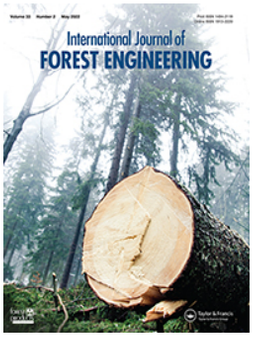

International Journal of Forest Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tife20

Use of vision and sound to classify feller-buncher operational state

Pengmin Pan, Timothy McDonald, Mathew Smidt & Rafael Dias

To cite this article: Pengmin Pan, Timothy McDonald, Mathew Smidt & Rafael Dias (2022) Use of vision and sound to classify feller-buncher operational state, International Journal of Forest Engineering, 33:2, 129-138, DOI: 10.1080/14942119.2022.2037927

To link to this article: https://doi.org/10.1080/14942119.2022.2037927

Published online: 20 Feb 2022.

INTERNATIONAL JOURNAL OF FOREST ENGINEERING
2022, VOL. 33, NO. 2, 129–138
https://doi.org/10.1080/14942119.2022.2037927

Taylor & Francis
Taylor & Francis Group

Use of vision and sound to classify feller-buncher operational state

Pengmin Pan [a], Timothy McDonald [a], Mathew Smidt [b], and Rafael Dias [c]

[a]Department of Biosystems Engineering, Auburn University, Auburn, AL, USA; [b]Research & Development, US Forest Service, Auburn, AL, USA; [c]Biosystems Engineering Department, University of Sao Paulo Piracicaba Campus, State of São Paulo, Brazil

ABSTRACT

Productivity measures in logging involve simultaneous recognition and classification of event occurrence and timing, and the volume of stems being handled. In full-tree felling systems these measurements are difficult to implement in an autonomous manner because of the unfavorable working environment and the abundance of confounding extraneous events. This paper proposed a vision method that used a low-cost camera to recognize feller-buncher operational events including tree cutting and piling. It used a fine K-nearest neighbors (fKNN) algorithm as the final classifier based on both audio and video features derived from short video segments as inputs. The classifier's calibration accuracy exceeds 94%. The trained model was tested on videos recorded under various conditions. The overall accurate rates for short segments were greater than 89%. Comparisons were made between the human- and algorithm-derived event detection rates, events' durations, and inter-event timing using continuously recorded videos taken during feller operation. Video results between the fKNN model and manual observation were similar. Statistical comparison using the Kolmogorov–Smirnov test to evaluate measured parameters' distributions (manual versus automated event duration and inter-event timing) did not show significant differences with the lowest P-value among all Kolmogorov–Smirnov tests equal to 0.12. The result indicated the feasibility and potential of using the method for the automatic time study of drive-to-tree feller bunchers.

ARTICLE HISTORY
Received 21 January 2021
Accepted 1 February 2022

KEYWORDS
Camera; machine learning; classifier; time study; cut; pile

Introduction

Many factors can be measured and used to characterize the performance of an industrial process such as a timber harvest. Machine productivity and cost per unit output are two important factors describing how well a logging system has been designed and adapted for its intended purpose on an individual job. Many full-tree loggers in the southeast US track their gross production but often lack either the incentive or the tools to track performance on a site-specific basis, despite the potential value of the data. This work was done by developing a low-cost system for continuous timber harvest productivity measurement, specifically for a full-tree feller-buncher, that could be used as a basis for long-term unattended monitoring of performance.

Productivity measures in logging involve simultaneous recognition and classification of (1) event occurrence and timing, and (2) the volume of stems being handled. In full-tree felling systems these measurements are difficult to implement in an autonomous manner because of the noisy and dangerous working environment and the abundance of confounding extraneous events, such as the presence of undergrowth or unmerchantable trees.

In traditional production studies the classification of time was accomplished using persons observing either the feller in real time or a video recording of the operation. Removing the human from the production study process requires a means of recognizing and classifying meaningful events occurring in the felling cycle. A minimal set of events would be those associated with cutting merchantable trees and with piling of bunches of stems. Felling events in full-tree systems can be detected using operator inputs to machine controls (e.g. McDonald et al. 2014) but the results can be ambiguous. Non-merchantable stems, for example, are felled during production cycles but do not accumulate harvested volume. Piling events are also the result of specific sequences of operator inputs and the occurrence of those could also be used for identification. But the same sequences of inputs are common to many felling operations and are not always easy to explicitly resolve as piling events.

Measurements to determine volumes of timber being felled also require significant effort when implemented in traditional production studies in full-tree harvesting. In many cases this involves scaling of trees by persons working in a stand, sometimes in proximity to dangerous, large equipment. Many other inventory automation systems have been proposed, in particular those using LiDAR sensors to estimate tree positions and sizes, both aerial (Holmgren 2004; Paris et al. 2016; Nelson et al. 2017) and ground-based (Maas et al. 2008; Murphy 2008; Murphy et al. 2010). These systems work well but require access to stands prior to harvest and their accuracy can be degraded if there is significant understory present.

Our purpose was to develop a system for performing production studies in full-tree logging that required little to no interaction with the machine operator and was ad hoc in the sense data are collected during the harvest itself. We felt such a system would provide performance feedback to logging crews

CONTACT Pengmin Pan pzp0014@auburn.edu Auburn University, 350 Mell St, Auburn, AL, 36849, USA

to assist in developing and implementing site-specific harvest planning, plus provide feedback to landowners and managers concerning site variability and growth potential. The main impediment in developing such a system has been the lack of robust, cost-effective sensors for accurate time and event classification, and for timber volume measurements.

