https://doi.org/10.2352/ISSN.2470-1173.2021.12.FAIS-026
© 2021, Society for Imaging Science and Technology

Recognition of manual forest work for time on task recording

Stefan Dilger; Fraunhofer SIT; Darmstadt, Germany

Abstract

In the manual forestry, the worker carries his equipment to a tree by foot. There the tree is felt and processed. Depending on the surrounding, this needs more or less time. This paper automatically analyses the needed time for different activities. Therefore, the worker gets an estimate of his time spends on different tasks and can estimate his productivity. The approach uses therefor a mobile phone and a smartwatch. This work could also be used to conduct general time studies for time and equipment comparison. The focus is on the frequency analysis, feature creation and the obstacles, such as asynchrony of the recordings and labeling errors during the annotation of the ground truth, which are described and calculated.

Introduction

Manual woodwork is a crucial part of ecological foresting. Usually, contracts are assigned and paid in a per tree or area base with a minor variation on the actual environment. The compensation salary is base on studies from around 1990. While the used machinery has improved over the years and ecological forest are mainly located on hard to access terrain, the fees for the worker stayed similar. The woodworker has to decide daily if an offered contract is financially feasible and therefore need a more reliable and easy way to determine their actual earnings. We propose a solution witch automatic recognizes the time needed for multiple processed in felling a tree. Additionally, this method only uses commercially available hardware like mobile phones and smartwatches. This not only helps the woodworker to get a better estimation of his spend time but also the possibility to compare different equipment without to conduct an expensive study.

Related Work

Hardware for Data collection

In the early stages of human activity recognition, the researchers used specially developed boards to classify gestures. Like in the much-cited work of Bao and Intille [3] and there five board equipment and placement of multiple sensors over the howl body. His lead to very unnatural environments for the probands. Same goes for other work like the recognizing of workshop activity by Lukowicz et al. [11]. Thanks to modern hardware like mobile phones this constraint is not given anymore. With the increasing number of integrated sensors, external sensors could be omitted [9]. But there are still the problems of placing the data logger on the body. Another problem is which data to log. This includes the acceleration which was shown to be the most crucial sensor for activity detection since it provides the most significant information content as concluded in [6]. However, the integration of further sensors can improve the detection rate. Therefore all available data is collected and later evaluated, which information is best used for the given activities.

Data labeling

Since there exists no ground truth or publicly available data set, the focus of this work is also the recording and annotation of training data in the field of semi-automated forest work. In addition to the pure recording of the acceleration/audio date the performed activity must also be recorded. This can be done under observation by the experimenter as in [13]. The subjects can also record the data themselves, such as by manually registering the times [3], pressing a start/stop button [6] or using the software on a mobile phone [8]. For this purpose, all activities were recorded with two cameras at the same time. This should simulate the annotation by an observing examiner and the worker himself. This was done because all other methods influence the workflow of the recorded person. To annotate the data, experts are needed which have to watch the recordings and label them. This is necessary for the training of the applied algorithms. In the process, it was tried not to influence the subject so that the recordings can reflect the same situation as in the real world. In long-term studies in which the subjects record their annotations afterward, it can come to recall bias and inconsistencies [5]. Despite the exact timing and synchronization of the recordings, it is still possible to get inaccurate timing. Also, confusion or erroneous classification of activities may occur. These errors and inaccuracies complicate further analysis of the data and influence the ground truth. Despite these obstacles, the annotation was created days after the actual work in the woods.

Classification

The use of pure raw data as a basis for machine learning results in rather poor recognition rates [7]. Instead, it is helpful to choose an alternative representation of the data with which the classifiers can work more efficiently. For this purpose, features are calculated from the raw data. Possible features include average, standard deviation, mean absolute deviation, the time between two peak values [9], working energy, entropy, correlation [3] and features in the frequency domain such as the frequency domain entropy [12]. In examining these features it quickly becomes apparent that many features contain redundant information and the detection rate is not necessarily negatively affected by omitting individual features [3] sometimes even positive [12]. A sliding window approach with an overlap of 50% and a window width in the range of one to two seconds has proved suitable for calculating these features [SVS08]. As a method of classification, k-Nearest Neighbors [14], decision trees [3], [1], and more models [4] are commonly used. K-Nearest Neighbor is a simple classification method that requires a lot of computing power with a large amount of training data and is therefore not suitable for more complex activity detection. Decision trees offer a high degree of accuracy while at the same time have a moderate computational effort [13]. Recognizing sensor data with machine learning algorithms usually requires a lot of computing time depending on

