Automatic pain estimation in equine faces: More effective uses for regions of interest

Jens Jan Ruhof, Albert Ali Salah
Dept. Information and Computing Sciences
Utrecht University
Utrecht, the Netherlands
j.j.ruhof@students.uu.nl
a.a.salah@uu.nl
https://orcid.org/0000-0001-6342-428X

Thijs J. P.A.M. van Loon
Department of Clinical Sciences
Utrecht University
Utrecht, the Netherlands
https://orcid.org/0000-0002-7828-191X

***Abstract*—Recognition of pain in equines is essential for their welfare. There are several tools, such as the Horse Grimace Scale, EquiFACS and EQUUS-ARFAP, developed for pain assessment in equines, and there are approaches to automate assessment, as training observers takes time, and disagreements between observers are common. In this work, we provide a system for pain assessment in equine faces based on the EQUUS-ARFAP scale. The proposed system consists of three steps, namely, automatic detection of the facial regions, automatic head orientation detection, and automatic pain detection for each facial region of interest separately. Our main contribution is a detailed analysis of the usage of regions of interest as the main representation of the assessment pipeline, instead of facial landmarks. We show improved pain classification on the publicly available UU Equine Pain Face Dataset and advance the state of the art in this problem.**[1]

***Index Terms*—Animal behaviour analysis; pain estimation; equines; horses; face analysis**

I. Introduction

Equine welfare is impacted by the recognition and quantification of pain [1], [2]. Therefore, the assessment and treatment of pain is vital in maintaining healthy and happy equines. Human pain assessment is facilitated through verbal examination, equines however do not possess verbal communication and are reliant on observers to locate and quantify their pain.

Several studies have shown that pain in equines can manifest as a change in behaviour, such as aggressiveness, reluctance to move, vocalisation and diminished socialisation [3]. However, pain can also be observed via subtle changes in their facial expressions [4].Several frameworks were developed to evaluate pain from facial expressions of equines, such as EquiFACS [5], the Horse Grimace Scale (HGS) [6], and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) [7]. Although the use of these scales to assess pain is proven to be efficient, it requires observer training and manual annotation of the pain score for each Action Unit (AU) or Action Descriptor (AD). It has also been shown that there is little consensus between veterinarians on the qualitative and quantitative pain in equines [8], [9], necessitating the need to automate this process.

The present work proposes a fully automatic system for pain estimation in horses based on the facial regions, and improves the state of the art in this problem. Our proposed system is based on the Equine Utrecht University Scale for Automated Recognition in Facial Assessment of Pain (EQUUS-ARFAP) [10], which scores different parts of a horse's face individually for pain indicators. The automatic assessment system is composed of a region of interest (ROI) extractor, a pose estimation step and a pain estimation step, which depends on the pose and ROI localisation.

The main contributions of this paper are as follows:

- We propose a pipeline for pain estimation in equines, incorporating several off-the-shelf tools.
- We implement a non-pose-aware ROI localisation model, robust to variations in horse breeds and poses.
- We illustrate that training a deep-neural network based pain model on the detected ROI's improves the pain estimation results, even beyond using manually selected ground truth regions.

II. Related Work

A. Facial expressions of pain

Objective analysis of pain starts by defining the type of pain we want to analyse, namely, acute pain. Acute pain is defined as pain that starts sharp or intense and serves as a warning sign of disease or threat to the body [11]. It is caused by injury, surgery, illness, trauma, or painful medical procedures and generally lasts from a few minutes to less than 6 months.

Objective pain assessment in humans is often achieved through the Facial Action Coding System (FACS) system [12]. FACS is designed to categorize facial movements by looking at the underlying muscles responsible for this movement. They divide these movements up into Action Units (AU), where an AU can be defined as the relaxation or contraction of a muscle group.

[1] This is the uncorrected author proof. Please cite as: J.J. Ruhof, A.A. Salah, T.J.P.A.M van Loon, "Automatic pain estimation in equine faces: More effective uses for regions of interest," 12th International Conference on Affective Computing and Intelligent Interaction, ACII 2024 - Workshops and Demos, Glasgow, UK, 2024.

979-8-3315-1645-1/24/$31.00 ©2024 IEEE

Pain estimation in animals can use indicators on the body or the face [13]. Face analysis based systems have been adopted for several species [5], [14]–[17].

Equines have a different facial and underlying muscle structure than humans and other primates. To get an overview of the underlying facial muscles, a horse head was dissected to map the connections and interactions of the facial muscles [5]. EquiFACS used the same AUs as FACS, if the same muscle groups were responsible for a similar facial movement, and created a new AU if there was no FACS equivalent. If a facial movement could not be defined using AUs, it was defined as an Action Descriptor (AD).

The complexity of these scales makes it difficult to train an observer to assess pain. There are also alternative scales focusing on the facial expressions themselves, and not on the underlying muscle structure. These are the so-called grimace scales. They have been used for several species (rats, mice, rabbits, sheep, piglets, cats, and horses) [18], [19]. Among these, the Horse Grimace Scale (HGS) focuses on the ears, orbital tightening, mouth strain, tension above the eyes, strained jaw, and the nostrils, and scores each category on a scale from 0 to 2, where 0 is “pain not present”, 1 is “pain moderately present” and 2 is “pain obviously present”. Other relevant assessment scales are the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) [7], which includes static as well as dynamic indicators (i.e. sound and video) and the Equine Utrecht University Scale for Automated Recognition in Facial Assessment of Pain (EQUUS-ARFAP) [10], which combines the static features of the HGS and the EQUUS-FAP scale into a single list of six features. While not a clinically validated scale, it is suitable for appearance-based automatic analysis.