In the previous work (Pan and McDonald 2019) we described a stem diameter estimation procedure for sawhead feller-bunchers using duration of the cut as the predictive variable. Duration was measured using sound, which for a sawhead feller-buncher severing a tree was distinctive and relatively simple to distinguish from background noise. In these experiments we extended the use of sound as a means of classifying the cutting cycle into production elements other than cutting. In initial experiments problems were identified with the accuracy of using sound alone as a general means of discriminating cutting cycle event occurrences. At the simplest level, the cutting cycle could be considered as being composed of movement, cutting, and piling types of events, each of which needed to be recognized in order to classify the time spent by the feller performing different functions. Sound, although reasonably accurate for detecting cutting events, was not successful in identifying piling events. It was decided, therefore, to pair sound and vision for more accurate recognition of felling activities.

The use of multiple sensors for classifying events in unstructured environments has been explored in many contexts, for example in classifying video streams (Brezeale and Cook 2008) or general monitoring of surroundings (Samaras et al. 2019). Sound is a rich source of environmental information and it has been used as a means of classification in many applications. Barchiessi et al. (2015) reviewed the state of the art in classifying environments based on sound for providing context-sensitive recommendations, for example, or audio archive management. The recognition of specific events through sound data has also been studied extensively (Mesaros et al. 2016; Sharan and Moir 2016; Chandrakala and Jayalakshmi 2019). The sound environment in proximity to an operating sawhead feller-buncher is a rich data stream, but is noisy and the information present is not always sufficient to accurately classify the state of the machine. Pairing these audio data with video, however, can lead to greater insight. The fusion of these separate data streams has most often been done using a deep learning framework, typically a convolutional neural net (Jiang et al. 2018). The same approach was chosen for application in this project.

The goal in this study was to evaluate integrated sound and vision data streams recorded from an operating feller-buncher as inputs to a classification model that would accurately classify the time spent by the machine in activities of importance in establishing its productivity. This required development of a system for post-processing recorded information that extracted relevant audio and video features and used a neural net model for classification of the time spent in various activities. The objective of the work was to evaluate this system relative to a traditional human-derived production study. Our hypothesis was the automated analysis would provide simple production study data that was not significantly different from that derived manually.

Materials and methods

For the purposes of this study, the input to the data processing system was a video, including sound, of feller-buncher operations and its output was a list of felling events and their time of occurrence. Event classes were chosen to reflect productive time and were kept to two for this preliminary analysis, including cutting of a merchantable stem(s), and piling of stems. A piling event was defined as beginning when the feller had begun tipping the head forward and simultaneously opening the accumulating arms. Classification of other productive time (i.e. moving between stems), or nonproductive time was not attempted.

The schematic shown in Figure 1 illustrates the steps used in converting raw video to time study element events. The first step in the analysis process was to partition the video stream into segments of uniform duration, T. Choice of T was a critical factor in design of the system, representing a trade-off between required processing, information content in the associated video and audio data streams, and accuracy in defining when an event occurred. Based on previous work, a sampling duration of $T = 0.8$ s was chosen for this study as being a good compromise between data processing burden, likelihood of capturing single events, and sufficient resolution of event timing. Outputs from the initial section of the classification system were single video frames extracted from the T-length segments, plus the audio stream.

The image and sound data were then fed separately into two feature extraction subsystems, one for each form of data. The final step in the time study system as illustrated in Figure 1 (labeled "Classifier") was a model identifying the most likely event category into which the current short video segment fell.

Video features

Video features used as input in the classification step were extracted using the deep learning tool "Alexnet" (Krizhevsky et al. 2012), a pre-trained convolutional neural network (CNN) general image classification tool. The system has been successfully applied in diverse applications for image feature extraction, for example in spectroscopic detection of counterfeit sesame oil (Wu et al. 2020) and in computer-aided diagnosis of eye pathology (Mansour 2018). The CNN itself is composed of numerous layers the lowest of which are associated with the image feature extraction function. These lower layers from the Alexnet model were retained in the current study up to the specific layer "fc7." This particular layer was chosen as the stopping point based on computation time and practicality for later processing, and the same approach has been successfully applied in other work, for example in detection of vehicles in images (Zhou et al. 2016). Output of the subsystem was a vector of 4096 image features that was combined with audio features for subsequent felling event classification.

Extracting features using Alexnet required input images of size 227 × 227 × 3. The process of resampling and scaling images to the specific size required was simplified using the built-in (MathWorks 1994-2020) augmented image datastore function. It took raw images from the cameras mounted on the feller-buncher cab (original size either 1920 × 1080 × 3 or

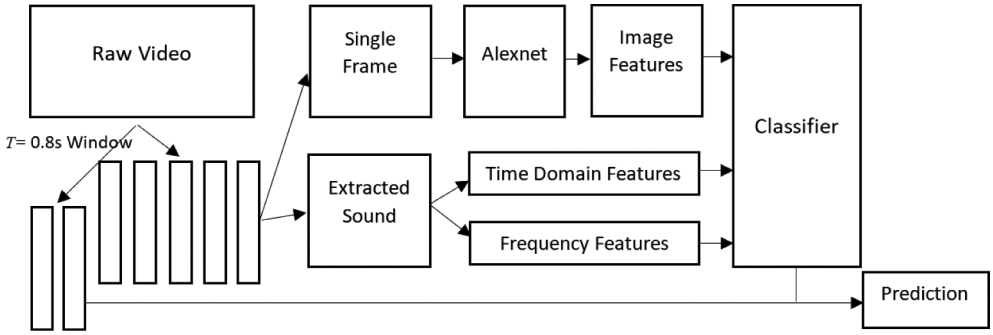

Figure 1. Schematic of data analysis system used to classify recorded video streams segmented into 0.8s windows into a set of important felling events.

