What should I annotate? An automatic tool for finding video segments for EquiFACS annotation

M. Rashid[1], S. Broome[2], P. H. Andersen[3], K. B. Gleerup[4], Y. J. Lee[1]

1 Department of Computer Science, University of California at Davis, Davis, US.
mhnrashid@ucdavis.edu

2 Department of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden

3 Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden

4 Department of Large Animal Sciences, Copenhagen University, Copenhagen, Denmark

Facial Action Coding System (FACS) is a system to define and name facial movements by their appearance on the face. Originally developed for humans, various mammal specific FACS have been developed (1-3), including Equine FACS (4) for describing horse facial action movements.

FACS coding of videos is a laborious but necessary process for the scientifically robust description and identification of facial expressions in mammals. Additionally, datasets with FACS labels are a necessary prerequisite for training machine learning methods for the automatic identification of action units as well as expressions, and have been successfully developed for human beings (6-8). While it is relatively easy to ensure that each frame in a human video dataset is easy to annotate with FACS coding - human subjects can be asked to face the camera during filming - the same is not easy, or in most settings, possible, to ensure for horses.

Therefore, EquiFACS annotators face an added challenge when annotating: finding video segments that are most suitable for annotation. The process of finding these segments is cumbersome, and presents extra time overhead to actual FACS coding and verification. We present a solution to this challenge by introducing a tool that can automatically find and mark segments in videos where horse faces are visible and in a pose ideal for annotation. The tool can reduce time spent on finding annotatable segments from half a day to 10 minutes for a 24 hour long video after frame extraction. Furthermore, it has a high recall rate of 94.92%.

In horses, video segments with the full horse face visible from a side or 45^{o} angle are ideal for EquiFACS annotation since the movement of facial muscles is clearly observable. Experienced annotators can identify such video segments while viewing a video at four times its real-time speed, with additional time spent to verify candidate segments. Overall, the process of finding video segments suitable for annotation can take an annotator the same amount of time as a quarter to half the length of a recorded video. This time overhead limits the number and length of footage that is collected as well as annotated.

Our proposed annotation tool can suggest candidate segments automatically. The software takes as input a video, and outputs a continuous confidence value between 0 and 1 for each time step in the video; 0 indicates that the time step is not usable at all, and 1 indicates high confidence in its usability. This confidence value allows users to prioritize their annotation efforts and improve productivity by going from most usable to least usable video segments. In addition, the tool also finds and localizes the horse head in each frame. The tool is extremely fast, and takes less than 13 seconds on a 25 minute video after frame extraction. See Figure 1 for an example output.

The tool works by first extracting a video frame every n seconds and running a deep convolutional neural network (CNN) that is trained to find front and side view horse faces on each extracted frame. The frequency of frame extraction, or n, can be set by the user and is set to one frame every 5 seconds for experiments in this paper. The confidence value of detections in each frame are recorded. Finally, detections with confidence less than 0.2 are ignored, averaged for every minute window, and plotted against video time.

Proceedings of Measuring Behavior 2018, (Manchester, UK, 6-8 June 2018).
Editors: Spink A.J. *et al.* www.measuringbehavior.org

The backbone of our tool is YOLO v.2 (9). YOLO is a CNN-based real-time object detection system (10). We adapt YOLO to detect two types of 'objects' - horse faces in side view, and horse faces in front view. Figure 2 shows an example of each of these categories. Training the detector to distinguish and identify these two types of horse faces correctly requires annotated training data: images of horses with front and side view faces marked. We used the dataset from (11) that had horse face bounding box annotation. We further marked each annotated horse face as either side view, front view, or neither if the face was mostly self-occluded. In addition we manually annotated and added frames from two twenty minute surveillance videos of horses. The addition of these frames was important to correct the domain difference between the dataset from (11), which comprised of images collected from the internet, and surveillance footage used by our collaborators for EquiFACS annotation. The addition of video frames improved performance from 87.01% recall to 94.92% recall.

While the tool has been developed for EquiFACS annotations, it can be easily used by researchers in other areas to find and label parts of videos most suitable for further analysis. It is currently freely available for EquiFACS annotators on our GitHub page (`https://github.com/menorashid/darknet`). Future plans include integrating the tool with previous work (11) to find and zoom in on parts of the face that are important for the identification of action units (eg. eyes) so that users can more easily observe and annotate action units.

