Estimation of Pain in Sheep Using Computer Vision

9

Marwa Mahmoud, Yiting Lu, Xijie Hou, Krista McLennan, and Peter Robinson

Introduction

The study of facial expressions in animals started with the work of Charles Darwin when he looked at expressions of emotions in man and animals (Darwin and Prodger 1998). Facial expressions are growing in popularity as a pain assessment tool for animals in research settings (Leach et al. 2012; Van Rysewyk 2016). Pain level assessment is critical to the welfare of sheep as severe pain in sheep often indicates diseases, such as foot rot (Dolan et al. 2003) and mastitis (Dolan et al. 2000). Recognising pain is essential to the subsequent treatment and pain alleviation (Flecknell 2008). Recognising and quantifying pain in sheep is particularly difficult due to their stoical nature. This difficulty can limit the use of pain-relieving drugs in these species, causing suffering and animal welfare problems (Flecknell 2008; Huxley and Helen 2006; Ison and Rutherford 2014; Lizarraga and Chambers 2012). Automating this process will facilitate early screening of large numbers of animals in a short period of time. Moreover, efficient and reliable pain assessment tools would help with early diagnosis.

The Sheep Pain Facial Expression Scale (SPFES) (McLennan et al. 2016) has recently been introduced as a standardised measure to assess pain level based on facial expressions of sheep and has been shown to identify and quantify pain in sheep faces with high accuracy. However, training of scorers and the scoring process can be time-consuming, and individual bias may lead to inconsistent scores (McLennan et al. 2016). In this chapter, we present how we can use computer vision techniques to automate the analysis of facial expressions in sheep. Our approach can improve efficiency and ensure consistency in estimation of pain. We have deployed techniques that are widely used in human facial expression recognition to address the problem of automatically assessing pain in sheep.

The overall pipeline of our sheep pain level estimation system is shown in Fig. 9.1. In the following sections we present:

This chapter is based on the work described in: “Estimating sheep pain level using facial action unit detection”, Yiting Lu, Marwa Mahmoud, Peter Robinson. IEEE Conference on Automatic Face and Gesture Recognition, Washington DC, May 2017.

M. Mahmoud (✉) · Y. Lu · X. Hou · P. Robinson
Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
e-mail: marwa.mahmoud@cl.cam.ac.uk

K. McLennan
Department of Biological Sciences, University of Chester, Chester, UK
e-mail: k.mclennan@chester.ac.uk

© Springer Nature Switzerland AG 2018
R. J. Moore (ed.), *Handbook of Pain and Palliative Care, Second Edition*,
https://doi.org/10.1007/978-3-319-95369-4_9

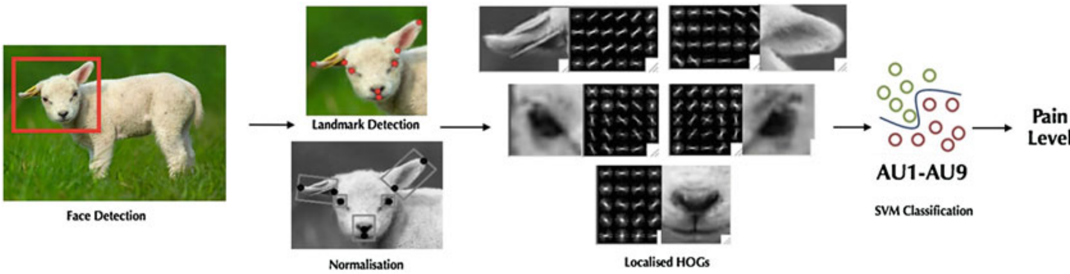

Fig. 9.1 The pipeline of our automatic approach to estimate pain level in sheep

1. our taxonomy for sheep facial Action Unit (AUs) based on the SPFES,
2. an automatic multilevel approach for estimating pain level in sheep by extending computer vision techniques that have been widely used in human facial expression recognition,
3. the evaluation results showing that we can successfully classify nine facial action units in sheep faces and can automatically estimate pain levels. We also show that our approach is generalisable across different datasets of sheep faces,
4. a simple user interface that integrates the full pipeline and can detect and analyse pain level for every individual sheep in an image of a large flock of sheep.

   Finally, we argue that – with their pain scales calibrated – the proposed automatic pain level estimation approach can be generalised to other animals, such as mice (Matsumiya et al. 2012; Sotocinal et al. 2011), rabbits (Keating et al. 2012) and horses (Dalla Costa et al. 2014).

Related Work

Analysing facial expressions of animals was first introduced by Langford et al. (2010) to facilitate detection of pain levels in mice. This approach has been advanced and generalised to many other animals. Yet manual scoring is the usual practice, and automatic assessment of pain level is still an underdeveloped area.