1080 × 720 × 3, depending on the specific camera model) and created properly sized images for the analysis. It was also used to randomly rotate, shift, or reflect the images about a vertical axis, a randomization process intended to prevent the final classifier from over-fitting and memorizing the exact features of images. The operation was also useful in this study because it minimized the effect of using different camera models and mounting angles between the machines and sites on which operations were filmed.

Audio features

Audio signals were extracted from the video recordings using the MATLAB "*audioread*" function at a sample rate and bit depth of 48 kHz and 16 bits, respectively. In the previous study on prediction of DBH from cut duration (Pan and McDonald 2019), the cutting sound itself was distinguished from other sounds using a Local Binary Pattern (LBP) feature in the frequency range between 50 and 450 Hz calculated from the spectrogram. The model successfully distinguished cutting sounds from most background noise, but had difficulty separating cuts made on merchantable trees from those on non-merchantable stems, mainly smaller trees and brush. To achieve better classification in this study more sound features were extracted, including mean, median, standard deviation, mean absolute deviation, quantiles (2 features), entropy, interquartile range, skewness, kurtosis from the time domain signal, and entropy, 2 largest peaks of the spectrogram and their frequencies, maximum of the power spectrum density and associated frequency, and 44 wavelet and 13 Mel frequency cepstral coefficient (MFCC) features all from the frequency domain. These were calculated in addition to the 59 LBP (texture) features used in the previous study. All features were calculated using MATLAB. The spectrogram was calculated using a fast Fourier transform (FFT) with window length of 1024 points, zero-padded to a total length of 4096, and overlap of 512 points. The wavelet features were computed with the MATLAB function "GETMSWTFEAT" (Khushaba 2020), with window size of 1024 points and spacing of 128. The MFCC features were obtained using "MFCC" function from Wojcicki (2011), with the frequency range between 50 and 450 Hz, number of filterbank channels 20, number of cepstral coefficients 13, and truncated using a hamming window. The total number of sound features was 133.

Event classifier model development

Data collection

Video/sounds from four different TigerCat (Tigercat International Inc., East Brantford, Ontario, Canada) drive-to-tree feller bunchers (models 718D, 720D, 720E) were sampled from a total of 4 sites in central Alabama, USA. Each site had a distinct set of stand and saw conditions, and terrain. Four experienced operators (over 3 years) were evaluated in this study. All harvests were clearcuts on planted pine stands in the 25–35 year-old range. Since the correlation between a tree's diameter and its cut duration has been studied previously (Pan and McDonald 2019), tree size was not included in this classification study. Amount and distribution of understory was variable and sites ranged from relatively clear conditions to very thick. Cameras were mounted as in Figure 2 on either, or both, sides of the machines' cabs depending on the day of testing. All cameras used (3 total, cameras "A," "B," and "C") were GoPro Hero or Hero+ models, the differences being in video resolution and sound (stereo or mono). A total of about 10 hours of timber harvesting video were recorded. Camera mounting points and angles on either side were slightly different between data collection dates (Figure 3), but this variability contributed to the robustness of the final event classifier predictions.

Model training and validation

Table 1 depicts a summary of the video (with included sound) data collected and used for training and validating the event classification model, along with recording dates and camera characteristics. The training videos were reviewed manually and event (cut/pile/other) times were manually recorded. Video segments (0.8 s in duration) covering the event occurrence were later extracted based on the recorded times. This process resulted in a total of 5,620 labeled short video segments. These were split roughly in half and assigned to either a training or validation data set.

Figure 2. A view of the position of the camera when mounted on the feller (a), and an example image from the camera at that location (b).

Figure 3. Prediction camera setup. Horizontal camera view on the left side (a), vertical camera view on the left side (b), horizontal upside down camera view on the left side (c) and horizontal upside down camera view on the right side (d).

The event classification model took as its input a composited set of audio and video features calculated in the feature extraction steps outlined above. The classification model itself was built using the MATLAB "*classificationLearner*" application, part of the "Machine Learning Toolbox" product. The tool allowed training of models using any of a number of different classification methods, including decision trees, discriminant analysis, support vector machines, and others (see Mathworks.com). Several models were built using an automated process included in the MATLAB toolbox. Output of the fitting was a list of models along with an estimate of their accuracy. The model resulting in the highest prediction accuracy (KNN, 94.1%) was selected for use in the time study analysis system.

Feller time study analysis

Automated time study

Additional video data were recorded 21 July 2020 to verify performance of the trained classification model in time study estimation. Cameras A (L) and C (R) were installed

Table 1. Summary of video characteristics and resulting manual event counts used in the event classifier prediction model training and validation. The "Other" category included movement, idle time (engine on and off), and brushing. Training was accomplished using 0.8 s clips that included the specific type of behavior from which both video and audio features were extracted. "Position" refers to the mounting point of the camera, on either the left (L, from the operators' perspective) or right (R) of the feller-buncher's cabin.

	Date	Camera				Event Count		
Dataset	Recorded	ID	Position	Resolution	Sound	Fell	Pile	Other
Training	05/04/20	A	R	1920 × 1080	Mono	341	537	1547
	06/18/20	A	L	1920 × 1080	Mono	5	10	35
	02/27/19	B	R	1920 × 1080	Mono	5	10	35
	05/04/20	C	L	1080 × 720	Stereo	5	10	35
Totals						356	567	1652
Validation	03/11/18	B	R	1920 × 1080	Mono	51	47	233
	02/27/19	B	R	1920 × 1080	Mono	131	279	879
	05/04/20	C	L	1080 × 720	Stereo	62	232	567
	08/18/20	C	R	1080 × 720	Stereo	67	89	417
Totals						311	638	2096

on opposite sides of the cab of a Tigercat 720E feller-buncher working on a site not included in the previous training and validation of the classifier. Recording was carried out over a period of about 2 hours, but the feller got stuck in a low spot at one point during filming and only about 90 minutes of productive time were observed. The harvest was a row thinning operation and was somewhat different in stand and operational characteristics than those used in training the classifier.

