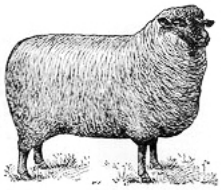
# Head pose estimation and facial landmark localisation for animals




Charles Hewitt
Trinity Hall


UNIVERSITY OF CAMBRIDGE

*A dissertation submitted to the University of Cambridge in partial fulfilment of the requirements for Part III of the Computer Science Tripos*


University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom

Email: cth40@cam.ac.uk
Supervisor: Dr Marwa Mahmoud


May 31, 2018

# Declaration

I, Charles Hewitt of Trinity Hall, being a candidate for Computer Science Tripos, Part III, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

**Signed**:

**Date**:

# Head pose estimation and facial landmark localisation for animals


## Abstract

This project investigates head pose estimation, facial landmark localisation and the deployment of these techniques for animals, specifically sheep. A new dataset of 850 sheep facial images, annotated with a 25 facial landmark scheme and occlusion information, is introduced: the Sheep Facial Landmarks in the Wild (SFLW) dataset. This provides a benchmark dataset for evaluation of animal facial landmark localisation techniques and includes a challenging range of images exhibiting large variations in head-pose and occlusion.

A novel data augmentation technique using thin-plate-spline warping is proposed to enhance the effectiveness of training on the SFLW dataset, along with the use of negatively correlated augmentation, similar to that proposed in [55], to boost representation of extreme head poses. These techniques are shown to be effective in improving performance for head pose estimation and facial landmark localisation.

An existing model for human head pose estimation from image data, without facial landmark locations, using a CNN is adapted for use on sheep. A pre-trained model is fine-tuned on the SFLW dataset with a resulting average absolute error for yaw, pitch and roll under 7 degrees. Correlation with ground truth pose information is also impressive, up to 0.91 for yaw.

A number of existing state-of-the-art methods used for human facial landmark localisation are evaluated on sheep using the SFLW dataset, the best achieving a success rate of 90% and a mean normalised error of 0.05. Analysis of the results of the highest performing technique motivates the introduction of a pose-informed localisation technique, incorporating the newly developed, landmark-free head pose estimation network.

This pose-informed localisation technique achieves a higher performance than the best existing method on the SFLW dataset; with 93% success rate and a mean normalised error of 0.045. Most significantly, error is markedly reduced for images with extreme head poses.

Finally, a near real-time demonstration of a full pipeline, incorporating automated face detection and pose-informed face alignment, is carried out for a number of pre-recorded videos. This serves as a proof-of-concept that a production system incorporating these technologies for automated health monitoring of livestock is eminently feasible.

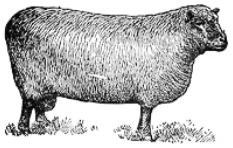
Total word count: 11,665

# Contents

# List of Figures

# List of Tables

# Chapter 1



# Introduction

## 1.1 Motivation

In recent years automation has become widespread in many industries. This has been facilitated by numerous technical advances in computer systems, as well as the significant reduction in the cost of deploying such systems. Agriculture is an industry where mechanisation has had a huge historical impact but is only just beginning to see the effects of computer-aided automation.

Global positioning system related technologies are some of the most established within agriculture [36]. GPS enables highly detailed monitoring of crops for assessment of yield, as well as targeted chemical application. Automated navigation by GPS has also become increasingly popular in recent years.

Most agricultural computer vision (CV) research to date has focussed on arable farming, with CV applied to asses quality of various fruits and vegetables, such as tomatoes [3], potatoes [59], as well as other quality control use cases [45, 50]. Some work has also been carried out aimed at pastoral farming, where CV has been used for feed and waste management systems [23]. An application making use of artificial intelligence to help dairy farmers has also recently been released [13], though instead uses accelerometer and GPS

sensors.

The focus of this project is also related to pastoral farming, with the aim of contributing towards automated pain recognition in livestock. Work in this area has shown that pain in sheep can be reliably predicted from a number of facial action units [33]. A key component of this process is the localisation of a number of landmarks—for example, the eyes, ears or nose—on the face of the sheep in order to identify these action units.

Facial landmark localisation is a well-explored problem in humans [38], but existing work tackling this problem for sheep [55] and horses [40] only considers very sparse landmarks and relatively minimal head pose variation, as well as achieving less than ideal accuracy. Denser landmarks would allow for more accurate action unit detection and increased head pose variation in training data would improve the resilience of models when deployed in-the-wild.

The primary aim of this work, then, is to investigate the problem of landmark localisation (also known as face alignment) for animals, primarily sheep. The end-goal being to demonstrate the feasibility of an automated system which can be used to detect medical issues that require further investigation as early as possible, rather than relying on infrequent veterinary evaluations of animals. This could be achieved through the use of CCTV monitoring of the animals, linked to a CV pipeline which detects sheep faces, localises the landmarks on the face, and determines pain levels from extracted action units.

## 1.2 Aims and Structure

This project has a number of key aims ranging from acquisition of appropriate data to end-to-end deployment of a video facial alignment system for sheep. These are broken into five distinct aspects:

- Consideration of available data and description of annotation procedure. Summary of augmentation techniques including the introduction of novel

image warping augmentation method.

- Exploration of landmark-free head pose estimation for animals.
- Evaluation of existing human facial landmark localisation methods applied to animal faces.
- Investigation into method providing improved facial landmark localisation performance for animals using head pose information.
- Application of animal facial landmark localisation in real-time on video data.

Each of these topics is covered in a separate chapter. Given the wide range of existing work relating to these different foci, a summary of relevant work is given at the beginning of each chapter where appropriate.

# Chapter 2

# Dataset Annotation and Augmentation

The critical issue faced in facial landmark localisation for animals is one of data sparsity. The sheep dataset used in [55] includes just 600 images (of which approximately 500 are available) annotated with only eight landmarks. The only other animal facial landmark dataset is that used in [40] for horses, which includes 3717 images annotated with only five landmarks. This is in stark contrast to facial datasets for humans; Multi-PIE [21] (750,000 images with 68 landmarks), Menpo [57] (over 10,000 images with 39 landmarks), AFLW [29] (25,993 images with 21 landmarks) and PUT [25] (9971 images with 30 landmarks).

As such, it is imperative to enhance the available data for animal landmark localisation as much as possible to enable improved performance. This chapter describes the annotation procedure used to extend the original sheep dataset annotations to produce the Sheep Facial Landmarks in the Wild (SFLW) dataset. This dataset contains 850 sheep images annotated with 25 facial landmarks and occlusion information. The augmentation techniques used to improve the data volume for training are also described; a novel thin-plate spline (TPS) [5] warping method and negatively correlated augmentation based on head pose [55]. These techniques are used to produce a series of variants of the raw SFLW dataset which are used during this dissertation, these are summarised at the end of the chapter.

## 2.1 Data Annotation

The SFLW dataset is composed primarily of images used in [55], with some new images also included to bring the total number of images to 850. Almost all of the original images feature unique animals, though it is difficult to confirm exactly how unique animals are contained in the dataset, it is certainly fewer than 850.

An annotation scheme containing 25 landmarks is devised based on the original eight-point annotation used in [55]. The original and updated schemes are shown in Figure 2.1a. This is approximately based on human annotation schemes, though with additional emphasis placed on the ears, which are typically excluded from human face alignment but are critical for most animals. The eyes, nose and mouth are represented by eight landmarks, with a further eight representing the ears and the remaining nine corresponding to the face boundary. This scheme also allows for effective extraction of the action units used for sheep pain estimation in [33]

Due to the shape of the sheep face, with an elongated snout, self-occlusion is very common, far more so than for human faces. As such, the SFLW dataset is also annotated with binary occlusion information for each landmark. Some face alignment methods incorporate occlusion prediction [10, 53] which can be exploited as a result of this additional annotation.

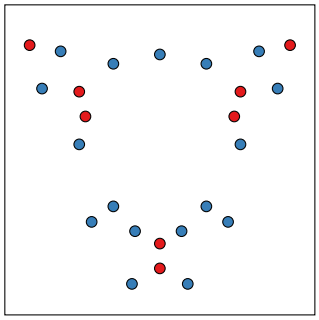
In order to more efficiently extend the original eight-point annotations from [55], a semi-automated annotation approach is used. This kind of method is not uncommon when attempting to unify annotations from various datasets [44], though typically relies on a large volume of existing annotations to inform automation.

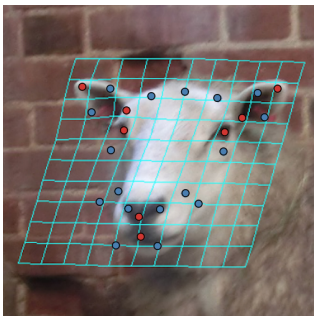
Instead, a purely shape-driven technique is employed; base shapes for both the original 8-point and new 25-point landmark schemes are defined (as shown in Figure 2.1a) and the thin-plate spline (TPS) [5] transformation from the 8-point base shape to the 8 annotated landmarks calculated. This transformation is then applied directly to the 25-point base shape to obtain

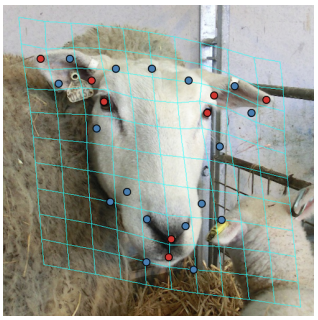
Figure 2.1: Annotation specification and procedure.

(a) Standard shapes for original (red) and new (blue) landmarks.

(b) Example images with original (red), predicted new (blue) landmarks and TPS grid shown.







an approximate prediction of the 25 landmark location.

The TPS transformation implementation is adapted from that used in [4], the resulting predicted annotations are visualised for two example images in Figure 2.1b. For demonstrative purposes, the TPS transformation is applied to a grid of points which are also rendered.

TPS warping is used due to its ability to incorporate both simple affine transformations to account for global effects caused by pose variation, as well as local deformations caused by variations in face shape or the relative position of the ears. The grids in the example images clearly show how rotations and shears are captured, as well as local effects such as in the area around the eyes and ears of the second example.

These initial predictions are then manually tuned to the correct image locations, and occlusions annotated as applicable. This semi-automated approach significantly increased the speed of annotation. The resulting SFLW dataset contains 850 images with 25 landmarks and occlusion annotations.

## 2.2 Data Augmentation

A number of data augmentation methods are utilised to increase the effective size of the SFLW dataset from the raw 850 images. Horizontal mirroring, rotation and translation are all well-established methods of data augmentation for machine learning applications. In this case, horizontally flipping the image (and transposing the landmark order as appropriate) is an effective way of improving pose invariance. Rotating the images and translating the facial bounding-boxes are also simple and effective augmentation methods.

This section introduces two additional techniques used for data augmentation; image warping using TPS transformations and negatively correlated augmentation (NCA), both described in detail below.

### 2.2.1 TPS Warping

In order to avoid repeating identical images when training localisation models, TPS warping [5] is used to generate slight variations on input image data. These variations are visually subtle but should allow for more general representations of features to be learnt, and hence over-fitting to be avoided.

TPS warping is able to simulate changes in ear position as well as providing low magnitude pose and face-shape variation. Affine warping of triangles from the Delaunay triangulation of the landmarks is a common technique for face morphing [19], but produces unrealistic results compared with TPS warping for this application.

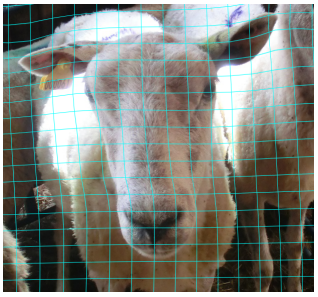
Somewhat similar warping techniques are highlighted in [12] with the goal of face frontalisation, rather than data augmentation. [34] presents a method of data augmentation using high-resolution 3D models of humans faces to aid with the largely unrelated problem of face recognition. The variation in the 3D shape of sheep faces levels of self-occlusion caused by the shape of sheep face make this method largely infeasible.

