ABSTRACT

In computer vision, it is often difficult to identify moving objects from videos recorded in real world environments where there can be ambiguity between the foreground and background. This issue becomes prominent in several applications, such as those related to the identification of unconstrained cattle in the farm premises. For example, videos recorded of cows are often against different background conditions, such as those involving external objects, equipment, boundary walls, fences, other animals, and farm personnel. To address background variability, removing background from recorded images and videos has been a well-studied area of research. Image matting is the task of removing background and estimating the foreground from images and videos which has drawn considerable attention from researchers in the field of computer vision. In recent years, deep image matting has evolved to address the ongoing limitations of conventional image matting.

In this thesis, I will explore the feasibility of deep image matting in the dairy industry; specifically, to explore deep image matting to remove background in cow videos for accurate pose estimation. Pose estimation has been widely studied on human images and videos to extract useful information and signatures from facial and body landmarks but has been largely unexplored in the dairy industry. In particular, I explored a popular pose estimation model (called DeepLabCut) in combination with deep image matting to improve the estimation of keypoints and model performance, which will be eventually beneficial in predicting the general health and well-being of the animals under study. This was made possible by collecting videos of cows in motion and training the deep neural network model with images with and without background to compare the differences between the keypoints obtained for posture detection. The geometrical configuration of multiple body parts is extracted by using transfer learning with deep neural networks. Training

the labeled dataset consisting of a small number of labeled frames using ImageNet, which is an image database consisting of more than 14 million images for pretraining the deep neural networks that act as the base for the feature detectors. The feature detector architecture has been derived from DeeperCut, which is a popular pose estimation algorithm for human pose detection but requires a large dataset (thousands of images) for training the neural network in order to achieve human-level accuracy.

My hypothesis is that pose estimation models (such as DeepLabCut, DeeperCut, and OpenPose) could benefit from background removal as a pre-processing step. While there are several algorithms for background removal and image matting algorithms, one common hurdle lies in accurately estimating the foreground in images having complicated textures or unclear demarcation between the foreground and background colors. Previous background removal and image matting algorithms were computationally simple because they used low-level features and did not incorporate high-level contextual information. Deep image matting is a recent development in computer vision that uses a more practical and efficient implementation and smaller datasets having complex background colors and textures with similar foreground and background colors.

As such, my proposed contribution lies in combining two distinct approaches (DeepLabCut and deep image matting) for pose estimation of dairy cows. Our research lab has access to videos of walking cows from Iowa State University Dairy Teaching Farm recorded by smartphones and camcorders. Around 20 videos will be used for training my deep learning model (with and without deep image matting). Each video will be split into 20 frames. The labeled frames from the videos will be split into train and test datasets. In this case, 95% of the labeled frames will be used as the train dataset and 5% will be used as the test dataset to train and evaluate the model. The trained deep learning model will be evaluated by measuring its performance in terms of mean Euclidean error (MAE), which is proportional to the average root mean square error. The MAE is calculated by measuring the difference between the labels (or keypoints) marked manually on frames and the labels applied by the model. New videos will be analyzed from the trained model through reviewing the output CSV file containing information about the keypoints predicted by the model. My results will show whether the combined approach leads to enhanced model performance which will be tested through pixel errors and the locations of keypoints. My results demonstrate that the background removal step may be beneficial in some situations to distinguish the foreground from the background, however this additional step requires extra time and effort.