Recently, a standardised sheep facial expression pain scale SPFES was developed by McLennan et al. (2016). They showed that their approach is able to recognise the sheep pain face with a high degree of accuracy. Since manual labelling was used, they found that for different scorers, the accuracy of the pain assessment ranged from 60% to 75%. Their work is the basis of our sheep AU taxonomy.

Sotocinal et al. (2011) attempted to automate animal pain assessment. They introduced a partially automated approach for pain level assessment on rats. A Haar feature cascade classifier is used for real-time eye and ear detection. The classifier served as a prescreening tool so that only frames detected with the key features are kept as candidates for manual assessment. They found such partially automated pain recognition largely solves the labour-intensive problem of manual scoring.

Yang et al. (2015) analysed sheep faces and proposed a novel approach to localise sparsely distributed facial landmarks, which uses triplet-interpolated feature (TIF) extraction scheme under the cascaded pose regression (CPR) framework (Dollar et al. 2010). They applied the TIF model on sheep and reported good results regardless of sheep breed, head pose, partial occlusion, etc. However, their work assumed that the bounding boxes of the sheep faces are known.

Data

Unlike human AU analysis, facial expression recognition of sheep is still an underdeveloped area. Very few datasets are available on sheep, and fewer include ground truth labels of facial expressions. In this section, we describe our main dataset and discuss the sheep facial AU taxonomy that is used in our experimental evaluation.

Dataset

We have used the same dataset which has been described by Yang et al. (2015). This dataset consists of a total of 480 images containing sheep faces. The face bounding boxes are given, but there are no labels for sheep facial expressions. Therefore, we labelled the facial expressions. The labelling criteria are discussed in the next section.

For the sake of our work, we divide the dataset into two subsets:

1. The sheep from farm (SFF) dataset: this subset includes 380 photos taken from a farm. This set includes sheep of different breeds, skin colour and head pose. The photos vary in lighting conditions with their background being either barn or fenced grassland. The image resolution is consistent throughout.
2. The sheep from the Internet (SFI) dataset: this subset contains 100 images collected from the Internet. This set is more diverse than SFF. Sheep of different breeds, skin colour and head pose are included. The lighting conditions, background and image resolution vary extensively.

AU Taxonomy and Labelling

Facial action units (AUs) have been widely used in human facial expression analysis (Freitas-Magalhes 2012; Reed et al. 2007). Human AUs have been indexed in the Facial Action Coding System (FACS) (Ekman and Friesen 1977), which forms the standard for automatic analysis of human facial expression and emotion recognition. In contrast, facial expressions in sheep are yet to be catagorised. We first discuss the sheep AU taxonomy and then present our labelling approach of SFF and SFI datasets accordingly.

The sheep facial AU taxonomy used in our work is based on the SPFES (McLennan et al. 2016). As a preliminary AU taxonomy, only frontal faces are considered. The key features considered are the ears, eyes and nose. Although cheek and lip profile are discussed in the SPFES, they are omitted in our work because those features can hardly be seen on a frontal face. The main differences between the SPFES and our AU taxonomy are illustrated as follows:

- **Ears**: In the SPFES, three pain levels are defined regarding the extent of the ear rotation with both profile and frontal faces taken into account. In our work, we map the three pain levels but only consider the frontal faces.
- **Nose**: In the SPFES, three pain levels are defined according to the nostril shape. In our work, we map the three pain levels as they are described in SPFES.
- **Eyes**: In the SPFES, three pain levels are defined in terms of the eye narrowing extent. In our work, we define only two pain levels, namely, pain and no pain, because the dataset is strongly biased towards the no pain case. We also define a separate class for nonclassifiable pain level (AU8) when not enough information can be drawn from the frontal face because of head pose deviation.

Figure 9.2 shows the detailed description of our taxonomy. Based on these rules, we labelled the SFF and SFI datasets with AU numbers. A mapping between AU numbers and feature-wise pain scores was developed based on the SPFES. Each frontal face is labelled with five features, namely, the left ear, right ear, left eye, right eye and nose. Although the SPFES scores for symmetric features are expected to be the same, our facial AU label might differ due to poor lighting, partial occlusion or head pose deviation. The overall pain rating was calculated from the feature-wise pain scores using the same rule-based approach adopted by experts.

Fig. 9.2 Sheep facial AU taxonomy with their description and sample. The taxonomy is based on the SPFES McLennan et al. (2016)

Methodology

In our work, we have developed a full pipeline for automatic detection of pain level in sheep. We first present face detection and facial landmark localisation. We then extract appearance descriptors from the normalised facial features, followed by the AU classification. The overall pain level is estimated based on the classification results of facial features. This pain assessment pipeline is not specific to sheep and can be generalised to other animals if the proper taxonomies are developed.