B. Automatic assessment of pain

While pain assessment scales are useful, their manual application can be time-consuming and open to subject interpretation [8], [9]. Recent efforts have focused on automating the assessment process in animals. Due to the limitations related to data scarcity, many works adopted transfer learning approaches. In mice, transfer learning was used to optimize an InceptionV3 neural network model, for the binary pain vs no-pain classification [20]. In [21], a model trained on sheep faces was used to bootstrap a model specialized for horses.

Previous work on assessing pain in animals has concentrated on three components: landmarking, pose estimation and pain estimation, respectively. Landmarking is used to isolate the ROIs of an animal’s face and for identifying morphometrically relevant points, such as eye corners, nostril boundaries, and mouth corners. Several machine learning approaches are frequently used for automatic landmarking, such as Ensemble of Regression Trees (ERT) models [21]–[25] and supervised descent models (SDM) [21], [25]. These models have been shown to work well on human faces as well. However, these models struggle with extreme poses and the wide variety of animal head shapes, necessitating the need for an improved localisation method. Some approaches have advocated detecting the head pose roughly first, and then using pose-specific landmarking models [21]. This is especially reasonable for horse faces, as the pose of the head significantly affects the visibility of landmarks.

The pose can be detected via traditional machine learning based approaches, such as those using Histogram of Gradients (HOG) features with Support Vector Machine (SVM) classifiers [21], [24], [26]. The detected pose is used to inform the ROI localisation, which needs to be accurate for pain estimation, which is the last step [21], [25].

III. Methodology

Our proposed approach uses an SVM-based pose classifier and a deep neural network based ROI classifier in parallel to determine a final set of ROIs on the facial image of a horse. The ROI’s are then processed individually by an SVM-based classifier, and each is assigned a pain score. The summation of all pain scores results in the overall pain score.

The main reason we use an SVM is that there is not enough data for end-to-end deep learning; there are very few samples with strong pain expressions. Previous works in the literature have explored various data augmentation methods [21], but since pain expressions are subtle, classical image augmentation and generative AI based image synthesis both have very limited usefulness for data augmentation and more innovative ways should be explored.

Our results are evaluated using both the macro and weighted metrics. The macro metric is calculated by taking the unweighted mean of all the per-class scores, whereas the weighted metric is calculated by taking the support of all per-class means into account. Equation 1 illustrates the macro F1-score, and Equation 2 shows the weighted F1-score, where C is the class and S the support.

$$MacroF1 = \frac{F1_{C_1} + F1_{C_2} + F1_{C_3}}{\textit{total number of classes}} \tag{1}$$

$$WeightedF1 = \frac{F1_{C_1} * S_{C_1} + F1_{C_2} * S_{C_2} + F1_{C_3} * S_{C_3}}{\textit{total number of instances}} \tag{2}$$

A. Pose estimation

Head pose variations cause evident changes in the facial appearance in equines due to self-occlusions. In this work, we use the pose to determine whether a particular ROI is supposed to be present in the image and whether or not it should be allocated to the left or right side of a horse. To this end, we implemented a 5-class pose classifier distinguishing between profile left, tilted left, frontal, tilted right, and profile right poses.

Based on results from the literature, we use Histogram of Oriented Gradients (HOG) features [27] with an SVM classifier for this purpose. All images were resized to (128×128) before the HOG features were extracted. We performed a nested 3-fold cross-validation within the training set to find the best HOG parameters (orientations, pixels per cell, and cells per block). We used a linear kernel SVM and balanced class weights, assuming no prior distribution information for any pose class. Furthermore, we used a one-vs-one decision function, which means that each sample gets a score for every pairwise class comparison the class with the most votes is the predicted class [28].

In our preliminary experiments, we observed that SVM made errors in images where the background was of a similar color as the horses coat, leading to HOG descriptors that included parts of the background. To mitigate this problem, we added an extra step of preprocessing, using the U^2-network for background removal [29] before resizing the images. The off-the-shelf REMBG Python library provides an implementation of this, which we employed without any adjustments.

B. ROI localisation

ROI localisation is a necessary step in the automatic assessment of pain. Previous works have focused on using transfer learning from landmarking models for human faces [30] and a pose-informed Ensemble of Regression Trees (ERT) [21], [24], [25]. In this work, we compare the performance of a pose-informed ERT model and a YOLOv8 (You Only Look Once) model [31] for ROI localisation. The baseline landmarking approach based on ERT uses three different models, one for each pose class, predicting 54, 44, and 45 points for the frontal, tilted and profile poses, respectively. The images are resized to a fixed height of 1000, with the width being calculated by the average aspect ratio for each pose, resulting in images of different size for frontal (582×1000), tilted (582×1000) and profile (706×1000) poses, respectively. The ERT models are implemented using the dlib library, with a tree depth of 5, 500 trees per cascade level and an oversampling rate of 30, which applies data augmentation jitter to the images. We assume that face detection is correctly performed, therefore we use the ground truth facial bounding boxes for estimation. The images in our dataset (see next section) contain prominent horse faces without much else, so face detection is not an issue, but a more challenging dataset may require the assessment of this step to determine its impact on the overall performance.