Figure 2.2: TPS warp augmentation for example image; note the rotated ears, the slightly wider spacing between the eyes, and the movement of the nose slightly towards the right of the image.

(a) Original image.

(b) Warped image.

(c) Warped image with TPS grid.



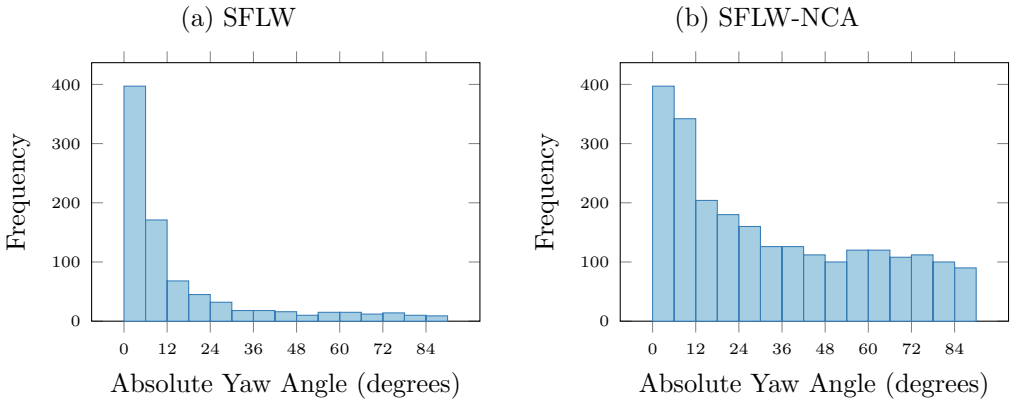
The visible landmarks of the original image are randomly permuted according to a hand-crafted set of rules, for example moving the eyes up or down, and closer together or further apart. The TPS transformation from the original landmarks to the permuted landmarks is then calculated and the inverse transformation applied to a grid over the image area. The grid is then linearly interpolated over the entire image area. This provides a set of coordinates which can be sampled from the original image to efficiently obtain a warped version of the image, with the correct image features now in locations matching the permuted landmarks. An example of this process is shown in Figure 2.2.

An alternative approach is to warp the landmarks of one image onto those of another randomly selected from the training data, providing the difference is not too great. Attempting this approach with such a small and varied dataset, however, provides unrealistic warped image and so is not explored further here.

### 2.2.2 Negatively Correlated Augmentation

One key issue with the SFLW dataset is the distribution of head poses included. Unsurprisingly, the majority of facial images of sheep available

Figure 2.3: Absolute yaw angle distributions, using 6 degree bins, for SFLW and SFLW-NCA datasets.




online—most of the original dataset was sourced from the internet—have frontal faces, or close to. Images collected specifically for the purposes of facial landmark localisation have a better distribution, but make up a smaller proportion of the SFLW dataset.

The distribution of absolute yaw angles (i.e., the angle away from frontal) for the raw dataset is given in Figure 2.3a. The imbalance is clear, with the most common range (0–6 degrees) having over 40 times more images than the least common range (84–90 degrees). Given the proposed application—using CCTV within an agricultural setting—large variation in head pose is expected to be common, and this imbalance is, therefore, a serious issue.

As such, a technique very similar to the negatively correlated augmentation (NCA) introduced in [55] is employed. Using the distribution of absolute yaw angles, a boosted augmentation factor for each pose bin is determined. The integer augmentation factor for pose bin $b$, $aug_b$, is given by:

$$aug_b = \left\lceil \left( \frac{count_{max}}{count_b} \right)^{\alpha} \right\rceil$$

Where $count_{max}$ is the maximum count for any pose bin, and $count_b$ is the count for pose bin $b$. The level of boosting is controlled by parameter $\alpha$, where $0 \leq \alpha \leq 1$, this allows for the underlying distribution to be somewhat

maintained, but with a greater representation of less common pose angles.

The SFLW-NCA dataset is generated using $\alpha = 0.6$ with pose bins 6 degrees wide, this results in augmentation factors between 0 (i.e., no augmentation for the most common bin) and 20, and an average augmentation number across the dataset of $\sim$3. The distribution of the resulting SFLW-NCA dataset is shown in Figure 2.3b. It is clear that extreme poses are much better represented using this method.

Augmented images are generated using the modifications described in the previous section: random rotation and TPS warping. Horizontally flipped versions of all images are also included, so the effective augmentation number of the final SFLW-NCA dataset ranges between 1 and 40, with average $\sim$6, totalling 4794 images.

## 2.3 Summary

This chapter has described the newly introduced Sheep Facial Landmarks in the Wild (SFLW) dataset, which includes 850 facial images of sheep annotated with 25 facial landmarks, as well as occlusion information. The semi-automated annotation procedure is described, along with the introduction of an image augmentation technique using thin-plate-spline warping [5] to emulate the effects of variation in face shape, head pose and local variations such as ear position. In addition, the motivation, implementation and effects of negatively correlated augmentation, as proposed in [55], are explored.

A number of variants of the SFLW dataset, employing various combinations of the augmentation methods introduced above, are used for comparative purposes throughout the remainder of the dissertation. These are summarised in Table 2.1. The SFLW dataset contains the raw 850 images, and SFLW-NCA is as described immediately above. In addition to these are the SFLW-flip and SFLW-warp datasets used for the comparative evaluation of individual image augmentation techniques.

Table 2.1: Summary of dataset variants.

| Dataset | #Images | Mirroring | TPS Warping | Rotation | NCA |
|---|---|---|---|---|---|
| SFLW | 850 | | | | |
| SFLW-flip | 1700 | ✓ | | | |
| SFLW-warp | 5100 | ✓ | ✓ | | |
| SFLW-AUG | 5100 | ✓ | ✓ | ✓ | |
| SFLW-NCA | 4794 | ✓ | ✓ | ✓ | ✓ |

The SFLW-AUG dataset is introduced for comparison to SFLW-NCA. SFLW-AUG also uses all forms of image augmentation but does not use NCA to determine balanced augmentation factors. Every image is instead augmented equally with augmentation factor 6 in order to approximately match the number of images in SFLW-NCA.

# Chapter 3

# Head Pose Estimation

Head pose estimation is a significantly less well-explored problem in computer vision than landmark localisation. This is primarily because use cases which require head pose information alone are not common compared with applications which require full facial alignment. Head pose can also be estimated by finding a solution to the perspective-n-point problem between 2D landmarks and a predefined 3D landmark model [31]. Given that facial landmark localisation is largely 'solved' for humans, this method is common.

This chapter first summarises work relevant to the problem of animal head pose estimation, then describes in detail the methodology used as part of this work; fine-tuning a human head pose estimation network. Finally, the proposed technique is evaluated, with both quantitative and qualitative results provided, and the effects of data augmentation are assessed.

## 3.1 Background

A number of classical techniques for human head pose estimation have been proposed [35], though, as described above, it is common in practice to estimate head pose indirectly using facial landmarks. Recently, deep learning has been applied to both landmark localisation and head pose estimation

for humans, often combined into a single network [39]. Specific networks have also been designed to determine only the head pose of humans from images [43], aiming to be more efficient than the often very large, multi-function networks and with impressive results.

In order to deal with the large amount of self-occlusion caused by variations in sheep head pose, the inverse procedure is considered. Rather than calculating head pose from localised landmarks, we instead aim to improve facial alignment performance by first predicting the head pose of the sheep directly from the input image, and then using the estimated pose to aid in the localisation process. Head pose is encoded as three angles: yaw, pitch and roll, representing left/right, up/down and clockwise/anticlockwise rotations in image space respectively. In this context, these angles are commonly known as Euler angles. For the purposes of sheep facial alignment, yaw is the most critical angle due to the resulting self-occlusions.

While inter-species transfer learning has shown to perform poorly for landmark localisation [40], presumably due to the significant difference in appearance between alike landmarks, there is significant scope for inter-species transfer learning for head pose estimation. Firstly, there is a greater degree of visual similarity between animals and humans when considering facial images holistically, rather than locally as in the case of landmark localisation. The task itself is also arguably simpler; the aim being to regress only three Euler angles (yaw, pitch and roll) as opposed to a large number of two-dimensional coordinates. As such, a landmark-free head pose estimation method for sheep is developed by fine-tuning a pre-trained CNN model for human head pose estimation.

## 3.2 Methodology

### 3.2.1 Data Preparation

In order to determine ground-truth head pose for the images in the SFLW dataset a 3D base landmark model is manually defined with neutral head pose (0 yaw, pitch and roll) and approximately average head shape. A RANSAC [20] based method for solving the perspective-n-point problem is then used to recover the approximate head pose using the 3D points of this landmark model and the 2D annotated landmarks for each image. The six landmarks representing the top, bottom and tip of both ears are excluded from this correspondence as they typically move significantly relative to the rest of the face. As the camera coefficients are unknown, the intrinsic parameters are estimated based on the image size and lens distortion is assumed to be negligible. While the generated ground-truth poses are not exact, they provide a very good approximation and are certainly sufficient for this application.

### 3.2.2 Hopenet

As described above, transfer learning from a deep, human head pose estimation network is employed to create a model capable of sheep head pose estimation. The Hopenet network from [43] is selected due to its design focus specifically for head pose estimation, it is the most performant of the networks trained in [43]. A Hopenet model pre-trained on the 300W-LP [62] is used as the base model.

Ruiz et al. [43] introduce the principle of multi-loss training for head pose estimation. Pose angles are sorted into 66 bins between $-99$ and $+99$ degrees, forming the basis of a classification problem, for which conventional soft-max loss is used. In addition to this, the expected continuous angle is calculated from the soft-max output and mean squared error (MSE) loss evaluated

against the raw ground truth angles. These two losses are then summed (with a significant weighting towards the cross-entropy loss). This is theorized to enable the network to learn first a course guess of the angle based on the classification problem, and then predict a more accurately tuned continuous value based on the MSE loss.

The Hopenet architecture uses a ResNet [22] bottleneck followed by three independent fully-connected layers, one for each of yaw, pitch and roll. Each of these is followed by a 66 node soft-max layer from which cross-entropy loss is calculated, along with the expected values used for the MSE loss.

### 3.2.3 Training Procedure

The base model is fine-tuned on the SFLW-NCA dataset with additional augmentation provided by randomly flipping the input images in the $x$-direction and translating the image by up to $\sim$7% in the $x$- and $y$-directions (as in [43]). Five-fold cross-validation is used, so a fifth of the dataset is isolated for testing of each fold. A tenth of each remaining training set is used for validation.

A similar training process as in [43] is employed, using the Adam optimizer [28] with default parameters. The model is trained in batches of 16 over 16 epochs, chosen as validation loss plateaus towards the end of this period. A low initial learning rate of 0.0001 is used as the model is only being fine-tuned and not trained end-to-end; a larger initial learning rate results in very poor performance. The learning rate is also reduced by a factor of ten halfway through training. The model with the lowest validation loss during training is selected for evaluation.

## 3.3 Evaluation

To evaluate the effectiveness of the sheep head pose estimation model a number of metrics are employed. Mean absolute error (MAE) is typically used—and is the metric presented in [43]—though can often be misleading.