the algorithm and amount of data recorded. The duration of such an algorithm can be in extreme cases several hours [2]. For simple activities such as walking or sitting there are detection rates of more than 90% achievable. In other activities too. High recognition rates are achieved if the activity is distinguishable from the moving sequences of other activities. Only in the case of very similar movements occur a lower recognition rate. But before a number of achievement should be celebrated the used performance metrics should be watched. With the wrong metrics, a bad model also can achieve a high score. But a predicted label can not only be measured by the means of occurrences but also by how many fragmentation or merging occurs. There are dozens of possibilities for creating a metric [15].

Hardware placement

To find the best placement for the hardware, the subject was equipped with 3 smartwatches and one mobile phone. The final recording used was from the right wrist (smartwatch) and chest pocket (mobile phone). Additionally, the subject was equipt with a 360 camera on its shoulder and also recorded form a save distance. The Video was used for annotation. Therefore, one video was annotated from a different person and one person annotated two videos. This results whre compared for ground truth and to determine the variation people annotate the given scene. Finale also the result of a different combination of sensors where compared. This lead to the answer if one ore more data sources were helpful. Also where best to place the mobile phone and smartwatch.

Evaluation Methods

This section summarizes the problems and used methods.

Offset

There are different methods to check the quality of the found offset. The first and simplest is to calculate the offset of a file that has been split with a known offset and then compares the deviation with the calculated. This is not optimal since in reality there not only exists an offset, but also a phase and amplitude difference between two files. An alternative is to determine the offset manual by looking at the wave produced from the audio and the hearing. This can then be compared with the automatically found offset.

Cross-correlation

To automatically find the right offset of the audio files the Cross-correlation was used. The following steps were taken to calculate the signal correlation:

1. Zero-pad the input signals to at least half of the wave
2. Take the FFT of both signals
3. Multiply the results together
4. Do the inverse FFT of the result
5. Find the position of the highest peak

The zero padding are needed since otherwise the audio overlaps and this way the signal acts like it zero out to infinity. Otherwise, there is a higher chance to find other peaks, which are not the right offset of the audio.

Ground Truth

The critical question is, how can optimally combine labels from multiple annotators used to form the estimate of a ground truth? Some simple heuristics for combining the labels are a majority vote, mean and median.In his paper, the majority vote was used. Still, there is the question for, how accurate was the result. Therefore the different ground truth was compared by calculation the confusion matrix with the F1 score.

F_1-Score

The F_1 score is defined as harmonic mean of precision and recall as seen in equation eq:F1Score.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1}$$

Cohen's Kappa Score

With the Kappa Score (as stated in equation eq:kappa), a better understanding of the labeling is given, especially if some labels are under and others are over-represented in the number of occurrences. This leads to a better score or a single number to compare how good less represented labels are detected.

$$\kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy} \tag{2}$$

Algorithm

Fourier Transform

The fast Fourier transform converts a signal from its original domain to a rep- resentation in the frequency domain. The FFT computes there for the discrete Fourier transform (DFT) and produces exactly the same result but is much faster For $x_0, \cdots, x_{N-1}$ complex numbers. The DFT is defined by equation eq:FFT

$$\overset{x}{k} = \sum_{n=0}^{N-1} x_n e^{\frac{-2i\pi kn}{N}} \qquad k = 0, \cdots N-1 \tag{3}$$

The evaluation needs $O(N^2)$[5]operations and there are N outputs. With the FFT this can be reduced to $O(NlogN)$. There are enough software libraries to calculate die coefficients. Since the signals are real, the real FFTs (RFFT/IRFFT) saves half of the computation time by only calculating half of the spectrum.

Energy of a Frequency Band

With the FFT coefficients it is possible to calculate the energy all frequency bands equation eq:FFTEnergie and also of a sub band equation eq:FFTSubEnergie

$$\sum_{n=0}^{N-1} |x[n]|^2 \tag{4}$$

$$\sum_{n=k}^{m} |x[n]|^2 \qquad n \in N, n \leq m \leq N-1 \tag{5}$$

Random Forest

Random Forests is a flexible and easy to use machine learning algorithm. It creates most of the time greats results even without hyper-parameter optimization. The forest is an ensemble of Decision Trees, which are trained with the bagging method. Thru

Recording	Hand	Automatic Offset
Shoulder Camera	41.472 s	41.5125 s
Stative Camera 1	86.791 s	86.4787 s
Stative Camera 2	1481.17 s	1481.2025 s
Watch Left	0.768 s	0.7975 s
Watch Upper wrist	1.024 s	0.7400 s

Table 1: Offset of Recording

this, randomness is added, which leads to splitting on the best feature of the subset. The result is a more diverse Tree and a better Random Forest. Therefore, the Random Forest incorporates a random subset of the features.