Face Detection

We experimented with two methods that have been widely used in human face detection. The first method is using the Viola-Jones object detection framework (Viola and Jones 2001) to implement the frontal face detection. The SSF dataset was used to provide the ground truth. The dataset does not contain very many ground truth images, so we adopted a booting procedure to achieve larger number of training samples. Sheep faces are clipped from the ground truth images with ears excluded, and then rotations and intensity deviation are applied to each sheep face. Finally, the processed sheep faces are put on top of some random background images. A fixed window size of 32×24 is used for positive samples. The final collection of positive images consists of 5000 image windows boosted from 250 ground truth images. This face detector gives an average accuracy of 71% using a tenfold cross validation approach.

The second method we used to sheep face detection is using a Dlib implementation of a DNN-based MMOD algorithm (Van Rysewyk 2016; Dolan et al. 2003), which was previously used for dog face detection. The training was done on the SSF dataset, augmented with 100 profile face photos collected from a local farm. MMOD is highly effective because it optimises over all sub-windows of each image, which boosts the performance of the face detector trained on our relatively small dataset. This model is robust in dealing with reasonable variations in capturing viewpoint and other variability in sheep appearance. It has a negligible false positive rate compared with the first method.

Facial Landmark Detection

Our method is based on the cascaded pose regression (CPR) (Dollar et al. 2010) scheme used for the facial landmark localisation. Given the sparsely distributed nature of sheep facial landmarks, the TIF (Yang et al. 2015) approach was adopted in our work. Compared with robust cascaded pose regression (RCPR) (Burgos-Artizzu et al. 2013), which accesses the features on the line segments between two landmarks by linear interpolation, the TIF model is able to draw features from a larger area. The shape indexed feature location is defined as:

$$\mathrm{p}(S, i, j, k, \alpha, \beta) = y_i + (\alpha \cdot \overrightarrow{v}_{i,j} + \beta \cdot \overrightarrow{v}_{i,k})$$

where S is the current shape; *i, j and k* are landmark indices; and $\alpha$and β are randomly generated constants. With $\overrightarrow{v}_{i,j}$ denoting the direction from landmark y_i to y_j ($\overrightarrow{v}_{i,k}$ from y_i to y_k), it can be shown that any feature is accessible on the area spanned by these two vectors. This approach is robust against large head pose deviation and sparsely distributed facial landmarks, which matches the sheep facial landmark localisation problem well. However, it requires a perfectly predefined bounding box around the sheep face as a prerequisite.

When the Dlib approach is used for the face detection, the resulted bounding box usually excludes the ears. Therefore, we employ a two-stage approach for landmark detection. First, we detect the four central landmarks (i.e. eyes, mouth and nostril) using a Dlib implementation of an ensemble of regression trees. After obtaining the coordinates of the four central landmarks, we recenter and resize the bounding box as follows:

$$\mathbf{x}_{new\ box\ centre} = \frac{\mathbf{x}_{left\ eye} + \mathbf{x}_{right\ eye} + \mathbf{x}_{mouth} + \mathbf{x}_{nostril}}{4}$$

$$height_{face} = \frac{|y_{left\ eye} + y_{right\ eye} - y_{mouth} - y_{nostril}|}{2}$$

$$height_{new\ box} = \alpha \cdot height_{old\ box} + \beta \cdot height_{face}$$

$$width_{new\ box} = k \cdot height_{new\ box}$$

where **x** is the position vector and y is the vertical coordinate. α, β and k are estimated empirically to 0.4, 1.3 and 1.25, respectively. Eight landmarks are then obtained by applying TIF to the refined bounding box.

The final localised sheep facial landmarks are both ear tips (p1, p6), both ear roots (p2, p5), both eyes(p3, p4), the crossing of the nostrils (p7) and the mouth (p8). See Figs. 9.3 and 9.4 for an illustration.

Feature-Wise Normalisation

Normalisation is commonly used in human face recognition Brunelli (2009) to ensure faces taken from various viewpoints are registered (Brown 1992) and comparable. In our work, feature-wise normalisation is applied on sheep faces. The ears, eyes and nose are extracted and normalised separately.