The ERT model's performance is measured using the Success Rate (SR) and the Mean Normalised Error (MNE). The SR refers to the percentage of landmark predictions with a distance lower than 6% of the eye-nostril distance from the manual ground truth (after [21]), and the MNE is the Euclidean distance between the prediction and ground truth normalised by the eye-nostril distance.

The YOLOv8 model is not pose-informed, meaning that the model uses a single model for each ROI under any pose. To maintain consistency across all pipeline components, we separate a test set, and the training set was further divided into training and validation sets using an 80-20 split within each pose group, resulting in 1187 training images, 297 validation images, and 371 test images. The images are resized using the average aspect ratio of all images with a fixed height of 1000, resulting in an image size of (615×1000). The bounding boxes are obtained by drawing rectangles around the landmarks of each ROI, with zero padding, and the classes of the ROI have been reduced by combining each left-right pair into a single category (e.g. left-eye and right-eye are both stored as the eye ROI). The YOLOv8 model was trained using the following parameters: image size of 512 (resizing all images to this dimension is done to reduce computational complexity), 100 epochs, a batch size of 10, and a random seed of 0. The YOLO model for bounding box localization is evaluated via precision, recall, mean Average Precision 50 (mAP50), mAP50-95 and the Intersection over Union (IoU) measures. For the latter, an overlap of 95% is considered as a very good agreement with the ground truth bounding box. mAP50 measures the mean average precision at an IoU threshold of 0.5, while mAP50-95 measures the mean average precision across IoU thresholds ranging from 0.5 to 0.95.

C. Pain estimation

To estimate the pain in horses we make use of a SVM trained on HOG features, following previous research [21], [25]. We trained a pose-and-ROI-specific SVM for every pose and ROI combination. Each ROI was resized to a size of (64,128) before the extraction of HOG features. We tested different values for the HOG parameters (orientations, cells per block, pixels per cell), as well as different SVM parameters (kernel, regularization parameter C) on the training partition, to find the optimal configuration per ROI and an optimal configuration for all ROIs on average. The best combination of parameters on average for all ROI was obtained by scoring all classification reports of a pose and ROI on highest macro avg recall, macro avg precision and last macro avg f1-score. The top 100 configurations for each pose and ROI were then given a score based on where they stood in the ranking, the configuration with the highest average ranking was then picked as the final configuration for the pain classifier and tested on the test set to produce the performance measurement results.

When only one ROI of a left-right pair is detected, but more are expected, the missing ROI is created by mirroring the detected one, assuming general symmetry between both. If there is no ROI detected, the expected ROI will be depicted as a zero array of the same size with the expected feature matrix size.

In total, we evaluate three pain classifiers with the optimal average hyperparameters. The first classifier is trained and tested on the ground truth ROIs. The second classifier is trained on the ground truth ROI and tested on the YOLO-predicted ROIs. The final classifier is trained and tested on the YOLO-predicted ROIs. We have chosen not to include the classifier that was trained on the ground truth and tested on the ERT-predicted ROIs, as the results for ERT-based ROI localisation were worse than the YOLO localisation results.

IV. Data and annotations

The UU Equine Pain Face Dataset consists has 1855 horse images and 531 donkey images and is publicly available,

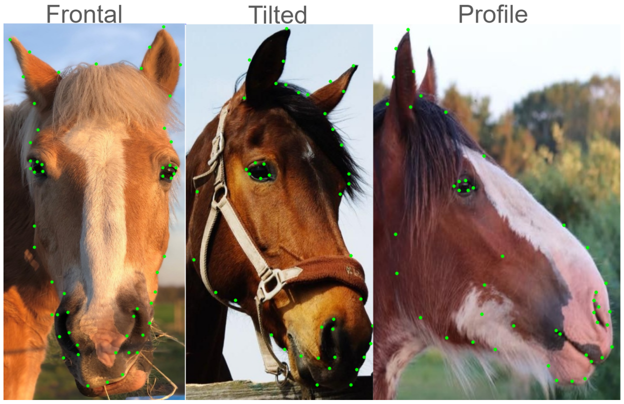

Fig. 1: Horse faces for each pose group with their respective landmarks in green.

including the pain scores per ROI [25]. The horses have different pose, illumination and background conditions, and some of them have bridles. Part of the data comes from horse owners, and part of it was collected by a veterinary medicine faculty during a clinical procedure. Ethics approvals were obtained in the preparation of the dataset.

The images used in this work are only those of horses and we focus on the facial region of these animals. Each image was annotated with pain labels and landmarks for each ROI in the EQUUS-ARFAP scale. The landmark schemes differ between the profile, tilted and frontal poses as different ROI are occluded per pose.

