Facial Image-Based Automatic Assessment of Equine Pain

Francisca Pessanha, Albert Ali Salah, *Senior Member, IEEE* Thijs van Loon, Remco Veltkamp

Abstract—Recognition of pain in animals is essential for their welfare. However, since there is no verbal communication, this assessment depends solely on the ability of the observer to locate visible or audible signs of pain. The use of grimace scales is proven to be efficient in detecting the pain visually, but the assessment quality depends on the level of training of the assessor and the validity is not easily ensured. There is a clear need for automating the pain assessment process. This work provides a system for pain prediction in horses, based on grimace scales. The pipeline automatically determines the quantitative pose of the equine head and finds facial landmarks before classification, proposing a novel scale-normalisation approach for equine heads. The pain estimation is achieved for each facial region of interest separately, following the clinical pain estimation procedure. We introduce a database of horse images, annotated by professional veterinarians for training and assessment. We also propose a data augmentation method to alleviate the data scarcity issues, which relies on generating realistic 3D equine face models based on 2D annotated images. We show that the data augmentation method improves the performance of both quantitative pose estimation and landmark detection. Our results establish a strong baseline for automatic equine pain estimation.

Index Terms—Pain estimation, animal behavior analysis, horses

1 Introduction

Recognition and quantification of pain in equines are essential to maintain their welfare and improve their convalescence [1]. However, contrary to humans, where verbal communication facilitates pain assessment, in animals, this process depends on the observer's ability to locate and quantify the pain, based on perceptible behaviour changes [2]. In particular, facial analysis is used for pain estimation in horses [3], but also in mice [4], rabbits [5] and sheep [6]. Enabling continuous monitoring of signs of pain in animals is useful for studying disease progression and effects of medication, for objective pain assessment, and to improve the time response of the care-givers, minimising animal suffering and the economic impacts of diseases.

Several frameworks were proposed for horse pain estimation from faces, the most important ones being the Horse Grimace Scale (HGS) [7] and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) [8], [9]. 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 region described. There is an evident necessity for automation, which can also help provide timely information about the animal.

- *F. Pessanha, A.A. Salah, R. Veltkamp are with the Department of Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands.*
  E-mail: f.pessanha@uu.nl
- *T. van Loon is with the Veterinary Faculty, Utrecht University, the Netherlands.*
- *A.A. Salah is also with the Department of Computer Engineering, Boğaziçi University, Turkey.*

The primary aim of this work is the development of an automatic equine pain assessment system based on facial expressions. The model we propose is robust to different coat colours (such as bay, chestnut, black) and markings, as well as to the existence or absence of a bridle in the equine's head.

This paper is an extension of previous work in equines [10] where a simple qualitative pose classifier was combined with an automatic landmark detection system for face based pain estimation. However, the data scarcity is a significant problem for automatic approaches. In this work, we improve the classification pipeline, and propose a 3D-based synthesis module to create larger training sets. Given a single horse face with pain indicators, we are able to synthesize a 3D face and produce many more 2D appearances with different textures (i.e. coat colouring), obtained from other horses in our database. Furthermore, we use transfer learning to leverage earlier work on sheep pain estimation. We focus on images, rather than videos, which are more informative in pain dynamics, but frame-level analysis is essential for video processing as well.

In sum, the main contributions of the present work are:

- We introduce a unique horse dataset with manually annotated landmarks and detailed, feature-level pain score ground truth, given by a veterinarian expert.
- We provide a hierarchical system for automatic pain prediction on equine faces from images.
- We introduce novel methods for accurate head pose estimation and scaling for equines.
- We show that multiple models should be trained in parallel for different poses of the animal's head.
- We improve both pose and landmark estimation with synthetic data generated with a 3D horse model.
- We illustrate the benefit of transfer learning for lever-

aging early work on sheep pain estimation for equine pain estimation.

2 Computer vision based assessment of pain from animal faces

Automatic assessment of subjective states in animals requires the recording of related behavioural and internal indicators via sensors, and computational modeling to link these observations to a target variable, which will ideally be a validated clinical measurement of the state. More elaborate models will incorporate more information, including for instance representations of the context of the behaviour, or acquire and integrate signals from multiple modalities. Pain is defined as "an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage," [11]. While it is a subjective experience, it causes visible signs of distress in animals, which can be used to infer its presence to a certain extent in the absence of verbal communication.

Manual pain assessment in animals requires clinical expertise, can be time-consuming, and human assessment can introduce biases. A computer vision based approach for pain assessment is appealing, because even if it is not as accurate as humans, it can be used for pre-screening animals, for long-term observations, and for quantification of certain observed indicators.

In this section, we focus on behavioural expressions of pain, and on computer vision based analysis methods that try to estimate whether an animal is in pain or not, automatically, from their facial appearance. Different species will pose different challenges for computer vision approaches, and consequently require different methods. For example, in horses, different skin colours and the possible presence of a bridle cause issues, which are absent for pain estimation in, say, mice. We first shortly describe some important related work on pain assessment from human faces, which is a much broadly studied problem in affective computing, and inspired models for animals. We then describe approaches for detecting and quantifying pain via facial expressions in other animals, like mice and sheep. Finally, we focus on specific challenges of assessing pain in horses.

2.1 Pain in human faces

A lot of works that look at animal faces for pain indicators are inspired by decades of work on human facial analysis. For objective measurement of facial expressions of humans, the Facial Action Coding System (FACS) was developed to describe movements of facial muscles, in terms of facial action units (AUs) [12]. In a similar spirit, coding systems were developed for other animals, such as EquiFACS for equines [13], and CatFACS for cats [14].

In humans, verbal communication facilitates the assessment of pain. Nevertheless, some circumstances like severe illness, speech impediments, or other communication issues (including deception) may hinder verbal reporting. These have motivated the design of pain assessment scales based on human facial expressions [15], [16], [17], [18]. Research found that closed eyes, raised cheeks, wrinkled nose, lowered brow, raised upper lip, and parted lips are some examples of facial expressions associated with pain [19].

Computer vision methods have been explored in the literature for automatic assessment of facial pain expressions. In one of the early works, neonatal facial expressions were used to detect pain [20]. Using computer vision based analysis of AUs, Bartlett et al. were able to automatically detect whether pain expressions were real or faked [21]. Videos are typically more informative than single images for pain estimation, as they provide the possibility to leverage spatio-temporal cues [22], [23]. However, many video-based approaches initially used image-based analysis at the frame level [24]. More recent deep neural network approaches can be directly trained on videos, but require much larger training sets [25].

One of the main challenges of this field is collecting large amounts of data, due to the ethical implications of recording pain. Lucey et al. introduced the influential UNBC-McMaster pain database of patients suffering from shoulder pain doing range-of-motion tests [17], which spurred a range of computer vision based approaches for pain estimation. This research also illustrated an important issue in automatic pain analysis; the participants were suffering from different causes of pain, including "arthritis, bursitis, tendonitis, subluxation, rotator cuff injuries, impingement syndromes, bone spur, capsulitis and dislocation", and over half of them were using medication. These differences are difficult to assess just from facial expressions, and often, the automatic analysis looks at a simple indicator, such as pain vs. no pain.

The most common practice found in "pain datasets" consists of inducing pain in healthy individuals [15], [16] or resorting to recording posed pain expressions [26]. These practices affect the generalisation of the resulting pain models, since the models end up being mainly trained on healthy adult faces. Furthermore, the characteristics of pain expression will be different when comparing acute pain due to a short stimulation and chronic pain. A comprehensive survey that collects over 100 methods for human facial pain estimation is given in [27].

2.2 Pain in animal faces

Grimace scales to analyse pain from animal faces were developed for several species. An early work for semi-automatic analysis of animal faces was [28], where rat faces were automatically detected and cropped from videos with a computer vision based approach inspired by face detection models for humans. However, pain assessment was done by humans, following the Rat Grimace Scale. Fully automatic approaches for pain detection on animal faces are fairly recent, and their target variable is either a binary annotation (e.g. "pain" vs. "no pain"), or based on a pain scale. An example is action unit based estimation of sheep pain [29], using Sheep Pain Facial Expression Scale (SPFES) [6].

