was synthetically generated by rendering the 3D models present in the BU-4DFE [31] and BP4D-Spontaneous [32] databases onto different backgrounds. The training data includes the face bounding box and the 3D coordinates of 66 facial landmarks, while the testing data only includes the face bounding box. Experiments were performed on the training and validation subsets only.

The landmark model is generated by a series of steps: 1. A near frontal face for calibration from the training subset (Figure 4.4(a) and 4.4(b)) is selected; 2. All other training

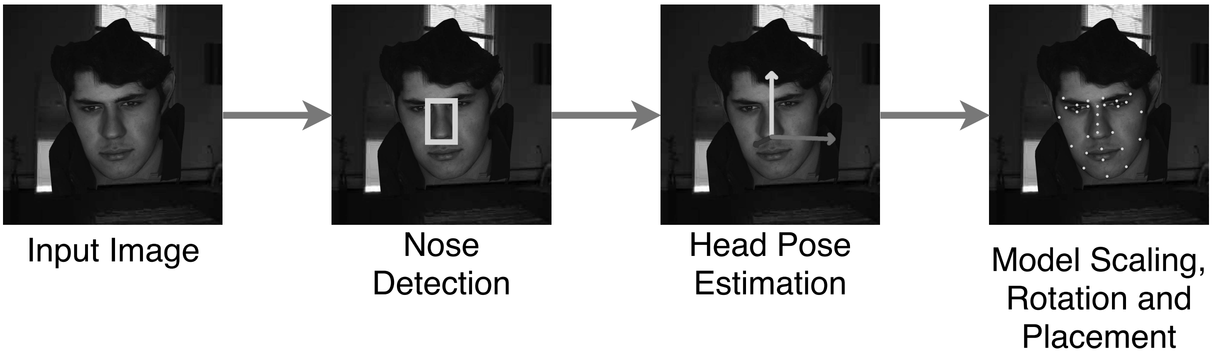


Figure 4.3: Overview of the 3D face alignment method

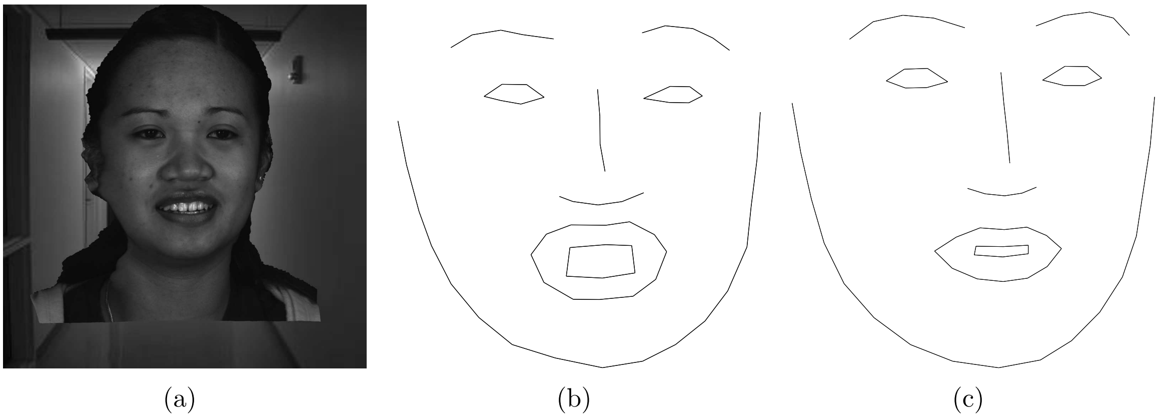


Figure 4.4: a) Calibration image; b) Calibration landmarks; c) Landmark model viewed with the calibration pose

facial landmark sets are then aligned using an affine transform on scale, translation and rotation; 3. The position of all landmarks is averaged (Figure 4.4(c)).

The head pose is estimated on both yaw and pitch in steps of 7.5 degrees and the model is aligned based on the detected nose region using a series of empirically determined parameters. If the face region is known, it can be used to better scale the landmark model, otherwise the size of the nose region is used as basis for the scaling. This process is detailed in Algorithm 4.

