Automatic Pain Detection on Horse and Donkey Faces

Hilde I. Hummel[1], Francisca Pessanha[1,2], Albert Ali Salah[1], Thijs J. P.A.M. van Loon[3] and Remco C. Veltkamp[1]

[1]Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

[2] Faculty of Engineering, Porto University, Porto, Portugal

[3] Department of Clinical Sciences, Utrecht University, Utrecht, The Netherlands

***Abstract*— Recognition of pain in equines (such as horses and donkeys) is essential for their welfare. However, this assessment depends solely on the ability of the observer to locate visible signs of pain since there is no verbal communication. The use of Grimace scales is proven to be efficient in detecting pain but is time-consuming and also dependent on the level of training of the annotators and, therefore, validity is not easily ensured. There is a need for automation of this process to help training. This work provides a system for pain prediction in horses, based on Grimace scales. The pipeline automatically finds landmarks on horse faces before classification. Our experiments show that using different classifiers for different poses of the horse is necessary, and fusion of different features improves results. We furthermore investigate the transfer of horse-based models for donkeys and illustrate the loss of accuracy in automatic landmark detection and subsequent pain prediction.**

I. Introduction

The recognition and quantification of pain in equines is essential to maintain their welfare and improve their convalescence [11]. However, contrary to humans, where pain assessment is facilitated through verbal communication, in animals, this process depends on the observer's ability to locate and quantify the pain, based on perceptible behaviour and physiological patterns.

Several studies have found a correlation between pain and behaviour changes in equines, such as aggressiveness, reluctance to move, vocalisation and diminished socialisation [2]. However, to study more subtle changes, it is useful to analyse the facial expressions of these animals [10]. This method has been extensively used in other species, such as mice [18], rabbits [17] and sheep [25] with promising results. Several frameworks have been proposed for horse pain estimation, the most important being the Horse Grimace Scale (HGS) [7] and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) [33], [34].

Although the use of grimace scales to assess pain is proven to be efficient, it requires the training of observers and the manual assessment of the pain score for each facial action unit (AU). There is a clear necessity for automation. Recent progress on action unit based estimation of sheep pain [24], using Sheep Pain Facial Expression Scale (SPFES) [25], illustrates the potential of this method. The foremost application is the development of training programs for recognizing pain in equines.

In this paper, we propose a hierarchical model for pain estimation in equines, with a focus on horses. The system is composed of a pose estimation step, followed by pose-informed landmark location, with further feature extraction of the regions of interest (ROI) and pose-specific pain score estimation. The last step integrates different appearance features to produce a more robust score. Additionally, preliminary experiments were conducted with donkey faces, evaluating the possibility of taking advantage of the higher number of horse faces available to produce a robust model for other equines.

The main contributions of this work are as follows:

- We present a horse and donkey dataset with manually annotated landmarks and feature-level, detailed pain score ground truth, given by a veterinarian expert.
- We implement a method for accurate head pose detection and automatic landmark detection, detecting either 44, 45, or 54 landmarks, depending on the head pose.
- We implement a hierarchical system for pose-specific automatic pain prediction on horse faces, and explore its extension for donkey faces.

II. Related Work

Objective analysis of the human face is frequently achieved using a Facial Action Coding System (FACS) approach [8]. In this system, action units (AUs) are defined based on the underlying facial muscles and can be used to evaluate changes in an expression, for instance, associated with pain [3], [22], [23]. The most accurate AU detection use video as input, as facial movements can be very subtle, and leveraging spatio-temporal cues seems essential [30]. Automatic pain detection in humans, using facial expressions, is based on automatic AU detection [19], [37], [22].

Using a similar approach, the EquiFACS system proposes the systematic characterization of horse facial movements [35]. These were identified based on the underlying muscles of the horse, and through behavioural studies.

The systematization of changes in AUs for pain estimation resulted in Grimace scales for multiple animals, including equines. However, considering the time-consuming aspect of manual pain assessment the idea of automating their application gained more popularity and explored in the literature for sheep [21] and mice [1], [32]. In mice, transfer learning was used to optimize an InceptionV3 deep neural network model, trained on ImageNet, for the binary pain vs no-pain classification in laboratory mice [32]. In sheep, a hierarchical system to estimate the pain levels based on

978-1-7281-3079-8/20/$31.00 ©2020 IEEE
DOI 10.1109/FG47880.2020.00114

the face was proposed by Mahmoud *et al.* [24]. Eight facial landmarks were detected for specifying regions of interest, using a modified version of Ensemble of Regression Trees (ERT) with triplet interpolated feature (TIF) extraction followed by a Histogram of Oriented Gradients (HOGs) based classifier [21]. Further work on the estimation of facial landmarks was developed [14] adding a pose estimation step to the pipeline and improving the detection for extreme poses. A complete pipeline was more recently proposed [27], combining a fine-tuned SSD-Mobilenet model for face detection, with a CNN-based pose estimation and pose-informed landmark location method. HOGs features, as well as geometric features and the quantitative pose values, were used to train a binary Support Vector Machine (SVM) classifier, adapted to different head rotations and consequent self-occlusion.