The videos were sectioned into the requisite 0.8 s segments then fed sequentially into the classifier, the output of which was a list of times and event classes (cut, pile, other). Events were ordered in time and sequential cut and pile occurrences were filtered/combined using the following rules:

(1) Eliminate short-duration cut (<1 sec) and pile (<2 sec) events.
(2) Merge adjacent sequential cut (and pile) events.
(3) Eliminate all but the longest of any sequential-merged pile events.

Once filtered and combined, the number of events were tabulated, resulting in what was essentially a work sampling-based time study at 0.8 sec intervals.

The rules for filtering were based on characteristic observed event times and sequences and the video short segment length (T = 0.8 s). Cuts lasted from less than 0.5 sec to about three, depending on tree size. Using a - one second threshold as a realistic cutting duration helped eliminate incidences of brushing and severing of small, non-merchantable stems. Piling events, at least as defined in this analysis, lasted somewhat longer, from two to more than 10 secs, depending on the number and size of trees collected in the bunch. Sound and video data were less definitive when identifying piling, particularly when the machine was articulated at a sharp angle during the occurrence. In some situations the head was offset from the fixed camera location to the point there was no available visual evidence for the piling event and the classifier was left with audio data alone. And, in general, the audio signature from a piling event was not as clearly definitive of its occurrence as it was in severing a tree. The classifier was simply not as effective in identifying piling events and it tended to label too many segments in that category, particularly when the machine was in motion with trees accumulated in the felling head. The effect of this ambiguity was evident as an excess of piling events when they should not have occurred. The filtering process was expected in removing these false piling events, particularly when multiple piling incidents were identified without an intervening tree cut.

Manual time study

A single human observer watched both (left- and right-side camera positions) of the recorded videos and was instructed to perform a manual work sampling analysis. A custom program read and displayed each video, pausing every 0.8 sec. The observer typed in a numeric code for what state the feller was currently in at that point in time, either cutting, piling, or other. The cutting category was defined as the saw being in contact with a stem while simultaneously accumulating. Felling of non-merchantable trees was not documented as there was a significant amount of understory present and the feller-buncher spent some time simply mowing down small stems without closing the grab or accumulation arms, or piling the wood. According to the definition of piling in this study, its conclusion was taken to be the first video segment after the bunch had been released and traveled toward, or hit, the ground.

Automated vs. manual time study

The human and algorithm result was compared in events' durations, the number of, and inter-event timing. The inter-arrival time was defined for these analyses as the time between successive events of the same class (cutting or piling). Times were calculated from the beginning of the event, e.g. the state change from "other" to "piling" classes, to the next transition of the same combination.

Results

Short segment classification

The classification model was trained using MATLAB classification learner toolbox across a broad range of potential techniques using pre-selected segments labeled by a human observer. The highest calibration accuracy achieved (94.1%) in various initial tests was obtained using the fine K-nearest neighbors (fKNN) classifier, comparing with cubic Support Vector Machine (SVM) (93.5%), quadratic SVM (90.7%), weighted KNN (89.4%), and many other methods generated by MATLAB. The fKNN technique was adopted as the default classification model in all subsequent analysis.

The training dataset included 2425 video segments from camera A plus 100 from the remaining two cameras. Initial training was carried out using data (sound and video) from only camera A. Results from training using only camera A data resulted in a 90% prediction accuracy on validation segments from camera A, but only 55% for segments from the remaining two cameras. The model was very likely trained with some unique features

resulting from camera A's setup and lost predictive power when the recording setup changed, when varying such factors as mounting angle and orientation. With this experience, a very small percentage of data from cameras B and C (2% of total videos in the training set) were included when retraining the classifier. The newly-trained model successfully predicted validation short segments from cameras B and C at rates ≥78.5% for cut, ≥79.0% for pile, and ≥89.4% for other events, without degrading accuracy for camera A. A summary of classification results by event type and date of data collection resulting from application of this final fKNN model were shown in Table 2.

Time study data comparisons

The two videos collected for verification of the method were split into segments of length T = 0.8 sec, fed through the classification model, and the resulting sequence of event classes were collected into a time series of discrete values representing event type. After filtering the model output using the rule-based approach outlined above, the sequences were compared to manually generated sequences coded using the same scheme. Short-duration examples of model, filtered model, and manually classified outputs are shown in Figure 4. The plotted values in Figure 4(a) represent the assigned class of the individual short segments in a two-minute video, and Figure 4 (b) the same sequence after filtering/combining. Figure 4(c) shows the human-assigned class for the same video. The correspondence between the human- and algorithm-defined sequences was very good for these data. Although there were some differences in events' durations, the number of events and inter-event timing matched well in this example. These data also illustrated the necessity of performing the filtering step as incorrect individual video segments classifications were common, even, for example, during a single cutting event. The filtering step was simple to perform and resulted in much closer correspondence between model and human analysis.

Measures of correspondence between the two event sequences (manual work sampling and the filtered model output) were calculated using event counts and also distributions of inter-event times and were summarized in Table 3. Manually derived event counts were different between camera angles, probably because of visibility of events and simple human error. A second human observer also performed work sampling on the A camera video as a check and the results were similar to the first observer (cuts/piles 188/87), but not exactly the same. Differences between model estimates for the two videos (between the left and right mounting positions), however, were about double those of the manual counts (Table 3).