A. ROI annotations

The landmarks of the images are annotated using the landmark annotation scheme from [10]. The head orientations frontal, tilted, and profile have their own landmark scheme using 54, 44, and 45 points, respectively (Figure 1). These landmarks came with the database distribution. We supplemented these annotations with bounding box annotations around the face, eyes, ears, nostrils, jaw, and the mouth. Some regions are occluded in different poses. The bounding boxes are formatted as (x centre, y centre, width, height) on a scale of 0 to 1. This allows convenient scaling for the different image resolutions present in the dataset.

B. Pain distribution

The database we have used is publicly available [25]. The images were annotated for potential signs of pain by three expert raters (one senior expert researcher and two graduate students) according to the EQUUS-ARFAP scale presented in Table I. All experts scored the entire dataset using the full images. In this work, we use the pain score annotations made by the senior expert researcher, following [21]. The distribution of these annotations can be found in Figure 2, where we observe that the severe pain label “2” is underrepresented for each ROI.

TABLE I: Score sheet for EQUUS-ARFAP

Data	Categories	Score
Ears	Both ears are turned forwards	0
	At least one ear lateral position or further to backwards	1
	Both ears turned backwards	2
Orbital tightening	Relaxed	0
	A bit of tightening of the eyelids	1
	Obviously tightening of eyelid/ eyelid closed	2
Angulated upper eyelid	Relaxed	0
	A bit more visible	1
	Obviously more visible	2
Visibility of the sclera	Sclera is not visible	0
	An edge of the sclera is visible	1
	Obviously more visible	2
Corners of mouth / lip	Relaxed	0
	Lifted a bit	1
	Obviously lifted/ strained	2
Nostrils	Relaxed	0
	A bit more opened	1
	Obviously more opened (dilated mediolaterally)	2
Total		.../12

C. Data preparation

We have split the 1855 images into a training and a test set with a roughly equal distribution of pain labels and following an eighty-twenty split. This is done by stratifying the data on pain scores. The distribution of the pain scores for the training and the test set can be seen in Figure 2.

V. Experimental Results

A. Pose estimation

The 3-fold cross-validation of the HOG parameters showed that the optimal parameters are nine orientations 4×4 pixels per cell and 4×4 cells per block Table II. When removing the background with REMBG before resizing the images and extracting the HOG features with the same parameters, we see a slight uplift in the overall performance (see Table III). The errors observed were often related to ambiguous head poses, or mistakes in the background removal process, being either too harsh or too subtle in some cases.

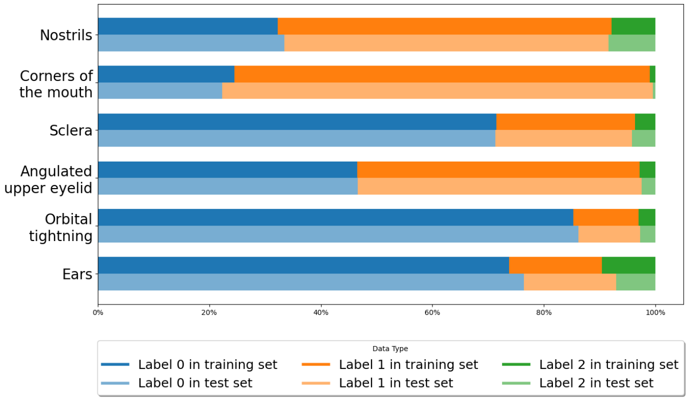

Fig. 2: Distribution of the pain scores in the dataset for the training and test sets.

TABLE II: Performance of the pose classifier trained on the HOG features with parameters: (9, 4x4, 4x4).

Orientation	Precision	Recall	F1-score	Support
Profile left	0.873	0.923	0.897	52
Tilted left	0.859	0.850	0.854	100
Frontal	0.818	0.750	0.783	72
Tilted right	0.892	0.938	0.915	97
Profile right	0.959	0.940	0.950	50
Macro avg	0.880	0.880	0.880	371
Weighted avg	0.875	0.876	0.875	371

TABLE III: Performance of the pose classifier using HOG parameters (9, 4×4, 4×4) with the background removed of the images using REMBG.

Orientation	Precision	Recall	F1-score	Support
Profile left	0.877	0.962	0.917	52
Tilted left	0.896	0.860	0.878	100
Frontal	0.836	0.778	0.806	72
Tilted right	0.901	0.938	0.919	97
Profile right	0.940	0.940	0.940	50
Macro avg	0.890	0.895	0.892	371
Weighted avg	0.889	0.889	0.889	371

B. ERT-based ROI localisation

The ERT model shows promising results for landmark localisation (see Table IV), which is in line with previous findings in the literature [21], [25]. However, the wide range of poses and horse breeds present in the dataset causes misclassifications for extreme poses and underrepresented breeds. [21] notes that incorporating the pitch, roll, and yaw improves the results. However, such a model would still struggle with breeds that are underrepresented or that have out of distribution appearances. Furthermore, some ROIs appear to perform well on paper when using the SR and MNE as evaluation, such as the eyes. However, it is important to place a caveat on these results. The eye ROI is a lot smaller and the landmark grouping is a lot closer to its peers, making it easier to pass the evaluation methods while still being off, leading to inaccurate ROI cropping. Finally, splitting the data over different pose-dependent classifiers reduces the quantity of available data for each classifier, possibly hindering performance.