Modern computer vision pipelines often use convolutional neural networks (CNN), which are powerful, but require large datasets for training. One CNN approach was used in [30] for detecting "pain" or "no pain" from faces of mice, on a dataset with 5771 images. Broomé et al. [31] used a Convolutional Long Short Term Memory (LSTM) method to simultaneously process the spatial and temporal features on horse faces from video. The model predictions surpassed

the expert performance, but the performance had a high variance across subjects. This was partially due to the small size of the dataset, which only contained six horses. More recently, hourglass-shaped models were assessed to provide self-supervision to deal with the sample size problem [32]. Further work explored the possibility of domain transfer between different pain types in horses, in particular, the potential of transferring features from a dataset of horses with acute pain (which is less ambiguous) to help the assessment of prolonged or more complex pain [33].

A complete pipeline for pain estimation for sheep faces was proposed in [34], combining a fine-tuned model for face detection, with a CNN-based pose estimation system, followed by facial landmarking, which is the detection of anchor points on faces to simplify subsequent analysis. Histogram of Oriented Gradient (HOG) features, as well as geometric features and the pose values were used to train a binary support vector machine (SVM) classifier, adapted to different head rotations for dealing with self-occlusions. In this paper, we follow a similar pipeline and add pose estimation as a step before facial landmarking.

Horse images pose specific challenges for visual processing, such as the high variations in colour and overall appearance between individual horses and between breeds. Following the approach introduced by Mahmoud et al. [29], previous work on horses suggested a classification model based on a combination of features, namely edges, colour histograms and HOG [35]. However, the extracted features were not sufficiently discriminative to achieve a satisfactory performance. Additionally, pose variations affected the performance significantly, with self-occlusion being an aggravating factor.

Our early work in pain estimation in equines investigated extracting HOG, scale invariant feature transform (SIFT), local binary pattern (LBP) and deep neural network based features, and combining them with SVM classifiers [10]. In this approach, a grimace scale was used to score pain levels of facial regions-of-interest (ROI) separately (see Table 2 for this image based assessment, which will also be used in this work). The total pain score was a combination of these indicators. Additionally, a three class HOG-SVM based head pose classifier ("frontal", "tilted," and "profile", respectively) was introduced. Knowing the head pose can improve both landmark detection and pain estimation, since the face appearance varies widely with the pose. In [36], three separate CNNs were used to assess three regions on the horse's face (ears, eye, and mouth and nostrils, respectively), which avoids pose related difficulties to a certain extent. Andersen et al. offer an extensive analysis of the challenges of machine recognition of facial expression of pain in horses [37].

Table 1 summarizes recent datasets on automatic facial pain estimation in different species. Some of these are based on images, and some on videos. As stated before, videos are more informative for pain estimation, but the processing of videos requires more computational resources, and typically makes use of image-based approaches at the frame level. Another advantage of videos is that a single frame may not contain many indicators of pain, and it may be necessary to extend the analysis to frames collected over a period to provide improved analysis [38]. Pain datasets are particularly challenging to create, due to the ethical issues associated with the induction of pain. In datasets where the pain stimulus is known, typically the pain is caused by a particular disease or surgical procedure, or it is induced moderately, without irreversibly damaging the animals. The first scenario is preferable, since the pain expressions will be comparable with the ones found in "real world" (i.e. in-the-wild) situations and it avoids hurting animals for experimental purposes. To the best of our knowledge, the dataset introduced in the present work (see Section 3) is the most extensive collection with grimace scale annotations for automatic pain estimation in animals.

2.3 Quantifying Facial Pain in Equines

Veterinarians can quantify the existence of facial pain indicators in equines using certain clinically validated scales. In these scoring systems, various pain states are described based on audio-visual cues. We base our automatic assessment in this paper on such scales.

Dalla Costa *et al.* proposed the Horse Grimace Pain Scale (HGS) for pain assessment in horses undergoing castration, based on still images extracted from video recordings [7]. This procedure is performed routinely, with studies showing evidence of acute and chronic pain after the procedure. The castration post-procedural pain is rated from mild to severe [43], which is reflected on the HGS. In contrast, abdominal disorders like obstipation or strangulation very often lead to severe pain, and are frequently diagnosed in horses [44]. Subsequently, creating tools for identifying colic pain is valuable for the quality of patient care and overall equine welfare.

van Loon *et al.* proposed the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) for horses suffering from acute colic [9]. While most of the indicators in this scale are based on images, sound and video dynamics are also used to identify certain indicators, such as the amount of head movements, focus of the horse, and the flehming behaviour, which is when the horse bares its upper front teeth for a short duration and inhales, showing a spiteful appearance. For behavioural assessments, EQUUS-FAP can be complemented with other indicators [45].

The dataset introduced in this paper is composed of still images, and subsequently excludes analysis of movement and sound-based features. This limits the usage of EQUUS-FAP. We employ here a practical grimace scale called Equine Utrecht University Scale for Automated Recognition in Facial Assessment of Pain (EQUUS-ARFAP) [35] that has been created by (some of) the original creators of the EQUUS-FAP, and combines all the static features of the EQUUS-FAP and HGS systems into a single list of six features (see Table 2 and Figure 1). While not a clinically validated tool per se, this scale produces useful indicators for pain states. The limitations are that deeper insight into pain is lacking, and it is not possible to infer duration and intensity of pain expressions, unless the method is applied to sequences.

3 The UU Equine Pain Face Dataset

The UU Equine Pain Face Dataset consists of a total of 1855 images of horses and 531 images of donkeys, respectively.

TABLE 1: Summary of datasets containing facial expressions of pain used for automatic animal pain estimation

Reference	*Animal*	*Pain stimulus*	*Data*	*Annotations (labels)*
Pessanha et al. [34]	Sheep	Mastitis and pregnancy toxaemia	86 individual frames; 4 sets of videos of pain evolution	Sheep Pain Facial Expression Scale (SPFES)
Noor et al. [39]	Sheep	Unknown	2350 images of sheep	”Normal” or ”Abnormal”
Tuttle et al. [30]	Mice	Laparotomy sham surgery	5771 unique images (2444 ”pain” and 3327 ”no pain”)	”Pain” or ”No Pain”
Andresen et al. [40]	Mice	Castration	Recordings of 124 unique animals over time	”Post-anesthetic/surgical effect” vs. ”No effect”
Broomé et al. [31]	Horses	Moderate induced pain	60 videos of 6 different horses	”Pain” or ”No Pain”
Ask et al. [41] and Broomé et al. [33]	Horses	Induced orthopaedic pain	90 videos of 7 different horses	Composite Pain Scale (0 - 39)
Hummel et al. [42]	Equine	Partly induced, partly unknown	1854 images of horses; 531 images of donkeys	Adapted EQUUS-FAP
Lencioni et al. [36]	Horses	Castration	3000 images from 7 different horses	“No pain present”, “Moderate pain”, “Obvious pain”

TABLE 2: Score sheet for facial pain score assessment in still images [46].

Data	*Categories*	*Score*
Ears	Both ears 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 tightening of the eyelids	1
	Obviously tightening of eyelid / eye 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 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***

Fig. 1: Example images of the pain score sheet used in the present work.

The images focus on the face region, but have different poses, with different facial landmark visibility. It is important to note that in horses, face pose has a greater effect on facial landmark visibility compared to human faces, and both the database and the processing methodology will reflect this. All images in the database have pain scores and landmarks annotations following the criteria described in Section 2.3. The data comes from three sources (see below, and Figure 2), but except for our preliminary work, these data were not published for computer vision analysis. We re-purpose the data, providing landmarks and ground truth annotations. We describe each subset separately, and only use the horse subsets in the present study.