Table 3.1: Quantitative head pose estimation results for network trained on SFLW-NCA and tested on SFLW compared with two baselines.

| | Mean Baseline | | | | Pre-trained Baseline | | | | Fine-tuned Model | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Yaw | Pitch | Roll | Ave | Yaw | Pitch | Roll | Ave | Yaw | Pitch | Roll | Ave |
| MAE | 15.73 | 11.88 | 8.36 | 11.99 | 27.77 | 19.44 | 11.38 | 19.53 | 6.04 | 7.58 | 6.13 | 6.58 |
| PCC | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.20 | 0.08 | 0.05 | 0.91 | 0.56 | 0.40 | 0.75 |
| SAGR | 0.50 | 0.57 | 0.50 | 0.52 | 0.51 | 0.36 | 0.45 | 0.44 | 0.78 | 0.83 | 0.77 | 0.80 |

For example, a model always predicting the mean of a dataset with little deviation will often perform well in this metric. As such, Pearson's Correlation Coefficient (PCC) is used to assess the correlation of predictions with the ground truth, arguably a better measure of a model's usefulness. In addition to these two metrics, Sign Agreement metric (SAGR) [37] is used to give a coarse indication of simply whether the prediction matches the general direction (left or right/up or down) of the head pose. This is a significant attribute when considering pose-informed landmark localisation. In all cases, the unaugmented test set is used for evaluation.

### 3.3.1 Overall Performance

Two baselines are included for comparison to the fine-tuned model; firstly taking the mean of the dataset as the prediction, and secondly using the estimates generated by the pre-trained human head pose estimation network when using sheep images as input. Results for these and the highest performing fine-tuned model are given in Table 3.1.

The fine-tuned model outperforms both baselines significantly in all metrics and for all of yaw, pitch and roll. The mean baseline achieves reasonable MAE, as described above, but has no correlation to the ground-truth angles and essentially random SAGR. The human baseline achieves some correlation but has very large MAE and worse than random SAGR. In contrast, the fine-tuned model achieves high PCC, perhaps the most critical metric, and also good values for SAGR. The MAE is also much lower than both baselines, with an average of approximately 6.5 degrees error. Critically, PCC for yaw

is particularly high at 0.91; accurate yaw prediction enables significantly improved landmark localisation, as highlighted in the following chapters.

It might be noted that SAGR is not as high as perhaps would be expected given the relative simplicity of this metric. It is important to note the effect of the magnitude of angle on the value of SAGR. Considering yaw, for example, SAGR for angles of magnitude less than 10 degrees is 0.67, while for angles with magnitude greater than 10 degrees is 0.96. This is not surprising given that for angles within 10 degrees of frontal there is relatively little visual difference in the resulting 2D image. Confusion in sign for these low magnitude angles is also not critical when considering application to landmark localisation.

Qualitative pose estimation results for some example images in the SFLW dataset are given in Figure 3.1, with pose visualised as a 3-dimensional axis at the centre of the image. The blue axis represents the gaze direction, and the red and green axes show the horizontal and vertical directions relative to the sheep's head respectively. As shown, pose predictions remain accurate across a variety of sheep breeds and ages, as well as in extreme poses.

### 3.3.2 Effects of Data Augmentation

To assess the effects of the various augmentation techniques introduced in Section 2.2 on head pose estimation, the network is also trained of the SFLW, SFLW-warp and SFLW-AUG datasets. The same training procedure as described above is used.

Table 3.2 shows the results for the raw SFLW-flip dataset compared with the SFLW-warp dataset, allowing evaluation of the effect of TPS warping augmentation in isolation. For individual angles results are inconclusive, but on average the model trained on SFLW-warp outperforms that trained on SFLW-flip significantly in terms of MAE. SAGR is also slightly improved, while PCC is equal for the two models. This indicates that for this TPS warping does have some positive impact on performance.

Figure 3.1: Qualitative head pose estimation results for network trained on SFLW-NCA and tested on SFLW. Head pose is visualised as a 3-dimensional axis at the centre of the image.

Table 3.2: Head pose estimation performance metrics for networks trained on the SFLW-flip and SFLW-warp datasets and tested on SFLW.

| | SFLW-flip | | | | SFLW-warp | | | |
|---|---|---|---|---|---|---|---|---|
| | Yaw | Pitch | Roll | Ave | Yaw | Pitch | Roll | Ave |
| MAE | 6.37 | 7.79 | 6.06 | 6.74 | 6.02 | 7.71 | 6.11 | 6.61 |
| PCC | 0.91 | 0.50 | 0.45 | 0.74 | 0.90 | 0.53 | 0.39 | 0.74 |
| SAGR | 0.77 | 0.81 | 0.75 | 0.78 | 0.80 | 0.84 | 0.77 | 0.80 |

Table 3.3: Head pose estimation performance metrics for networks trained on the SFLW-AUG and SFLW-NCA datasets and tested on SFLW.

| | SFLW-AUG | | | | SFLW-NCA | | | |
|---|---|---|---|---|---|---|---|---|
| | Yaw | Pitch | Roll | Ave | Yaw | Pitch | Roll | Ave |
| MAE | 5.95 | 7.70 | 6.10 | 6.59 | 6.04 | 7.58 | 6.13 | 6.58 |
| PCC | 0.91 | 0.55 | 0.43 | 0.75 | 0.91 | 0.56 | 0.40 | 0.75 |
| SAGR | 0.79 | 0.83 | 0.78 | 0.80 | 0.78 | 0.83 | 0.77 | 0.80 |

Table 3.3 shows the performance of the network when trained on the fully augmented SFLW-AUG dataset, in comparison to the SFLW-NCA dataset. Performance for both the AUG and NCA variants is slightly improved over the SFLW-warp dataset, indicating the rotation of training images also provides a somewhat positive effect, and significantly improved over the unaugmented SFLW dataset. Over-fitting during the training process was markedly lower for SFLW-AUG and SFLW-NCA.

There is very little difference between the results for the AUG and NCA variants. It seems that, for this application, NCA has little effect on resulting performance. This is likely because the SFLW-AUG dataset already encapsulates enough variation in images with large pose angles for the network to learn sufficiently. SFLW-NCA also contains slightly fewer images than the AUG variant, but achieves near identical performance, suggesting that NCA may have a small positive impact.

The model trained on the SFLW-NCA dataset has a smoother distribution of MAE across the range of output angle magnitudes—as discussed for SAGR

above—so is used for head pose prediction in the following chapters. The results of this model are therefore also presented in Table 3.1 and Figure 3.1 above.

## 3.4 Summary

This chapter has demonstrated that, through fine-tuning of a CNN trained for human head pose estimation, accurate estimation of sheep head pose from images can be achieved. The Hopenet CNN [43] was fine-tuned on the SFLW-NCA dataset, resulting in an average absolute error of 6.6 degrees and average correlation of 0.75, with a maximum 0.91 for the yaw angle. TPS warping augmentation was shown to have a positive impact on the effectiveness of training, though NCA was found to provide no significant improvement over the model trained on SFLW-AUG. Qualitative results demonstrated that the network produces visually convincing head pose predictions for a variety of head poses, as well as for different breeds and ages of sheep.

# Chapter 4

# Evaluation of Existing Facial Landmark Localisation Methods

Face alignment for humans is a long-standing problem in computer vision and has been tackled in a number of ways over the past decade or more. Hand-crafted methods were initially popular, followed by techniques utilising cascades of regression trees, which were largely deemed to have solved the 2D facial landmark localisation problem. More recently deep learning approaches have provided even more impressive results, and improved resilience to variation in head pose, along with 3D landmark localisation.

This chapter briefly summarises some of the existing methods for human facial alignment, as well as some relevant work for animals. A number of the described methodologies are then evaluated on the SFLW dataset with no domain-specific modification. The effects of the data augmentation described in Section 2.2 are also considered as part of this evaluation.

## 4.1 Background

### 4.1.1 Classical Approaches

There are a large variety of classical approaches to facial alignment, though these typically tackle the problem only in 2D. Older approaches form separate shape and appearance models from training data which are matched to a test image by solving a non-linear least squares problem. These approaches do not deal well with large variance in pose and are typically quite slow, they have almost entirely been superseded by cascaded regression-based techniques.

Recent regression-based methods instead learn an ensemble, or cascade, of regressors based on local image features in order to iteratively refine an estimate of landmark locations. These typically obtain higher accuracy than shape models and are more resilient to variation in pose, but can still struggle with significant pose variation and occlusion. With the exception of some very recent deep learning based methods, modern regression techniques are considered state-of-the-art for 2D human facial alignment.

Some extensions to these methods have been proposed, primarily focussing on smart initialisation or the use of higher accuracy sparse localisation followed by refinement. Many of these rely on large volumes of data or additional information (e.g., 3D) so are not directly applicable, but provide an interesting source of inspiration for the problem of animal facial alignment.

#### Shape Based Techniques

The concept of matching a shape model to a deformable object in an image was first introduced by Cootes and Taylor in 1992 in the form of Active Shape Models [15] (ASM). The same authors later introduced Active Appearance Models [14] (AAM), a more performant refinement of ASM. AAMs make use of a statistical deformable model of the shape constructed during training, taking the form of a Point Distribution Model constructed using principal

component analysis (PCA). In addition, an appearance model is learnt either using holistic image features, or patch-based features taken from regions around the landmarks. These features are based on image intensities, and in modern applications typically use algorithms such as HOG [17] or SIFT [32]. Images are warped to match the reference shape of the training data and a generative statistical model is built as to what image intensities are expected in the region of each landmark. The solution is then found by solving a non-linear least squares problem in order to fit the shape model to the test image based on the appearance model using the Gauss-Newton optimization method.

Constrained Local Models [16] (CLM) were proposed as an improvement over AAM by instead forming the appearance model to generate likely feature templates, instead of trying to approximate image pixels directly. Other more recent developments include Active Pictorial Structures [2] (APS) which instead formulate the shape model as a number of pairwise distributions based on the edges of the graph of edges between landmarks. This is demonstrated to outperform the use of PCA in typical applications of AAMs. Tree Structured Part Models [63] (TSPM) use a similar alternative shape model and demonstrate impressive resilience to large variations in head pose, but rely on images annotated with precise pose information. The Supervised Decent Method [52] (SDM) was also proposed as an alternative to the Gauss-Newton method for optimization, this is shown to improve the results obtained from methods such as AAM.

#### Cascaded Regression Techniques

More recently, cascaded regression was introduced as a method for facial landmark localisation, initially in 2010 as Cascaded Pose Regression [18] (CPR). Rather than constructing shape and appearance models and fitting by optimization, a cascade of regressors is trained to iteratively refine an estimate of landmark locations starting from a rough initial guess. Results are good, but most importantly this method is very fast compared with classical

approaches, a key development being that features are sampled directly from the image based on the current estimate.

This technique was built upon directly to form Robust CPR [10] (RCPR) which can also regress an occlusion value (binarised from a continuous prediction between 0 and 1) and uses multiple initialisations and alternative feature extraction to improve performance. Project-Out Cascaded Regression [49] (PO-CR) instead solves the optimization problem faced by the shape based techniques described above using a cascade of regressors, this provides a major speed-up, but accuracy is not as high as more recent approaches. Explicit Shape Regression [11] (ESR) and Regression viva Local Binary Features [42] (LBF) are two other efficient regression-based approaches, though with slightly different training and feature extraction methods.

The current widely accepted state-of-the-art in terms of classical facial alignment, however, is a technique using an Ensemble of Regression Trees introduced in 2014 [26]. This approach is extremely efficient (up to 10000 fps) and highly accurate. A cascade of regressors are learnt via gradient boosting with a squared error loss function and, unlike other techniques, features are extracted directly from the image using an exponential prior. This method is certainly applicable to animal facial alignment and has been used for this purpose effectively as explored below.

### Extensions

A number of methods which take completely different approaches to the problem of face alignment, or augment the functionality of the techniques described indirectly, have also been proposed. [60] first predicts sparse landmarks using a cascaded regression model, then using a nearest-neighbours method to search the training data for a similar example. This then serves as an improved initialisation for another cascaded regression alignment step. [56] also predicts a sparse group of landmarks, but then aligns a 3D landmark model with sparse predictions in order to obtain a good initialisation. [51] instead uses a classifier to select an appropriate initialisation from a prede-

termined set which then serves as a basis for face alignment by cascaded regression. [61] uses an entirely different approach, treating facial alignment instead as a search problem using image features to asses the quality of a prediction.