A random forest for the average sum of decision trees as seen in equation eq:RandomForest with $f_b(x)$ is a decision tree result and B is the number of decision trees.

$$f = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{6}$$

Convolutional Neural Network

A Convolutional Neural Network (CNN) is a learning algorithm which can take in an input and assign importance by learnable weights and biases. With enough training, CNN can learn filters and predict classes. The kernel convolution is a key element of CNNs. In its process, it takes a small matrix of numbers (kernel or filter) and process it over the input and transform the input based on values from the filter. The feature map values are calculated according to equation eq:CNN where the input matrix is f and the filter is h. The indexes of rows and columns of the resulting matrix are n and m.

$$G_{n,m} = \sum_j \sum_k h_{j,k} f_{m-j,n-k} \tag{7}$$

Result

Synchronization

With the Cross-correlation, there is usually a peak at the position with the biggest overlap of the audio file. Compare is the result of the manual and automatic synchronization. The maximal difference between the two results in 0.43 seconds.

Ground Truth

The next results are in regards to the reliability of the ground truth form a different person and cameras. Therefore, several people have viewed the footage of one video source. Additionally, one person also viewed and annotated the footage of different sources (shoulder and stative camera) within a week of pause.

The κ value of 0.8 is a good indicator, that there is a large mismatch just between different camera sources. This can be contributed to the fact that the 360 shoulder camera a better view of the task has. But also when the different person one source annotate then the result just gets a κ of only 0.87. This leads to the problem that the task separation can be arbitrary in some cases. like when starts the Wedging or is the task part of the preparation of the tree? Or where is the lin between complementary work like equipment moving and of preparation for the tree? For example, is determining where a tree is falling while moving the equipment already part of the preparation and when it is, how much?

Random Forest

The first comparison of an automatic method is the random forest against the manual labeled dataset.

The random forest is being trained with all data and labels with increasing the number of available sources. First only the audio data of the mobile phone was used. This leads to a terrible result with a negative κ score Table 3. From this, it can be concluded that the audio data alone is not sufficient to train the classifier.

With Table 4 the classifier improves since also acceleration data are now used but only from the mobile phone. The F_1 score is now for felling and complementary work over 0.5 but the entire F_1 score and κ are only at around 0.32. Please note that the F_1 score has increased for nearly all labels.

In the last result table 5 is not only the mobile phone but also the data from the right wristwatch (dominant hand) have been used. All other comparison resulted in a lower κ and F_1 score. This has led to a further significant improvement in all labels. This results in a F_1 score of 0.598 and a κ score of 0.507.

Convolutional Neural Network

The CNN was only trained on the right smartwatch and mobile phone. This leads to an even better result than the random forest. The final κ score was 0.710 and an F_1 score of 0.800 as the confusion matrix table 6 shows. The result is overall good and usable.

Conclution

Thru the mismatch in the ground truth is shown that even experts can not fully agree on the labeling of one video. Additionally, the mismatch between two different recording sources shows that a κ and F_1 score of 0.8 results in human performance. One source of data like mobile phone or smartwatch is not enough to predict the right label and audio with the acceleration of two different recording devises result in the best result. Still, the random forest result was with a κ of 0.50 and F_1 score of 0.59 not really good. There where many outliers on the hydraulic and mechanical wedging. As a comparison, the CNN resulted in a near-human performance with a κ of 0.71 and F_1 score of 0.80.

Acknowledgment

Special thanks to Felix Rinderle[1] from the Albert-Ludwigs-Universität Freiburg of the department "Professur für Forstliche Verfahrenstechnik" which supervised the data collection and annotation.

And also to Philipp M. Scholl[2] from the Albert-Ludwigs-Universität Freiburg of the Computer Science Department for supervising my master thesis.