- **Horses from the Netherlands - HFN** *(1520 images)*: Images provided by horse owners all over the country. The photographs are very diverse, showcasing an extensive set of backgrounds, breeds, and image

resolution. 873 images have bridle, 647 do not.

- **Horses with clinically induced injuries - HWI** *(334 images)*: Images collected as part of a project running in the Faculty of Veterinary Medicine of the Utrecht University, where clinical procedures were applied to create reversible lameness and pain [47]. The study design and experimental protocol to induce acute orthopaedic pain on a set of horses under controlled conditions were approved by the Ethics Committee on the Care and Use of Experimental Animals in compliance with Dutch legislation on animal experimentation (permission number: AVD1080020153307WP16), and pictures were taken in different time-periods. The images have similar backgrounds with comparable illumination and resolution. There are multiple pictures of the same horse with different head poses and with several timestamps.
- **Donkeys from a Donkey Sanctuary - DFS** *(531 images)*: Images provided by a donkey sanctuary in the UK, with multiple photographs per donkey. These are not used in the present work.

During the following sections, the HWI and HFN subsets will be used as a combined dataset for horse pose estimation and landmark detection.

3.1 Landmark ground truth annotations

We follow the landmark annotation scheme from [35], which described the head shape and facial features in great detail. In this annotation method, three different landmark schemes were established considering a qualitative evaluation of the head pose. These landmarking schemes use 54, 44 and 45 points for frontal, tilted and profile views, respectively (see Figure 2). The landmarking was completed with a follow up work [10] and the landmarks are publicly available.

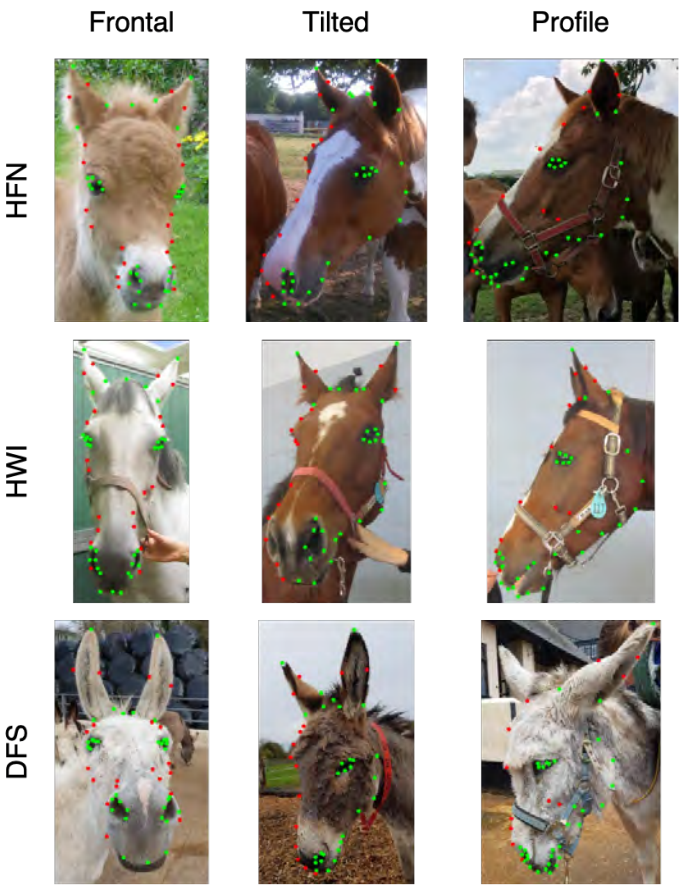


Fig. 2: Faces extracted from each subset with marked points of interest. First column - Frontal view; Second column - Tilted view; Third column - Profile view; Green points indicate the stable landmarks; Red points indicate the relative landmarks. HFN - Horses from the Netherlands; HWI - Horses with clinically induced injuries; DFS - Donkeys from a Donkey Sanctuary.

3.2 Pose distribution

We used a weak perspective projection method to define the quantitative head pose ground truth. We divided the landmark annotations into two sets. The first set is called "stable landmarks", and contains the points whose position doesn't change with rotation, and the second set is called "relative landmarks", which are outline landmarks that will change with pose variations (Figure 2). We only used the first set to estimate the head pose. Since the images are cropped, the camera intrinsic parameters have limited use in solving the pose automatically from landmark positions. For this reason, all observations were checked manually, and images with uncertain poses were excluded from training. The resulting dataset has 370 frontal images, 952 tilted images and 348 profile images.

The *yaw* values are restricted to the $[-25, 25]$ degrees range for the "frontal" class, and have an absolute value predominantly in $[50, 90]$ degrees range for the "profile" class. The "tilted" class overlaps with these classes, with *yaw* values ranging from an absolute value of 10 to 75 degrees. The *roll* and *pitch* distributions are similar in all classes, with higher variance in the profile class. Some of this variance may be due to noise, as we have less landmarks for pose assessment in the profile images, where the 2D-3D correspondence is more difficult to assess.

3.3 Pain annotations distribution

The images were annotated for potential signs of pain by expert raters according to the adapted EQUUSFAP scale presented in Table 2. Three distinct raters (one senior expert researcher, and two graduate students in the veterinary masters program, trained by the senior expert) scored the entire dataset according to the previously mentioned scale, using full images. In developing our pipeline, we used the pain score annotations of the senior research and we present the distribution of these annotated pain scores in Figure 3, where we observe that the dataset is not completely balanced, with very few instances of class "2" in all regions-of-interest. The predominant class will alternate between "0" and "1". We further note a few important issues in this figure. The corners of the mouth are not always visible, and subsequently have fewer annotations than the rest of the facial areas. Most importantly, the distribution of the scores

per area are not completely aligned. This points out to a fundamental difficulty in equine pain assessment.

Note that the used grimace scale (adapted EQUUS-FAP) is not explicitly head-pose dependent, except in the case of occluded regions of interest. As a result, a 3-fold cross-validation set and a separate test set were defined, maintaining the proportion of pain score combinations and quantitative *yaw* values in each subset. The complete training set (i.e. all three validation folds) contains 259 frontal faces, 666 tilted faces, and 243 profile faces, and the test set contains 111 frontal faces, 286 tilted faces, and 105 profile faces.

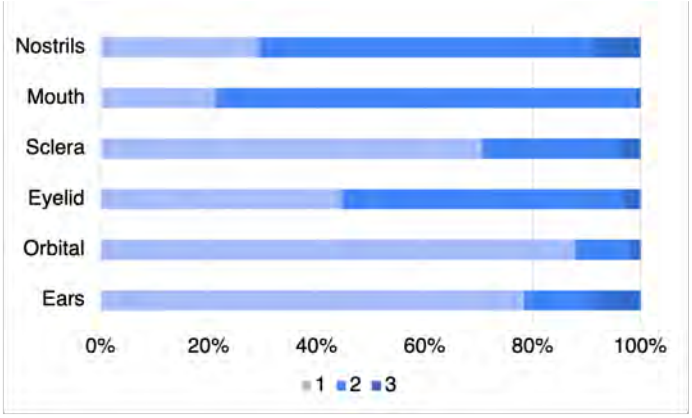


Fig. 3: Distribution of the pain scores in the dataset.

4 METHODS

Figure 4 illustrates the automatic pain estimation pipeline. One of the contributions of the present work is a data augmentation approach that we use in training. We first discuss this approach, followed by the individual steps of the estimation pipeline, namely, pose estimation, landmarking, and pain estimation, respectively.

4.1 Data augmentation

Part of the difficulty of pain estimation in equines comes from data scarcity. In this section, we propose an approach for data augmentation, based on 3D modeling of the horse face, to address this issue, where each existing horse face in the training set is used to provide more images in different poses.