Our results obtained on the validation dataset are presented evaluated on the Ground Truth Error (GTE) metric (Equation 4.1). We do not compare our results against any other methods as there are currently no publications on this dataset. However, we present a set of examples for visual inspection of the results (Figure 4.5).

$$E(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\frac{\|x_k - y_k\|_2}{d_i} \tag{4.1}$$

The achieved alignment precision, of 11.113 pixels, and robustness to challenging environments, including extreme head pose and face expression indicates that our estimation has the potential to be used as initialization for finer alignment methods that rely on such step [30, 33].

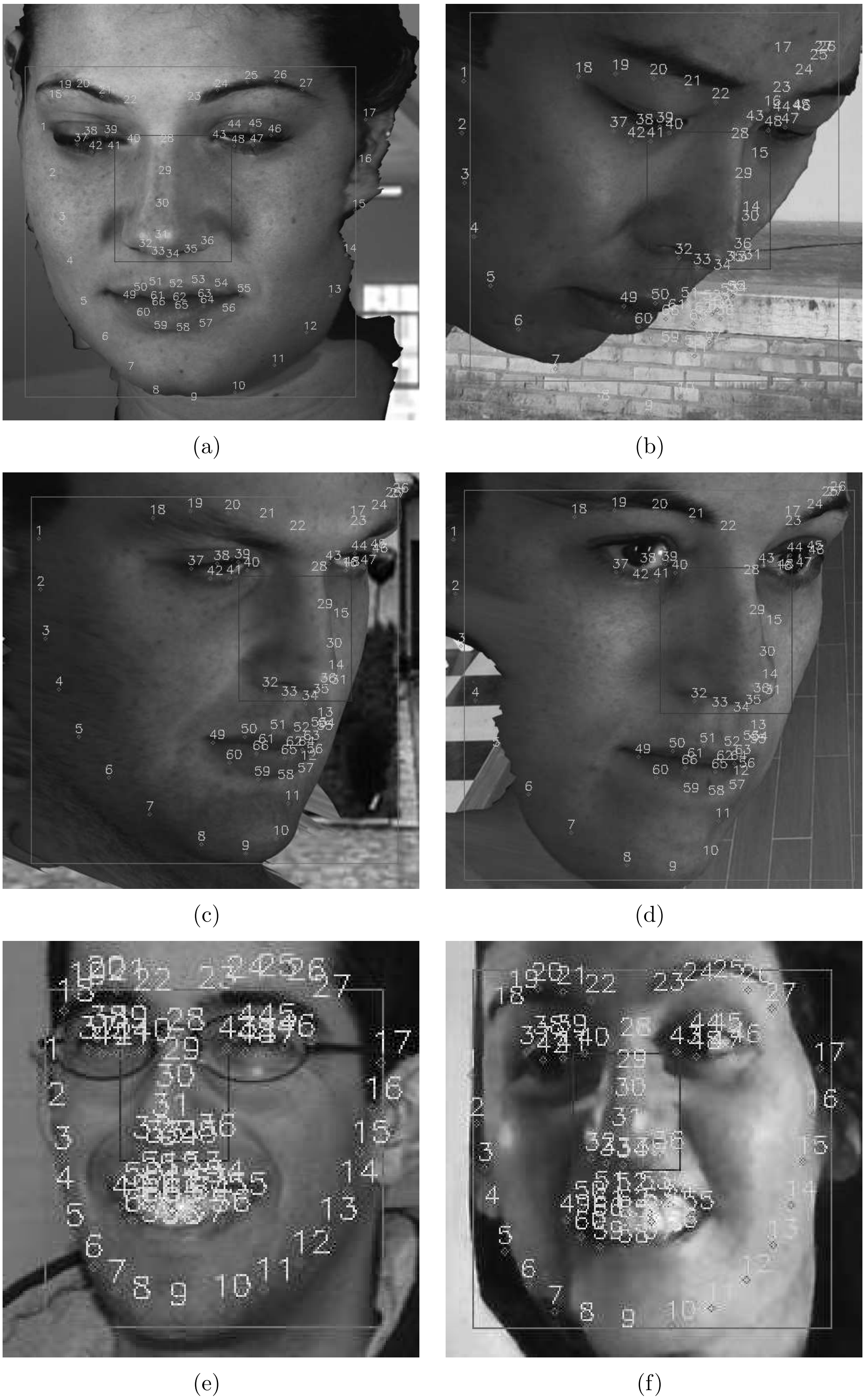


Figure 4.5: Example results of the model fitting stage, showing the face bounding-box in red, the detected nose region in blue and the estimated position of the landmarks in green. a) Near frontal good fit; b) Bad fit caused by bad head pitch estimation; c) and d) Half-profile good fit; e) Good fit in an image sourced from the MultiPIE dataset; f) Modest fit in one of the most challenging images in the dataset

Algorithm 4 Model Fitting Algorithm

```
function FITMODEL($model, noseBB, headPose, faceBB$)
    $modelNoseBase \leftarrow average(model.noseBaseLeft, model.noseBaseRight)$
    $rotate(model, modelNoseBase, headPose)$
    if $isDefined(faceBB)$ then
        $xScale \leftarrow .975 * faceBB.width/model.width$
        $yScale \leftarrow .975 * faceBB.height/model.height$
        $zScale \leftarrow (xScale + yScale)/2 * .95$
        $scale(model, xScale, yScale, zScale)$
    else
        $modelNoseWidth \leftarrow l2Norm(model.noseBaseLeft, model.noseBaseRight)$
        $scale(model, .6 * nose.width/modelNoseWidth)$
    end if
    $noseBase \leftarrow \{nose.x + nose.width * .5, nose.y + nose.height * .9, 0\}$
    $translate(model, modelNoseBase - noseBase)$
    $translate(model, \{0, 0, -average(model).z\})$
    return model
end function
```

CHAPTER 5

FINAL REMARKS

We performed head pose estimation based on the nose region using our approach, SVM-NosePose. Our method was tested on six different publicly available datasets.