Figure 3 left shows an example of a wrong ERT prediction, where the background misleads the detector to predict a wrong pose class, and leading to wrong HOG features further down the pipeline.

C. YOLO-based ROI localisation

The YOLO model similarly shows promising results for landmark localisation (see Table V). It performs well on the box precision, recall, and mAP50, showing the model is capable of detecting and selecting the right bounding boxes for an image. However, the IoU and mAP50-95 scores are a bit lower, showing that the model is not perfectly accurate, which can be seen when looking at the pain estimation results trained on the ground truth ROI and tested on the YOLO ROI (see Table VII). However, it must be noted that the IoU is 0 if there are either too many predictions or too few predictions. Figure 3 shows an example of the YOLO model still providing accurate ROI localisation, where the predicted pose is different from the ground truth pose, due to the models' independence of pose when detecting ROIs. It also shows the shortcoming of the YOLO-based approach, as not every detected ROI is allocated, due to the pose-informed allocation logic.

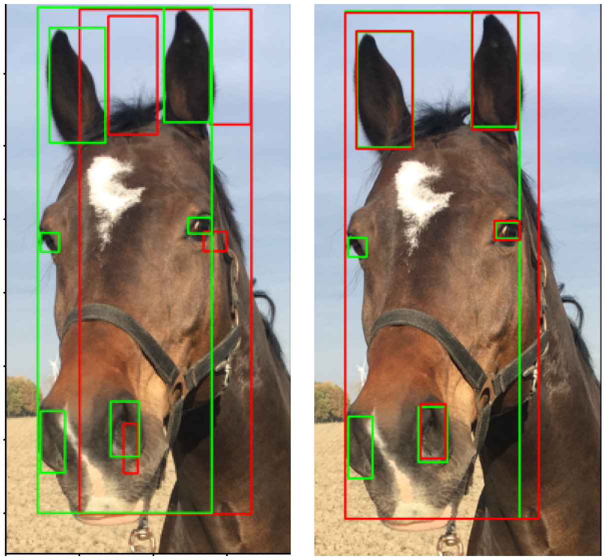

Fig. 3: Example showing the impact of wrong pose estimation on ERT-based (left) and YOLO-based (right) ROI detection. The green boxes represent the ground truth for each ROI, while the red boxes indicate the predicted ROI.

D. Pain estimation

The best-performing pain classifier configuration (on the three-fold validation set) averaged over all ROIs consists of the following HOG parameters: 9 orientations, 8×8 pixels per cell, and 3×3 cells per block, and C regularization of 1 with a linear kernel for the SVM parameters. Table VI shows the classification scores for each ROI trained and tested on the ground truth bounding boxes with the best average pain classifier mentioned above. The model performs well on the ears and mouth, but the results for the other ROIs are less accurate. This table will serve as a baseline for the pain classifiers.

Table VII shows the performance of a classifier trained on the ground truth ROI and tested on the YOLO bounding boxes. The results are worse, suggesting that the bounding boxes identified with YOLO have discrepancies compared to the ground truth bounding boxes. This may be an effect of how closely these are cropped, and therefore it may make sense to use automatically detected ROIs for both training and testing to remove the discrepancy in these two conditions.

To verify the hypothesis of discrepancies between ground truth bounding boxes and YOLO bounding boxes, a third

TABLE IV: The Success Rate (SR) and Mean Normalised Error (MNE) using the Ensemble of Regression Trees (ERT) models on the landmarks for each region-of-interest (ROI) and pose. Missing values indicate that the ROI is not defined in that class. Results are obtained with the ground truth pose.

ROI	Frontal SR	Tilted SR	Profile SR	Frontal MNE	Tilted MNE	Profile MNE	Average SR	Average MNE	Average IoU
Left Ear	0.669	0.715	0.519	0.061	0.063	0.080	0.673	0.065	0.693
Right Ear	0.678	0.652	0.610	0.055	0.067	0.075	0.651	0.066	0.712
Left Eye	0.766	0.918	0.772	0.053	0.033	0.049	0.834	0.043	0.514
Right Eye	0.792	0.844	0.783	0.049	0.040	0.046	0.813	0.044	0.507
Left Nostril	0.616	0.557	0.821	0.068	0.060	0.041	0.637	0.058	0.491
Right Nostril	0.646	0.558	0.837	0.075	0.076	0.039	0.651	0.067	0.474
Mouth	-	-	0.868	-	-	0.037	0.868	0.037	0.634
Jaw	-	-	0.579	-	-	0.057	0.579	0.057	0.777

TABLE V: The performance metrics for the ROI prediction of the YOLO model on the test set.

ROI	Images	Instances	Box precision	Box recall	Box mAP50	Box mAP50-95	IoU
Face	371	371	0.999	1.000	0.995	0.977	0.964
Eyes	371	443	0.946	0.907	0.953	0.502	0.665
Nostrils	371	443	0.950	0.936	0.963	0.532	0.701
Ear	371	639	0.988	0.962	0.984	0.732	0.799
Mouth	102	102	0.922	0.929	0.935	0.543	0.666
Jaw	102	102	0.942	0.841	0.967	0.644	0.753
All	371	2100	0.958	0.946	0.966	0.655	0.767

TABLE VI: Pain prediction scores per ROI, classifiers trained and tested on the ground truth ROIs.