Table 2. Best results applying the short-segment classification model. The "N" column referred to the number of video clips evaluated and "Correct" lists the number of those clips correctly classified. The model used a fine KNN classification step with inputs derived from an Alexnet-trained CNN for video data and a custom neural net for audio feature calculation.

	Cut			Pile			Other		
Date	N	Correct	%	N	Correct	%	N	Correct	%
03/11/18	51	47	92.2	47	45	95.7	233	216	92.7
02/27/19	279	226	81.0	131	108	82.4	879	786	89.4
05/04/20	232	182	78.5	62	49	79.0	567	512	90.3
08/18/20	89	74	83.1	67	53	79.1	417	402	96.4

Differences between human and model-derived counts within a single video were smaller, except for piling events on the A camera. Orientation of the two cameras was not the same for the validation tests, with the C camera being mounted upside down. This may have negatively affected the accuracy of those results. Although at least four video orientations had been included in the training data set, relatively few upside down images had been used and this may have limited prediction accuracy for the model results. Another difference between the two cameras was the fact nearly 90% of all training data had come from camera A, which resulted in 90% or greater accuracy for videos from camera A in every instance and much lower accuracy for other cameras. For example, video recorded on 05/04/20 with camera C achieved only 78.5%, 79.0%, and 90.3% accuracy rates for cut, pile, and others, respectively. The two cameras in this comparison were from the same manufacturer but were not the same model. Video and sound capture resolution were different between the two, perhaps affecting accuracy. Overall, however, errors relative to human assessments of activities were similar to what had been found in validation tests on the trained model.

Inter-arrival times of cutting and piling events were tabulated from the work sampling and model-derived results. Distributions of the observed values were plotted in Figure 5. The Kolmogorov–Smirnov test, a nonparametric means of evaluating if samples of two continuous independent random variables were drawn from the same distribution (Baumgartner et al. 1998), was used to evaluate statistically the similarity of the results. Values for the statistic were calculated using the *scipy.stats.ks_2samp* function (The SciPy Community, 2020) in Python and results were summarized in Table 4. All P-values for the tests were greater than $\alpha = 0.1$ meaning the null hypothesis could not be rejected and the underlying distributions were indistinguishable. The model-based automated work sampling procedure, therefore, produced results that were not statistically different from the human-derived version.

The lowest P-value (most nearly significantly different) among all Kolmogorov–Smirnov tests, 0.12, was from the comparison between cameras of cut inter-event times derived from the model analysis, indicating there was some difference in identifying tree cuts between the two recordings when applying the method. As mentioned above, this was possibly due to differences in resolution (audio and video) between the two cameras and indicated the method as developed was sensitive to how the model was trained.

Table 3. Comparison between model and manually derived felling event counts. Manual referred to counts derived from human observation of videos. Values are total counts, observed and predicted, along with percent differences within videos for the same method and within video source between prediction method (human and model)

Source	Cut		Difference (%)	Pile		Difference (%)
	Camera A	Camera C		Camera A	Camera C	
Manual	188	179	5	88	91	3
Model	189	171	11	77	87	13
Difference (%)	-0.5	5		13	4	