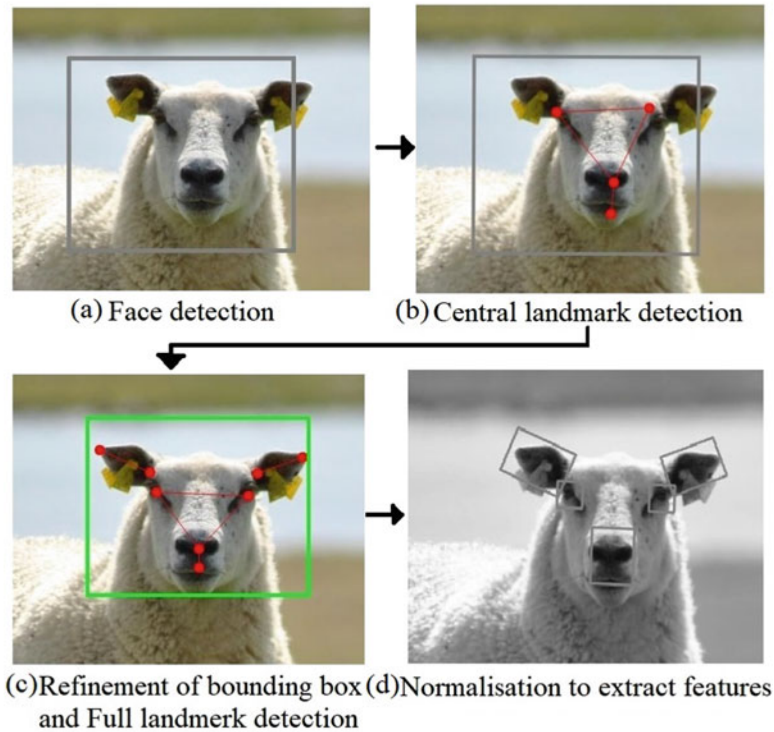

Fig. 9.3 Face detection pipeline: (**a**) initial face detection, (**b**) central landmark detection using Dlib, (**c**) full set of landmarks detected, then (**d**) normalisation step

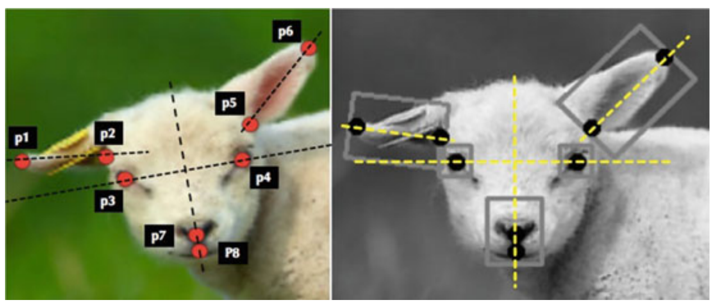

Fig. 9.4 Left: final set of landmarks detected are both ear tips (p1, p6), both ear roots (p2, p5), both eyes(p3, p4), the crossing of the nostrils (p7) and the mouth (p8). Right: face normalisation

Eye normalisation is achieved by rotating the image to keep the two eyes (p3, p4) aligned horizontally. The nostril crossing and the mouth (p7, p8) are then automatically aligned vertically since they are inherently at a right angle to the eye alignment regardless of the head pose. The scaling factor for both eyes and nose is defined as the interpupillary distance. The feature bounding boxes (see Fig. 9.3) can then be drawn according to their dominant directions. The optimal box size is determined by optimising the AU classification accuracy. The bounding box sizes we have used are listed as follows: eyes, 50×50 pixels, and nose, 100×80 pixels (all are $rows \times cols$), with 172 pixel interpupillary distance.

Unlike human ears, sheep ears vary greatly in size depending on their breed and are able to show large rotations regardless of the head pose. The dominant direction of each ear is defined as the alignment of the ear tip and the ear root (p1 with p2 and p5 with p6). The scaling factor for each ear is the distance between the paired-up tip and root. The normalised bounding box size for ears is 56×80 pixels.

Feature Descriptor

Histogram of oriented gradients (HOG) (Dalal et al. 2006) has been widely used as an appearance feature descriptor for human facial expressions. We used the Dlib (King 2009) implementation of HOGs to analyse facial features. As proposed by Felzenswalb et al. (2010), each block of HOG stands for a 31-dimensional vector: 4 normalisation masks are applied on top of the 9-orientational histogram, followed by PCA dimensional reduction (Felzenszwalb et al. 2010). Each HOG descriptor spans (*total number of blocks*) $\times$ 31 dimensions. In Fig. 9.1, HOG descriptors are visualised, showing the block dimensions for the ear, eye and nose. It can be seen that HOGs are able to depict the shape and texture of each feature.

Pain Level Estimation

With HOGs extracted and AUs labelled, we use support vector machines (SVMs) (Cortes and Vapnik 1995) to train separate classifiers for each facial feature. The overall pain level estimation approach can be described as follows: we first map the predicted AUs to feature-wise pain scores. Then we average the scores for symmetric features (i.e. eyes, ears) and average all three feature-wise scores (ear, eye, nose) to get the overall pain score. Finally, we define two thresholds (0.4, 0.8) to generate the overall pain score.

Experimental Evaluation

In this section, we evaluate the approach presented in the previous section. We compare 3-class and 2-class AU classification approaches. We also discuss the effect of data rebalancing as well as the generalisability of our AU classifiers.

