Pose-Informed Face Alignment for Extreme Head Pose Variations in Animals

Charlie Hewitt
Department of Computer Science and Technology
University of Cambridge
Cambridge, UK
ac@chewitt.me

Marwa Mahmoud
Department of Computer Science and Technology
University of Cambridge
Cambridge, UK
marwa.mahmoud@cl.cam.ac.uk

***Abstract*—Landmark localisation is a vital step in automatic analysis of facial expressions of animals. Head motion is one of the most challenging problems for face alignment for humans and animals. For animals this is exacerbated by the increased amounts of self-occlusion resulting from variations in head pose. In this paper, we present a novel model for detection of an extensive set of facial landmarks for sheep. A 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, including a wide range of variations in head-pose and occlusion. Data augmentation techniques are introduced using thin-plate-spline warping and negatively correlated augmentation to boost representation of extreme head poses. We then present a novel pose-informed landmark localisation method based on a fine-tuned CNN model for human head pose estimation. This method is shown to significantly outperform the existing state-of-the-art approach on the introduced SFLW dataset and the viability of the technique for real-world use is demonstrated through the implementation of a near real-time video pipeline.**

***Index Terms*—animal face alignment, facial landmark localisation, head pose estimation, ovine affect**

I. Introduction

Automatic analysis of facial expressions of animals can assist in early diagnosis of many diseases and help to improve animal welfare. Automated systems can be used to detect medical issues that require further investigation as early as possible, rather than relying on infrequent veterinary evaluations of animals. Previous work in this area has shown that distress in sheep can be reliably predicted from a number of facial action units [19]. 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 the changes in these facial features.

Facial landmark localisation is a well-explored problem in humans [22], but existing work tackling this problem for sheep [26] and horses [24] only considers very sparse landmarks and relatively minimal head pose variation, as well as achieving less than ideal accuracy. For human face alignment, landmark localisation has progressed mostly because of the availability of large annotated datasets that have allowed for better training of models. As a new field, animal facial expression analysis is yet to benefit from such large and varied datasets. As such, pose variation is a much more significant problem in animal face alignment due to the limited amount of data.

In this work, we present a model for improving 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. We demonstrate that this, given the relatively small dataset, improves face alignment results significantly. We also detect a denser set of landmarks (25 landmarks compared to 8 in [26]) which allows for more accurate action unit extraction and subsequent facial expression analysis and emotion detection. We also consider increased head pose variation in training data to improve the resilience of the proposed model when deployed in-the-wild.

The main contributions of this paper are as follows:

- Presenting the SFLW dataset and annotation and augmentation techniques used to boost the dataset variation and improve representation of less common and more challenging pose variations. This is the first sheep dataset that includes annotation for 25 facial landmarks.
- Implementing a novel pose-informed landmark localisation approach for 25 landmarks on the sheep face.The proposed approach is demonstrated to outperform state-of-the-art methods when applied to sheep.
- Describing an in-depth experimental evaluation of the proposed pose estimation and landmark localisation approaches, showing that our approach outperforms the current state-of-the-art on the SFLW dataset and can be deployed in a near-real-time pipeline.

II. Related Work

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. 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 [7]–[10], [18]. Later, Regression methods were introduced. Cascaded Pose Regression [11] (CPR) was the first of these methods, in which a cascade of regressors is trained to iteratively refine an estimate of landmark locations starting from a rough initial guess. Robust CPR [5] (RCPR) is a variation that can also regress an occlusion value (binarised from a continuous prediction between 0 and 1) and uses

978-1-7281-3888-6/19/$31.00

multiple initialisations and alternative feature extraction to improve performance. Explicit Shape Regression [6] (ESR) is another technique very similar to CPR aimed at high efficiency.

The current widely accepted state-of-the-art in terms of classical facial alignment is a technique using an Ensemble of Regression Trees [14] (ERT). This approach is 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. Modern deep learning methods [4], [23] achieve extremely impressive performance for human face alignment, but are not directly applicable in our case given the limited availability of data.

There has so far been little work looking at facial landmark localisation for animals. Eight facial landmarks are detected from sheep faces in [26] using a modified version of ERT with triplet interpolated feature (TIF) extraction. Pain recognition from sheep faces is analysed in [19] based on the same landmark localisation method.

Transfer learning is explored in [24], which focuses primarily on horses, but also utilises the dataset of [26]. 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 method 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 [3] with application to cat and dog faces, as well as humans with limited comparison to any baselines to verify the effectiveness of the approach.