On Multi-PIE [21], we achieved better performance than Zhu and Ramanan [10] and similar performance to Demirkus *et al.* [13]. On the Pointing'04 dataset, our accuracy surpassed Pawelczyk and Kawulok's [12] when estimating the yaw, the pitch and both at the same time. After annotating SFEW [27], we were able to evaluate our method and achieved a hit rate of 83.6%.

On the publicly available subset of the McGillFaces dataset [26], we were able to train our algorithm and evaluate all images independently, we achieved 59.24% accuracy, lower than Demirkus *et al.* [13], who used temporal information and almost three times as many images. We noticed the ground-truth annotations on the McGillFaces dataset [26] were inconsistent, and produced a filtered version of the dataset, with only four fifths of the total images, the ones with consistent labels, and were able to achieve 70.71% accuracy on it.

We achieved 70.09% hit rate on AFW [10] dataset, considering the weak protocol. This can be considered poor when compared to Zhu and Ramanan's [10] result, however our method was not developed for training on a controlled dataset and testing on an in-the-wild dataset. We also present our achieved accuracy of 86.91% when training and testing on a large in-the-wild dataset, PaSC [22].

A state-of-the-art object detection method was applied and evaluated for finding the nose region. Our results show that with proper training, high detection rates can be achieved while keeping the number of false positives low.

Two applications for NosePose were suggested and tested: tracking initialization and 3D face alignment. The results indicate great potential when integrating our method due to it being robust even in challenging scenarios.

In future work, we also intend on using temporal information to improve our accuracy and allow better comparisons against other published methods. This can be done in datasets where the images were extracted from video sequences, such as McGillFaces [26] and PaSC [22].

BIBLIOGRAPHY

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[2] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 6, pp. 1124–1133, 2006.

[3] Y. Lee, P. Phillips, J. Filliben, J. Beveridge, and H. Zhang, "Generalizing face quality and factor measures to video," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pp. 1–8, 2014.