AU Classification Results

We first evaluated our AU detection approach using a 3-class classifiers for each feature. The SFF dataset was used for both training and testing. Each face is given five labels (left ear, right ear, left eye, right eye and nose), and each label is associated with three AUs (ears, AUs 1, 2, 3; nose, AUs 4, 5, 6; eyes, AUs 7, 8, 9). Altogether 15 SVM classifiers were trained for all 5 features using linear kernel (LNR), radio basis function (RBF) and sigmoid function (SIG). A tenfold cross validation approach was used in all of our experiments.

Table 9.1 shows the evaluation results, with the distribution of the ground truth and the corresponding majority vote classifier accuracies. The accuracy is defined as true positives divided by the total number of samples. With most AUs achieving more than 60% detection rate, our experimental evaluation confirms that the presented AU taxonomy is reasonable and that our proposed AU detection approach are able to classify AUs of sheep.

It can be seen that SVM with LNR outperforms RBF and SIG for most AUs (as highlighted in Table 9.1). Moreover, the overall accuracy of the LNR model is the highest among the three – achieving a 67% detection rate in average. We therefore used LNR SVM model for the rest of our experiments due to its good performance as well as high computation speed.

Among all three features, the ears appear to be the strongest pain level indicator. Our approach achieved high accuracy on ear action units: AU1 (SPFES: no pain) and AU3 (SPFES: great pain). This is expected as AU1 and AU3 classes can be unambiguously differentiated.

Confusion Reduction

As seen in our classification results, our challenge is to map evolutionary features into a fixed number of AU classes.

Sheep facial expressions change gradually as their pain level increases. In manual scoring, the decision boundaries are inherently soft due to human nature and can easily be recalibrated to fit in more pain levels (tighter decision boundary) or fewer pain levels (wider decision boundary) assuming the human scorer knows the trend of

Table 9.1 Classification accuracy of our 3-class AU classifiers compared with majority vote classifier. We compare SVM linear, RBF kernel and sigmoid function. As shows, linear outperforms RBF and Sigmoid for most AU's. Linear SVM also has the highest overall detection rate. [trained on SFF, tested on SFF]

Feature	Ear(left)			Ear(right)			Nose			Eye(left)			Eye(right)			
AUnumber	AU1	AU2	AU3	AU1	AU2	AU3	AU4	AU5	AU6	AU7	AU8	AU9	AU7	AU8	AU9	Mean
Samplesize	210	**80**	**40**	200	**80**	**50**	100	160	**70**	230	90	**10**	220	100	**10**	-
Majorityvote	0.64	0.24	0.12	0.61	0.24	0.15	0.30	0.48	0.21	0.70	0.27	0.03	0.67	0.30	0.03	0.33
LNRSVM	0.80	**0.61**	0.83	0.85	**0.65**	0.72	**0.64**	0.49	**0.63**	0.72	**0.82**	**0.50**	0.77	**0.88**	**0.20**	**0.67**
RBFSVM	**0.96**	0.60	0.80	0.94	0.58	0.76	0.58	0.71	0.59	**0.91**	0.68	0.10	**0.93**	0.85	0.00	0.66
SIGSVM	**0.96**	0.55	**0.88**	**0.97**	0.35	**0.82**	0.47	0.64	0.36	0.85	0.60	0.30	0.82	0.60	0.10	0.62

the evolution. However, in computer vision, this sense of trend is missing when those evolutionary features are simply split into different classes and used in a one-vs-all training approach. In this case, the number of classes, the taxonomy and the labelling of the training samples become crucial.

In this section, we attempt to reduce the confusion by reducing the number of AUs. As a sample feature, we focus on ear-related AUs because the intermediate state (AU2: 61%) shows obvious confusion compared with the AU1(82%) and AU3 (78%).

AU Reduction by Relabelling

Training samples labelled as AU2 are relabelled and split into AU1 and AU3. The rationale here is that the facial symptom indicating pain would progressively become more obvious as the pain level deteriorates. Since there is no solid boundary between two consecutive pain levels, by splitting up AU2 (SPFES: slight pain) into AU1 (SPFES: no pain) and AU3 (SPFES: great pain), we are simply recalibrating the pain scale by making each pain level cover a wider range of facial expressions. We then train a binary classifier for each ear. The resulting accuracy (see Table 9.2) exceeds our 3-class approach accuracy by 6% on average.

AU Reduction by Exclusion

In this section, we change the way of AU reduction: we exclude the confusing AUs and the associated samples from the classification stage. AU2 training samples are excluded. Using this approach, we managed to get a 15% increase in detection rate (see Table 9.2). Such increase indicates that by excluding the confusing intermediate class, a more reliable classifier can be trained.

The results are reasonable since we are mapping between a continuous scale of feature changes to a set of discrete AU's.