Head pose estimation is a significantly less well-explored problem in computer vision than landmark localisation. A number of classical techniques for human head pose estimation have been proposed [17], [20]. Recently, deep learning has been applied to both landmark localisation and head pose estimation for humans, often combined into a single network [23]. Specific networks have also been designed to determine only the head pose of humans from images [25], aiming to be more efficient than the often very large, multi-function networks and with impressive results. Despite this, it is most common in practice to estimate head pose indirectly based on the results of classical facial landmark localisation techniques.

Because sheep landmark localisation methods are not as mature and due to data scarcity, we consider the inverse procedure. Rather than calculating head pose from localised landmarks, we 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.

III. The Sheep Facial Landmarks in the Wild (SFLW) Dataset

One of the main issues faced in facial landmark localisation for animals is that of data sparsity. The sheep dataset used in [26] includes 600 images annotated with only eight landmarks. Another animal facial landmark dataset is used in [24] for horses and it includes 3717 images annotated with just five landmarks. This is in stark contrast to humans facial datasets, such as Multi-PIE [13] (750,000 images with 68 landmarks), Menpo [27] (over 10,000 images with 39 landmarks) and AFLW [16] (25,993 images with 21 landmarks).

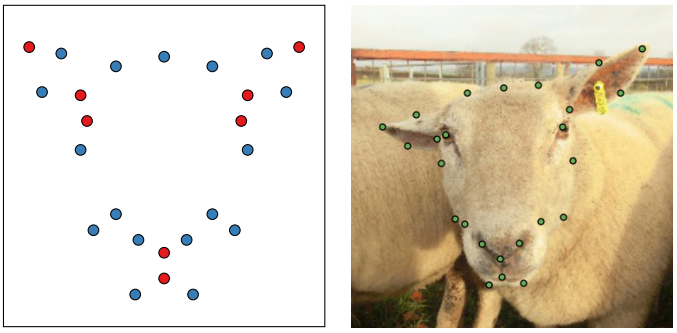

Fig. 1. Original (red) and new (blue) sheep facial landmark annotation schemes and ground-truth annotations for example image from the SFLW dataset.

The SFLW dataset is composed primarily of images used in [26], with an additional 250 photos collected from the internet (total of 850 images). An annotation scheme containing 25 landmarks is devised based on the original eight-point annotation used in [26]. The original and updated schemes are shown in Figure 1, along with an example annotated image showing which facial features the landmarks relate to. The new scheme is approximately based on human annotation schemes 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.

Annotation was performed in a semi-automated manner, with landmark position predicted by calculating the thin-plate-spline transformation [2] between the neutral eight-point positions and the existing annotations from [26] and applying the same transformation to the neutral 25-point positions. These predictions were then manually fine-tuned to ensure high accuracy. The proposed denser landmark scheme should enable more precise extraction of action units for use in affect determination algorithms, such as that introduced in [19].

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. We utilise two additional techniques: image warping using thin plate spline (TPS) transformations and negatively correlated augmentation (NCA).

In order to avoid repeating identical images when training localisation models, TPS warping [2] is used to generate slight variations on input image data by translating the annotated landmarks using hand-crafted rules and warping the associated image accordingly. These variations are visually subtle

but should allow for more general representations of image features to be learnt. TPS warping is able to simulate changes in ear position as well as providing low magnitude pose and face-shape variation.

In addition to this, negatively correlated augmentation (NCA), as introduced in [26], is used to deal with the imbalance in head poses present in the SFLW dataset. Many images were sourced from the internet, so include primarily frontal faces, but in real-world applications of this technology it is expected that a roughly uniform distribution of head poses will be observed. NCA boosts the augmentation factor for images with extreme head poses in order to reduce the imbalance within the dataset, and is parameterised to somewhat retain the underlying distribution.

In the experimental evaluation below, two variants of the base SFLW dataset are used. SFLW-flip contains horizontally mirrored versions of every image, and SFLW-NCA uses horizontal mirroring, TPS warping and rotation augmentations, with the number of times each original image is used weighted by NCA. The SFLW dataset (and augmented variants) are available upon request.

IV. Pose-Informed Facial Alignment

Landmark-free head pose estimation enables accurate yaw, pitch and roll to be determined from arbitrary images of sheep faces. This information can be used to improve the performance of facial landmark localisation by specialising a number of models to specific head poses. This is particularly beneficial for sheep in the case of variable yaw as, due to the elongated nose, self-occlusion completely alters the 2D projection of the sheep's face recorded by a camera. We therefore propose a pipeline where head pose is first determined from an image, and then face alignment is performed by a tailored model, enabling more accurate localisation performance.