There is a vast amount of 3D resources for synthesising human faces, with data collected via multi-view stereo cameras and commercial depth sensors. Several approaches addressed how these resources can be leveraged for facial landmark detection on 2D images [48], [49]. These approaches typically require 3D ground truth or a pre-trained 3D morphable model. Unfortunately, when working with animal faces, such 3D resources are not readily available. Collecting 3D-scans from an animal is more difficult and the applications are perhaps more restricted, compared to the work on human faces. 3D models available in datasets such as TOSCA [50] have limited realism. At the moment, there are no realistic, parametric, publicly available and flexible 3D horse face models that can be used to synthesize large amounts of data.

In this paper, we follow an approach that combines a single 2D image and a generic 3D model. Cashman *et al.* used images for 3D modeling of animals [51]. In their work, they defined a 3D morphable model based on a set of images, with silhouette and landmark annotations. Given an initial 3D model and a set of images of the same class, a deformed 3D model was created. Building on these ideas, Kanazawa *et al.* improved the deformation strategy by considering the local stiffness of each area, specific for the class [52]. The final model tried to minimize the deformation energy, the location variation between the points in the 3D and 2D and the local stiffness, only distorting the less stiff "tets" (i.e. a tetrahedron of the mesh). More recent work by Zuffi *et al.* produced a small dataset based on 3D scans of toy figures in arbitrary poses, and, after pose normalisation, learned a statistical shape model to fit a combination of 2D keypoints and 2D silhouettes [53]. Texture transfer was not implemented in any of the previous studies. Furthermore, the final shape was a rough estimate of the silhouette and not a direct point-to-point match between the image and the projection.

In this paper, we propose a textured 3D horse head generation system to augment the training data for pose estimation and landmarking. Our approach works with a simple 3D horse head model, which is not adequate for generating pain expressions, which are too subtle. In future, if more detailed 3D models become available, the proposed method can be extended to generate pain expressions as well.

In our approach, we use profile images and corresponding landmarks to deform a pre-existing 3D horse head model from the TOSCA dataset (see Figure 5). Assuming that the horse head is approximately symmetrical, the occluded side will have a similar shape and texture as the visible side.

For this purpose, we annotated the 3D model with the same landmarking system as the profile faces, using the stable landmarks. The training data we augmented already had manually annotated landmarks, but matching the 3D model automatically to these 2D images was not straightforward. The contour landmarks were further processed to prevent alignment issues in the texture transfer approach. First an edge detector was used to correct the position of the contour landmarks, which were replaced with the closest edge points. Next, we estimated the quantitative head pose based on the 3D-2D point correspondence. Considering a field-of-view of 60 degrees to define the focal length, we used an iterative approach based on Levenberg-Marquardt optimization [54], [55] to solve the Perspective-n-Point problem. The resulting rotation matrix, R, and translation vector, t, allowed the projection of the 3D model points onto the image plane:

$$\begin{pmatrix} wx \\ wy \\ w \end{pmatrix} = K[R|t]\mathbf{X} = K\mathbf{X}_{cam} \tag{1}$$

with (x, y) corresponding to the pixel coordinates of the world point $\mathbf{X}$ projected onto the image.

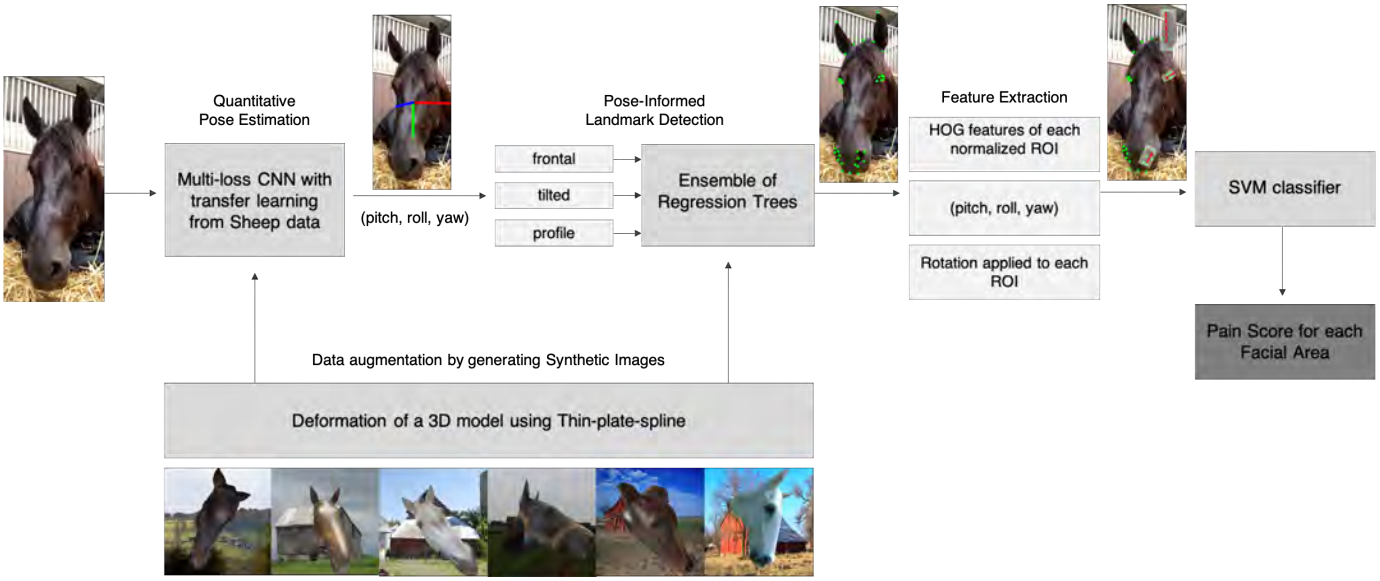


Fig. 4: Proposed pipeline for horse pain estimation based on facial features.

Fig. 5: Full horse and head model with respective axes [50] (red: *x axis*, green: *y axis*, blue: *z axis*).

Assuming the w parameter of each ground truth landmark projection in the image is the same as the one of the corresponding projected 3D point, the image point in the camera frame is:

$$X_{cam} = K^{-1} \begin{pmatrix} wx \\ wy \\ w \end{pmatrix} \tag{2}$$

Next, we projected the landmarks according to the symmetry plane and applied a Thin Plate Spline approach [56] to deform the 3D model initialised according to the ground truth landmarks and their projections. This method interpolates surfaces over scattered data by having a fixed set of nodes in the plane (in this case, the landmarks), and minimizes the bending energy by warping the surface to fit the ground truth. The landmark points are warped exactly to fit the targets, and the rest of the points are interpolated according to their distance to the landmarks.

After deforming the 3D face model, we obtain the correspondence between the vertices in both sides of the symmetry plane for further texture transfer from the 2D image to the 3D model. By projecting the final 3D model onto the image, we define a mean colour for each triangle. This value corresponds to the average between the vertices and the centroid colour, and we attribute the same colour to symmetrical triangles that are not visible in the 2D image. We save the resulting colour map, allowing the interchange of texture between the different horse head shapes in the subset (see Figure 6). In total, we collected 29 different colour maps, from which horse faces can be synthesized in any pose.

To generate synthetic images with a common reference, the 3D points were converted to the world frame:

$$\mathbf{X} = [R|t]^{-1}\mathbf{X}_{cam} \tag{3}$$

The rotation matrix, R, was then defined as:

$$R(\alpha, \beta, \gamma) = R_z(\alpha)R_y(\beta)R_x(\gamma) \tag{4}$$

with the *yaw* being the counterclockwise rotation of α about the *y*-axis, the *pitch* being the counterclockwise rotation of β about the *x*-axis and the *roll* being the clockwise rotation of γ about the *z*-axis.

Finally, the background was replaced with images extracted from *Flickr* under the tags "farm", "field", "barn" and "stable" (see Figure 7) to complete the 2D image synthesis.

4.2 Quantitative pose estimation

Head pose variations cause evident changes in the facial appearance of equines due to self-occlusion. For this reason, head pose-specific methods for landmark detection and further pain estimation should be preferred, allowing for a better description of the areas of interest visible from a specific pose. We divide the training set into pose bins, and train a regressor for estimating the pose.