Table 9.2 Comparison between the classification accuracies of our 2-class and 3-class classifiers for action units AU1 and AU3 [trained on SFF, tested on SFF]

AU number	AU1(L)	AU1(R)	AU3(L)	AU3(R)
3-class	0.80	0.85	0.83	0.72
2-class(relabelling)	0.83	0.83	0.87	0.84
2-class(exclusion)	0.84	0.86	0.98	0.98

Training Sample Rebalancing

Some AUs perform worse than the others, such as AUs 2, 5 and 9. We have further explored our data to check if the exceptionally low accuracies resulted from the imbalance in training samples. In this experiment, we enforced training sample rebalancing and investigated its effect on accuracy. The eyes were not examined because there are only ten samples labelled as AU9 in SFF dataset. Three 3-class linear kernel SVM classifiers were trained (Table 9.3). AQ1

By reducing the samples for AUs 4 and 5, the detection rate of AU6 improves by 17% (see Table 9.4), and the detection rates of AUs 4 and 5 increase by about 4%. Note that the accuracy of AU5 is the lowest among AUs 4, 5 and 6 despite having a large number of samples, while among AUs 1, 2 and 3, AU3 has the highest accuracy even with the smallest number of samples. The accuracy of AU2 is about 30% lower than AU1 and 3 in both the imbalanced and balanced cases. These results suggest that data rebalancing would, to some extent, improve the accuracy of the AU with the lowest majority vote accuracy. However, the nature of the AU definition still takes the leading role in affecting its detection rate.

Generalisation

We have also evaluated the generalisability of our approach. In this experiment, we tested to see if a classifier trained on a specific dataset is generalisable to another dataset. Five 3-class classifiers were trained using the SFF dataset and then tested on the SFI dataset. The SFI dataset varies a lot in resolution and is strongly unbalanced and biased towards AUs 1, 5 and 7. This makes the cross-dataset testing a challenging task.

Table 9.3 Comparison between the classification accuracies of our 3-class linear SVM classifiers before and after data rebalancing [trained on SFF, tested on SFF]

Feature	Ear (Left)			Ear (Right)			Nose		
AU number	AU1	AU2	AU3	AU1	AU2	AU3	AU4	AU5	AU6
Sample size before rebalancing	210	80	**40**	200	80	**50**	100	160	**70**
Majority vote accuracy	0.64	0.24	0.12	0.61	0.24	0.15	0.30	0.48	0.21
Accuracy	0.80	0.61	0.83	0.85	0.65	0.72	0.64	0.49	0.63
Sample size after rebalancing	40	40	40	50	50	50	70	70	70
Accuracy	0.85	0.53	0.73	0.84	0.60	0.76	0.66	0.51	**0.74**

Table 9.4 Cross-dataset testing, showing the classification accuracy of our 3-class AU classifiers. We can see that our approach is generalisable across different datasets [trained on SFF, tested on SFI]

Feature	Ear (left)			Ear (right)			Nose			Eye (left)			Eye (right)			
AU	AU1	AU2	AU3	AU1	AU2	AU3	AU4	AU5	AU6	AU7	AU8	AU9	AU7	AU8	AU9	Mean
Sample size	96	8	13	102	7	8	24	77	16	80	33	4	91	20	6	-
SVM LNR	0.65	0.63	0.62	0.77	0.43	0.63	0.54	0.65	0.31	0.60	0.39	0.00	0.37	0.10	0.67	0.49

The test results are shown in Table 9.4. The detection rates of AUs 1, 2 and 3 show strong correlation with their test results on the SFF set, while the accuracies of AUs 4, 5, 6, 7, 8 and 9 are affected by the data distribution. This result suggests that the ear is a strong pain indicator and its classifier generalises well, whereas for noses and eyes, more data is needed in order to achieve better classification results.

The overall pain level estimation was also tested on the SFI, and the confusion matrix is shown in Table 9.5. We adopted the same rule-based method as used in manual scoring: we used five classifiers to predict five feature-wise pain scores, and then those scores were averaged and compared with two thresholds separating the three pain levels. No extra error was introduced during the estimation stage; therefore, the overall pain level is expected to be a fair measure of our overall performance. An obvious trend favouring higher pain levels can be seen from the confusion matrix. It suggests that our automated pain level estimation approach is able to detect, though exaggerate, the pain level based on the five features. The inherited softness in human decision boundary is expected to be achieved by adjusting the two thresholds of the three pain levels.