A. Head Pose Estimation

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 [25] is selected due to its design focus specifically for head pose estimation. It also has the best performance of the networks trained in [25]. A Hopenet model pre-trained on the 300W-LP dataset [28] is used as the base model.

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 average head shape. A RANSAC [12] 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 edges of the ears are excluded from this correspondence as they typically move relative to the rest of the face. The intrinsic camera 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 sufficient for this application.

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 [25]). A similar training process to that in [25] is employed, using the Adam optimiser [15] 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. 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.

B. Landmark Localisation

The introduced Pose-Informed ERT (PI-ERT) method is based on the highest performing currently available method (ERT [14]), as explored in the following section. The training set is first partitioned based on the absolute yaw angle for that image, with right facing images mirrored horizontally to ensure all sheep face in the same direction. A separate ERT model is then trained for each of these partitions.

For an arbitrary test image, the head pose is first predicted and the image mirrored if the sheep is facing right. The appropriate ERT model is then selected based on the magnitude of the yaw angle and executed to localise the facial landmarks. The image (and landmarks) are then mirrored back if required.

There is some trade-off in choosing the number of partitions; with fewer partitions the pose estimation task can be easier and each partition is likely to have enough training data. With a greater number of partitions, each fitter should be able to achieve higher performance as it is more closely tailored to a specific range of angles, but this relies on highly accurate head pose predictions and each partition may lack training data. For the SFLW dataset three was found to be the most effective number of partitions.

V. Experimental Evaluation

To evaluate our techniques we use five-fold cross-validation for all experiments and report the mean. When augmented datasets are used, the augmented training fold is selected such that it contains variants of training images only. The resulting models are then evaluated on the unaugmented test fold with identical folds used for head pose estimation and landmark localisation. That is to say that no test image (or a variation of) has been used in the training phase of any model within the pipeline.

A. Head Pose Estimation

To evaluate the sheep head pose estimation model a number of metrics are employed. Mean absolute error (MAE) is typically used [25] but it is not sufficient on its own. 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 also

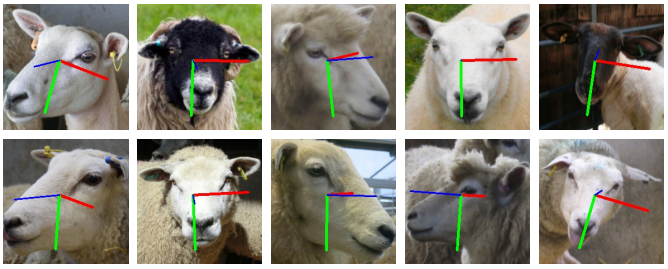

Fig. 2. Qualitative head pose estimation results for network trained on SFLW-NCA and tested on SFLW. Head pose is visualised as a 3D axis at the centre of the image.

TABLE I
HEAD POSE ESTIMATION RESULTS FOR NETWORK TRAINED ON SFLW-NCA AND TESTED ON SFLW COMPARED WITH BASELINE.

	Dataset Mean				Fine-tuned Model			
	Yaw	Pitch	Roll	Ave	Yaw	Pitch	Roll	Ave
MAE	15.73	11.88	8.36	11.99	6.04	7.58	6.13	6.58
PCC	0.00	0.00	0.00	0.00	0.91	0.56	0.40	0.75
SAGR	0.50	0.57	0.50	0.52	0.78	0.83	0.77	0.80

used to assess the correlation of predictions with the ground truth, arguably a better measure of a model's usefulness. In addition, Sign Agreement metric (SAGR) [21] 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.

Qualitative results are shown in Figure 2 and quantitative results given in Table I. The overall performance is very good considering the relatively small dataset and clear visual difference between human and sheep faces. An average MAE of under 7 degrees is achieved, with PCC as high as 0.9 for a single angle (yaw) and 0.75 overall; qualitative results for extreme head poses are also accurate. Results are reported for SFLW-NCA as this dataset achieved the best performance.

B. Landmark Localisation

For all our experiments on facial landmark localisation, three evaluation metrics are used: 1) 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, 2) Success rate: the proportion of the dataset with an MNE under 10%, and 3) AUC: the normalised area under the cumulative MNE distribution. Clearly, for MNE lower is better and for success rate and AUC higher is better.