Model	Macro precision	Macro recall	Macro f1	Weighted precision	Weighted recall	Weighted f1
Ear	0.756	0.740	0.744	0.878	0.883	0.879
Nostril	0.566	0.539	0.549	0.639	0.644	0.640
Sclera	0.750	0.493	0.532	0.724	0.721	0.711
Orbital	0.637	0.485	0.528	0.821	0.834	0.825
Angulated	0.379	0.387	0.382	0.555	0.565	0.559
Mouth	0.667	0.703	0.679	0.807	0.780	0.791

classifier was trained and tested on YOLO-detected bounding boxes, the results of which can be seen in Table VIII. The results of this classifier improved for each ROI compared to the previous model, but did not perform as well as the ground truth model, aside from the nostril ROI, which seems to perform better.

The demonstrated improvement in ROI classification shows the potential of fully automating equine facial pain assessment using YOLO. However, it is important to note that the predicted YOLO bounding boxes with which we have trained the pain classifier are more likely to be closer to the ground truth bounding boxes than the predicted bounding boxes of the test set. This potential difference in accuracy could explain the slight reduction in performance when comparing this classifier to the ground truth baseline.

TABLE VII: Pain prediction scores per ROI, classifiers trained on ground truth ROIs and tested on the YOLO-detected ROIs.

Model	Macro precision	Macro recall	Macro f1	Weighted precision	Weighted recall	Weighted f1
Ear	0.702	0.695	0.696	0.856	0.862	0.857
Nostril	0.582	0.550	0.560	0.639	0.639	0.636
Sclera	0.406	0.419	0.412	0.670	0.701	0.685
Orbital	0.568	0.467	0.500	0.809	0.826	0.816
Angulated	0.352	0.359	0.355	0.515	0.525	0.520
Mouth	0.582	0.592	0.586	0.746	0.730	0.737

TABLE VIII: Pain prediction scores per ROI, classifiers trained and tested on the YOLO-based ROIs.

Model	Macro precision	Macro recall	Macro f1	Weighted precision	Weighted recall	Weighted f1
Ear	0.725	0.735	0.717	0.875	0.873	0.869
Nostril	0.582	0.544	0.566	0.633	0.639	0.635
Sclera	0.616	0.483	0.515	0.698	0.715	0.700
Orbital	0.626	0.457	0.497	0.822	0.845	0.829
Angulated	0.364	0.370	0.367	0.533	0.542	0.537
Mouth	0.639	0.609	0.619	0.769	0.790	0.777

TABLE IX: Pain prediction scores per ROI, classifiers trained on the ground truth ROIs and tested with ERT ROIs.

Model	Macro precision	Macro recall	Macro f1	Weighted precision	Weighted recall	Weighted f1
Ear	0.684	0.660	0.669	0.819	0.832	0.824
Nostril	0.466	0.401	0.385	0.543	0.533	0.520
Sclera	0.537	0.416	0.438	0.640	0.670	0.647
Orbital	0.388	0.389	0.387	0.788	0.790	0.789
Angulated	0.371	0.377	0.370	0.543	0.548	0.539
Mouth	0.561	0.565	0.562	0.731	0.720	0.725

Table IX shows the performance of a classifier trained on the ground truth and tested on ERT-predicted ROIs. The results are similar to the Table VII, albeit with lower macro precision and macro recall for every ROI, except for sclera, indicating that the ERT detected ROIs are less accurate than the YOLO detected ROIs. Table X shows the performance of a pain classifier trained and tested on the ERT-predicted ROIs. Table X shows a different improvement pattern compared to Table VIII, as the Angulated upper eyelid, Ears, and Mouth ROI macro precision and macro recall values do not improve over Table IX. The detected ROIs of the test set are not as accurate as those of the training set, and thus provide poorer boundaries to aid decision making.

Table XI and Table XII show the fully automated pipeline with predicted pose. When looking at the macro recall and macro F1, we see a small reduction in Table XI in comparison to Table VIII indicating that the YOLO model adapts well to mistakes in pose classifications. Similarly, Table XII shows an

TABLE X: Pain prediction scores per ROI, classifiers trained and tested on ERT ROI.

Model	Macro precision	Macro recall	Macro f1	Weighted precision	Weighted recall	Weighted f1
Ear	0.655	0.644	0.647	0.811	0.821	0.815
Nostrils	0.567	0.489	0.505	0.614	0.617	0.609
Sclera	0.706	0.421	0.450	0.667	0.682	0.656
Orbital	0.469	0.442	0.453	0.800	0.807	0.803
Angulated	0.361	0.368	0.364	0.529	0.537	0.531
Mouth	0.536	0.548	0.534	0.719	0.660	0.684

identical reduction pattern in comparison to Table X, which is against our expectations that ERT is more prone to make pose-informed mistakes. A possible explanation for this is that the images with a wrongly predicted pose are often related to ambiguous head poses that are similar to the predicted pose.