Previous work in horses used a combination of thinning, color histograms and HOG features to predict the pain level of the horse [15]. This work used a smaller dataset than ours but a similar set of landmarks and pain features. However, the extracted features were not sufficiently discriminative, and SVM-based classification did not produce a high accuracy.

There are several important challenges in horse pain estimation. There is much less data available for studying animal faces, compared to the vast amounts of data on human faces. Furthermore, the appearance of a horse's head is affected much more (compared to humans) by the variance in color of horse coats, potentially including patches of different colors, such as a white blaze often found on the head. Also, the bounding box of a horse's head is typically not square. Most human-face analysis approaches assume a scaled image with equal width and height, this is not possible with horse faces unless we distort the faces severely for larger pose angles. Another important challenge is that there is no verbal ground truth for the pain level of the horse, whereas humans can rate their own pain, albeit subjectively.

III. Data and Annotations

The dataset used in this study consists of 1854 images of horse heads and 531 images of donkey heads. A veterinarian specialist scored the pain level of the horses and donkeys based on six pain features. These pain features are derived from both the HGS [7] and the EQUUS-FAP [33], scoring the stand of the ears, the visibility of the sclera, the angulation of the upper eyelid, tightening of the orbital, the widening of the nostrils, and the corners of the mouth. A score is given separately for each feature, in a 0-2 scale: "0" (no pain), "1" (mild pain) or "2" (severe pain). The label distribution is imbalanced, most pain indicators are on nostrils, eyelid, and mouth (Fig. 1).

The horse images we use in this study come from three different sources. The first one was gathered from a clinical study, where the pain was clinically induced and the images were taken over time [15]. The second subset was gathered from a home housing older horses. Within each of these subsets, the recording conditions are similar. The third subset consists of images provided by horse owners. This subset

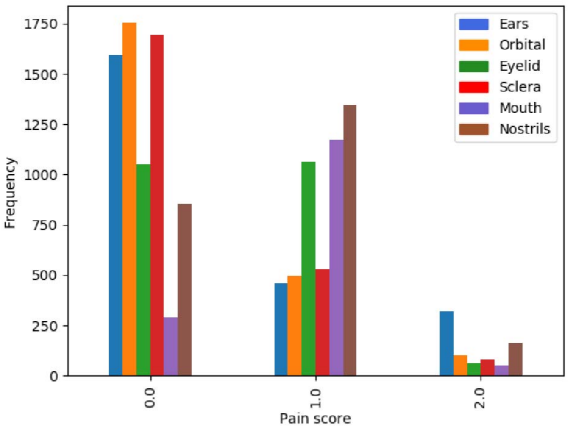

Fig. 1. The distribution of the pain scores per relevant feature

(a) (b) (c)

Fig. 2. Manually annotated landmarks of (a) a frontal head pose containing 54 landmarks, (b) a side head pose containing 45 landmarks and (c) a tilted head pose containing 44 landmarks. The red dots indicate the landmarks of interest, while the magenta dots indicate the contour landmarks. The image width is fixed, but the width-height ratio is not changed to illustrate the size differences. Best viewed in color.

is more diverse compared to the first and second subsets, containing a variety of horses in different settings. Additionally, a set of donkey images were provided by a donkey sanctuary. The characteristics of the images are similar and multiple images are present per donkey. Part of the dataset and annotations can be obtained by request from the authors.

To provide a solid ground truth, the landmarks of every image were manually annotated. These landmarks depend on the head pose, which we grouped into three: a frontal pose (54 landmarks), a profile pose (45 landmarks), and a tilted pose (44 landmarks), respectively. The landmarks of interest describe the ear(s), the eye(s), the nostril(s) and the mouth of the face. Additional landmarks describe the contour of the face.

IV. Methodology

The complete processing pipeline is visualized in Fig. 3. We describe each stage separately. The horse dataset was split into a training set (1496 images) and a test set (357),

maintaining the head pose distribution of full dataset in each subset. One horse may appear in multiple images, so we have ensured that none of the horses are present in both the training and the test set. The hyperparameters of all the models are optimized using a 5-fold cross-validation within the training set, and the test set is kept separate.

A. Pose Estimation

Head pose variations cause evident changes in the facial appearance in equines due to self-occlusion. For this reason, using the pose information to define the areas of interest visible from a specific pose would be pertinent for both landmark detection and further pain estimation.

We implemented a 5-class pose classifier to distinguish tilted and frontal faces, as well as the sign of their rotation. Based on results from the literature, we decided to use Histogram of Oriented Gradients (HOG) features [6] with a support vector classifier (SVM) for this purpose. All the horse faces were cropped according to the location of the manually given landmarks, with an added margin considering the original image size. Additionally, both tilted and profile faces were augmented by flipping the image vertically, reducing the data imbalance between different poses. For the individual performance metrics, the class of interest was labeled as "positive" and all the others as "negative". We have calculated Precision, Recall, and the F1-Score based on this labeling. A higher weighted F1-score was achieved with 9 orientations, 8×8 pixels per cell and 4×4 cells per block, associated with a weighted SVM classifier with a linear kernel.