To estimate the quantitative head pose, we use a multi-loss convolutional neural network for head pose estimation [57]. This approach combines a ResNet50 architecture [58] with the Mean Squared Error and Cross Entropy Loss.

Fig. 6: Examples of deformation and further texture transfer of the 3D model. The images on the first row are the originals, and the images on the second row are the synthesized image.

Fig. 7: Examples of synthetic images produced using the described method. Examples of different head shapes with the different head pose with the same texture are presented, showing the effects of texture transfer.

We used transfer learning by initialising our models with models pre-trained on more extensive data collections. Since early layers of deep neural networks learn basic features, using pre-trained models from other tasks helps in reducing the required samples for training. We experimented with 1) a model trained on the 300W-LP dataset on human faces, and 2) a model trained on a sheep facial dataset. Although horses and sheep have significant anatomical differences, they pose similar challenges for pose estimation, such as an elongated nose, which led us to expect a better performance from the second approach.

For data augmentation, we pay attention to the distribution of images for different ranges of yaw, as this is the most important parameter that affects landmark visibility. Different amounts of data augmentation were tested, introducing synthetic images on the training set. We performed 3-fold cross-validation with an un-augmented validation set. Since there are some differences in appearance between the synthetic and real images, it is desirable to maintain a high percentage of real images in the training set to avoid overfitting to the visual appearance of synthetic images. Subsequently, the number of synthetic images generated per yaw angle bin (n_{aug}) was defined as the number of images in the training set in that bin (n_{bin}), multiplied by an augmentation factor determined as a function of the maximum number of images in a pose bin in the training set (n_{max}). The pose bins with the maximum number of images will not be augmented. The data was augmented according to the *yaw* representation in the training set:

$$n_{aug} = n_{bin} \times [(\frac{n_{max}}{n_{bin}})^{\alpha} - 1] \quad (5)$$

4.3 Facial landmarking

For facial landmarking, we compared two state-of-the-art landmark detection algorithms, namely, Ensemble of Regression Trees (ERT) [59] and Supervised Descent Model (SDM) based on SIFT features [60], respectively. These were formerly proposed for detecting landmarks on human faces. Additionally, a mean shape model was calculated for each pose class based on the training set shapes.

Measuring landmarking accuracy is difficult because of scale differences in the images. In the human facial landmark localisation literature, automatically located landmarks within 10% of the inter-ocular distance to the ground truth location are considered to be accurate [61], [62]. Since there is not a standard normalisation factor for landmark analysis in horses, we proposed the use of the distance between the centre of an eye and the centre of the underlying nostril. These two features are present in every pose of the horse face (unlike the two eyes), and their distance is sufficiently long to make it robust against errors. The 10% threshold in human facial landmarking designates an error margin that will result in a good alignment, and no landmark overlaps. With the same concern, we empirically define a threshold of 6% for the eye-nostril distance [10]. The performance measures we use for landmarking are the Mean Normalised Error (MNE), corresponding to the Euclidean distance between the prediction and the ground truth normalised by the eye-nostril distance, and the Success Rate (SR), referring to the percentage of predictions with a distance lower than 6% of the eye-nostril distance. Still, variations in pose (especially, changes in *pitch*) will distort the eye-nostril distance, and if dealing with a dataset with

diverse breeds, the nose length in proportion to the face will vary.

To assess the landmarking algorithms, we assume that the face detection is correctly performed. The horse faces were cropped based on the ground truth landmark locations and resized according to the face proportions in each of the three classes (i.e. "frontal", "tilted" and "profile") of the training set. For each pose bin, the height of the face was scaled to 600 *px*, with the width being defined according to the training set aspect ratio. This resulted in 600×270 "frontal" images, 600×330 "tilted" images and 600×380 "profile" images. Furthermore, we rotate all faces according to the absolute *yaw* angle.

We use the Mempo project's python package for implementing the ERT and SDM models and to introduce uniform perturbations in each bounding box for data augmentation [63]. We performed 3-fold cross-validation to adjust the number of perturbations to apply in each pose class. We use only the "stable landmarks" to determine the optimal number of perturbations. These landmarks do not change with the head position (i.e. the outline landmarks are excluded, as rotating the head changes the absolute position of the head contour). After the simple geometric data augmentation, we also use synthetic images to augment the training set further, and to have a balanced set of poses across the training set. This is particularly important, because we will train pose-specific pain estimators next.

4.4 Pain score estimation

To observe the potential of both appearance and geometric features, we use SVM models trained with the quantitative pose, the local rotation angle of each region-of-interest (ROI), and HOG features [34]. Since there is no clear division between the different qualitative head pose bins, we proposed a unique pose specific model per ROI.

Firstly, we normalised each ROI by rotating the ears and nostrils into a vertical position and the eyes and mouth into a horizontal position. The rotation angles were also used to train the pain classification model. Each ROI was resized based on the mean ratio in the training set, with the longer side set to 128 *px*. We tested different values for the HOG parameters (orientations, cells-per-block, pixels-per-cell) in 3-fold cross-validation, as well as different kernels for the SVM model (linear, RBF and polynomial). Each model was trained with HOG features, angles and head pose from both right and left ROI. In case a ROI was not visible, it was replaced by a $(200, 200, 3)$ zero array, with rotation equal to zero.

As shown in Section 3, the dataset is highly unbalanced. In addition to training SVMs with a balanced set, the performance measures are calculated for each class separately and the average weighted value is reported.

We trained the final model in the complete training set (1168 images) and evaluated it on the test set (502 images). The number of occurrences of each ROI is variable.

5 Experimental Results

5.1 Pose estimation

The performance measures used for pose estimation are the Mean Absolute Error (MAE), calculated in degrees, Pearson's Correlation Coefficient (PCC), which measures the correlation between predictions and the ground truth, and the Signal Agreement (SAGR), which is defined for two vectors x and y of equal length n as [64]:

$$SAGR(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(sign(x_i), sign(y_i)) \tag{6}$$

with $\delta(x, y)$ denoting the Kronecker delta. It is desirable to achieve low MAE and high PCC and SAGR values.

We contrast two versions of the Hopenet [57] model for head pose estimation. The original model was trained on the 300W-LP dataset, and it needs to be fine-tuned for equine heads. We propose to use an initial fine-tuning on the Sheep dataset [34], which is more similar to equine faces than human faces. A mean model is used as the baseline, predicting for each sample the mean angle of the training set.

As expected, there is a significant improvement in the model's performance after transfer learning from the sheep-based model (see Table 3), particularly in the *yaw* values, as the Sheep dataset was augmented for different yaw values. The sheep faces have similar appearance changes as horse faces when changing pose, with similar problems related to the elongated nose and consequent self-occlusion. We suggest that the lower performance for the *roll* and *pitch* angles can be justified by the reduced diversity of these values in the dataset.

Next, we have evaluated the effect of data augmentation. A 3-fold cross-validation was performed to define the ideal number of epochs and α (see Eq. 5). When compared to the results in Table 3 there is an improvement in performance for the target *yaw* angle, with a decrease of MAE and an increase of PCC. Although the MAE for the *roll* and *pitch* angles is similar to the *yaw*, their PCC is significantly lower. This observation can be explained by the smaller range of values for the *pitch* and *roll* (mainly between $[-25, 25]$ degrees instead of $[-90, 90]$ degrees). Considering the diversity of poses in the dataset, and the error associated with the ground truth pose estimation, the presented results are satisfactory, with a high signal agreement and a PCC of 97%.

5.2 Landmarking

The Ensemble of Regression Trees (ERT) model has shown promising results for the landmark localisation, outperforming both the Mean Shape model and the Supervised Descent Model (Table 4). Overall, extreme angles, combined with a lack of representation of these angles in the dataset, resulted in incorrectly located landmarks. Additionally, not all outline landmarks are associated with strong changes in appearance, which leads to deviations in their prediction. It's also important to note that there is a clear performance improvement when applying a train-test split based on the *yaw* angle values, with a decrease in the MNE of around 2.5% in both the ERT and SDM models and an improvement of 0.15 in the success rate of the ERT classifier compared with previously published work [42]. This fact reflects a decrease of the MNE for regions-of-interest (ROIs).