TABLE XI: Pain prediction scores per ROI, classifiers trained and tested on YOLO ROI with predicted pose.

Model	Macro precision	Macro recall	Macro f1	Weighted precision	Weighted recall	Weighted f1
Ear	0.704	0.724	0.706	0.867	0.864	0.863
Nostrils	0.559	0.525	0.539	0.617	0.625	0.620
Sclera	0.558	0.472	0.492	0.701	0.721	0.707
Orbital	0.626	0.457	0.497	0.822	0.845	0.829
Angulated	0.371	0.379	0.375	0.543	0.553	0.548
Mouth	0.645	0.618	0.629	0.785	0.802	0.792

TABLE XII: Pain prediction scores per ROI, classifiers trained and tested on ERT-based ROI with predicted pose.

Model	Macro precision	Macro recall	Macro f1	Weighted precision	Weighted recall	Weighted f1
Ear	0.646	0.632	0.638	0.811	0.821	0.815
Nostrils	0.535	0.474	0.485	0.602	0.609	0.600
Sclera	0.697	0.415	0.443	0.660	0.673	0.649
Orbital	0.478	0.451	0.462	0.805	0.812	0.809
Angulated	0.355	0.363	0.357	0.520	0.528	0.522
Mouth	0.520	0.528	0.515	0.722	0.653	0.682

TABLE XIII: Pain prediction using aggregated ROIs for the ground truth (gt), YOLO, and ERT models, all using the gt pose. Bold numbers indicate statistical significance (5×2 CV F-test, α = 0.05) between the gt and the model.

Model/Metric	Macro precision	Macro recall	Macro F1
GT (baseline)	0.626	0.558	0.572
YOLO (YOLO ROIs with gt pose)	0.592	0.533	0.547
ERT (ERT ROIs with gt pose)	0.549	**0.485**$^{-}$	**0.492**$^{-}$

Table XIII shows the 5×2 CV F-test results [32] comparing the ground truth model with models trained and tested on YOLO and ERT detected ROIs. All models use the ground truth pose, as the pose classifier is trained on the data used in the significance test and would therefore not make wrong predictions. Bold results indicate rejection of the null hypothesis, showing significant performance differences. The ERT model performs significantly worse in macro recall and macro F1 compared to the ground truth model, while the YOLO model does not show significant differences. This suggests our proposed method outperforms the previously used ERT model.

VI. Conclusion

The pose of a horse's head greatly affects the visibility of facial landmarks and regions of interest, and earlier approaches for facial pain analysis for horses employed multiple models, each trained for a different pose class. In this paper we have provided an alternative way for region of interest (ROI) localisation, circumventing the need for pose-informed ROI localisation models. We contrasted an Ensemble of Regression Trees (ERT)-based landmark detection approach with a YOLO-based region detection approach for obtaining the ROIs. Our experiments showed that training the pain classifier on the automatically detected ROIs improved the results for all ROIs when using a YOLO-based model, compared to manually selected ROIs (i.e. ROI ground truth). The results show that the automatic ROI detection for both training and testing provides consistency, and can be preferable to using a manual ground truth in the training stage.

Our model is not pose-informed, but uses a pose logic. One advantage of our approach over the traditional pose-informed ROI localisation lies in its handling of pose prediction errors. In the pose-informed methods, if the pose prediction is incorrect, such as the detection of a frontal pose when the actual pose is a profile pose, the model would falsely detect non-existent ROIs for the second eye, nostril, and ear. In contrast, our model mitigates this issue by not having a fixed set of ROIs it needs to detect. If the pose prediction is incorrect, the worst-case scenario would be that some ROIs are either not detected or are misallocated to the wrong side of the face. This avoids wrong HOG features being fed into the pain classifier. When comparing the YOLO-based ROI detector to the ERT-based detector, we see similar results, but in favour of the YOLO-based classifier when using the classifier trained on the ground truth ROI, except for the Sclera. However, when training the pain classifiers on the predicted ROIs, only the YOLO-based model saw an improvement for all ROIs, whereas the ERT-based model worsened for three of the six ROIs. This indicates that the ERT-based model is not as consistent in ROI detection as the YOLO-based model. This point is also confirmed by a 5×2 CV F-test, which shows no statistical difference between YOLO and GT models, but significantly worse performance for the ERT model. In conclusion, our move from a landmark estimation based approach to a patch based representation of the regions of interest improved the downstream pain estimation results, and pushed the state of the art in the publicly available UU Equine Pain Face Dataset benchmark.

VII. Ethical Impact statement

This paper concerns itself with the automatic pain estimation in equine faces, with a focus on more effective region of interest localisation and usage. We acknowledge that our dataset comes from an ethics-board approved study which made the dataset and pain annotations available publicly.

Automatic pain estimation for animal welfare is a task that is potentially useful for detecting health issues with animals, and for monitoring. The main risk of such automated approaches is that people may rely too much on these systems and reduce the actual human oversight in animal welfare. The technology for monitoring has not progressed to the point of replacing professional caregivers, and this needs to be clearly realized in any actual use of such systems.

References

[1] K. M. Rutherford, "Assessing pain in animals," *Animal Welfare*, vol. 11, no. 1, pp. 31–53, 2002.