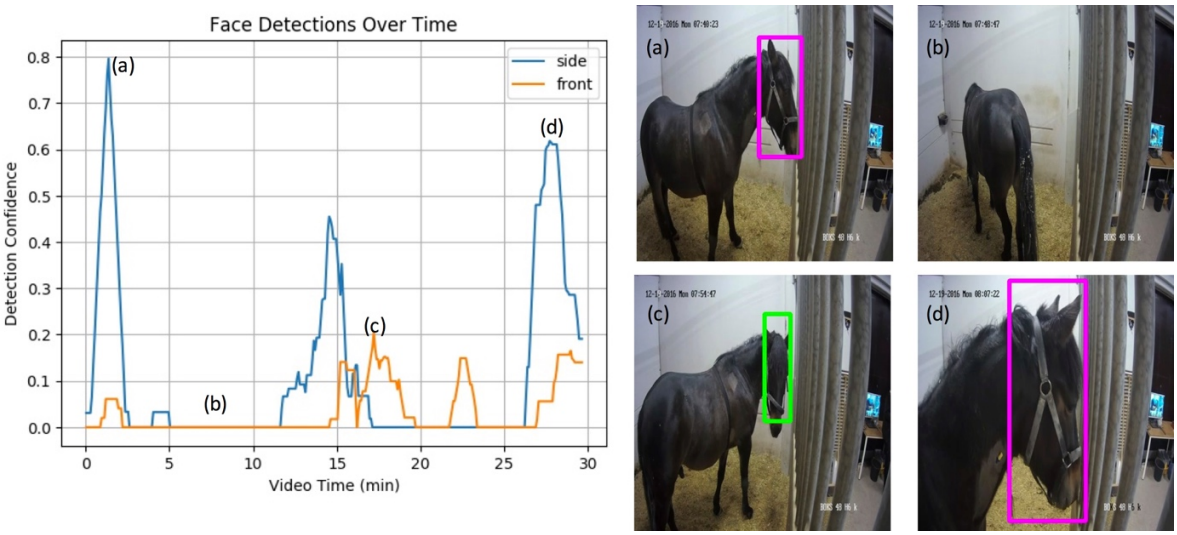

Figure 1 The output of our tool over a 30 minute video with sample frames with no, side and front view horse head detections.

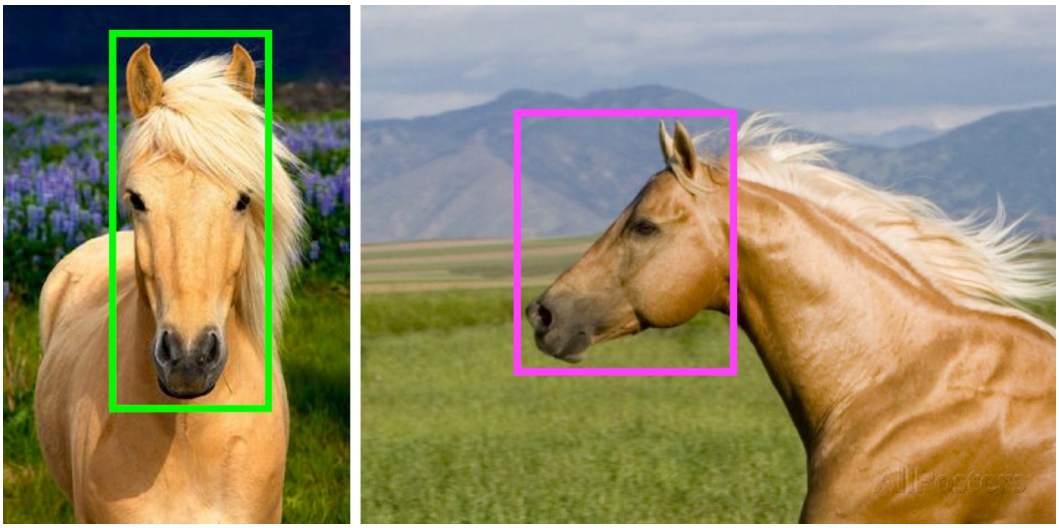

Figure 2 Examples of front and side view images used for training the horse head detector.

Acknowledgements

This research was supported in part by a Hellman Fellowship to Yong Jae Lee.

References

[1] Caeiro, Cátia C., et al. "OrangFACS: A Muscle-based facial movement coding system for orangutans (Pongo spp.)." International Journal of Primatology 34.1 (2013): 115-129.

[2] Caeiro, C. C., A. M. Burrows, and B. M. Waller. "Development and application of CatFACS: Are human cat adopters influenced by cat facial expressions?" Applied Animal Behaviour Science 189 (2017): 66-78.

[3] Langford, Dale J., et al. "Coding of facial expressions of pain in the laboratory mouse." Nature methods 7.6 (2010): 447.

[4] Wathan, Jen, et al. "EquiFACS: the equine facial action coding system." PLoS one 10.8 (2015): e0131738.

[6] Lucey, Patrick, et al. "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression." Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010.

[7] Mavadati, S. Mohammad, et al. "DISFA: A spontaneous facial action intensity database." IEEE Transactions on Affective Computing 4.2 (2013): 151-160.

[8] Zhang, Xing, et al. "BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database." Image and Vision Computing 32.10 (2014): 692-706.

[9] Redmon, Joseph, and Ali Farhadi. "YOLO9000: Better, Faster, Stronger." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017.

[10] Everingham, Mark, et al. "The PASCAL visual object classes (VOC) Challenge." International Journal of Computer Vision 88.2 (2010): 303-338.

[11] Rashid, Maheen, Xiuye Gu, and Yong Jae Lee. "Interspecies knowledge transfer for facial keypoint detection." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017.