B. Landmarking

Precise landmark detection is essential for further extraction of the areas of interest and accurate pain prediction. Although little work has been done for landmark detection in animals, several methods have emerged for transfer learning from landmarking models for human faces. Hewitt *et al.* proposed a pose-informed Ensemble of Regression Trees (ERT) approach [14] and Rashid *et al.* [28] explored the fine-tuning of a network implemented for human faces, by correcting the differences in shapes of equines and human faces.

In this work, we compare the performance of two state-of-art landmark detection algorithms originally developed for human faces, namely ERT [16] and Supervised Descent Model (SDM) based on SWIFT features [36], as well as a mean shape model for baseline.

The horse faces were cropped based on the landmark location and resized according to the face proportions in each of the pose bin. This reflected in an image of 265×500 for frontal faces, 500×447 for tilted faces and 500×363 for profile faces. Further, all faces were rotated according to their absolute pose.

The amount of augmentation to apply for each pose bin was evaluated, and a performance plateau was observed in the ERT model when increasing the number of perturbations above 30, for frontal and profile faces, and above 10, for tilted faces. This observation is coherent with the number of images per bin. Additionally, a mean shape model was calculated for each pose bin based on the training set landmarks. The results presented later refer to the performance in the un-augmented test set, with a total of 90 frontal faces, 177 tilted faces, and 90 profile faces.

C. Facial alignment and augmentation

We have a small number of annotated training faces for training our models. To create a robust pain classification model, the training set needs to be augmented and properly aligned.

Images can be aligned using either a rigid alignment or a non-rigid alignment. The non-rigid alignment allows deformation of the image, while the rigid alignment only allows rotation, scaling, and translation [9]. Since we want to preserve local detail for pain estimation, we use a rigid alignment approach, namely, Procrustes Analysis (PA). PA states that the shape of I: $l \times p$ is the same as the shape I': $l \times p$ after the following transformation:

$$I' = I\Gamma + 1_N \gamma^T. \tag{1}$$

Where Γ is the rotation matrix, 1_N is an N vector of ones and γ is the translation matrix of the size $p \times 1$ [12]. A mean shape per head pose was generated by using the Generalized Procrustes Analysis (GPA) [13]. This algorithm creates a consensus shape from the input set of shapes by optimizing the residual-sum-of-squares error after alignment. During aligning each face to the mean shape, the faces are scaled and reflected if necessary.

We augment the training set by injecting some noise to the landmarks before the alignment, which results in different aligned images per original image. The random noise was determined by adding maximally 2% and maximally 6% of the eye-nostril distance to the landmark coordinates, separately (Fig. 4). The noise added could either be positive or negative. If the noisy coordinate exceeds the image boundary, the original coordinate was chosen.

D. Pain Estimation

The difference in pain scores is caused by variations in the AU appearance. For this reason, each face was cropped based on several Regions of Interest (ROIs); the eyes, ears, nostrils, and the mouth, respectively. Previous work on estimating pain in sheep extracted HOG features from the ROIs and used an SVM for pain classification [21].

In this paper, we experimented with HOG, Local Binary Pattern (LBP) [26], Scale Invariant Feature Transform (SIFT) [20], as well as with features created by a VGG16 deep neural network model [31] with weights trained on the ImageNet database.

Before any features were extracted, the background of the image was subtracted and the faces were aligned to a consensus shape. The consensus shapes have a bounding box dimension of 1358×2424 for the tilted head pose, 243×542 for the frontal head pose and 363×403 for the profile head pose. Next, the ROIs were determined by adding a bounding

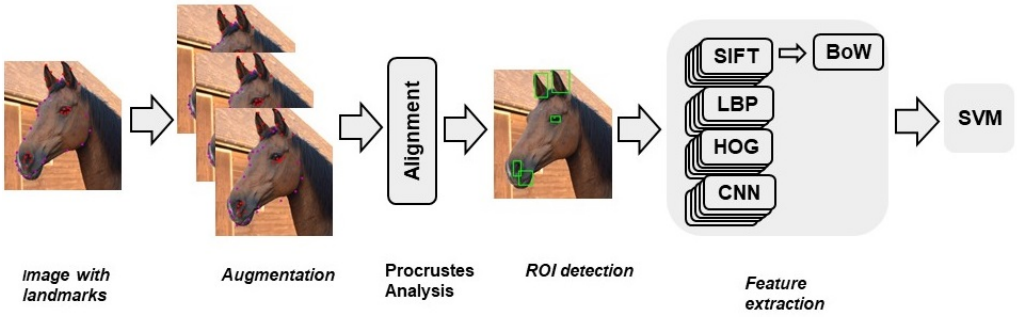

Fig. 3. The training pipeline for the pain classification approach. The output is a set of SVM models, which are used to predict pain scores per feature.

box around the landmarks of interest which extends the box described by the landmarks by 2% of the normalization distance. The number of ROIs differed per head pose; the profile head pose has four ROIs, the frontal head pose has six ROIs and the tilted head pose has five.