The concept of smart initialisation is certainly applicable to animal facial landmark localisation, but these techniques often rely on large training datasets in order to obtain good results—something which is not available here.

### 4.1.2 Deep Learning Approaches

Recently, deep learning has become a popular approach to tackling a huge variety of computer vision problem, facial alignment included. Technological improvements and increased availability of large annotated datasets have enabled very accurate landmark localisation models to be produced which outperform all classical methods.

A number of applications tackle 2D landmark localisation specifically with impressive results [47, 58, 30, 48]. Increasingly common is the use of 3D data to further enhance performance and provide more detailed 3D output. Networks producing landmark heat-maps [7] have been developed, along with 3D morphable models [24] and 3D dense morphable models [62] which are able to deal with extreme variations in head pose while retaining very high accuracy. [54] instead predicts head pose using a CNN and projects a 3D landmark model to provide improved initialisation for classical face alignment techniques (an idea explored later in this dissertation).

Over the last year, very large datasets with comprehensive annotations (pose, gender, age, 2D/3D landmarks) have enabled all-in-one tools for the analysis of human faces. Highly accurate, unified networks such as 3DFAN [8] and hyperface [39] are able to predict many forms of output data with very high accuracy.

One might think, then, that 3D deep learning solutions are the obvious answer to the complex task of animal face alignment, given the large variations

in head pose and high levels of self-occlusion. However, there is insufficient data to consider these types of approaches. Transfer learning could be a solution to lack of data, but others have shown that the inter-species divide is too great between humans and horses/sheep for this to prove effective for the task of facial alignment [40].

### 4.1.3 Landmark Localisation for Animals

There has so far been little work looking at facial landmark localisation for animals. Sheep are targeted in [55] using a modified version of ERT with triplet interpolated feature (TIF) extraction. This paper largely forms the basis for the current work, contributing towards pain recognition in sheep [33] and providing the start point for the dataset used here.

Transfer learning is explored in [40], which focusses primarily on horses, but also utilises the dataset of [55]. Standard transfer learning is demonstrated to have limited effectiveness for inter-species facial alignment and a two-stage pipeline proposed which involves first warping the input image to more closely match human proportions, then localising landmarks with a fine-tuned CNN. This proves more effective than the TIF method proposed in [55], but can only deal with landmarks defined within standard human annotation schemes (i.e., not ears, which are critical for sheep).

An alternative deep learning approach is presented in [6] with application to cat and dog faces, as well as humans. The dataset used in this case is reasonably small, but there is no comparison to any meaningful baselines to verify the effectiveness of the approach.

## 4.2 Evaluation

To assess the performance of facial landmark localisation techniques three metrics are used. Firstly, mean normalised error (MNE); the mean across the dataset of the average normalised error, that is the Euclidean distance

of a predicted landmark to the ground truth landmark divided by the mean edge length of the bounding-box, averaged across all landmarks. Success rate; the proportion of the dataset with an MNE under 10%, and AUC; the normalised area under the cumulative MNE distribution (such as those pictured in Figure 4.1). Clearly for MNE lower is better and for success rate and AUC higher is better.

As the SFLW dataset is small, five-fold cross-validation is used; the dataset is split into five folds with localisation models tested on each one of the five sequentially. Performance metrics are then calculated using the data from all test folds. In cases where augmented datasets are used, the augmented training set is selected such that it contains variants of the raw training images only. The resulting models are then evaluated on the raw, rather than the augmented, test set.

This section first provides a comparison between a number of the techniques described above applied to the SFLW dataset. Then the effects of the data augmentation introduced in Section 2.2 are explored.

### 4.2.1 Overall Comparison

Existing implementations of state-of-the-art classical facial alignment methods—ESR [11] (C++), (R)CPR [18, 10] (Matlab), ERT [26] (C++) and TIF [55] (C++)—are modified to incorporate the 25 landmark annotation scheme used for the SFLW dataset. The modified implementations are then trained and tested on the applicable data and predicted landmarks exported to a common format. These predictions are then loaded within the Menpo framework [1] allowing for an identical evaluation procedure.

As the apparent state-of-the-art technique for animals [40] relies on transfer learning from human datasets, it is unable to localise the full SFLW annotation scheme (human annotation schemes do not incorporate ear landmarks). It is therefore excluded from this analysis, and comparison instead focussed on the state-of-the-art method which is able to localise all land-

Table 4.1: Quantitative performance metrics for existing landmark localisation methods trained on SFLW-flip and tested on SFLW.

| | Baseline | ESR | CPR | RCPR | TIF | ERT |
|---|---|---|---|---|---|---|
| MNE | 0.139 | 0.090 | 0.065 | 0.061 | 0.058 | **0.054** |
| Success Rate | 0.46 | 0.73 | 0.86 | 0.86 | 0.85 | **0.88** |
| AUC | 0.858 | 0.907 | 0.932 | 0.937 | 0.939 | **0.943** |

marks; ERT/TIF [26, 55].

All methods are trained with augmentation by horizontal mirroring (i.e., on the SFLW-flip dataset). Where appropriate, image repetitions with bounding-box perturbation are used with augmentation factor 30.

To provide a crude baseline, the mean shape of the SFLW dataset is calculated and projected into the bounding-box of each image, from this the error to the ground truth shape is calculated. This gives an indication of the level of performance expected when no facial alignment model, or *fitter*, is used at all.

Table 4.1 gives the performance metrics for this baseline, along with those for the existing methods being evaluated. Most achieve a reasonable accuracy, with the success rates for CPR, RCPR, TIF and ERT all over 85%. As expected RCPR improves slightly over CPR, oddly TIF performs markedly worse than unmodified ERT in contrast to the results presented in [55]. This may be due to the increased density of the SFLW landmark scheme (25 versus the 8 in [55]). Cumulative MNE distributions for these fitters are given in Figure 4.1.

ERT achieves the highest performance and is also the fastest to localise landmarks for an image, significantly so compared with (R)CPR which is the only other methods to achieve similar performance. As such ERT is selected as the base fitter for the remainder of this work.

Figure 4.2 shows the errors for each landmark individually using the ERT fitter trained on the SFLW-AUG dataset. The four landmarks corresponding

Figure 4.1: Cumulative MNE distributions for existing landmark localisation methods trained on SFLW-flip and tested on SFLW.

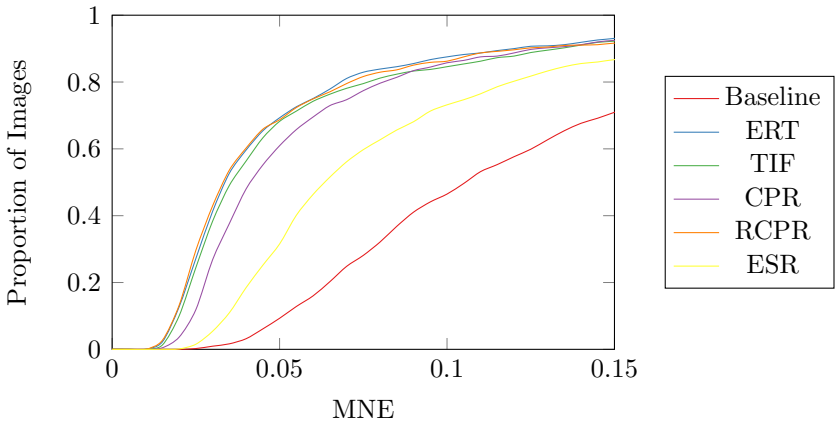
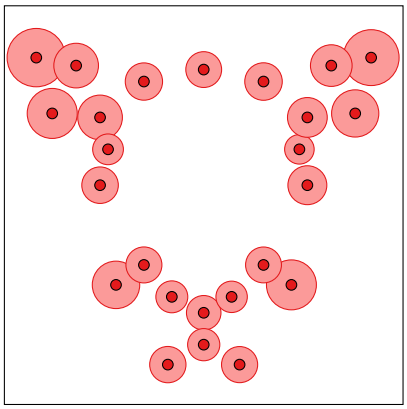



Figure 4.2: MNE for each landmark for the optimal ERT fitter tested on SFLW.



to the ears of the sheep display the highest error rate, which is not surprising given the large variation in the position of the ears relative to the rest of the head. The two landmarks on the sides of the face also exhibit high error rates, likely because these are the most often occluded of all landmarks. The most successfully localised landmarks are the eyes and those around the nose and mouth. This is probably due to the distinguishable visual nature of these elements, typically being much darker than the surrounding areas of the image.

Table 4.2: Localisation performance for ERT fitters trained on SFLW, SFLW-flip and SFLW-warp datasets and tested on SFLW.

| | SFLW | SFLW-flip | SFLW-warp |
|---|---|---|---|
| MNE | 0.0619 | 0.0581 | **0.0579** |
| Success Rate | 0.834 | 0.847 | **0.856** |
| AUC | 0.933 | 0.939 | **0.940** |

### 4.2.2 Effects of Data Augmentation

As described above, the ERT fitter is selected due to its high performance and low latency; all further investigation uses the ERT fitter only. In this section the effects of the newly introduced TPS warping technique are considered, followed by the impact of NCA and the number of bounding-box perturbations—repeats of a training image with slight variations in the position and size of the facial bound-box—referred to as $\#pert$.

#### TPS Warping

To evaluate the impact TPS warping has on performance, three ERT fitters are trained. First on the SFLW dataset with $\#pert = 30$, then SFLW-flip dataset with $\#pert = 15$, and finally on the SFLW-warp dataset with $\#pert = 5$. The augmentation factor of the SFLW-warp dataset is 6 (i.e., there are six TPS warped images for every raw image) and is 2 for SFLW-flip, so in total every image is repeated an equal number of times (30) in all cases.

Table 4.2 shows the localisation performance for these two fitters tested on the SFLW dataset. The fitter trained on SFLW-warp outperforms that trained on the dataset with no TPS warping, though the impact is slight.

#### NCA

Table 4.3 shows quantitative results for the ERT fitter trained on the unaugmented SFLW dataset, the SFLW-AUG dataset—with uniform augmentation—

Table 4.3: Localisation performance for ERT fitters trained on SFLW, SFLW-AUG and SFLW-NCA datasets ($\#pert = 30$).

| | SFLW | SFLW-AUG | SFLW-NCA |
|---|---|---|---|
| MNE | 0.062 | **0.050** | 0.063 |
| Success Rate | 0.83 | **0.90** | 0.85 |
| AUC | 0.933 | **0.945** | 0.932 |

and the SFLW-NCA dataset with negatively correlated augmentation; the results are perhaps surprising. As expected, the fitter trained on SFLW-AUG significantly outperforms that for SFLW, with a 7% increase in success rate, but the fitter trained on the SFLW-NCA dataset performs significantly worse. In fact, it only marginally improves over that trained on the unaugmented dataset.

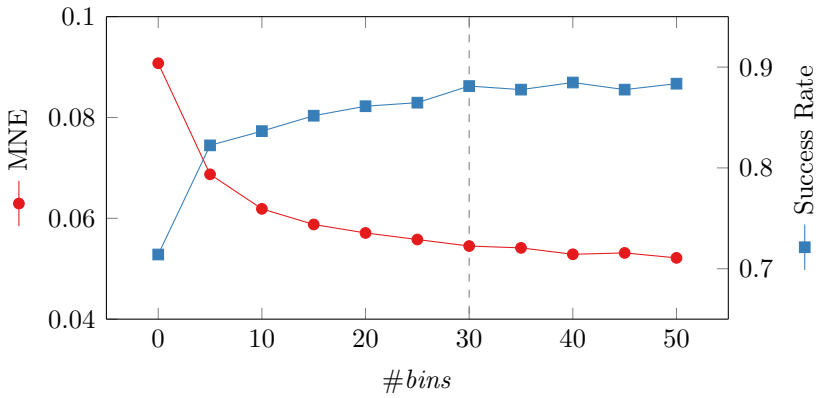
This is because the raw dataset used for testing is significantly skewed towards lower yaw angles (as described in Section 2.2.2). As such, providing additional augmentations for angles with high variation in pose has little impact on the performance when averaged over the whole test set. As there is less augmentation for images with low absolute yaw angle—common in the test set—it is actually not surprising that performance is much worse than for SFLW-AUG. This is explored further in the Chapter 5, motivating the use of a pose-informed facial alignment technique.