[1]Albert-Ludwigs-Universität Freiburg, Professur für Forstliche Verfahrenstechnik, Werthmannstraße 6,79085 Freiburg i.Br. , Germany

[2]Albert-Ludwigs-Universität Freiburg, computer Science Department ,Georges-Köhler-Allee 010, 79110 Freiburg i.Br. , Germany

		shoulder Camera						
		comp Work	prepar	felling	mech Wedg	proc	hy Wedg	Score
Stative Camera	comp Work	80823	54302	0	0	10839	1357	
	prepar	24	46205	2484	0	0	0	
	felling	440	149	130311	371	0	967	
	mech Wedg	0	0	0	9507	0	1334	
	proc	8705	0	0	0	108634	0	
	hy Wedg	1273	0	262	119	0	57769	κ = 0.800
	F_1 Score	0,678	0,619	0,982	0,912	0,917	0,956	F_1 = 0.844

Table 2: Ground truth of different cameras

		Ground Truth						
		prepar	comp Work	proc	felling	hy Wedg	mech Wedg	Score
Predicted	prepar	35159	14155	23939	25369	4348	573	
	comp Work	52813	30171	24800	61509	7254	3785	
	proc	19560	18000	18185	67219	4574	2389	
	felling	69341	34278	57821	66693	9051	753	
	hy Wedg	2188	5970	474	849	181	40	
	mech Wedg	0	0	0	0	0	0	κ = -0.021
	F_1 Score	0,249	0,213	0,142	0,29	0,01	0	F_1 = 0,181

Table 3: Random forest trained on only audio files of mobilephone

		Ground Truth						
		prepar	comp Work	proc	felling	hy Wedg	mech Wedg	Score
Predicted	felling	1279	124	130	12	850	0	
	prepar	16	152	398	63	12	7	
	comp Work	19	231	554	119	17	22	
	hy Wedg	4	8	21	4	1	2	
	proc	360	29	10	1	571	0	
	mech Wedg	0	0	0	0	0	0	κ = 0.327
	F_1 Score	0,628	0,255	0,534	0,033	0,472	0	F_1 = 0.320

Table 4: Random forest trained on only audio and acceleration data of mobile phone

		Ground Truth						
		prepar	comp Work	felling	proc	hy Wedg	mech Wedg	Score
Predicted	prepar	409	287	18	3	66	10	
	comp Work	159	574	10	24	68	8	
	felling	129	97	1306	581	4	0	
	proc	17	2	343	843	0	0	
	hy Wedg	0	3	1	0	61	0	
	mech Wedg	0	0	0	0	0	15	κ = 0,507
	F_1 Score	0,543	0,636	0,688	0,635	0,462	0,625	F_1 = 0,598

Table 5: Random forest trained on audio and acceleration data of mobile phone and left wristwatch

		Ground Truth						
		prepar	comp Work	felling	proc	hy Wedg	mech Wedg	Score
Predicted	prepar	546	163	13	2	47	2	
	comp Work	151	698	10	24	23	6	
	felling	58	27	1306	187	4	0	
	proc	17	2	343	1050	0	0	
	hy Wedg	0	3	1	0	223	0	
	mech Wedg	0	0	0	0	0	24	κ = 0.710
	F_1 Score	0,707	0,773	0,802	0,785	0,851	0,857	F_1 = 0,800

Table 6: Random forest trained on audio and acceleration data of mobile phone and left wristwatch

References

[1] Akhavian, Reza; Behzadan, Amir H.: Smartphone-based construction workers??? activity recognition and classification. Automation in Construction, 71(Part 2):198–209, 2016.

[2] Berchtold, Martin; Budde, Matthias; Gordon, Dawud; Schmidtke, Hedda R; Beigl, Michael: ActiServ: activity recognition service for mobile phones. ISWC '10: Proceedings of the 14th IEEE International Symposium on Wearable Computers, S. 1–8, 2010.

[3] Bao, Ling; Intille, Stephen S: Activity Recognition from User-Annotated Acceleration Data. In (Ferscha, Alois; Mattern, Friedemann, Hrsg.): Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings, S. 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[4] Bijak, Szymon; Sarzyński, Jakub: Accuracy of smartphone applications in the field measurements of tree height. Folia Forestalia Polonica, Series A, 57(4):240–244, 2015.

[5] Csikszentmihalyi, Mihaly; Larson, Reed: Validity and Reliability of the Experience-Sampling Method. In: Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi, S. 35–54. Springer Netherlands, Dordrecht, 2014.

[6] Gyorbiró, Norbert; Fábián, Ákos;Ományi, Gergely: An Activity Recognition System For Mobile Phones. Mobile Networks and Applications, 14(1):82–91, 2008.