SIFT keypoints are extracted from each ROI. A codebook was prepared by an 3000-means clustering using all the keypoints in the training set. This codebook is then used to create a 3000-bin histogram to serve as a Bag of Words for the image, using all the keypoints of all the ROIs in that specific image. The histogram was then normalized using L2-normalization.

To extract the HOG features, first, the ROIs were resized to 64×128, then the HOG features were extracted using 8×8 pixels per cell and nine orientations. The histograms were normalized using four cells and L2-normalization. LBP features were extracted using a cell radius of eight pixels and the number of points was set to 24. Finally, for features generated from VGG16, the ROI image was resized to 244×244 and directly used as input for the VGG16 model. The features were extracted by removing the softmax layer of the model, so the output of the VGG16 is a vector of the length of 4096.

First, the dataset was divided per head pose. Next, a single SVM regressor was optimized and trained per pain scoring metric, seperately. To classify the pain, all the ROIs extracted were included for each pain prediction. Sample weights were added, using higher weights for the class '2', since the cost of not correctly classifying a horse in pain is higher than the cost of not correctly classifying a horse with no pain. We use rounding to the nearest value to convert the regression prediction to a classification. We experimented with linear kernels, histogram intersection kernels [4] and the generalized histogram intersection kernels [5] for each pain model, separately. The models are trained and tested using solely the manual ground truth landmarks.

E. Classifier Fusion

Combining models trained with different types of features could improve the pain detection. For this purpose, simple fusion and weighted fusion approaches are implemented. The results generated by feature-specific predictors (using SIFT, HOG, LBP, and VGG16, respectively) are fused using both methods. The simple fusion extracts the minimum predicted value or the mean of all the predicted values. The weighted fusion assigns weights to each score separately as described in Eq.2, where W_x is the weight and S_x is the predicted score of a model.

$$S_{fused} = W_1 * S_1 + W_2 * S_2 \quad (2)$$

The weights for fusion are optimized using a 5-fold cross-validation.

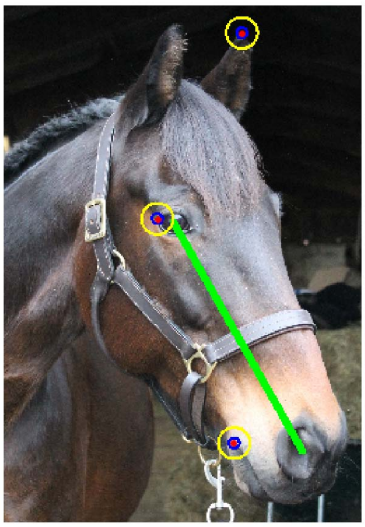

Fig. 4. Illustration of the eye-nostril distance used for normalization (green line). Limits of the noise surrounding a ground truth landmark (red dot) for augmentation are given in blue (minor - 2%) and in yellow (severe - 6%) to illustrate the perturbation in the augmentation. Best viewed in color.

F. Donkey classification

For donkey classification, both donkey and horse images are grouped based on their head poses and were aligned to a single mean shape per head pose before pain estimation. All the horse images were used in the training set, 80% of the donkey images were used as the validation set, while

the remaining donkey images were used as a test set. An SVM regressor was trained on the training set and the hyper-parameters of the model were optimized using the validation set. The final model was trained on the combination of the validation set and the training set and predicted the test set containing exclusively donkey images. Due to a lack of data, the frontal head pose of the donkeys was excluded from the predictions.

G. Performance Measures

The performance of different stages of our model are measured separately. For landmark localization on human faces, the inter-ocular distance is typically used as a normalization factor, and landmarks automatically located within 10 per cent of inter-ocular distance are considered to be accurate [30], [22]. In horses, there is no accepted normalization factor for facial landmark analysis. In this paper, we propose to use the distance between the center of an eye and the center of the underlying nostril for this purpose (see Fig. 4). These two features are present in every pose of the horse face, as opposed to two eyes required for the inter-ocular distance, thus the distance can be measured independent of the head pose position. Furthermore, the eye-nostril distance is sufficiently long to make it robust against errors. However, some poses of the horse (especially, changing pitch) will distort this distance. The error of the automatic landmarking is calculated by computing the Euclidean distance of the prediction to the ground truth landmark. The opposed normalization distance is used to normalize the error.

For the assessment of pain scores, we use the F1-scoring metric. The F-scoring equation is described in Eq. 3:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}, (0 \leq \beta \leq \infty), \tag{3}$$

where P is the precision, R is the recall and β is a parameter, set to 1.

V. Experimental Results

A. Pose Estimation

The implemented pose classifier proved to be efficient in distinguishing the five pose classes, especially when one considers the wide variety of head poses represented in our dataset. Table I summarizes the results in terms of precision, recall, and F1-Score. The errors observed were often related to ambiguous head poses or extreme angles.