### Number of Bounding-Box Perturbations ($\#pert$)

As mentioned previously, zooming and horizontal and vertical image translation are common methods of data augmentation for machine learning applications. In the case of facial alignment, this is achieved by repeating an image while permuting the position and size of the facial bounding-box. For the ERT fitter the number of repetitions can be controlled, here referred to as parameter $\#pert$.

The effects of varying the value of $\#pert$ are explored for the standard ERT fitter trained on the SFLW-flip dataset, with results presented in Figure 4.3.

Figure 4.3: Effect of *#pert* on MNE and Success Rate for ERT fitter.




As expected, increasing the number of perturbations generally results in reduced MNE and higher success rate, as the fitter is able to achieve a more generalisable solution. The improvement in performance is drastic for low values of *#pert*, but performance gains reduce exponentially as *#pert* increases. MNE continues to decrease very slightly, but Success Rate appears to plateau at around 0.88 for $\#pert > 30$ with this training setup.

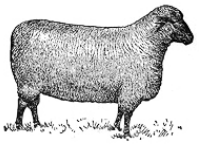
Increasing the number of perturbations also has a large impact on training speed. Higher values of *#pert* result in significantly longer training times, and larger memory requirements during the training process. As such, very large values of *#pert* are generally infeasible. $\#pert = 30$ is selected for the majority of experiments in this work as it provides near-optimal performance while maintaining reasonable time and memory requirements for training. A lower value is used in some comparative evaluations to facilitate faster training.

## 4.3 Summary

This chapter has summarised classical and deep learning approaches for the task of human facial alignment, as well as describing existing work focussing on landmark localisation for animals. A number of the state-of-the-art meth-

ods were modified to for use with the sheep facial landmark annotation scheme and evaluated on the SFLW dataset. ERT [26] is shown to be the most highly performing method, with 0.05 MNE and 90% success rate when trained on the SFLW-AUG dataset. This method is able to accurately localise the eyes, nose and mouth, though struggles particularly with the ears. TPS warping is shown to be effective for the task of facial alignment, but NCA is demonstrated to have a negative impact on results in this case. The effect of the number of bounding-box perturbations is also assessed for the ERT fitter, with an optimal value of 30 selected.

# Chapter 5



# Pose-Informed Face Alignment

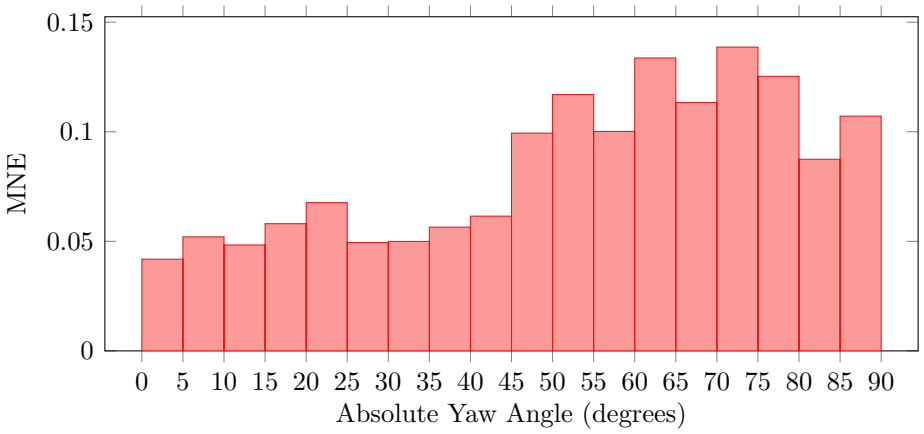
Informed by the results of landmark localisation using existing methods designed for human faces, this chapter introduces a novel, pose-informed technique for facial alignment of animals. First, the context and motivation of the technique are explained, followed by a description of the methodology used. Finally, the proposed pose-informed landmark localisation method is evaluated and compared with existing methods, along with an investigation into the effects that the choice of parameters and data augmentation have on results.

## 5.1 Motivation

In order to devise an improved method of landmark localisation for sheep, we must first consider what it is that causes conventional approaches to fail. Head pose variation is a common problem in human facial alignment [8], but is mitigated by having access to very large and heterogeneous datasets. Large datasets for animals are not available and, given the significant variations in pose and resulting self-occlusions within the SFLW dataset, this seems an obvious cause for error.

Figure 5.1 shows the MNE for images grouped by absolute yaw angle using

Figure 5.1: MNE distribution, using 5 degree bins, for ERT fitter trained on SFLW-AUG and tested on SFLW.




5 degree bins. It is clear that head pose has a significant impact on the accuracy of facial alignment. For angles less than 45 degrees from frontal, the error is relatively low, but for angles between 45 and 90 degrees the error is significantly higher, sometimes up to almost 15% of the average side length of the facial bounding-box. This supports the assertion that large head pose variation is the primary cause of failure.

Looking qualitatively, Figure 5.2 includes example images which have some of the highest MNE of the dataset. All images exhibit extreme poses, with many of the sheep facing almost 90 degrees away from frontal. It is unsurprising that a fitter initialised using the mean shape of the dataset would struggle to adapt to these poses.

Consequently, the concept of better initialisation is considered; rather than initialising from the mean shape, instead the fitter is initialised with a shape that is closer to the ground-truth, theoretically making the task of localisation simpler. A number of methods were investigated: re-projection of a 3D model of the sheep facial landmarks using predicted pose (as used for humans in [54]), selection of pre-defined initialisations based on predicted pose bin (similar to [51]), and nearest neighbour search (inspired by [60]). None of these methods, however, produced significant improvements. It is possible that due to high levels of self-occlusion and inter-breed variation that the

Figure 5.2: Failure cases for standard ERT fitter. The top row shows ground truth landmarks and the bottom row shows predicted landmarks.

image data is simply too varied for a single fitter to encompass all poses.

So instead of a smart initialisation technique using a single fitter, a multi-fitter approach is introduced. This increases storage space requirements and can increase training time, but for the proposed application this is not an important consideration. The proposed method is similar to initialisation using pose bins, but a separate fitter is instead trained for each bin. The correct fitter is then selected at test time based on the predicted head pose. This technique is dubbed Pose-Informed ERT (PI-ERT) and described in detail in the following section.

## 5.2 Methodology

### 5.2.1 Training

Training images are first partitioned into a number of bins based on their associated yaw angles. Images with negative yaw angles (i.e., those with sheep facing to the right of the image) are mirrored horizontally and the positive yaw angle is taken for training. This means that all images used for training contain left-facing sheep, and the number of bins (hereafter referred

Figure 5.3: Example images for series of seven pose bins. As right facing images are ordinarily flipped, this is equivalent to $\#bins = 4$.

to as $\#bins$) can essentially be halved. Flipping all images to be in the same direction also has the effect of doubling the number of training images per bin, if the effects of horizontal mirroring augmentation are ignored. Examples of images partitioned into pose bins are given in Figure 5.3.

Pose bin divisions are evenly sized within the range 0–90 degrees, so, for example, $\#bins = 3$ would result in the three bins: 0–30, 30–60 and 60–90. A separate fitter is then trained for each of these bins. Additional augmentation is provided by repeating training images with slightly perturbed bounding-boxes (as in the previous chapter, the number of perturbations used for training of a given model is referred to as $\#pert$). Image sizes are also normalized relative to the bounding-box size prior to training.

The ERT localisation algorithm [26] is selected due to its high performance and very fast localisation time, as highlighted in the previous chapter, with the aim of retaining real-time performance for the PI-ERT localisation method. The dlib [27] implementation of ERT is again used within the Menpo framework [1] for data handling.

As throughout, 5-fold cross-validation is used. It is also important to note that the cross-validation sets used here are identical to those used for pose estimation. Thus no test images for the PI-ERT system have been used for training of either the pose estimation step or the landmark localisation step. As such, ground-truth pose annotations are used for training, and predicted pose from the network described in Chapter 3 for testing.

### 5.2.2 Testing

Given a test image with an annotated bounding-box, the face is cropped and resized then run through the pose estimation network of Chapter 3. This provides an estimate of yaw, roll and pitch on a continuous scale from $-90$ to $+90$ degrees. As for training images, if the angle is below 0 (the sheep is facing to the right) the image is flipped horizontally. The absolute value of the predicted yaw angle is then used to determine which pose bin the image falls into, and therefore which of the pre-trained fitters should be applied for the image. The fitter is applied as usual to localise the facial landmarks, the image (and predicted landmarks) can then be mirrored back if required.

It is clear that there is some trade-off in choosing the value of $\#bins$. With fewer bins, it is more likely that the pose estimation will be accurate enough to select the correct bin, and there are sure to be enough training images in each bin. With a greater number of bins it is more likely there will be an error in bin selection and there might be insufficient training images in each bin, but each fitter should be able to achieve higher performance as it is more closely tailored to a specific range of angles. The choice of $\#bins$ and the overall performance of the PI-ERT approach are explored in detail in the following section.

## 5.3 Evaluation

This section summarises and evaluates the performance of the proposed PI-ERT facial landmark localisation technique. The evaluation procedure and performance metrics used for PI-ERT are the same as described for the existing landmark localisation techniques in the previous chapter. Five-fold cross-validation is used, and results reported in terms of MNE, Success Rate, and AUC.

Table 5.1: Baselines and optimal PI-ERT localisation performance. (SFLW-NCA, $\#pert = 30$ and $\#bins = 3$).

| | Mean Shape | ERT (SFLW-AUG) | PI-ERT |
|---|---|---|---|
| MNE | 0.139 | 0.050 | **0.045** |
| Success Rate | 0.46 | 0.90 | **0.94** |
| AUC | 0.858 | 0.942 | **0.949** |

### 5.3.1 Overall Performance

#### Comparison to Existing Methods

The proposed PI-ERT method is compared with the crude mean shape baseline used previously, as well as the standard ERT model (as presented in Chapter 4) trained on the SFLW-AUG dataset, which represents the current state-of-the-art for animal facial landmark localisation.

Table 5.1 includes quantitative results for these baselines, along with the optimally performing PI-ERT fitter ($\#bins = 3$). Figure 5.4 shows the cumulative MNE distributions for the same. The impact of the pose-informed method is clear; MNE is reduced significantly and Success Rate and AUC are markedly increased. The MNE curves also clearly demonstrate the improvement that PI-ERT makes over the standard ERT models.

Figure 5.5 shows qualitative fitting results for the images features previously in Figure 5.2. It is clear that the PI-ERT makes significant improvements in most of these cases. The right-most image also fails using the PI-ERT fitter, estimated yaw angle for this image is inaccurate likely leading to this failure.

Further qualitative results for the PI-ERT fitter are shown in Figure 5.6. Accurate facial alignment is achieved across a wide variety of poses, image resolutions and breeds of sheep.

Figure 5.4: Baselines and optimal PI-ERT cumulative MNE distributions. (SFLW-NCA, $\#pert = 30$, $\#bins = 3$).