[4] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[5] Z. Zhang, P. Luo, C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), vol. 8694 of *Lecture Notes in Computer Science*, pp. 94–108, Springer International Publishing, 2014.

[6] Z. Hammal, J. F. Cohn, C. Heike, and M. L. Speltz, "Automatic measurement of head and facial movement for analysis and detection of infants' positive and negative affect," *Frontiers in ICT*, vol. 2, p. 21, 2015.

[7] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] S. Tulyakov, R.-L. Vieriu, S. Semeniuta, and N. Sebe, "Robust real-time extreme head pose estimation," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pp. 2263–2268, 2014.

[9] C. Papazov, T. Marks, and M. Jones, "Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 4722–4730, 2015.

[10] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2879–2886, 2012.

[11] X. Peng, J. Huang, Q. Hu, S. Zhang, and D. Metaxas, "Three-dimensional head pose estimation in-the-wild," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1, pp. 1–6, 2015.

[12] K. Pawelczyk and M. Kawulok, "Head pose estimation relying on appearance-based nose region analysis," in *Computer Vision and Graphics* (L. Chmielewski, R. Kozera, B.-S. Shin, and K. Wojciechowski, eds.), vol. 8671 of *Lecture Notes in Computer Science*, pp. 510–517, Springer International Publishing, 2014.

[13] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, "Probabilistic temporal head pose estimation using a hierarchical graphical model," in *Computer Vision–ECCV 2014*, pp. 328–344, Springer, 2014.

[14] K. Chang, W. Bowyer, and P. Flynn, "Multiple nose region matching for 3d face recognition under varying facial expression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 10, pp. 1695–1700, 2006.

[15] N. Zehngut, F. Juefei-Xu, R. Bardia, D. K. Pal, C. Bhagavatula, and M. Savvides, "Investigating the feasibility of image-based nose biometrics," in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, 2015.

[16] Z. Lubing and W. Han, "Local gradient increasing pattern for facial expression recognition," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 2601–2604, 2012.

[17] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in *IEEE CVPR*, 2016.

[18] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross, “Design and evaluation of photometric image quality measures for effective face recognition,” *IET Biometrics*, vol. 3, no. 4, pp. 314–324, 2014.

[19] T. Ojala, M. Pietikainen, and D. Harwood, “Performance evaluation of texture measures with classification based on kullback discrimination of distributions,” in *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, pp. 582–585 vol.1, 1994.

[20] B. Jun and D. Kim, “Robust face detection using local gradient patterns and evidence accumulation,” *Pattern Recognition*, vol. 45, no. 9, pp. 3304 – 3316, 2012. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA’2011).

[21] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-pie,” *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[22] J. Beveridge, P. Phillips, D. Bolme, B. Draper, G. Givens, Y. M. Lui, M. Teli, H. Zhang, W. Scruggs, K. Bowyer, P. Flynn, and S. Cheng, “The challenge of face recognition from digital point-and-shoot cameras,” in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pp. 1–8, 2013.

[23] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems 28*, pp. 91–99, Curran Associates, Inc., 2015.

[24] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, “A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 7, pp. 1788–1794, 2013.

[25] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures*, vol. 6, 2004.

[26] M. Demirkus, J. Clark, and T. Arbel, "Robust semi-automatic head pose labeling for real-world face video sequences," *Multimedia Tools and Applications*, vol. 70, no. 1, pp. 495–523, 2014.

[27] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 2106–2112, IEEE, 2011.

[28] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 673–689, 1996.

[29] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *IEEE ICCV Workshop*, pp. 1003–1011, 2015.

[30] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d video for real-time use," *Image and Vision Computing*, 2016.

[31] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pp. 1–6, 2008.

[32] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692 – 706, 2014. Best of Automatic Face and Gesture Recognition 2013.

[33] X. Peng, S. Zhang, Y. Yang, and D. N. Metaxas, “Piefa: Personalized incremental and ensemble face alignment,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3880–3888, 2015.