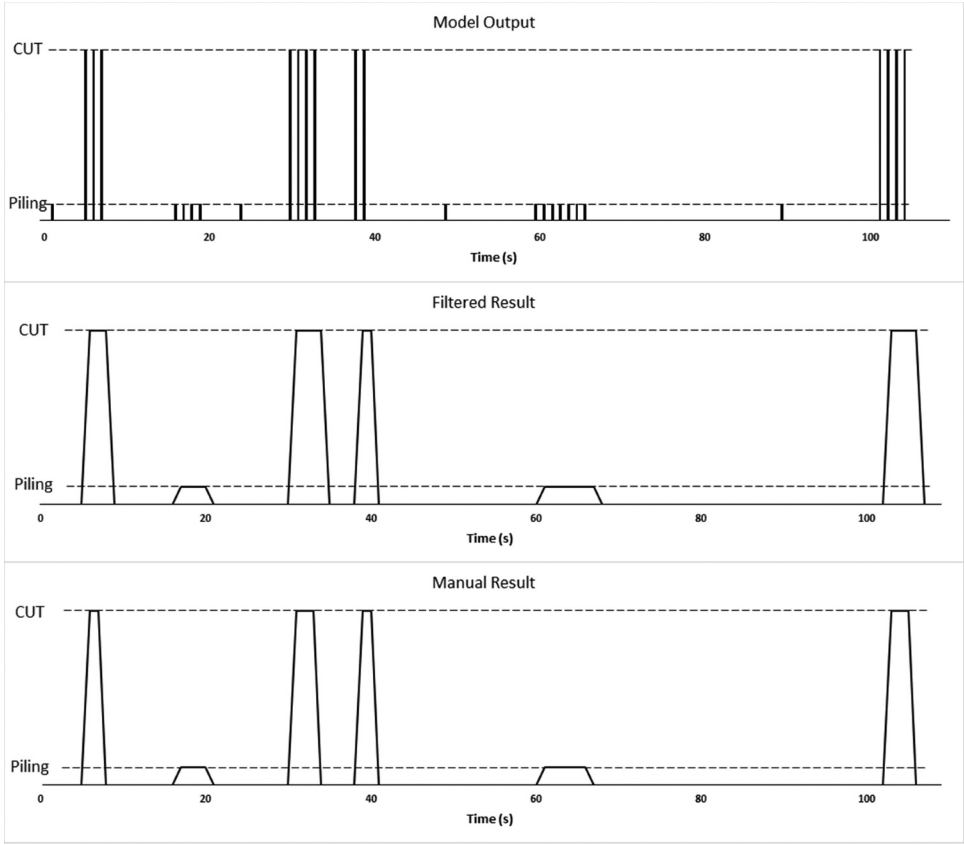

Figure 4. An example of filtering the raw output from the combined audio/video classification tool. The values plotted in (top) represent the sequence of assigned classes for a contiguous stretch of video lasting for two minutes. Assigned class for a particular segment was represented as a numeric value, with 100 assigned to segments identified as "cut," 10 to those as "pile," and 0 to those labeled as "other." (middle) shows the same sequence after filtering using the rule-based approach described in the text. (bottom) shows labels assigned by human observers to the same video data.

Discussion

The misclassification of cutting events resulted mainly from falsely identifying felling of non-merchantable trees as being merchantable stems. There was significant understory present in the stand and the feller-buncher spent some time clearing smaller hardwood stems out of the way. These brushing cuts were sometimes made with crop trees already accumulated in the head of the machine, which contributed essentially false visual information to the classification model. This result emphasized the necessity of at least two streams of information, auditory and visual, to adequately analyze machine function. When both sources were not present, as when brushing smaller stems or when vision was occluded by overhanging canopy, false cutting events were more likely. A third source of information on machine function that could have been helpful in interpreting results would have been monitored via CAN bus data of the machine, but those data were not included in this analysis. A further analysis of tree size from cut sound duration might also have been useful in distinguishing cuts of merchantable stems, but was not attempted in this study.

Misclassification of piling events resulted principally from acute articulation angles while dropping the trees. Sound data were not as important in identifying piling operations so when visual data were not available the classifier tended to miss the events. Although not a common issue, it happened at least four times in these experiments and accounted for about 80 percent of the piling errors observed.

There were limited opportunities to acquire data from a variety of felling equipment in this study as all crews tested were running Tigercat machines. And of those machines, the felling heads used varied slightly in age and model but were, more or less, about the same. No information was available, therefore, about the effect of machine type or condition on classifier results, although it seemed likely sound and visual profiles of other manufacturers/models would be different and the classifier would require retraining for individual circumstances. It was apparent, for example, that over 80% percent of audio features important in the tree cut event classification related closely with changes in frequency spectra, which could be expected to relate to, for example, tree species or saw teeth condition. Accounting for these factors might require retraining the classification model. Retraining was not an onerous process, but required time to be spent on collecting raw data and extracting and labeling events of interest. If the model was not robust to variations in these conditions and required frequent retraining, it might be difficult to maintain accuracy. Training the model with greater breadth of camera models and locations, plus different machines and site conditions, might have improved the performance.