B. Landmarking

The Ensemble of Regression Trees model has shown promising results for the landmark location, outperforming both the Mean Shape model and the Supervised Descent Model (Table II). As expected, considering the qualitative head pose is unknown and the quantitative pose is only based in the yaw angle, the variations in the normalization distance with the pitch angle introduced variation in the final mean error. Overall, extreme angles, combined with a lack of representation of these angles in the dataset, are influential in the incorrectly located landmarks. Additionally, not all outline landmarks are associated with strong changes in appearance, which leads to deviations on their prediction, visible in Figure 5.

TABLE I

Pose estimation evaluation on the test set containing 357 images of horse faces. The performance is presented separately for each pose bin and the weighted average of each performance measure is given with weights according to the pose distributions in the test set.

	Precision	***Recall***	***F1-Score***
Profile (-)	0.96	0.88	0.92
Tilted (-)	0.82	0.95	0.88
Frontal	0.94	0.84	0.89
Tilted (+)	0.88	0.92	0.90
Profile (+)	0.89	0.82	0.85
Weighted Average	**0.90**	**0.89**	**0.89**

Analyzing the individual results for each region of interest (ROI) (Table III) in combination with the examples presented in Figure 5, additional sources of error are observed. Ear position will vary greatly from horse to horse and with the pain level, so, although they commonly look straight and facing forwards, there are several examples were this ROI is rotated, in a low position near the head. These variations will make the precise location of all the landmarks in the ears more challenging. Consequently, a higher error in comparison with the other ROIs was expected for ears, in particular in the profile pose. Regarding the nostrils, the subtle outer contour, when compared to the nostrils' dark interior, leads to landmark deviations, visible in the second column of Figure 5. Similar to what was described in the nostrils, eye landmark detection is affected by the clear contrast between the outer eye and the pupil when compared with the outer eye landmarks. Lastly, the lack of a clear definition in the mouth landmarks results in deviations in the landmark location, similar to the outline landmarks.

Due to the difference between the normalization factor used and the ones applied in previous work, it's not possible to do a direct comparison with the literature. To evaluate the performance of our model in comparison to previous work, the errors were normalized by the bounding box edge length, similar to what was done in [14]. The average normalized error achieved for the ERT model was 0.05, comparable to the ones presented in related work.

TABLE II

Weighted average results for the test set, with three pose-informed landmark location models, based on the pose annotation of each image. The success rate indicates the ratio of landmarks with a location error less than 0.06 of eye-nostril distance.

	ERT	***SDM***	***Mean Shape***
Mean Error	**0.09**	0.12	0.15
Success Rate	**0.51**	0.32	0.20

TABLE III

EVALUATION OF THE MEAN LOCATION ERROR PER AREA OF INTEREST IN THE TEST SET. THE HIGHEST ERROR FOR EACH REGION OF INTEREST IS HIGHLIGHTED.

	Frontal	*Tilted*	*Profile*
Ear(s)	0.06	0.10	**0.12**
Nostrils	0.09	**0.11**	0.08
Left eye	0.06	0.07	**0.09**
Right eye	0.06	-	-
Mouth	-	0.08	**0.08**

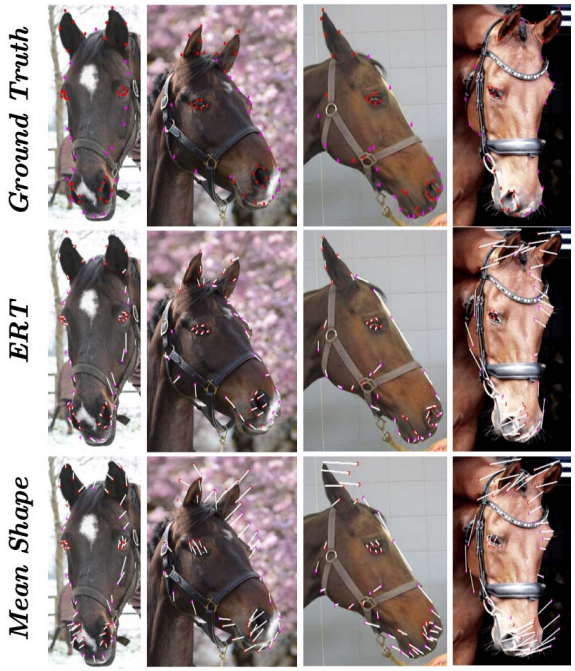

Fig. 5. Examples of the landmarks located in the test set by the ERT model compared to the baseline mean shape model. The three pose bins are represented from left to right: frontal, tilted and profile, respectively. The last column represents a image with a big location error for the landmarks due to an extreme pitch angle. The white lines connect the predicted point with the ground truth landmark location.

C. Classification Experiments

Table IV shows the classification rates for the 3-class problem in terms of micro F1 scores, and across features. Standard implementations were used per SIFT, HOG and LBP.