Table 9.5 Cross-dataset testing, showing the confusion matrix of the estimated overall pain level [trained on SFF, tested on SFI]

Truth/label	P = 0	P = 1	P = 2	Sample size	Accuracy
P = 0	35	27	18	80	0.44
P = 1	3	12	11	26	0.46
P = 2	1	1	9	11	0.82

Concatenated Features

Finally, we tried an alternative approach to detect pain level in sheep. Instead of training five separate AU classifiers, we trained a single pain level classifier. We concatenated all five feature descriptors into a 3568 dimensional whole face descriptor and labelled the training samples with the overall pain levels. A 3-class pain level classifier was trained with linear SVM model on the SFF dataset. The classifier was tested on both the SFF and SFI datasets. The confusion matrices are presented in Table 9.6. In the generic dataset (SFF) test, the pain level classifier shows high accuracy on low pain ($Pain = 0'$) and high pain ($'Pain = 2'$) classes despite of the small number of samples of the high pain class. Yet, in the cross-dataset (SFI) test, the detection rate approaches the majority vote accuracy. A

Table 9.6 Confusion matrices of the pain level estimation (m.v.: majority vote accuracy) [trained on SFF, tested on SFF & SFI]

Truth/label	P = 0	P = 1	P = 2	Total	M.V.	Accuracy
Concatenated feature – trained on SFF, tested on SFF						
P = 0	94	19	27	140	0.41	0.67
P = 1	30	67	53	150	0.44	0.45
P = 2	3	7	40	50	0.15	0.80
Concatenated feature – trained on SFF, tested on SFI						
P = 0	43	43	5	80	–	0.54
P = 1	17	9	0	26	–	0.35
P = 2	5	4	2	11	–	0.18

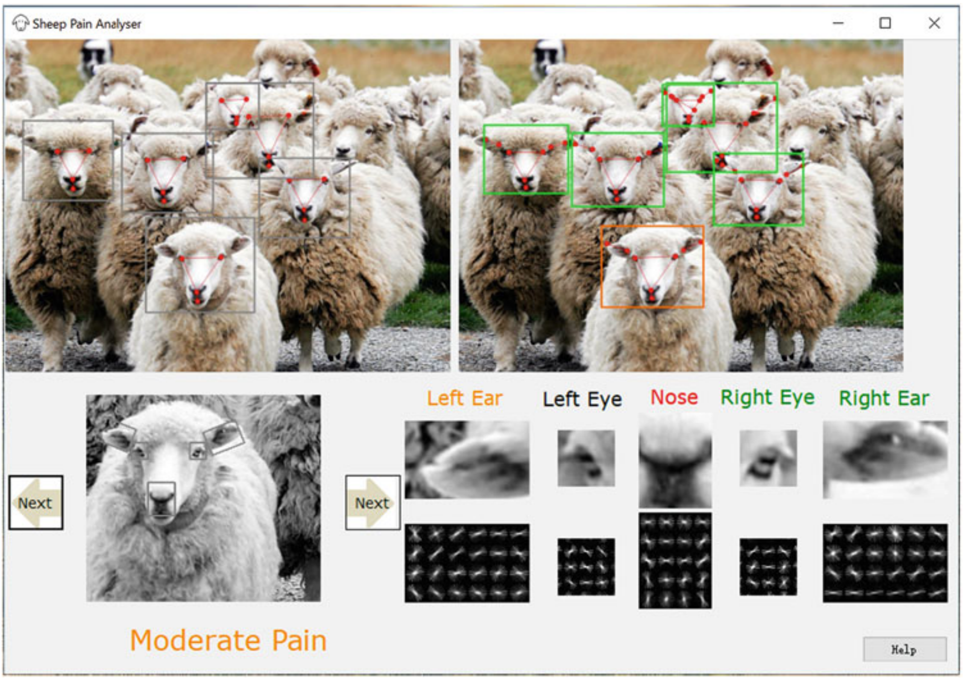

Fig. 9.5 A screenshot of the sheep pain analyser interface. Top-left: initial bounding box and four central landmarks detected. Top-right: refined bounding box and eight landmarks detected. Colour of a bounding box indicates the overall pain level estimated (green, no pain; orange, moderate pain; red, severe pain). Bottom-left: the sheep of interest is shown with arrow buttons to move from one sheep face to the next. Bottom-right: facial features segmented with their HOG features extracted. Colour indicates the pain level of each feature (green, orange or red). Black indicated that pain level cannot be determined

larger balanced dataset is required for further exploration of this method.

Pain Level Estimation Tool

As a proof of concept, we developed a tool that implements the previously described pipeline and estimates the pain levels of individual animals in an image of a large number of sheep faces. The simple user interface can automatically process an image of sheep to recognise whether a sheep is in pain and estimate the severity of that pain based on changes in facial expressions. The severity of the pain is indicated by the colour of the displayed pounding box. The same colour scheme is used for individual features on the face. Figure 9.5 shows the system in use.

Conclusions

In this chapter, we have presented a multilevel approach to automatically estimate pain levels in sheep. We have developed a preliminary sheep facial AU taxonomy based on the SPFES. We automated the assessment of facial expressions in sheep by adopting the techniques for human facial expression recognition. We have demonstrated that our approach can successfully detect facial AUs and assess pain levels of sheep. Our experiments also show that our AU classifiers are generalisable across different datasets.