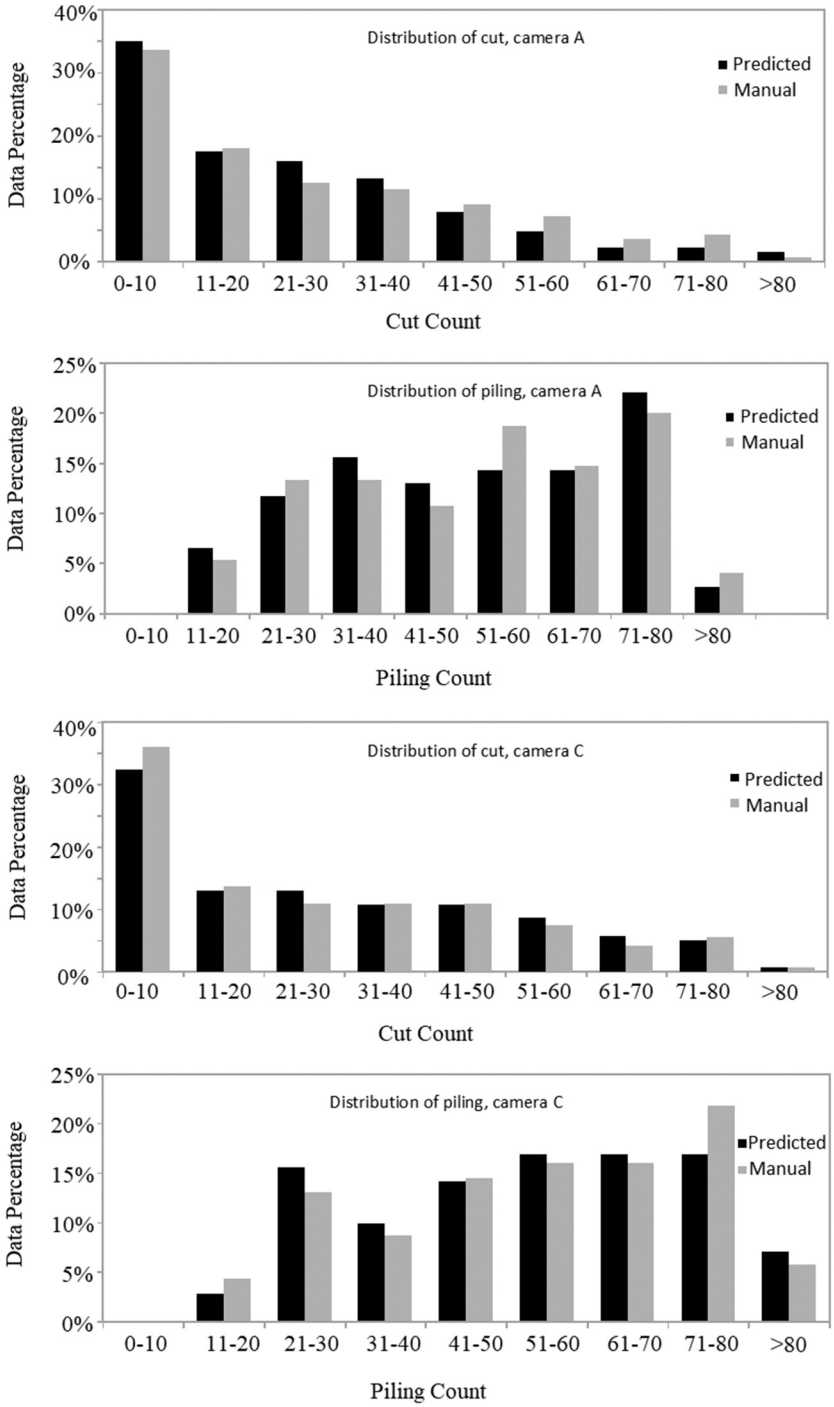

Figure 5. Inter-event time frequency distributions for cutting and piling events, grouped according to the view (camera) from which the data were derived and the method by which they identified (human work sampling vs. model). x axis: time (second, s), y axis: frequency (fraction of observations).

The system developed was not a complete stand mapping package for drive-to-tree feller-bunchers. In particular, the mapping portion remained to be integrated. This would require incorporating a positioning component along with a means of aligning sound, video, and position data streams in time. Given these data could be collected and processed in real time, the system could be used to inform skidder operations, perhaps even be used in development of a driver-less skidder.

Similar studies have been reported in cut-to-length harvesting systems, with time study estimates derived from machine-based sensing sources. Timing data from StanForD files have been parsed and filtered to develop feller productivity estimates by Brewer and others (Brewer et al. 2018), Strandgard et al. (2013), and Palander et al. (2013). In each case productivity models derived from the automatically collected data were compared to, and/or supplemented with, manually prepared models. Some differences were noted in

Table 4. Kolmogorov–Smirnov test results comparing the inter-arrival time distributions of cutting and piling events. Null hypothesis in the test was that the distributions were the same. The "Between" factor indicated the distributions being compared by the "Within" factor. For example, the first line of the Table compares distributions assembled from work sampled data derived from the two videos available.

Between	Within	Event	KS Test Statistic	P-Value
Cameras: A-C	Method: Work Sample	Cut	0.030	0.99
		Pile	0.094	0.80
	Method: Model	Cut	0.12	0.12
		Pile	0.078	0.95
Methods: Work Sample – Model	Camera: A	Cut	0.061	0.85
		Pile	0.071	0.97
	Camera: C	Cut	0.12	0.16
		Pile	0.16	0.18

Strandgard et al. (2013), but quality of model predictions based on the StanForD data were generally acceptable. Olivera et al. (2016) cited the development of a feller productivity model based on harvester-derived data collected at a large scale and concluded it was valid as long as sufficient rules were applied to discard data not meeting validity criteria. Similar work in full-tree harvesting systems was reported in McDonald et al. (2012). Their data collection system was based on selecting relevant operator input sequences to the machine control system from the CAN bus and using a rule-based approach to parse functional elements of the felling cycle. They reported the number of cycle elements was typically within 10% of values obtained using manual time study, the result of which was similar to the presented automatic extracted result from camera A in every instance, and better than results from other cameras, especially in cut and pile event classifications (the lowest accurate rates were 78.5% and 79.0%, respectively). However, timing results for those elements were not reported.

Making large-scale economic decisions based on data drawn from a system such as presented in this study would not be practical given the observed accuracy. The hardware simplicity of the system, however, makes it a viable candidate to serve as the basis for long-term unattended productivity analysis in full-tree felling systems, similar to analytical possibilities available via StanForD file parsing in CTL harvesters. With additional data sources, such as is typically available on a feller-buncher's CAN bus, accuracy of estimates could likely be improved, perhaps to the point real-time, stand-level decision-making would become feasible and a practical form of `precision forestry' developed.

Conclusions

A machine learning process has been developed to classify combined video/audio data streams of recorded activity for a drive-to-tree feller-buncher into events meaningful in time study analysis. Classification accuracy of the model relative to human-derived assignments was sensitive to mounting location and camera model, plus the level of diversity included in training data.

Comparison of work sampling on two sets of video data, model- and human-derived, results were comparable in counting events of classes "cut" and "pile." Piling events, however, were subject to slightly higher variation in counts compared with human results. The largest observed difference in piling event counts was 13%. Distributions of inter-event initiation times were not statistically significant from manually-derived data.

Acknowledgements

This material is based upon work supported by the Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE), under Award Number DE-EE0006639. Authors would also thank Industree Companies (Wetumpka, AL) for providing the opportunities in data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the U.S. Department of Energy [EE0006639].

ORCID

Pengmin Pan http://orcid.org/0000-0003-2653-3836
Timothy McDonald http://orcid.org/0000-0002-2907-4017

References

Barchiessi D, Giannoulis D, Stowell D, Plumbley MD. 2015. Acoustic scene classification: classifying environments from the sounds they produce. IEEE Signal Processing Magazine. 32(3):16–34. doi:10.1109/MSP.2014.2326181.

Baumgartner W, WeiB P, Schindler H. 1998. A nonparametric test for the general two-sample problem. Biometrics. 54(3):1129–1135. doi:10.2307/2533862.

Brewer J, Talbot B, Belbo H, Ackerman P, Ackerman S. 2018. A comparison of two methods of data collection for modelling productivity of harvesters: manual time study and follow-up study using on-board-computer stem records. Annals of Forest Research. 61(1):109–124. doi:10.15287/afr.2018.962.