[2] European Commission, "Animal welfare," 2023, 20-01-2023. [Online]. Available: https://food.ec.europa.eu/animals/animal-welfare_en

[3] F. Ashley, A. Waterman-Pearson, and H. Whay, "Behavioural assessment of pain in horses and donkeys: application to clinical practice and future studies," *Equine veterinary journal*, vol. 37, no. 6, pp. 565–575, 2005.

[4] K. B. Gleerup, B. Forkman, C. Lindegaard, and P. H. Andersen, "An equine pain face," *Veterinary Anaesthesia and Analgesia*, vol. 42, no. 1, pp. 103–114, 2015.

[5] J. Wathan, A. M. Burrows, B. M. Waller, and K. McComb, "Equifacs: the equine facial action coding system," *PLoS one*, vol. 10, no. 8, p. e0131738, 2015.

[6] 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*, vol. 9, no. 3, p. e92281, 2014.

[7] 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*, vol. 220, no. January, pp. 88–90, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.tvjl.2017.01.006

[8] J. Price, J. M. Marques, E. M. Welsh, and N. K. Waran, "Pilot epidemiological study of attitudes towards pain in horses," *Veterinary Record*, vol. 151, no. 19, p. 570–575, Nov 2002.

[9] N. C. N Waran, VM Williams and I. Bridge, "Recognition of pain and use of analgesia in horses by veterinarians in new zealand," *New Zealand Veterinary Journal*, vol. 58, no. 6, pp. 274–280, 2010, pMID: 21151212. [Online]. Available: https://doi.org/10.1080/00480169.2010.69402

[10] B. Jonkers, "Equine Utrecht University scale for automated recognition in facial assessment of pain-EQUUS-ARFAP," master's thesis, Utrecht University, 2018, available at https://studenttheses.uu.nl/handle/20.500.12932/29723.

[11] International Association for the Study of Pain, "Acute pain," "https://www.iasp-pain.org/resources/topics/acute-pain/", Jan 2023, 26-01-2024. [Online]. Available: https://www.iasp-pain.org/resources/topics/acute-pain/

[12] P. Ekman, W. v. Friesen, and J. C. Hager, *FACS Manual*, The Ekman group, 2002.

[13] S. Broome, M. Feighelstein, A. Zamansky, G. Carreira Lencioni, P. Haubro Andersen, F. Pessanha, M. Mahmoud, H. Kjellström, and A. A. Salah, "Going deeper than tracking: A survey of computer-vision based recognition of animal pain and emotions," *International Journal of Computer Vision*, vol. 131, no. 2, pp. 572–590, 2023.

[14] L. A. Parr, B. M. Waller, S. J. Vick, and K. A. Bard, "Classifying chimpanzee facial expressions using muscle action." *Emotion*, vol. 7, no. 1, p. 172, 2007.

[15] B. M. Waller, K. Peirce, C. C. Caeiro, L. Scheider, A. M. Burrows, S. McCune, and J. Kaminski, "Paedomorphic facial expressions give dogs a selective advantage," *PLoS one*, vol. 8, no. 12, p. e82686, 2013.

[16] C. C. Caeiro, A. M. Burrows, and B. M. Waller, "Development and application of catfacs: Are human cat adopters influenced by cat facial expressions?" *Applied Animal Behaviour Science*, vol. 189, pp. 66–78, 2017.

[17] M. Feighelstein, Y. Ehrlich, L. Naftaly, M. Alpin, S. Nadir, I. Shimshoni, R. H. Pinho, S. P. Luna, and A. Zamansky, "Deep learning for video-based automated pain recognition in rabbits," *Scientific Reports*, vol. 13, no. 1, p. 14679, 2023.

[18] M. C. Evangelista, B. P. Monteiro, and P. V. Steagall, "Measurement properties of grimace scales for pain assessment in nonhuman mammals: a systematic review," *Pain*, vol. 163, no. 6, pp. e697–e714, 2022.

[19] 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*, vol. 9, no. 3, pp. 1–10, 2014.

[20] 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*, vol. 14, p. 1744806918763658, 2018.

[21] F. Pessanha, A. A. Salah, T. van Loon, and R. Veltkamp, "Facial image-based automatic assessment of equine pain," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2064–2076, 2023.

[22] 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*. Springer, 2018, pp. 145–157.

[23] 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 (FG 2020)*. IEEE Press, 2020, p. 387–393. [Online]. Available: https://doi.org/10.1109/FG47880.2020.00107

[24] 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)*. IEEE, 2019, pp. 1–6.

[25] H. I. Hummel, F. Pessanha, A. A. Salah, T. J. van Loon, and R. C. Veltkamp, "Automatic pain detection on horse and donkey faces," in *Proc. IEEE FG*. IEEE, 2020, pp. 793–800.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2980–2988.

[27] 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)*, vol. 1. IEEE, 2005, pp. 886–893.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," vol. 106, 2020, p. 107404.

[30] 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*, 2017, pp. 6894–6903.

[31] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[32] E. Alpaydin, "Combined 5 × 2 cv F Test for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 11, no. 8, pp. 1885–1892, 11 1999. [Online]. Available: https://doi.org/10.1162/089976699300016007