Existing implementations of state-of-the-art classical facial alignment methods—ESR [6], (R)CPR [5], [11], ERT [14] and TIF [26]—are modified to incorporate the 25 landmark annotation scheme used for the SFLW dataset. The modified implementations are then trained and tested on our dataset and predicted landmarks are exported to a common format. These predictions are then loaded within the Menpo framework [1] allowing for an identical evaluation procedure. The mean shape of the dataset is projected into the facial bounding box with no fitting to serve as a baseline.

TABLE II
QUANTITATIVE PERFORMANCE METRICS FOR EXISTING LANDMARK LOCALISATION METHODS TRAINED ON SFLW-FLIP AND TESTED ON SFLW.

	Mean Shape	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**

TABLE III
BASELINE, ERT AND PI-ERT LOCALISATION PERFORMANCE FOR MODELS TRAINED ON SFLW AND SFLW-NCA, BOTH TESTED ON SFLW.

		No Augmentation		Full Augmentation	
	Mean Shape	ERT	PI-ERT	ERT	PI-ERT
MNE	0.139	0.062	0.053	0.050	**0.045**
Success Rate	0.46	0.83	0.89	0.90	**0.94**
AUC	0.858	0.933	0.942	0.942	**0.949**

Results for localisation performance of these existing method for the SFLW dataset are given in Table II. ERT [14] achieves the highest results in all metrics and consequently forms the basis of our method. Interestingly, the TIF method which was developed previously specifically for sheep [26] does not perform as well, this is likely due to the higher density of landmarks (25 vs 8) used in the SFLW dataset.

Quantitative results for the PI-ERT method are compared with the baseline and current state-of-the-art (ERT) in Table III. The proposed method improves notably over the state-of-the-art both with and without data augmentation, with an increase in success rate of 4 and 6% respectively and significant reductions in both MNE and AUC. Qualitative results shown in Figure 3 support this, showing clear improvements for the PI-ERT method, with landmark locations approaching the ground-truth positions. The data augmentation techniques described above also provide notable improvements in performance for both ERT and PI-ERT.

Figure 4 shows the distribution of MNE for the ERT and PI-ERT methods. PI-ERT reduces MNE across almost the entire range of yaw angles, but notably results in a flatter distribution of error demonstrating that extreme head poses are dealt with significantly better by this pose-informed approach.

C. Application

As a proof of concept, the trained PI-ERT model was deployed in a complete pipeline to process videos of sheep. As a pre-processing step, a simple HOG-based face detection model [10] is used to detect the sheep face. For successfully detected faces, head pose estimation took on average 13ms per frame, and facial alignment 16ms for a 700×400 pixel image on a consumer laptop[1] with external NVIDIA 1080Ti

[1]MacBook Pro 2.8 GHz Intel Core i7 16 GB 1600 MHz DDR3

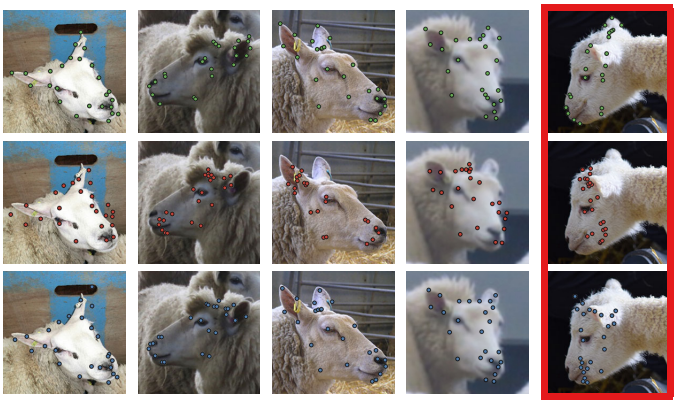

Fig. 3. Qualitative examples of landmarks localisation improvements made by PI-ERT. The right-most column shows an example where PI-ERT struggles with some landmarks due to inaccurate head pose estimation. Rows (from top to bottom): ground-truth, standard ERT and PI-ERT.