Brezeale D, Cook DJ. 2008. Automatic video classification: a survey of the literature. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews. 38(3):416–430. doi:10.1109/TSMCC.2008.919173.

Chandrakala S, Jayalakshmi S. 2019. Environmental audio scene and sound event recognition for autonomous surveillance: a survey and comparative studies. ACM Computing Surveys. 52(3):34.

Holmgren J. 2004. Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. Scandinavian Journal of Forest Research. 19(6):543–553. doi:10.1080/02827580410019472.

Jiang YG, Wu Z, Tang J, Li Z, Xue X, Chang SF. 2018. Modeling multimodal clues in a hybrid deep learning framework for video classification. IEEE Transactions on Multimedia. 20(11):3137–3147. doi:10.1109/TMM.2018.2823900.

Khushaba R 2020. Feature extraction using multisignal wavelet transform decom. [Accessed 2020 Dec 12]. https://github.com/RamiKhushaba/getmswtfeat .

Krizhevsky A, Sutskever I, and Hinton G 2012. ImageNet classification with deep convolutional neural networks. NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems; December 3 - 6; Lake Tahoe Nevada. 1: p. 1097–1105.

Maas H-G, Bienert A, Scheller S, Keane E. 2008. Automatic forest inventory parameter determination from terrestrial laser scanner data. International Journal of Remote Sensing. 29(5):1579–1593. doi:10.1080/01431160701736406.

Mansour RF. 2018. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. Biomedical Engineering Letters. 8(1):41–57. doi:10.1007/s13534-017-0047-y.

Mathworks. 1994–2020. Train classification models in classification learner app. [Accessed 2020 Nov 01]. https://www.mathworks.com/help/stats/train-classification-models-in-classification-learner-app.html

McDonald T, Fulton J, and Pan P 2012. Mapping machine productivity and tree size of a feller-buncher harvesting biomass in pine plantations. In Proceedings of the 2012 International COFE Meeting; September 9-12; New Bern, North Carolina. p. 10.

Mesaros A, Heittola T, Virtanen T. 2016. Metrics for polyphonic sound event detection. Applied Sciences. 6(162):17. doi:10.3390/app6060162.

Murphy G. 2008. Determining stand value and log product yields using terrestrial lidar and optimal bucking: a case study. Journal of Forestry. 106(6):317–324.

Murphy GE, Acuna MA, Dumbrell I. 2010. Tree value and log product yield determination in radiata pine (Pinus radiata) plantations in Australia: comparisons of terrestrial laser scanning with a forest inventory system and manual measurements. Canadian Journal of Forest Research. 40(11):2223–2233. doi:10.1139/X10-171.

Nelson R, Margolis H, Montesano P, Sun G, Cook B, Corp L, Anderson HE, deJong B, Pellat FP, Fickel T, et al. 2017. Lidar-based estimates of aboveground biomass in the continental US and Mexico using ground, airborne, and satellite observations. Remote Sensing and Environment. 188:127–140. doi:10.1016/j.rse.2016.10.038.

Olivera A, Visser R, Acuna M, Morgenroth J. 2016. Automatic GNSS-enabled harvester data collection as a tool to evaluate factors affecting harvester productivity in a Eucalyptus spp. harvesting operation in Uruguay. International Journal of Forest Engineering. 27 (1):15–28. doi:10.1080/14942119.2015.1099775.

Palander T, Nuutinen Y, Kariniemi A, Väätäinen K. 2013. Automatic time study method for recording work phase times of timber harvesting. Forest Science. 59(4):472–483. doi:10.5849/forsci.12-009.

Pan P, McDonald TP. 2019. Tree size estimation from a feller-buncher's cutting sound. Computers and Electronics in Agriculture. 159:50–58. doi:10.1016/j.compag.2019.02.021.

Paris C, Valduga D, Bruzzone L. 2016. A hierarchical approach to three-dimensional segmentation of lidar data at single-tree level in a multilayered forest. IEEE Transactions on Geosciences and Remote Sensing. 54(7):4190–4203. doi:10.1109/TGRS.2016.2538203.

Samaras S, DIamantidou E, Ataloglou D, Sakellariou N, Vafeiadis A, Magoulianitis V, Lalas A, Dimou A, Zarpalas D, Votis K, et al. 2019. Deep learning on multi sensor data for counter UAV Applications—A systematic review. Sensors. 19(22):4837. doi:10.3390/s19224837.

The SciPy community. 2008–2020. [accessed 2021 Jan 02]. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html

Sharan RV, Moir TJ. 2016. An overview of applications and advancements in automatic sound recognition. Neurocomputing. 200:22–34. doi:10.1016/j.neucom.2016.03.020.

Strandgard M, Walsh D, Acuna M. 2013. Estimating harvester productivity in Pinus radiata plantations using Stanford stem files. Scandinavian Journal of Forest Research. 28(1):73–80. doi:10.1080/02827581.2012.706633.

Wojcicki K 2011. HTK MFCC Matlab. [accessed 2020 Sept 11]. https://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab

Wu X, Zhao Z, Tian R, Shang Z, Liu H. 2020. Identification and quantification of counterfeit sesame oil by 3d fluorescence spectroscopy and convolutional neural network. Food Chemistry. 311:7. doi:10.1016/j.foodchem.2019.125882.

Zhou Y, Nejati H, Do T, Cheung N, and Cheah L 2016. Image-based vehicle analysis using deep neural network: a systematic study. 2016 IEEE International Conference on Digital Signal Processing (DSP); October 16-18; Beijing, China. 276–280.