Overall, the classifier shows high micro F1-scores. LBP and SIFT features show overall low results. Since the SIFT keypoints are automatically extracted over the entire face (using the SIFT descriptor), the features are probably not able to distinguish the small changes of the Grimace scale. This could be solved by extracting SIFT features using the ground truth landmarks as keypoints. Another issue is that the ROIs were not resized separately before any SIFT or LBP features were extracted. Even though the full faces are scaled to the same size (depending on the pose) after alignment, this could have affected the ability of the feature extraction to define scale independent distinguishable features. Finally, lowering the overall resolution of the ROIs before LBP extraction may improve the results as well. This is a trick used in robust landmarking to fit more discriminating features into a small feature window [29].

Combining all the head poses into a single model leads to a model that is not able to learn to distinguish between pain and no pain in horses (results not shown). The appearance differences are too great in this case, and the increased complexity can only be countered by greatly increasing the training set size.

D. Fusion Experiments

Fusion of the model's predictions increases the F1-scores slightly (Table V). The improvements are mainly seen in the tilted and frontal head poses. The largest increase is seen in the prediction of the sclera in the tilted head pose. More elaborate fusion schemes should be explored in future work.

E. What about Donkeys?

The present dataset has around three times more horse faces than donkey faces, which makes the idea of applying a horse-based model in donkeys very appealing, especially considering the similarities between their Action Units and the common grimace scale (EQUIFAP), described for equines. To evaluate the potential of extrapolating a horse-based model to donkey images, some preliminary tests were made by applying the trained models on the donkey dataset.

The steep drop in results for both pose detection (Table VI) and landmarking (Table III), reveals clear differences in the face proportions, caused by the distinct ear sizes and nose length. To improve the performance of the model for donkeys, it is possible to "donkify" the horse faces by defining a Thin Plate Spline-based deformation field from horse to donkey, based on the facial landmarks. Alternatively, experiments can be made with mixed training.

TABLE IV

EVALUATION OF THE MICRO F1 SCORE OF THE PREDICTION OF THE PAIN SCORE IN HORSES. THE HIGHEST F1 SCORE PER PAIN FEATURE IS HIGHLIGHTED.

	Pain feature	*SIFT*	*HOG*	*LBP*	*CNN*
Frontal pose	Ears	0.65	0.79	0.63	**0.82**
	Eyelid	0.56	**0.62**	**0.62**	0.56
	Mouth	-	-	-	-
	Nostrils	0.56	**0.63**	0.56	0.60
	Orbital	0.83	**0.84**	0.80	**0.84**
	Sclera	0.62	0.72	0.63	**0.81**
Tilted pose	Ears	0.53	**0.88**	0.77	0.74
	Eyelid	0.44	0.46	**0.51**	0.50
	Mouth	0.65	0.68	**0.76**	0.72
	Nostrils	0.57	0.66	**0.68**	0.64
	Orbital	0.77	**0.85**	0.79	**0.85**
	Sclera	0.51	**0.66**	0.62	**0.66**
Profile pose	Ear	0.46	0.81	0.56	**0.83**
	Eyelid	0.45	**0.55**	0.49	0.54
	Mouth	0.77	**0.83**	0.81	0.77
	Nostril	0.50	0.62	0.55	**0.69**
	Orbital	0.51	0.70	0.55	**0.74**
	Sclera	**0.75**	0.68	0.63	0.73

TABLE V

EVALUATION OF THE FUSION OF THE MODELS TRAINED ON ALL TYPES OF EXTRACTED FEATURES. THE F1 SCORES THAT ARE IMPROVED BY FUSION COMPARED TO THE BEST PERFORMING MODEL TRAINED WITH JUST A SINGLE EXTRACTION METHOD ARE HIGHLIGHTED.

	Fusion	*Ears*	*Eyelid*	*Mouth*	*Nostrils*	*Orbital*	*Sclera*
Frontal pose	Simple Min	0.67	0.47	-	0.37	**0.89**	**0.85**
	Simple Mean	0.66	0.47	-	0.57	**0.89**	**0.85**
	Weighted	**0.83**	**0.64**	-	**0.64**	0.84	0.77
Tilted pose	Simple Min	0.88	0.43	0.23	0.22	**0.87**	**0.74**
	Simple Mean	0.88	**0.53**	**0.77**	0.68	**0.87**	**0.74**
	Weighted	0.76	0.49	0.69	0.65	0.85	**0.67**
Profile pose	Simple Min	0.56	0.49	0.17	0.46	**0.75**	0.74
	Simple Mean	0.26	0.47	0.81	0.46	**0.75**	0.74
	Weighted	0.81	0.55	0.77	0.64	**0.76**	0.69

Pain prediction in donkeys shows a overall decrease in performance compared to horses (Table IX). Aligning the horse and donkey faces with PA improves the comparability between the species, especially for the tilted head pose. The ear-based pain prediction goes noticeably down, which is expected, since the shapes of the donkey ears are very different from the shapes of the horse ears. The profile head has the biggest decrease in performance. These results show the difference in face morphology between horses and donkeys.