Since the outline landmarks are highly variable and do not have a direct appearance correlation, we trained an ERT

TABLE 3: Quantitative pose estimation results in the test set transfer learning from the model trained on the 300W-LP [49] and the Sheep datasets [34]. Low MAE and high PCC and SAGR values are preferred.

	Model	*Yaw*	*Pitch*	*Roll*
MAE	Baseline	37.24	11.15	9.41
	300W-LP	23.41	12.10	9.63
	Sheep	9.30	9.35	7.17
	Sheep + data aug	8.95	9.83	7.55
PCC	Baseline	0.00	0.00	0.00
	300W-LP	0.76	0.30	0.24
	Sheep	0.96	0.61	0.58
	Sheep + data aug	0.97	0.60	0.60
SAGR	Baseline	0.51	0.70	0.71
	300W-LP	0.82	0.70	0.70
	Sheep	0.85	0.79	0.81
	Sheep + data aug	0.85	0.76	0.83

TABLE 4: Mean Normalised Error (MNE) and Success Rate (SR) using ERT, SDM and a baseline mean shape model for both landmarking systems. Presented values are weighted average results for the test set for the three qualitative pose classes. SR indicates the ratio of landmarks with a location error less than 6% of eye-nostril distance.

	Landmark system	*ERT*	*SDM*	*Mean Shape*
MNE	Relative + Stable	0.061	0.067	0.116
	Stable	0.060	0.063	0.115
SR	Relative + Stable	0.629	0.577	0.236
	Stable	0.637	0.604	0.232

TABLE 5: Mean Normalised Error per region-of-interest (ROI) in the test set with a model trained on stable landmarks. The highest error for each ROI is highlighted. Missing values indicate that the ROI is not defined for that pose class.

ROI	*Data aug.*	*Frontal*	*Tilted*	*Profile*	*Average*
Ears	no	0.067	0.062	0.083	0.067
	yes	0.067	0.62	0.083	0.068
Nose	no	0.069	0.073	0.039	0.065
	yes	0.071	0.071	0.040	0.065
Left Eye	no	0.049	0.031	0.047	0.039
	yes	0.049	0.031	0.043	0.038
Right Eye	no	0.046	-	-	0.046
	yes	0.044	-	-	0.044
Mouth	no	-	0.069	0.037	0.061
	yes	-	0.065	0.037	0.058

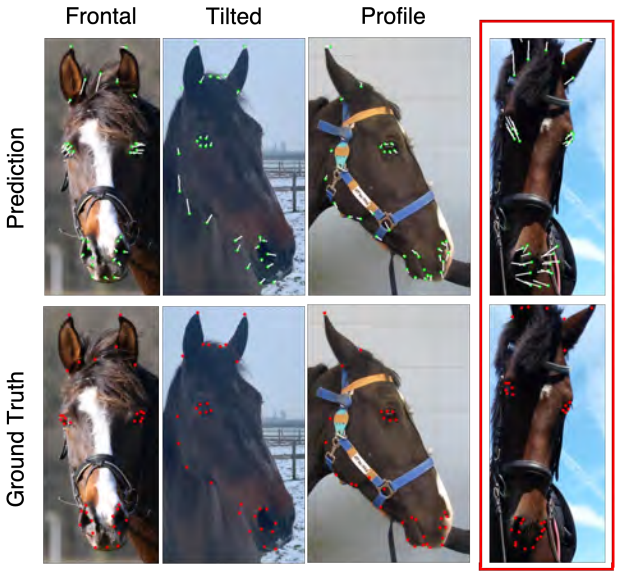


Fig. 8: Examples of ERT-based landmark predictions compared to the ground truth. The last column shows an image with large error due to the ears being cropped in the original image. The white lines connect the predicted point with the ground truth landmark location.

model with solely the "stable landmarks". The results obtained with this landmark scheme were similar to the ones presented for the full landmark scheme (Table 4). This also enabled the use of synthetic images for data augmentation. There is significant MNE difference between poses. We used a 3-fold cross-validation to define the augmentation factor, applying inverse data augmentation based on the MNE in each fold. Then, a full model was trained using the complete training set and augmenting the data according to the average MNE per *yaw* bin in the cross-validation (Table 8). Small to moderate performance improvement is observed for most of the ROIs, illustrating the potential of the data augmentation method. Having three separate landmarking systems based on the qualitative pose annotation is not ideal. There are errors related to manual pose annotation, and ambiguity of images in pose transition areas is an issue. However, a single landmarking system that will work on all poses, and with varying numbers of landmarks, will be more complex to design and train. The strong structural constraints of the ERT models may affect the performance when landmarking less represented horse breeds, for which the facial proportions can vary widely. Lastly, regarding the ears, the lack of "anatomical" points associated with the annotations for the base of the ears, make these landmarks particularly difficult to assess, especially for the "tilted" and "profile" poses.

5.3 Pain scores estimation

In the dataset, the score "2" has much fewer examples compared to "1" and "0". We present the performance of the best model for each ROI in both the 3-class pain estimation task and the binary pain estimation task combining class "1" and "2" into one class, in Table 6. Note that the problem of unbalanced data is not solved entirely with binarisation.

TABLE 6: Performance of the 3-class and binary pain estimation models. The performance metrics are weighted according to the number of samples of each class. The last column corresponds to the weighted F1-score for a majority class classifier (baseline). The highest F1-score value for each classification is highlighted.

	n. classes	*Precision*	*Recall*	*F1-Score*	*Baseline*
Ears	3	0.65	0.74	**0.68**	0.66
	2	0.72	0.74	**0.72**	0.66
Nostrils	3	0.50	0.53	**0.52**	0.48
	2	0.58	0.59	**0.58**	0.58
Orbital	3	0.79	0.83	0.81	**0.83**
	2	0.81	0.85	**0.83**	0.83
Eyelid	3	0.44	0.46	**0.45**	0.33
	2	0.49	0.50	**0.50**	0.37
Sclera	3	0.59	0.62	0.60	**0.61**
	2	0.60	0.60	0.60	**0.61**
Mouth	3	0.61	0.65	**0.62**	0.61
	2	0.65	0.68	**0.66**	0.63

Furthermore, the F1-score of the underrepresented class is significantly lower than that of the majority class, suggesting that there is room for improvements.

While the pain score estimation results are very promising, it is clear that significant challenges still exist. There is an overall ambiguity in classifications, with significant disagreement among experts (Figure 9), and a noticeable data imbalance, even after combining the pain classes. Lastly, the dataset has a lot of different breeds with different face morphology and proportions, which makes the comparison between facial features more difficult. As an example, different breeds can have very distinct nostril shapes, which makes it difficult to assess whether they are "relaxed" or "open" when using a mixed dataset.

6 Conclusion

In this paper, we have provided a unique image based equine pain dataset with feature-level expert annotations, and implemented a complete system to provide a strong baseline for automatic estimation of pain indicators.

Automatic landmark detection is an important step to identify the regions of interest presented in the grimace scale. Yet, variations in the head pose, in particular, the *yaw* angle, will lead to significant changes in visibility and overall head silhouette. We show that a CNN-based quantitative pose estimation system can be used to deal with this issue. For dealing with scale normalisation of horse facial landmarks, we have proposed a novel head-nostril distance.

To deal with data sparsity, as well as varying coat coloring in horses, a novel data augmentation system was proposed, deforming a simple 3D-horse head model according to 2D landmarks with texture transfer from the images. This allowed the creation of diverse synthetic images with precise landmarking and known pose and after data augmentation, the CNN-based pose estimator achieved a high performance and decreased the error in the majority of regions-of-interest. Lastly, a pain estimation system was developed, introducing an SVM model for each region-of-interest trained based on geometric features, the head pose and the HOG features extracted from the bounding box defined by the landmarks.