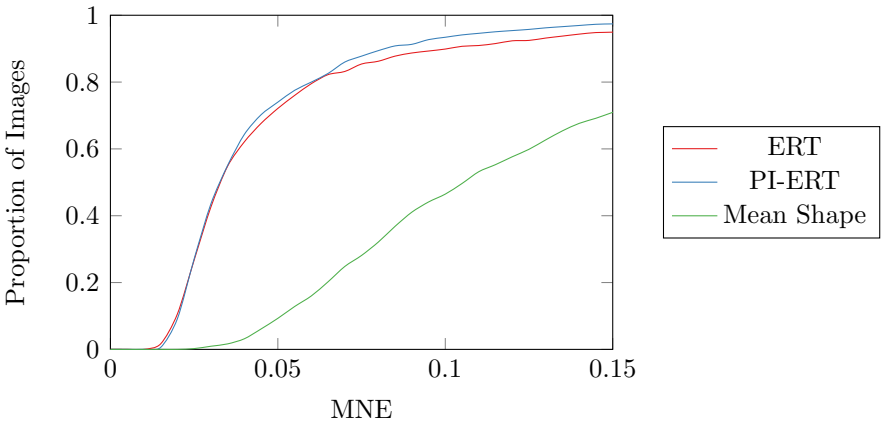



Figure 5.5: Example failure cases for the standard ERT fitter and improvements made by PI-ERT. Rows show ground-truth, standard ERT fitting results and PI-ERT fitting results from top to bottom. The right-most column shows an example where PI-ERT also fails.

Figure 5.6: Qualitative results for optimal PI-ERT fitter. (SFLW-NCA, $\#pert = 30$, $\#bins = 3$).

Table 5.2: 3-bin PI-ERT with horizontal mirroring and 6-bin PI-ERT without horizontal mirroring localisation performance metrics. (SFLW-NCA, $\#pert = 10$).

| | 3-bin PI-ERT | 6-bin PI-ERT (no mirroring) |
|---|---|---|
| MNE | **0.050** | 0.054 |
| Success Rate | **0.91** | 0.90 |
| AUC | **0.948** | 0.943 |

### Impact of Horizontal Mirroring

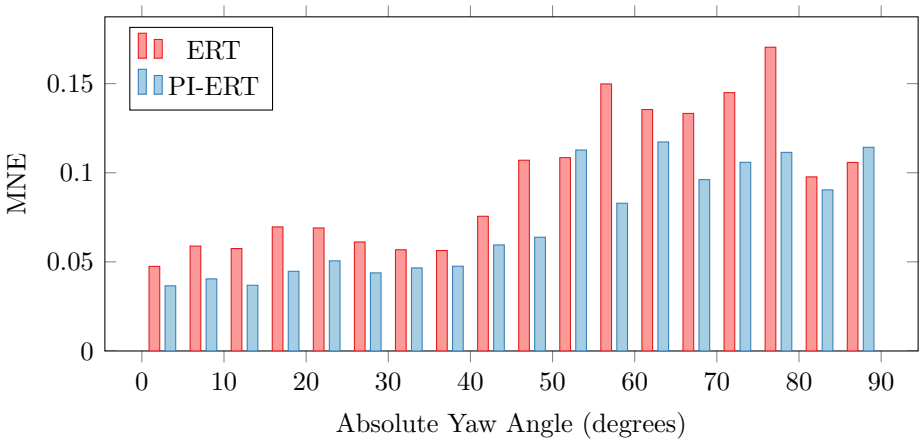
Table 5.2 shows the results for 3-bin PI-ERT (with mirroring) compared with 6-bin PI-ERT (without mirroring), both trained using the same dataset and parameters. For all metrics 3-bin PI-ERT with mirroring outperforms the equivalent setup without mirroring, this can only be a result of the increased number of training images per fitter obtained via the mirroring process. Mirroring images horizontally is therefore demonstrated to not only reduce storage requirements, but also result in improved performance.

### Error Distribution

Next, the impact of PI-ERT on the distribution of error relative to head pose is considered. Figure 5.7 shows the MNE distribution by absolute yaw angle for the standard ERT fitter (as shown previously in Figure 5.1), along with that for the PI-ERT fitter. MNE is reduced across almost the entire range of angles, though is still larger for high magnitude angles. For angles between 45 and 80 degrees, however, there is a noticeably more significant reduction in error than for lower magnitude angles. This demonstrates that the PI-ERT method does prove effective in tackling large pose variation.

Finally, MNE for individual landmarks is considered, as shown in Figure 5.8. Error for all landmarks is lower using the PI-ERT fitter compared with the standard ERT fitter. The three landmarks on the top of the head show the least change, with only very marginally reduced error. Most notably reduced

Figure 5.7: MNE distribution, using 5 degree bins, for optimal PI-ERT and standard ERT fitters. ($\#pert = 30$, $\#bins = 3$)



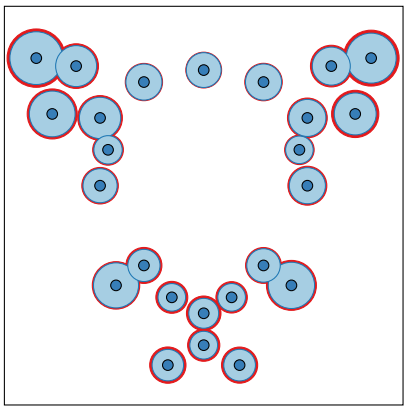

are the landmarks representing the nose, mouth and chin and those for the ears. Though the error for ear landmarks is significantly reduced, these still show the highest error of the landmarks, along with those on the side of the face.

### 5.3.2 Effect of Number of Bins ($\#bins$)

As mentioned above, the selection of the value of $\#bins$ can have a large impact on the performance of the technique due to the trade-off between the training set size and bin selection accuracy, and the fitter pose specificity. In this section, the impact of varying the value of $\#bins$ for PI-ERT with fixed $\#pert$ is evaluated.

Figure 5.9 shows the variation in MNE and Success Rate for values of $\#bins$ from 1 to 8. As expected, values towards the middle of this range achieve the lowest MNE and highest success rate. For higher values of $\#bins$, e.g., 8, the bin width is only around 10 degrees. Given the MAE of the head pose estimation network is around 6 degrees, the number of bin selection errors in this case is likely to be high. Likewise, the number of training images in each of these narrow bins for higher pose angles is very low (order of tens of

Figure 5.8: MNE for each landmark using PI-ERT fitter (blue) overlaid on errors for standard ERT fitter (red). The larger the red area visible for a landmark, the greater the improvement made by PI-ERT.



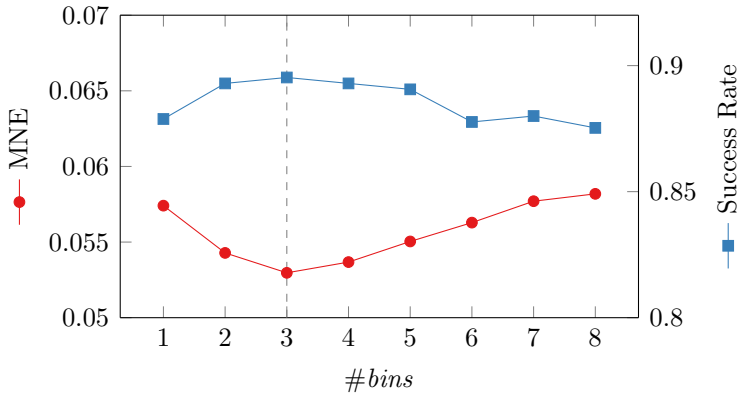
images), even for augmented datasets.

At the opposite end of the range, there are plenty of training images per bin and bin selection accuracy is likely to be high, but there is little gain in terms of fitter specificity. A single fitter is still having to deal with the difference in image data between a head pose of 0 and a head pose of 45 or more degrees, for example.

As Figure 5.9 shows graphically, $\#bins = 3$ is found to be the best setting for this use case. With a significant reduction in performance for lower $\#bins$, and a more gradual reduction in performance for increasing $\#bins$. The results for $\#bins = 1$ and $\#bins = 8$ are almost equivalent, it is probable that any higher values of $\#bins$ are even less effective. Models used for evaluation of PI-ERT throughout this section therefore use $\#bins = 3$.

It is important to note that selection of this value for $\#bins$ is driven by the characteristics of the SFLW dataset. For other datasets with a greater representation of extreme head poses, or for which a more accurate head pose estimation technique is available, it is likely that a higher value of $\#bins$ will result in improved performance.

Figure 5.9: Effect of $\#bins$ on MNE and Success Rate for PI-ERT. Optimal performance for both metrics achieved at $\#bins = 3$.




### 5.3.3 Effects of Data Augmentation

The effects of most augmentation methods are evaluated in detail for the standard ERT fitter in the previous chapter. For the PI-ERT fitter the effects are largely unchanged, so are not discussed in detail here. There is one exception, however; the effect of NCA on the PI-ERT fitter is drastically different from that described for the ERT fitter above.

#### NCA

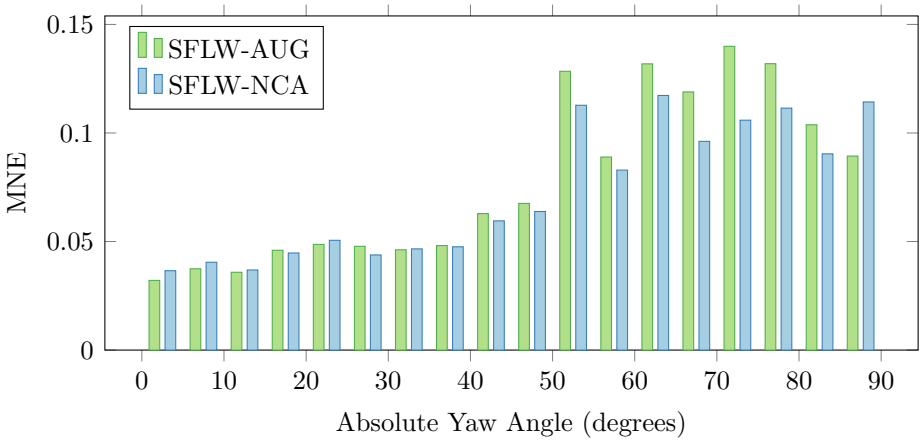
In the previous chapter, NCA was found to provide little improvement over the unaugmented dataset for the standard ERT fitter, in contrast to the SFLW-AUG dataset which significantly improved performance. Table 5.3 shows the results for the PI-ERT fitter trained on the same three datasets; here a different trend is observed.

As before, the SFLW-AUG dataset provides a significant improvement in performance over the unaugmented SFLW dataset, but for PI-ERT this improvement is equalled when trained on SFLW-NCA. In fact, training on the SFLW-NCA dataset results in a slightly higher success rate, though also a marginally higher MNE and lower AUC. We are now faced with the question

Table 5.3: Localisation performance for PI-ERT fitters trained on SFLW, SFLW-AUG and SFLW-NCA datasets ($\#pert = 30$, $\#bins = 3$).

| | SFLW | SFLW-AUG | SFLW-NCA |
|---|---|---|---|
| MNE | 0.053 | **0.044** | 0.045 |
| Success Rate | 0.89 | 0.93 | **0.94** |
| AUC | 0.942 | **0.950** | 0.949 |

Figure 5.10: MNE distributions, using 5 degree bins, for PI-ERT fitter trained on SFLW-AUG and SFLW-NCA. ($\#pert = 30$, $\#bins = 3$)




as to which of these datasets is actually better for training the PI-ERT fitter given their near equivalent overall performance.

Figure 5.10 shows the distribution of MNE with respect to absolute yaw angle for the PI-ERT fitter trained on the SFLW-AUG and SFLW-NCA datasets. For lower yaw angles there is little apparent difference in error, with SFLW-NCA resulting in marginally higher errors in some cases. However, for higher magnitude yaw angles the difference is more pronounced; the SFLW-NCA trained fitter produces notably lower MNE in these cases, with the exception of only the 85–90 degree bin.

As such, the PI-ERT fitter trained on the SFLW-NCA achieves optimal performance when tested on SFLW compared with all other fitters.