TABLE VI

POSE ESTIMATION ON THE DONKEY FACE DATASET CONTAINING 534 IMAGES, USING A HOG-SVM MODEL TRAINED WITH THE HORSE FACE DATASET. THE PERFORMANCE IS SEPARATELY PRESENTED FOR EACH POSE BIN AND THE WEIGHTED AVERAGE OF EACH PERFORMANCE METRIC, ACCORDING TO THE POSE DISTRIBUTION IN THE TEST SET.

	Precision	*Recall*	*F1-Score*
Profile (-)	0.54	0.39	0.45
Tilted (-)	0.50	0.29	0.36
Frontal	0.07	0.67	0.12
Tilted (+)	0.49	0.39	0.43
Profile (+)	0.54	0.39	0.45
Weighted Average	**0.48**	**0.37**	**0.41**

TABLE VII

LANDMARKING RESULTS FOR THE DONKEY FACE DATASET CONTAINING 531 IMAGES, USING AN ERT MODEL TRAINED ON THE HORSE FACE DATASET. THE SUCCESS RATE INDICATES THE RATIO OF LANDMARKS WITH A LOCATION ERROR BELOW 0.10 OF THE EYE-NOSTRIL DISTANCE. DONKEY FACES ARE SMALLER, SO THE THRESHOLD OF ERROR CAN BE INCREASED.

	Mean Error	*Success Rate*
Frontal	0.27	0.07
Tilted	0.32	0.07
Profile	0.37	0.08
Weighted Average	***0.33***	***0.08***

VI. CONCLUSIONS

We extended several methods for automatic facial landmarking and pain estimation in horse and donkey faces. Our experiments showed that multiple models should be trained in parallel for different poses of the animal's head. We achieved 0.89 F1 score on pose estimation, and 0.51-0.88 F1 score on pain estimation on tilted poses (with different face regions), and 0.53-0.87 after decision fusion of classifiers based on different features. The lack of balance between the train and test sets affected the eventual F1-score of the pain prediction. We have tested a single model instead of three pose-specific models, but the appearance variations are great, and the single model approach did not work. We have shown the difficulties of transferring models to donkey faces. Neither automatic landmarking, nor pain estimation is directly transferable.

TABLE VIII

EVALUATION OF THE MEAN LOCATION ERROR PER AREA OF INTEREST IN THE DONKEY FACE DATASET CONTAINING 531 IMAGES, USING ERT MODELS TRAINED ON THE HORSE FACE DATASET. THE HIGHEST ERROR FOR EACH REGION OF INTEREST IS HIGHLIGHTED.

	Frontal	*Tilted*	*Profile*
Ear(s)	0.30	0.35	**0.44**
Nostrils	0.23	0.30	**0.33**
Left eye	0.32	0.38	**0.38**
Right eye	0.29	-	-
Mouth	-	0.24	**0.26**

TABLE IX

THE MICRO F1 SCORES OF THE MODEL TRAINED ON THE HORSE DATASET AND OPTIMIZED TO PREDICT THE PAIN SCORE OF DONKEYS. THE HIGHEST F1 SCORES ARE HIGHLIGHTED.

	Pain feature	*SIFT*	*HOG*	*LBP*	*CNN*
Tilted pose	Ears	0.33	**0.75**	0.65	0.67
	Eyelid	0.29	**0.59**	0.47	0.47
	Mouth	0.47	0.78	**0.84**	0.81
	Nostrils	0.43	0.61	**0.63**	0.61
	Orbital	0.43	0.57	**0.61**	0.56
	Sclera	0.48	0.62	0.63	**0.71**
Profile pose	Ear	**0.40**	0.32	**0.40**	0.36
	Eyelid	**0.53**	0.33	**0.53**	0.40
	Mouth	0.67	0.63	**0.76**	0.71
	Nostril	0.48	0.56	0.54	**0.66**
	Orbital	**0.55**	0.45	0.51	**0.55**
	Sclera	0.53	0.63	**0.73**	0.67

REFERENCES

[1] N. Andresen, M. Wöllhaf, K. Hohlbaum, L. Lewejohann, O. Hellwich, C. Thöne-Reineke, and V. Belik. Towards a fully automated surveillance of well-being status in laboratory mice using deep learning. *BioRxiv*, page 582817, 2019.

[2] F. H. Ashley, A. E. Waterman-Pearson, and H. R. Whay. Behavioural assessment of pain in horses and donkeys: application to clinical practice and future studies. *Equine Veterinary Journal*, 37(6):565–575, 2010.

[3] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful face–pain expression recognition using active appearance models. *Image and vision computing*, 27(12):1788–1796, 2009.

[4] A. Barla, F. Odone, and A. Verri. Histogram intersection kernel for image classification. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, volume 3, pages III–513. IEEE, 2003.