Potential sources of error are the subtle appearance associated with several landmarks, especially near strong edges and the limitations coming from the 3D model. Additionally, shape constraints of the model may be too strong for the variations in head morphology observed in the dataset due to different breeds. Clearly, more labeled data will help to improve the image-based system, and going to video analysis will provide more visual evidence, along with possibilities of evaluating sounds and dynamics. The results presented in this paper advance the state of the art in horse pain estimation.

References

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

[2] FH Ashley, AE Waterman-Pearson, and HR Whay. Behavioural assessment of pain in horses and donkeys: application to clinical practice and future studies. *Equine veterinary journal*, 37(6):565–575, 2005.

[3] Karina B Gleerup, Björn Forkman, Casper Lindegaard, and Pia H Andersen. An equine pain face. *Veterinary anaesthesia and analgesia*, 42(1):103–114, 2015.

[4] Dale J Langford, Andrea L Bailey, Mona Lisa Chanda, Sarah E Clarke, Tanya E Drummond, Stephanie Echols, Sarah Glick, Joelle Ingrao, Tammy Klassen-Ross, Michael L LaCroix-Fralish, et al. Coding of facial expressions of pain in the laboratory mouse. *Nature methods*, 7(6):447–449, 2010.

[5] Stephanie CJ Keating, Aurelie A Thomas, Paul A Flecknell, and Matthew 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):e44437, 2012.

[6] Krista M McLennan, Carlos JB Rebelo, Murray J Corke, Mark A Holmes, Matthew C Leach, and Fernando 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.

[7] Emanuela Dalla Costa, Michela Minero, Dirk Lebelt, Diana Stucke, Elisabetta Canali, and Matthew C Leach. Development of the horse grimace scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS one*, 9(3):e92281, 2014.

[8] Johannes PAM van Loon and Machteld 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. *The Veterinary Journal*, 206(3):356–364, 2015.

[9] Johannes PAM van Loon and Machteld C Van Dierendonck. Monitoring equine head-related pain with the Equine Utrecht University Scale for Facial Assessment of pain (EQUUS-FAP). *The Veterinary Journal*, 220:88–90, 2017.

[10] Hilde Hummel. *Analysing Horse and Donkey Faces for Measuring Pain Expressions (MSc Thesis)*. Universiteit Utrecht, 2020.

[11] Srinivasa N Raja, Daniel B Carr, Milton Cohen, Nanna B Finnerup, Herta Flor, Stephen Gibson, Francis Keefe, Jeffrey S Mogil, Matthias Ringkamp, Kathleen A Sluka, et al. The revised IASP definition of pain: Concepts, challenges, and compromises. *Pain*, 161(9):1976, 2020.

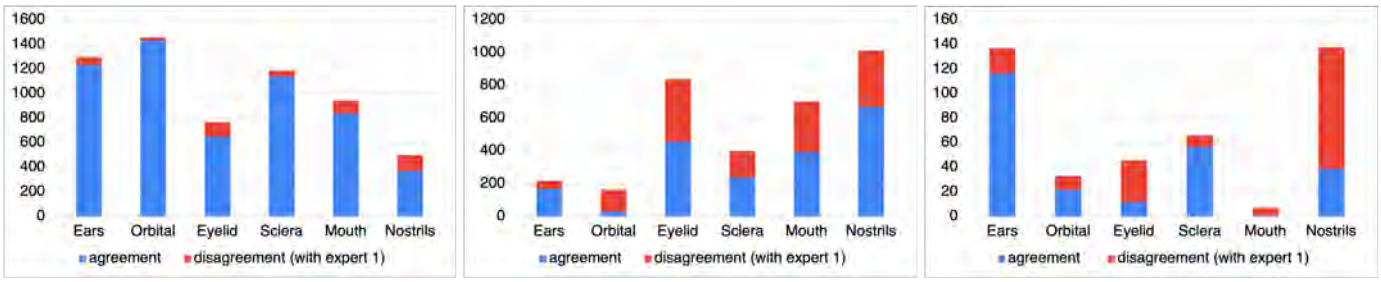


Fig. 9: Agreement in the annotations made by three specialists in 1655 images of horse faces. The annotations from "expert 2" and "expert 3" were compared with the ones from "expert 1", used to train the models. From left to right: agreement for score "0", agreement for "1" and agreement for "2".

[12] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[13] J. Wathan, A. M. Burrows, B. M. Waller, and K. McComb. EquiFACS: The equine facial action coding system. *PLoS ONE*, 10(8):1–35, 2015.

[14] Catia Correia Caeiro, Anne M Burrows, and Bridget M Waller. Development and application of catfacs: Are human cat adopters influenced by cat facial expressions? *Applied Animal Behaviour Science*, 189:66–78, 2017.

[15] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *Proc. IEEE CYBCO*, pages 128–131, 2013.

[16] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[17] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Proc. IEEE FG*, pages 57–64, 2011.

[18] Thomas Hadjistavropoulos, Keela Herr, Kenneth M Prkachin, Kenneth D Craig, Stephen J Gibson, Albert Lukas, and Jonathan H Smith. Pain assessment in elderly adults with dementia. *The Lancet Neurology*, 13(12):1216–1227, 2014.

[19] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Trans. on Affective Computing*, 2019.

[20] Sheryl Brahnam, Chao-Fa Chuang, Frank Y Shih, and Melinda R Slack. Machine recognition and representation of neonatal facial displays of acute pain. *Artificial intelligence in medicine*, 36(3):211–222, 2006.

[21] Marian Bartlett, Gwen Littlewort, Esra Vural, Kang Lee, Mujdat Cetin, Aytul Ercil, and Javier Movellan. Data mining spontaneous facial behavior with automatic expression coding. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pages 1–20. Springer, 2008.

[22] Joy Egede, Michel Valstar, and Brais Martinez. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In *Proc. IEEE FG*, pages 689–696, 2017.

[23] Ramin Irani, Kamal Nasrollahi, Marc O Simon, Ciprian A Corneanu, Sergio Escalera, Chris Bahnsen, Dennis H Lundtoft, Thomas B Moeslund, Tanja L Pedersen, Maria-Louise Klitgaard, et al. Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition. In *Proc. IEEE CVPRW*, 2015.

[24] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *International Symposium on Visual Computing*, pages 368–377. Springer, 2012.

[25] Ghazal Bargshady, Xujuan Zhou, Ravinesh C Deo, Jeffrey Soar, Frank Whittaker, and Hua Wang. Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert Systems with Applications*, 149:113305, 2020.

[26] Bogdan J Matuszewski, Wei Quan, and Lik-Kwan Shark. High-resolution comprehensive 3-d dynamic database for facial articulation analysis. In *Proc. IEEE CVPRW*, pages 2128–2135, 2011.

[27] Teena Hassan, Dominik Seuß, Johannes Wollenberg, Katharina Weitz, Miriam Kunz, Stefan Lautenbacher, Jens-Uwe Garbas, and Ute Schmid. Automatic detection of pain from facial expressions: a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2019.

[28] Susana G Sotocina, Robert E Sorge, Austin Zaloum, Alexander H Tuttle, Loren J Martin, Jeffrey S Wieskopf, Josiane CS Mapplebeck, Peng Wei, Shu Zhan, Shuren Zhang, et al. The rat grimace scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. *Molecular pain*, 7:1744–8069, 2011.

[29] Marwa Mahmoud, Yiting Lu, Xijie Hou, Krista McLennan, and Peter Robinson. Estimation of pain in sheep using computer vision. In *Handbook of Pain and Palliative Care*, pages 145–157. Springer, 2018.

[30] Alexander H Tuttle, Mark J Molinaro, Jasmine F Jethwa, Susana G Sotocinal, Juan C Prieto, Martin A Styner, Jeffrey S Mogil, and Mark J Zylka. A deep neural network to assess spontaneous pain from mouse facial expressions. *Molecular pain*, 14:1744806918763658, 2018.