[7] Golding, A.R.; Lesh, N.: Indoor navigation using a diverse set of cheap, wearable sensors. Digest of Papers. Third International Symposium on Wearable Computers, S. 29–36, 1999.

[8] Kern, Nicky; Schiele, Bernt; Schmidt, Albrecht: Multi-sensor activity context detection for wearable computing. Ambient Intelligence, S. pp 220–232, 2003.

[9] Kwapisz, Jennifer R.; Weiss, Gary M.; Moore, Samuel A.: Activity recognition using cell phone accelerometers. ACM SIGKDD Explorations Newsletter, 12(2):74, 2011.

[10] Lakshminarayanan, Balaji; Teh, Yee Whye: Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. arXiv preprint, S. 1–19, 2013.

[11] Lukowicz, Paul; Ward, Jamie A; Junker, Holger; Stäger, Mathias; Tröster, Gerhard; Atrash, Amin; Starner, Thad: Recognizing Workshop Activity Using Body Worn Microphones and Accelerometers. PERVASIVE, S. 18–34, 2004.

[12] Long, Xi Long Xi; Yin, Bin Yin Bin; Aarts, R.M.: Single-accelerometer-based daily physical activity classification. 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, S. 6107–6110, 2009.

[13] Maurer, U.; Smailagic, A.; Siewiorek, D.P.; Deisher, M.: Activity recognition and monitoring using multiple sensors on different body positions. International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06), S. 4–7, 2006.

[14] Van Laerhoven, K.; Cakmakci, O.: What shall we teach our pants? Digest of Papers. Fourth International Symposium on Wearable Computers, (c):77–83, 2000.

[15] Ward, Jamie a.; Lukowicz, Paul; Gellersen, Hans: Performance metrics for activity recognition. ACM Transactions on Intelligent Systems and Technology (TIST), 2(1):1–23, 2011.

Author Biography

Stefan Dilger received his MS in Embedded System Engineering from the University of Freiburg (2018). Since 2019 he worked as research associate at Fraunhofer Institut Secure Information Technology (SIT) in Darmstadt.

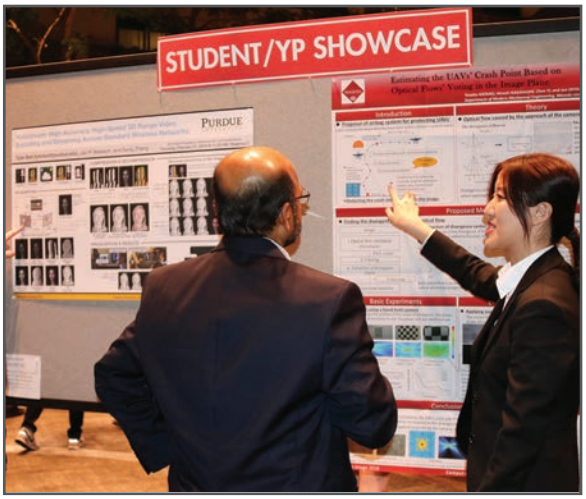
JOIN US AT THE NEXT EI!

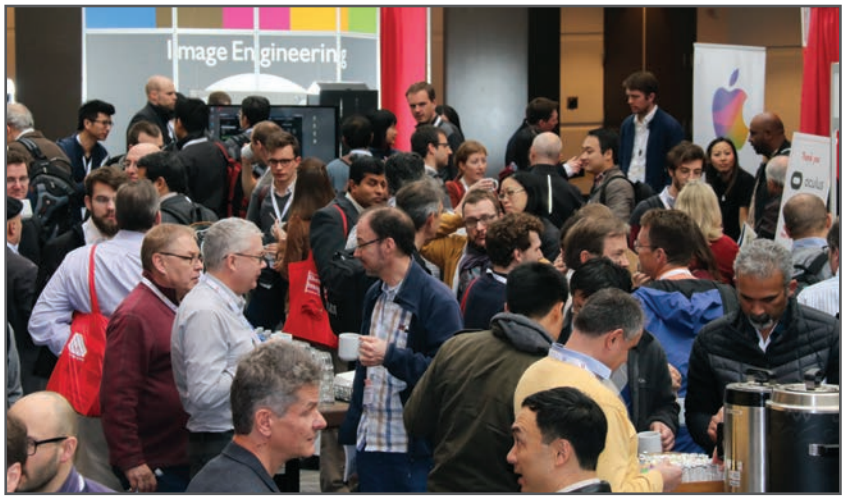
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org