[5] S. Boughorbel, J.-P. Tarel, and N. Boujemaa. Generalized histogram intersection kernel for image recognition. In *IEEE International Conference on Image Processing 2005*, volume 3, pages III–161. IEEE, 2005.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer*

vision and pattern recognition (CVPR'05), volume 1, pages 886–893. IEEE, 2005.

[7] E. Dalla Costa, M. Minero, D. Lebelt, D. Stucke, E. Canali, and M. C. Leach. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS ONE*, 9(3):1–10, 2014.

[8] P. Ekman and W. Friesen. Facial coding action system (facs): A technique for the measurement of facial actions, 1978.

[9] C. B. Fookes and M. Bennamoun. *Rigid and non-rigid image registration and its association with mutual information: A review*. Queensland University of Technology, 2002.

[10] K. B. Gleerup, B. Forkman, C. Lindegaard, and P. H. Andersen. An equine pain face. *Veterinary Anaesthesia and Analgesia*, 42(1):103–114, 2015.

[11] K. B. Gleerup and C. Lindegaard. Recognition and quantification of pain in horses: A tutorial review. *Equine Veterinary Education*, 28(1):47–57, 2016.

[12] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):285–321, 1991.

[13] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[14] C. Hewitt and M. Mahmoud. Pose-informed face alignment for extreme head pose variations in animals. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–6. IEEE, 2019.

[15] B. Jonkers. Equine Utrecht University scale for automated recognition in facial assessment of pain-EQUUS-ARFAP. Master's thesis, 2018.

[16] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[17] S. C. Keating, A. A. Thomas, P. A. Flecknell, and M. C. Leach. Evaluation of EMLA Cream for Preventing Pain during Tattooing of Rabbits: Changes in Physiological, Behavioural and Facial Expression Responses. *PLoS ONE*, 7(9):1–11, 2012.

[18] D. J. Langford, A. L. Bailey, M. L. Chanda, S. E. Clarke, T. E. Drummond, S. Echols, S. Glick, J. Ingrao, T. Klassen-Ross, M. L. Lacroix-Fralish, L. Matsumiya, R. E. Sorge, S. G. Sotocinal, J. M. Tabaka, D. Wong, A. M. Van Den Maagdenberg, M. D. Ferrari, K. D. Craig, and J. S. Mogil. Coding of facial expressions of pain in the laboratory mouse. *Nature Methods*, 7(6):447–449, 2010.

[19] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Face and gesture 2011*, pages 298–305. IEEE, 2011.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[21] Y. Lu, M. Mahmoud, and P. Robinson. Estimating sheep pain level using facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 394–399. IEEE, 2017.

[22] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(3):664–674, 2010.

[23] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Face and Gesture 2011*, pages 57–64. IEEE, 2011.

[24] M. Mahmoud, Y. Lu, X. Hou, K. McLennan, and P. Robinson. Estimation of pain in sheep using computer vision. In *Handbook of Pain and Palliative Care*, pages 145–157. Springer, 2018.

[25] K. M. McLennan, C. J. Rebelo, M. J. Corke, M. A. Holmes, M. C. Leach, and F. Constantino-Casas. Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Applied Animal Behaviour Science*, 176:19–26, 2016.

[26] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[27] F. Pessanha, K. McLennan, and M. Mahmoud. Towards automatic monitoring of disease progression in sheep: A hierarchical model for sheep facial expressions analysis from video. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2020.

[28] M. Rashid, X. Gu, and Y. Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6894–6903, 2017.

[29] A. A. Salah, H. Cinar, L. Akarun, and B. Sankur. Robust facial landmarking for registration. In *Annales des télécommunications*, volume 62, pages 83–108. Springer, 2007.

[30] A. A. Salah, N. Sebe, and T. Gevers. Communication and automatic interpretation of affect from facial expressions. In *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, pages 157–183. IGI Global, 2011.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] A. H. Tuttle, M. J. Molinaro, J. F. Jethwa, S. G. Sotocinal, J. C. Prieto, M. A. Styner, J. S. Mogil, and M. J. Zylka. A deep neural network to assess spontaneous pain from mouse facial expressions. *Molecular pain*, 14:1744806918763658, 2018.

[33] J. P. van Loon and M. C. Van Dierendonck. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): A scale-construction study. *Veterinary Journal*, 206(3):356–364, 2015.

[34] J. P. van Loon and M. C. Van Dierendonck. Monitoring equine head-related pain with the Equine Utrecht University scale for facial assessment of pain (EQUUS-FAP). *Veterinary Journal*, 220(January):88–90, 2017.

[35] J. Wathan, A. M. Burrows, B. M. Waller, and K. McComb. Equifacs: the equine facial action coding system. *PLoS one*, 10(8):e0131738, 2015.

[36] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[37] X. Xu, K. D. Craig, D. Diaz, M. S. Goodwin, M. Akcakaya, B. T. Susam, J. S. Huang, and V. R. de Sa. Automated pain detection in facial videos of children using human-assisted transfer learning. In *International Workshop on Artificial Intelligence in Health*, pages 162–180. Springer, 2018.