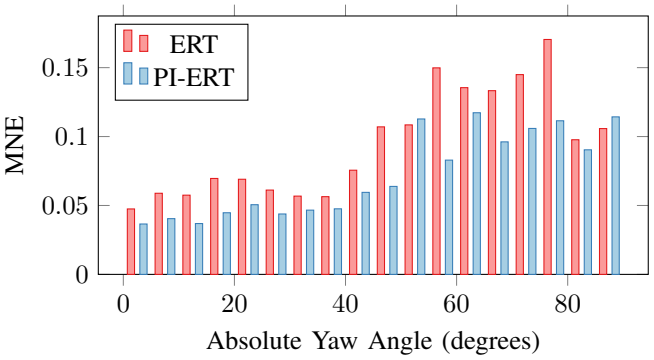

Fig. 4. MNE distribution for optimal PI-ERT and ERT fitters.

GPU. Face detection was a bottleneck in this pipeline, with an execution time >50ms per frame. This could be improved significantly using alternative face detection techniques and a larger dataset of images annotated with facial bounding boxes. Developing better sheep face detection models is one of our directions for future work.

Despite the low performance of the face detection step, the full pipeline ran at near real-time (∼15fps) for prerecorded videos on this modest hardware. This is suitable for most surveillance and monitoring applications, and real-time performance could easily be achieved on currently available commercial hardware. Figure 5 shows results for a sample frame extracted one of our test videos. Qualitative results showed the resilience of the method to motion blur, which is likely to occur in real-world videos.

VI. Conclusions and Future Work

This paper has introduced the Sheep Facial Landmarks in the Wild (SFLW) dataset, containing 850 in-the-wild images of sheep faces annotated with 25 landmarks and binary occlusion information. Driven by the small volume of data available, image warping and negatively correlated augmentation techniques were used to augment the dataset, with evaluation

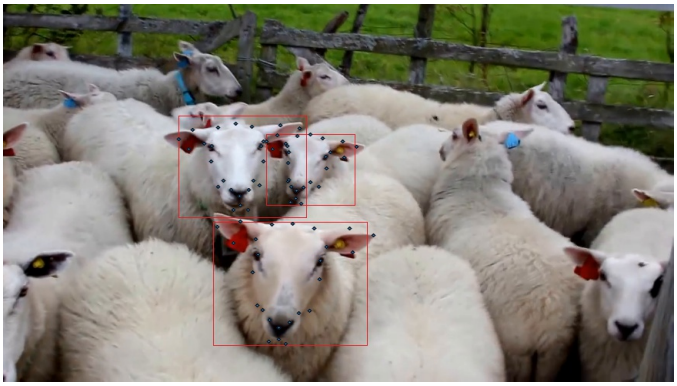

Fig. 5. Sample frame from example video run through landmark localisation pipeline showing the robustness of the approach for real-world data. Although detection performance is limited, landmark localisation results are accurate for successfully detected faces.

showing a significant improvement in landmark localisation performance as a result.

A novel pose-informed landmark localisation method (PI-ERT) based on a fine-tuned CNN head pose estimation network pre-trained for human head pose estimation was also introduced. The proposed approach achieved an average head pose estimation error of under 7 degrees and improved landmark localisation performance significantly over the current state-of-the-art, especially for sheep faces exhibiting extreme head pose variations. In addition, the PI-ERT method was demonstrated in conjunction with sheep face detection in near real-time as a proof-of-concept for surveillance applications on video feeds.

Analysis of per-landmark localisation error showed ears to have generally the lowest localisation results. Thus, future work might focus on improving the ability of models to deal with these landmarks which vary significantly in position relative to the rest of the face and which are critically important given the application domain. Face detection was also a limiting factor in deployment of the pipeline for real-world use, improved sheep face detection would, therefore, also be a useful avenue of investigation. Since this is the first model to detect an extensive set of landmarks (25) on the sheep face, ultimately we would like to use it to enhance facial action unit detection and pain detection models, extending on the work in [19].

References

[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] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(6):567–585, 1989.

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

[4] 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.

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

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

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Comput. Vision Image Understanding*, 61(1):38–59, 1995.

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

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

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

[12] 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 CV*, pages 726–740. Elsevier, 1987.

[13] 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.

[14] 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.

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

[16] 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.

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

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

[19] 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.

[20] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, 2009.

[21] 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 Comp*, 2(2):92–105, April 2011.

[22] H. Ouanan, M. Ouanan, and B. Aksasse. Facial landmark localization: Past, present and future. In *Proc. of Int'l Coll. of Info. Science and Tech.*, pages 487–493. IEEE, 2016.

[23] 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. Pattern Anal. Mach. Intell.*, 2017.

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

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

[26] 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.

[27] 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.

[28] 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.