### 5.3.4 Application to Human Faces

In order to assess the generalisability of the proposed pose-informed facial alignment method, standard ERT and PI-ERT fitters are trained for facial landmark localisation using a common human dataset, AFLW [29]. This dataset is selected in lieu of an animal dataset or other human datasets for two reasons. Firstly, the only other animal dataset [40] does not contain head pose annotations or locations for occluded landmarks, so is not viable for use with PI-ERT. Secondly, AFLW contains in-the-wild facial images with head pose annotations and a much larger range of head poses than most other human facial datasets.

However, the original AFLW dataset does not contain complete landmark annotations for images with occlusions. So the annotations generated by Smith and Zhang [46] are used instead. This provides full 21 landmark annotations for all images of the AFLW dataset. Ground-truth head pose annotations from the AFLW dataset are used for the purposes of the evaluation, as Ruiz et al. [43] have already demonstrated that highly accurate, landmark-free head pose estimation is possible for humans, hence so is PI-ERT.

A semi-randomly selected subset of the full AFLW dataset is used, containing 1750 images, with 5-fold cross-validation used for evaluation. Sampling is performed in a way that somewhat mimics the effects of NCA, with images selected such that the distribution of absolute yaw angles for the resulting subset is less skewed than the distribution of the original AFLW dataset. No augmentation is performed, apart from perturbation of bounding-boxes ($\#pert = 10$). $\#bins = 3$ is used due to its effectiveness on the SFLW dataset as described above.

Table 5.4 shows the results for ERT and PI-ERT fitters trained on the described subset of AFLW. PI-ERT outperforms the standard ERT fitter in all metrics, indicating that this technique does generalise to other types of subjects where head pose variation is significant—in this case, humans.

Table 5.4: Standard ERT and PI-ERT localisation performance on a subset of the AFLW dataset. ($\#pert = 10$ and $\#bins = 3$).

| | Standard ERT | PI-ERT |
|---|---|---|
| MNE | 0.059 | **0.056** |
| Success Rate | 0.87 | **0.90** |
| AUC | 0.939 | **0.941** |

## 5.4 Summary

This chapter first provides an analysis of the performance of the current state-of-the-art method, ERT, and demonstrates that head pose variation is the key factor in reducing effectiveness. Based on this, a novel pose-informed method is proposed—Pose-Informed ERT, or PI-ERT—making use of the head pose prediction network introduced above. PI-ERT is demonstrated to improve significantly on the performance of the ERT fitter, achieving 0.045 MNE and 94% success rate. In contrast to the standard ERT fitter, NCA is shown to improve the performance of the PI-ERT method, particularly in reducing errors for more extreme head poses. An exploration of the effect of the number of pose bins used shows 3 to be the most performant setting. The PI-ERT fitter is also demonstrated to be effective in improving performance for human facial alignment under large head pose variation using a subset of the AFLW dataset [29] when compared with the same baseline method.

# Chapter 6

# Deployment Pipeline

As described in Chapter 1, the end-goal of this research—and motivation for improving facial alignment for animals—is for application to pastoral agriculture. Pain levels in sheep have been estimated effectively based on facial imagery [33]. Integrating this into a CCTV monitoring system would allow for detection of medical issues requiring further attention as early as possible, rather than relying on the infrequent visits of veterinary professionals.

Based on this concept, a proof-of-concept pipeline is developed taking a pre-recorded video as input. Sheep faces are detected automatically, and facial landmarks are localised using the PI-ERT technique introduced in Chapter 5. The example pipeline, therefore, demonstrates both the overall feasibility of such a system, as well as the deployability of the proposed PI-ERT method within a real-world application. This chapter describes the implementation details of this pipeline and provides an evaluation of its effectiveness.

## 6.1 Methodology

### 6.1.1 Face Detection

Face detection is an important aspect of the pipeline which has not been explored already within this work. The dlib [27] HOG-based SVM object detector [17] is used with three separate detectors trained: one for sheep facing the camera, one for left-facing sheep and one for right-facing sheep. This should enable improved detection accuracy across a range of head poses where a single detector typically struggles to capture such large variation in image data; a similar principle to that which motivates PI-ERT.

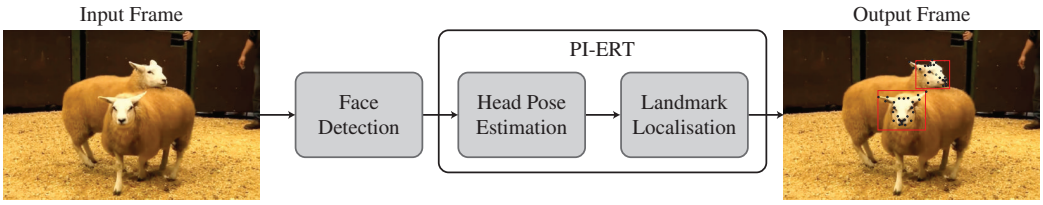
The images in the SFLW are annotated with additional bounding-boxes for sheep in the background of the images which were not deemed suitable for full landmark annotation. This is required as the training procedure treats non-annotated areas of images as negative examples, so training is compromised if unannotated sheep faces are present.

One-tenth of the resulting dataset is set aside for testing, the remaining training set is manually split into three subsets: one containing left-facing sheep, one for right-facing sheep and one for sheep facing the camera. The left-facing sheep are also mirrored and included in the right-facing subset and vice-versa.

### 6.1.2 Pipeline Architecture

A simple diagram of the pipeline is shown in Figure 6.1. The input video frame is first run through the multi-pose face detector and any resulting detections cropped from the full frame. Each detection is then resized to the correct dimensions and passed into the head pose estimation network. This provides a yaw angle between $-99$ and 99 degrees, if this is below 0 then the frame is mirrored horizontally. The applicable fitter is then selected based on the absolute predicted yaw angle and executed using the full input frame.

Figure 6.1: Basic pipeline structure with example input and output images.




Finally, the resulting landmark locations are mirrored back, if required, and rendered onto the frame for output.

Within the pipeline, rudimentary tracking of bounding-boxes is used to smooth transitions between frames and prune overlapping boxes. The proportional overlap area of each pair of boxes in a frame is used to remove any boxes with overlap factor greater than 0.6. Between frames, any two boxes with an overlap factor greater than 0.8 are considered to correspond to the same sheep face, so the positions of the two are averaged. This results in a smoothed detection path where frames further in the past has ever decreasing impact on the current bounding-box position.

As described above, the full PI-ERT localisation procedure is performed for each detection in every frame, with no tracking of landmarks. This is clearly an area of potential for improvement. It is likely that a production system would use the head pose estimation network for fitter selection every few frames as a form of checkpoint and rely on only the ERT fitter coupled with some tracking method, such as [9], for intermediary frames.

## 6.2 Evaluation

In order to assess the effectiveness of the proposed pipeline, the following evaluative section is broken into three parts. First, the performance of the face detection method is considered for a test set composed of images from the original SFLW dataset. Secondly, the overall qualitative performance of the pipeline when applied to two example videos is considered, including

Table 6.1: Face detection performance metrics.

| | Left | Centre | Right |
|---|---|---|---|
| Precision | 0.32 | 0.96 | 0.56 |
| Recall | 0.11 | 0.45 | 0.14 |

face detection, as well as the resulting landmark localisation using PI-ERT. Finally, the feasibility of the pipeline for real-time deployment is evaluated, focussing on the run-time of each component when executed on an example video.

### 6.2.1 Face Detection

The test set used for the sheep face detector described above is selected randomly from the raw dataset and not split by pose as the training set is. As such, it is composed primarily of frontal images with only a small number of left- and right-facing images. The performance metrics of the three detectors evaluated on this test set are given in Table 6.1. We would expect recall values to sum to approximately 1 across the three, assuming detections are unique, and for the central detector to have higher recall than the left and right. The very low recall values in Table 6.1 are therefore not as bad as they first appear. That said, the overall recall achieved is just 0.7, clearly not as high as might be desired. Precision is also quite low for the left and right detectors, indicating that false positives are common.

### 6.2.2 Overall Effectiveness

No ground-truth annotations for video files are available, so evaluation of face alignment in this context is solely qualitative. Figure 6.2 includes some example frames from two videos for which the full pipeline has been used. As shown, facial landmarks of multiple sheep are localised effectively across a number of different frames and for videos with very different lighting con-

ditions.

Missing detections are relatively common and, to a lesser extent, false positives; detection is definitely one of the less successful elements of the pipeline as suggested by the quantitative results above. Motion blur might be a large factor in reducing the effectiveness of face detection, as well as landmark localisation. Very few training images incorporate motion blur, though it is common in the demonstration videos. Adding video frames with motion blur to the training set would likely improve performance significantly. For a fixed camera, such in the case of CCTV, motion blur might also be a less common problem.

As shown quantitatively for the SFLW dataset in Figure 5.8 above, ears are localised much less accurately in many cases. This might inform the focus of future work in improving facial landmark localisation techniques for sheep and similar animals.

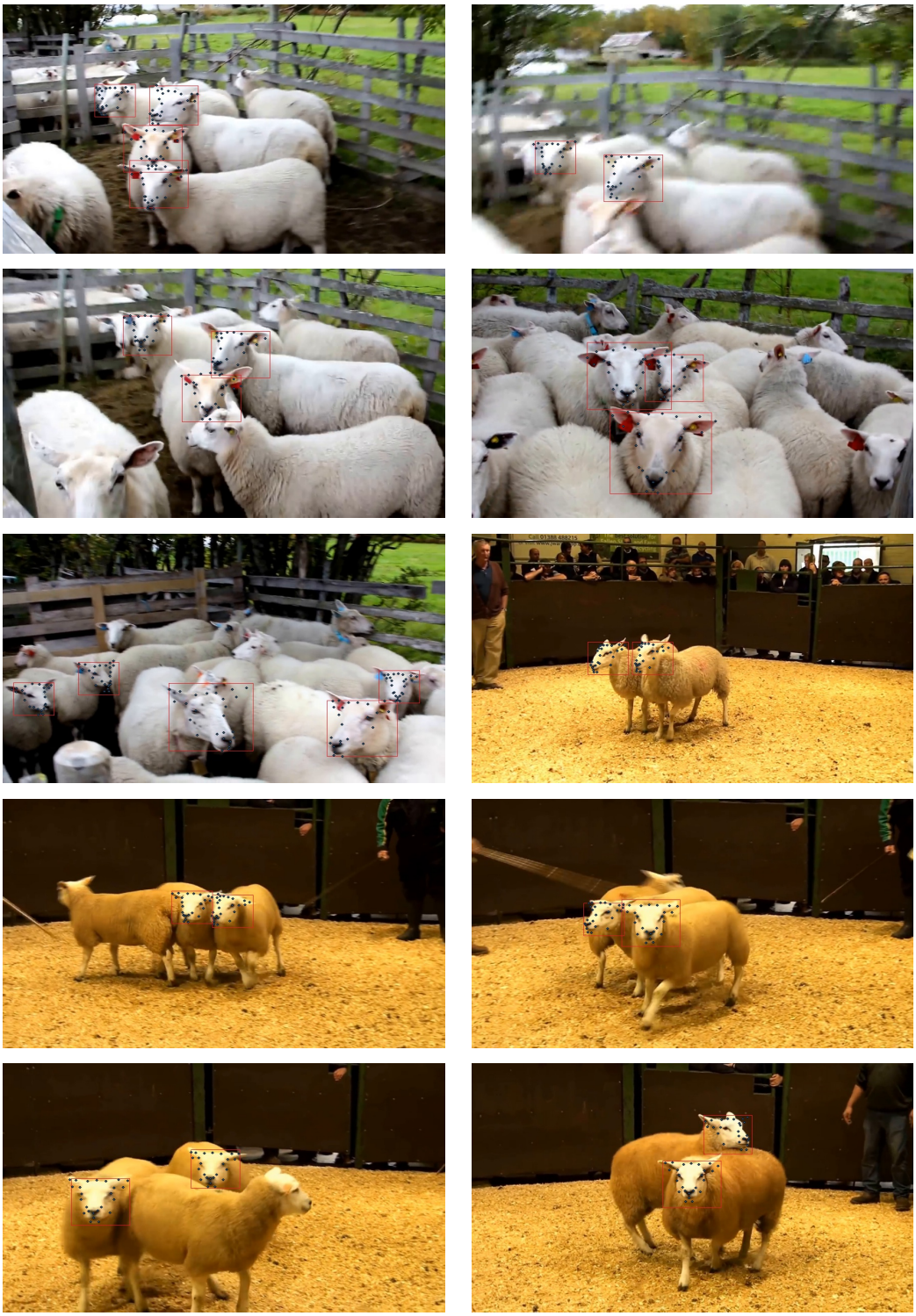
### 6.2.3 Computational Feasibility

When applied to live video feeds the speed of operations within the pipeline is a key concern. There are a number of steps in the pipeline: face detection, head pose estimation, and landmark localisation, so real-time performance is not a given. The times taken for the execution of each stage of the pipeline are recorded across 500 frames of one of the videos used in Figure 6.2, at approximately $700 \times 400$ pixel resolution. Measurements are made on a 2013 MacBook Pro (2.8 GHz Intel Core i7 16 GB 1600 MHz DDR3) with an external NVIDIA GeForce GTX 1080Ti.