For future work, we would like to explore training the classifier with the concatenated feature descriptor to map facial feature directly to pain levels. We would also like to add geometry features – such as distances between facial landmarks – as well as appearance features. This will help our AU classifier to be more robust to head pose deviation as well as breed variation. Larger number of labelled data is needed to further investigate data balancing and generalisation. Ultimately, we would like to test our automatic pain assessment approach on different animals. However, this will again require more efforts in data collection and labelling.

Acknowledgements The authors would like to thank the help of the Department of Veterinary medicine at University of Cambridge.

References

Brown, L. G. (1992). A survey of image registration techniques (abstract). *ACM Computing Surveys Archive, 24*, 325–376.

Brunelli, R. (2009). *Template matching techniques in computer vision: Theory and practice*. Hoboken: Wiley.

Burgos-Artizzu, X. P., Perona, P., & Dollar, P. (2013). Robust face landmark estimation under occlusion. In *ICCV*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273.

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*.

Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E., et al. (2014). Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS One, 9*, e92281.

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*. New York: Oxford University Press.

Davis E. K. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research, 10*, 1755–1758.

Davis E. K. (2015). Max-Margin Object Detection. CoRR abs/1502.00046. http://arxiv.org/abs/1502.00046. AQ2

Dolan, S., Field, L. C., & Nolan, A. M. (2000). The role of nitric oxide and prostaglandin signalling pathway is spinal nociceptive processing in chronic inflammation. *Pain, 86*(3), 311–320

Dolan, S., Kelly, J. G., Monteiro, A. M., & Nolan, A. M. (2003). Up-regulation of metabotropic glutamate receptor subtypes 3 and 5 in spinal cord in a clinical model of persistent inflammation and hyperalgesia. *Pain, 106*(3), 501–512

Dollar, P., Welinder, P., & Perona, P. (2010). Cascaded pose regression. In *CVPR*.

Ekman, P., & Friesen, W. V. (1977). *Manual for the facial action coding system*. Palo Alto: Consulting Psychologists Press.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminative trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1627–1645

Flecknell, P. (2008). Analgesia from a veterinary perspective. *British Journal of Anaesthesia, 101*, 121–124. AQ3

Freitas-Magalhes, A. (2012). Microexpression and macroexpression. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior*. Oxford: Elsevier.

Huxley, J., & Helen R. W. (2006). Cow based assessments Part 2: Rising restrictions and injuries associated with the lying surface. *UK Vet Livestock, 11*, 33–38.

Ison, S. H., & Rutherford, K. M. D. (2014). Attitudes of farmers and veterinarians towards pain and the use of pain relief in pigs. *The Veterinary Journal, 202*, 622–627.

Keating S. C. J., Thomas, A. A., Flecknell, P. A., & Leach, M. C. (2012). Evaluation of EMLA cream for preventing pain during tattooing of rabbits: Changes in physiological, behavioural and facial expression responses. *PLoS One, 7*, e44437.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *JMLR, 10*, 1755–1758.

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

Leach, M. C., Klaus, K., Miller, A. L., Scotto di Perrotolo, M., Sotocinal, S. G., & Flecknell, P. A. (2012). The assessment of post-vasectomy pain in mice using behaviour and the mouse grimace scale. *PLoS One, 7*, e35656.

Lizarraga, I., & Chambers, J. P. (2012). Use of analgesic drugs for pain management in sheep. *New Zealand Veterinary Journal, 60*, 87–94.

Matsumiya, L. C., et al. (2012). Using the Mouse Grimace Scale to reevaluate the efficacy of postoperative analgesics in laboratory mice. *Journal of the American Association for Laboratory Animal Science, 51*, 42–49.

McLennan, K. M., et al. (2016). Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Applied Animal Behaviour Science, 176*, 19–26.

Reed, L. I., Sayette, M. A., & Cohn, J. F. (2007). Impact of depression on response to comedy: A dynamic facial coding analysis. *Journal of Abnormal Psychology, 116*, 804–809.

Sotocinal S. G., et al. (2011). The rat grimace scale: A partially automated method for quantifying pain in the laboratory rat via facial expressions. *Molecular Pain, 7*, 1–10.

Van Rysewyk, S. (2016). Nonverbal indicators of pain. *Animal Sentience: An Interdisciplinary Journal on Animal Feeling*. AQ4

Viola, P. A., & Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR*, Issue 1.

Yang, H., Zhang, R., & Robinson, P. (2015). Human and sheep facial landmarks localisation by triplet interpolated features. AQ5

AUTHOR QUERIES

AQ1. Please check if inserted citation for “Table 9.3” is okay.
AQ2. Please cite “Davis (2009, 2015)” in text.
AQ3. Refs. [20] and [28] are identical. So, the duplicate reference has been deleted and the remaining references are renumbered accordingly. Please check if okay.
AQ4. Please provide volume number and page range for “Van Rysewyk (2016)”.
AQ5. Please provide publisher details for “Yang et al. (2015)”.