[31] Sofia Broomé, Karina Bech Gleerup, Pia Haubro Andersen, and Hedvig Kjellstrom. Dynamics are important for the recognition of equine pain in video. In *Proc. IEEE CVPR*, pages 12667–12676, 2019.

[32] Maheen Rashid, Sofia Broomé, Katrina Ask, Elin Hernlund, Pia Haubro Andersen, Hedvig Kjellström, and Yong Jae Lee. Equine pain behavior classification via self-supervised disentangled pose representation. In *Proc. IEEE/CVF WACV*, pages 1646–1656, 2022.

[33] Sofia Broomé, Katrina Ask, Maheen Rashid, Pia Haubro Andersen, and Hedvig Kjellström. Sharing pain: Using domain transfer between pain types for recognition of sparse pain expressions in horses. *arXiv preprint arXiv:2105.10313*, 2021.

[34] Francisca Pessanha, Marwa Mahmoud, and Krista McLennan. Towards automatic monitoring of disease progression in sheep: A hierarchical model for sheep facial expressions analysis from video. In *Proc. IEEE FG*, 2020.

[35] Bram Jonkers. *Equine Utrecht University Scale for Automated Recognition in Facial Assessment of Pain – EQUUS-ARFAP (MSc Thesis)*. Universiteit Utrecht, 2018.

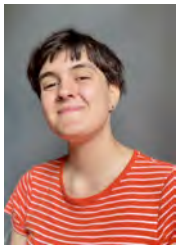
[36] Gabriel Carreira Lencioni, Rafael Vieira de Sousa, Edson José de Souza Sardinha, Rodrigo Romero Corrêa, and Adroaldo José Zanella. Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling. *PloS one*, 16(10):e0258672, 2021.

[37] Pia Haubro Andersen, Sofia Broomé, Maheen Rashid, Johan Lundblad, Katrina Ask, Zhenghong Li, Elin Hernlund, Marie Rhodin, and Hedvig Kjellström. Towards machine recognition of facial expressions of pain in horses. *Animals*, 11(6):1643, 2021.

[38] Maheen Rashid, Alina Silventoinen, Karina Bech Gleerup, and Pia Haubro Andersen. Equine facial action coding system for determination of pain-related facial responses in videos of horses. *Plos one*, 15(11):e0231608, 2020.

[39] Alam Noor, Yaqin Zhao, Anis Koubâa, Longwen Wu, Rahim Khan, and Fakheraldin YO Abdalla. Automated sheep facial expression classification using deep transfer learning. *Computers and Electronics in Agriculture*, 175:105528, 2020.

[40] Niek Andresen, Manuel Wöllhaf, Katharina Hohlbaum, Lars Lewejohann, Olaf Hellwich, Christa Thöne-Reineke, and Vitaly Belik. Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *Plos one*, 15(4):e0228059, 2020.
[41] Katrina Ask, Marie Rhodin, Lena-Mari Tamminen, Elin Hernlund, and Pia Haubro Andersen. Identification of body behaviors and facial expressions associated with induced orthopedic pain in four equine pain scales. *Animals*, 10(11):2155, 2020.
[42] Hilde I Hummel, Francisca Pessanha, Albert Ali Salah, T van Loon, and Remco C Veltkamp. Automatic pain detection on horse and donkey faces. In *Proc. IEEE FG*, pages 717–724, 2020.
[43] J Price, JM Marques, EM Welsh, and NK Waran. Pilot epidemiological study of attitudes towards pain in horses. *Veterinary record*, 151(19):570–575, 2002.
[44] Sakha Mehdi and Vatandost Mohammad. A farm-based prospective study of equine colic incidence and associated risk factors. *Journal of Equine Veterinary Science*, 26(4):171–174, 2006.
[45] Johannes PAM van Loon and Lucia Macri. Objective assessment of chronic pain in horses using the horse chronic pain scale (hcps): A scale-construction study. *Animals*, 11(6):1826, 2021.
[46] WWAJ Zwitserloot. *Quantification of pain from facial expression characteristics assessed from photos of horses and donkeys (MSc Thesis)*. Universiteit Utrecht, 2019.
[47] FM Serra Bragança, Christoffer Roepstorff, Marie Rhodin, Thilo Pfau, PR Van Weeren, and Lars Roepstorff. Quantitative lameness assessment in the horse based on upper body movement symmetry: The effect of different filtering techniques on the quantification of motion symmetry. *Biomedical Signal Processing and Control*, 57:101674, 2020.
[48] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *Proc. IEEE CVPR*, pages 4188–4196, 2016.
[49] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3D solution. In *Proc. IEEE CVPR*, pages 146–155, 2016.
[50] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.
[51] Thomas J. Cashman and Andrew W. Fitzgibbon. What shape are dolphins? building 3D morphable models from 2D images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):232–244, 2013.
[52] Angjoo Kanazawa, Shahar Kovalsky, Ronen Basri, and David Jacobs. Learning 3D deformation of animals from 2D images. *Computer Graphics Forum*, 35(2):365–374, 2016.
[53] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proc. IEEE CVPR*, pages 6365–6373, 2017.
[54] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.
[55] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.
[56] Fred L. Bookstein. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.
[57] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *Proc. IEEE CVPRW*, volume 2018-June, pages 2155–2164, 2018.
[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
[59] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. IEEE CVPR*, pages 1867–1874, 2014.
[60] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE CVPR*, pages 532–539, 2013.
[61] Albert Ali Salah, Nicu Sebe, and Theo 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.
[62] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. Automatically detecting pain in video through facial action units. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(3):664–674, 2010.
[63] Joan Alabort-i-Medina, Epameinondas Antonakos, James Booth, Patrick Snape, and Stefanos Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proc. ACMMM*, pages 679–682, 2014.
[64] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.



Francisca Pessanha is a PhD candidate at the Information and Computer Sciences Department of Utrecht University. They obtained an integrated M.Sc degree in Bioengineering with a specialization in Biomedical Engineering at Porto University in 2020. During this period, Francisca was a guest researcher at both Cambridge and Utrecht University, working in facial pain estimation assessment in animals. Their current research focuses on affective computing, particularly computer vision and paralinguistics for affective state analysis.

Albert Ali Salah (M08, SM17) is professor and chair of Social and Affective Computing at the Information and Computing Sciences Department of Utrecht University and adjunct professor at the Computer Engineering Department of Boğazici University. He has co-authored over 200 publications on pattern recognition, multimodal interaction, and computer analysis of human behavior. He currently serves as a Steering Board member of ACM ICMI and IEEE FG conferences, and as an associate editor of IEEE Trans. Affective Computing, IEEE Trans. on Cognitive and Developmental Systems, Int. Journal on Human-Computer Studies, and Pattern Recognition journals.

Thijs van Loon graduated from vet school in the Netherlands in 2000. He worked in a mixed private practice for 3 years, did a residency in veterinary anaesthesia and critical care and completed a Ph.D. with the topic of local anaesthetic techniques and objective pain assessment in horses. After 18 years at University, working mostly with horses, he joined the Altano Gruppe since September of 2021 and will be focusing on improving the standards of equine anaesthesia and analgesia in all Altano clinics. Thijs has dedicated most of his time to clinical anaesthesia, pain management and intensive care treatment of horses, and together with fellow researchers and two non-profit foundations (Friends of Vet Med and resting home for horses “de Paardenkamp”), launched the freely available Equine Pain and Welfare App.

Remco C. Veltkamp is professor and chair of Multimedia at the Information and Computing Sciences Department of Utrecht University. He obtained a M.Sc. degree in computer Science at Leiden University, and a Ph.D. degree in computer science at Erasmus University Rotterdam, the Netherlands. His research interests include the analysis, recognition and retrieval of, and interaction with, music, images, and 3D objects and scenes. He has authored over 300 refereed papers, and supervised 28 Ph.D. theses. He is the scientific director of the national research school ASCI - Advanced School for Computing and Imaging, and editor of Computers & Graphics, International Journal of Serious Games, and Graphical Models.