Detection takes on average 64ms per frame with around 1.1 bounding-boxes found per frame. Head pose estimation requires 13ms and facial alignment 16ms per bounding-box. Therefore, each frame takes approximately 100ms to process, equivalent to $\sim$10fps. This is lower than conventional video frame-rate but is not unreasonable for a surveillance system.

It is also clear from these results that the detection step is the major bottle-

Figure 6.2: Sample frames from two example videos run through the demonstration pipeline.



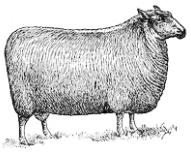
Videos from YouTube, released under Creative Commons by Tim Berglund and Robert Smith.

neck, taking four times as long as the landmark localisation step. There are more time-efficient detection methods available, such as YOLO [41], which could be exploited here given sufficient training data. Real-time performance (30fps) is certainly realistic for such a system, provided slightly more powerful hardware and time for fine-tuning the pipeline were available.

## 6.3 Summary

This chapter has demonstrated the applicability of the introduced PI-ERT landmark localisation technique to real-time monitoring of livestock. An example pipeline was implemented which detected sheep faces, estimated the head pose and, using this information, localised the facial landmarks. Implementation included the training of a multi-pose face detector to find sheep faces within a video frame. Evaluation showed that there is significant room for improvement in this aspect of the pipeline. Execution of the full pipeline on two example videos proved to be effective qualitatively, in the absence of ground-truth data to make a quantitative assessment. Timing analysis showed the pipeline to execute at approximately 10fps, which is close to real-time. This certainly serves as a proof-of-concept that such a pipeline is feasible given adequate resources and an improved face detection technique.

# Chapter 7



# Conclusion

This project has provided an investigation into facial landmark localisation and its application for animals, with a specific focus on sheep, for which automated pain detection has been shown to be effective [33]. A new dataset containing 850 images of sheep was annotated using a 25 landmark scheme, much denser than previous schemes for animals [55, 40]. The dataset, Sheep Facial Landmarks in the Wild (SFLW), incorporates a challenging mix of images, with sheep exhibiting large variations in head pose and occlusion. Per landmark occlusion information is also annotated, though not explored in detail as part of this project. Robust landmark localisation for animals, like RCPR [10] for humans, is definitely an area that might be of interest for future work.

Due to the small volume of data, augmentation was of critical importance. A novel image warping augmentation technique using TPS warping [5] was introduced and shown to be effective in improving the performance of both head pose estimation and landmark localisation. In addition to this, a negatively correlated augmentation (NCA) technique based on head pose was used for certain applications, similar to that proposed in [55]. These augmentation methods could prove useful in similar scenarios where data is sparse.

In order to improve head pose invariance for landmark localisation, landmark-free head pose estimation for animals was explored based on a CNN-oriented

technique for humans [43]. Fine-tuning this model on the SFLW dataset proved effective, with an average absolute error of less than 7 degrees. Pose-informed landmark localisation is a relatively unexplored area that could prove fruitful given the high accuracy of current head pose estimation techniques and increasing computational feasibility of such techniques.

Using this sheep head pose estimation network, a pose-informed landmark localisation method based on the ERT [26] algorithm was developed, dubbed Pose-Informed ERT (PI-ERT). This technique was shown to be effective in reducing localisation error on the SFLW dataset compared with existing localisation methods, also evaluated as part of this project. An overall success rate of 93% and mean normalised error of 0.045 were achieved, with performance particularly improved for sheep with more extreme head poses. PI-ERT also proved effective for human facial alignment using a subset of the AFLW dataset [29]. Ears were localised quite poorly due to the large range of positions they can take relative the rest of the head. Future work might focus on this as an area to most effectively reduce localisation error.

In order to demonstrate the feasibility of applying this form of localisation in a real-world setting—for example, as part of a surveillance system for monitoring livestock health on a farm—a pipeline including face detection, head pose estimation and pose-informed landmark localisation was constructed and tested on a number of pre-recorded videos. Near real-time performance was achieved on relatively modest hardware, with visually successful fitting results. Sheep face detection requires significant improvement to produce a truly deployable application, though this is a problem that could be readily solved through increased experimentation and greater data volume. The demonstrated pipeline certainly served as a proof-of-concept that this form of application is eminently feasible using techniques such as those proposed in this project.

# Bibliography


[1] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proc. of the ACM int'l conf. on Multimedia*, pages 679–682. ACM, 2014.

[2] E. Antonakos, J. Alabort-i Medina, and S. Zafeiriou. Active pictorial structures. In *Proc. of CVPR*, pages 5435–5444, 2015.

[3] M. P. Arakeri et al. Computer vision based fruit grading system for quality evaluation of tomato in agriculture industry. *Procedia Computer Science*, 79:426–433, 2016.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE trans. on pattern analysis and machine intelligence*, 24(4):509–522, 2002.

[5] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE trans. on pattern analysis and machine intelligence*, 11(6):567–585, 1989.

[6] A. Bulat and G. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. University of Nottingham, 2016.

[7] A. Bulat and G. Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *Proc. of ECCV*, pages 616–624. Springer, 2016.

[8] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proc. of ICCV*, volume 1, page 8, 2017.

[9] X. P. Burgos-Artizzu, D. C. Hall, P. Perona, and P. Dollár. Merging pose estimates across space and time. In *Proc. of BMCV*, 2013.

[10] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proc. of ICCV*, pages 1513–1520. IEEE, 2013.

[11] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *Int'l Journal of Computer Vision*, 107(2):177–190, 2014.

[12] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *Proc. of CVPR*, pages 3386–3395, 2017.

[13] ConnecTerra. Connecterra: The intelligent dairy farmers assistant, 2017.

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE trans. on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[15] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[16] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, volume 1, pages 886–893. IEEE, 2005.

[18] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *Proc. of CVPR*, pages 1078–1085. IEEE, 2010.

[19] A. Efros. Lecture slides on image warping and morphing, 2011. Carnegie Mellon University.

[20] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Readings in computer vision*, pages 726–740. Elsevier, 1987.

[21] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *Proc. of Automatic Face & Gesture Recognition*. IEEE Computer Society, September 2008.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778, 2016.

[23] B. Hu, Q. Tian, Z. Chen, G. Xiong, X. Wang, and Q. Wang. Intelligent farming control system based on computer vision. In *Proc. of SOLI*, pages 202–205. IEEE, 2014.

[24] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proc. of CVPR*, pages 4188–4196, 2016.

[25] A. Kasinski, A. Florek, and A. Schmidt. The put face database. *Image Processing and Communications*, 13(3-4):59–64, 2008.

[26] V. Kazemi and S. Josephine. One millisecond face alignment with an ensemble of regression trees. In *Proc. of CVPR*, pages 1867–1874. IEEE Computer Society, 2014.

[27] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[29] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE Int'l Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[30] A. Kumar, R. Ranjan, V. Patel, and R. Chellappa. Face alignment by local deep descriptor regression. *arXiv preprint arXiv:1601.07950*, 2016.

[31] K. Lee. Real-time head pose estimation built with opencv and dlib, 2017.

[32] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of ICCV*, volume 2, pages 1150–1157. IEEE, 1999.

[33] Y. Lu, M. Mahmoud, and P. Robinson. Estimating sheep pain level using facial action unit detection. In *Proc. of Automatic Face & Gesture Recognition*, pages 394–399. IEEE, 2017.

[34] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. G. Medioni. Do we really need to collect millions of faces for effective face recognition? *CoRR*, abs/1603.07057, 2016.

[35] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE trans. on pattern analysis and machine intelligence*, 31(4):607–626, 2009.

[36] N. National Coordination Office for Space-Based Positioning and Timing. GPS.gov: Agricultural applications, 2018.

[37] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE trans. on Affective Computing*, 2(2):92–105, April 2011.

[38] H. Ouanan, M. Ouanan, and B. Aksasse. Facial landmark localization: Past, present and future. In *Proc. of Int'l Colloquium of Information Science and Technology*, pages 487–493. IEEE, 2016.

[39] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 2017.

[40] M. Rashid, X. Gu, and Y. J. Lee. Interspecies knowledge transfer for facial keypoint detection. In *Proc. of CVPR*, volume 2, 2017.

[41] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[42] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proc. of CVPR*, pages 1685–1692, 2014.

[43] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. *CoRR*, abs/1710.00925, 2017.

[44] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proc. of CVPRW*, pages 896–903. IEEE, 2013.

[45] J. Siswantoro, A. S. Prabuwono, A. Abdullah, and I. Bahari. Hybrid neural network and linear model for natural produce recognition using computer vision. *ICT Research and Applications*, 11(2):184–198, 2017.

[46] B. M. Smith and L. Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In *Proc. of ECCV*, pages 78–93. Springer, 2014.

[47] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proc. of CVPR*, pages 3476–3483. IEEE, 2013.

[48] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proc. of CVPR*, pages 4177–4187, 2016.

[49] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proc. of CVPR*, pages 3659–3667, 2015.

[50] N. VG and K. Hareesh. Quality inspection and grading of agricultural and food products by computer vision-a review. *Int'l Journal of Computer Applications*, 2(1), 2010.

[51] Z. Wang and X. Yang. Joint face detection and initialization for face alignment. In *Int'l Conf. on Multimedia Modeling*, pages 164–175. Springer, 2017.

[52] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. of CVPR*, pages 532–539. IEEE, 2013.

[53] H. Yang, X. He, X. Jia, and I. Patras. Robust face alignment under occlusion via regional predictive power estimation. *IEEE trans. on Image Processing*, 24(8):2393–2403, 2015.

[54] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. *arXiv preprint arXiv:1507.03148*, 2015.

[55] H. Yang, R. Zhang, and P. Robinson. Human and sheep facial landmarks localisation by triplet interpolated features. In *Proc. of WACV*, pages 1–8. IEEE, 2016.

[56] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proc. of ICCV*, pages 1944–1951. IEEE, 2013.

[57] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *Proc. of CVPRW*, pages 2116–2125, 2017.

[58] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proc. of ECCV*, pages 94–108. Springer, 2014.

[59] G. Zheng, Y. Tan, J. Zhang, W. Li, et al. Automatic detecting and grading method of potatoes with computer vision. *trans. of the Chinese Society for Agricultural Machinery*, 40(4):166–156, 2009.

[60] H. Zhu, B. Sheng, Z. Shao, Y. Hao, X. Hou, and L. Ma. Better initialization for regression-based face alignment. *Computers & Graphics*, 2017.

[61] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proc. of CVPR*, pages 4998–5006, 2015.

[62] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proc. of CVPR*, pages 146–155, 2016.

[63] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. of CVPR*, pages 2879–2886. IEEE, 2012.