Crowdsourcing Video-based Workface Assessment for Construction Activity Analysis

Kaijian Liu and Mani Golparvar-Fard

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, USA
E-mail: kliu15@illinois.edu, mgolpar@illinois.edu

ABSTRACT

Today, the availability of multiple cameras on every jobsite is reshaping the way construction activities are being monitored. Research has focused on addressing the limitations of manual workface assessment from these videos via computer vision algorithms. Despite the rapid explosion of these algorithms, the ability to automatically recognize worker and equipment activities from videos is still limited. By crowd-sourcing the task of workface assessment from jobsite videos, this paper aims to overcome the limitations of the current practice and provides a large empirical dataset that can serve as the basis for developing video-based activity recognition methods. As such, an intuitive web-based platform for massive marketplaces such as Amazon Mechanical Turk (AMT) is introduced that engages the intelligence of non-expert crowd for interpretations of selected group of frames from these videos and then it automates remaining workface assessment tasks based on the initial interpretations. To validate, several experiments are conducted on videos from concrete placement operations. The results show that engaging AMT non-experts together with computer vision algorithms can provide assessment results with an accuracy of 85%. This minimizes the time needed for workface assessment, and allows the practitioners to focus their time on the more important task of root-cause analysis for performance improvements. This platform also provides significantly large datasets with ground truth for algorithmic development purposes.

Keywords – Construction Productivity, Workface Assessment, Crowdsourcing, Computer Vision

1 Introduction

Activity analysis, the process of analysing and improving the time proportions craft workers spend on various construction activities, provides a plausible solution for monitoring onsite operations [1]. It also supports root-cause analysis on the issues that adversely affect productivity. Nevertheless, a wide-spread adoption of activity analysis has been challenging due to several inefficiencies such as the large-scale of the manual on-site observations needed to guarantee statistically significant workface data. The necessary judgments of the observers may also produce erroneous data due to the over-productiveness phenomenon caused by workers being under direct observations, prompt reaction of the observers to benchmarking activity categories, the required distance limits to construction workers, and finally observers' partiality and fatigue [2]. Hence, current labour-intensive processes requires the field engineers to spend most of their time analysing current worker activities. This takes away time from the more important task of studying the root causes of low productivity or how improvements can be planned and implemented.

To address the limitations of manual workface assessment, a large body of research has focused on automating current practices. These methods range from application of Ultra Wide Band [3,4], RFID tags [5,6], and GPS sensors [7] to vision methods using jobsite videos [8–10]. A large group of these methods leverage non-vision sensors to track the location information of the workers and equipment. However, without interpreting the activities, and purely based on location, deriving workface data is challenging [11].

Meanwhile, the growing number of cameras on jobsites has provided a unique opportunity for automated interpretation of onsite operations. Yet, computer vision methods are not advanced enough to enable detailed assessments from jobsite videos. This is because the methods for detecting and tracking equipment and workers (especially when workers interact with tools [12–14]), and the methods for interpreting activities from long sequences of videos [15,16] are not well established. Beyond CII-defined activities, the taxonomies of construction activities are also not fully developed to enable visual activity recognition at the operational level. Training and testing computer vision models for activity analysis requires large amount of empirical data which is not yet available to the research community. Without

robust methods for video interpretation, extracting workface assessment information from videos still has to rely on tedious manual review process. This will take away time from the more important task of root-cause analysis.

In this paper, a workface assessment framework is introduced to collect accurate workface assessment information by interpreting jobsite videos. Through crowdsourcing the task of workface assessment from jobsite videos on a web-based platform on Amazon Mechanical Turk (AMT), and with the supports of several automated methods and the intelligence of non-expert crowds, this framework aims to overcome the limitations of the current practices of activity analysis. It also collects significantly large empirical datasets together with their groundtruth that can serve as the basis for developing automated computer vision activity recognition methods. The preliminary experiments from using platform shows that engaging non-experts on AMT to interpret and annotate construction activities on jobsite videos can achieve accurate activity analysis results. In the following, the related works, methods and the developed prototype tools, together with experimental results are introduced and discussed in detail.

2 Related Work

A large body of literature in vision-based analytics has focused on developing methods to infer worker and equipment activities from jobsite videos. Methods such as [2,12,13,17,18] treat detecting and tracking workers and equipment and activity recognition as two mutually independent tasks. A few studies have also focused on the end-to-end activity analysis problem [10,15,19]. Instead of assuming strong priors on the relationships between activities and locations as in the above works, other vision-based methods such as [15,16,20] recognize atomic activities of construction workers and equipment. Despite the explosion of these methods, the ability to automatically recognize and understand worker and equipment activities is still limited. The key challenges are the large variability in execution of construction operations, and the lack of formal taxonomies for construction activities in terms of expected worker roles, and the expected sequence of activities. The complexity of the visual stimuli in activity recognition in terms of camera motion, occlusions, viewpoint changes, and background clutter are among these challenges as well. The lack of datasets together with ground truth is also another barrier to more extensive research on automated activity recognition methods.

To overcome these limitations, the computer vision community has initiated several projects to investigate the potential of Crowdsourcing. Crowdsourcing refers to the collaborative participation of a crowd of non-experts to help solve a specific problem and typically devises a rewarding mechanism. While crowdsourcing annotations of images has been very successful (see [21,22]), yet videos and their dynamic nature make crowd-sourced their annotation very challenging [23]. Particularly, crowd-sourced video annotation requires cost-aware and efficient methods instead of frame-by-frame labelling. The large number of frames in a video requires a more intelligent mechanism for propagating annotations from a subset of keyframes [23,24]; otherwise video crowd-sourcing methods will not be scalable.

Despite the benefits and popularity of crowdsourcing computer vision tasks, directly applying it to the task of video-based construction workface assessment can be challenging. Site videos exhibit different number of crew members involved in onsite operations. Workers continuously interact with tools, and exhibit changing body postures even when the same activity is being performed. These issues beyond the typical challenges in the task of activity recognition and could negatively affect the optimal length of annotation tasks or the number of necessary keyframes for annotation. Due to the complexity of construction operations and the lack of formal activity categories beyond CII defined activities, crowdsourcing workface assessment necessitates a new taxonomy to describe construction activities. Until now, the reliability of crowdsourcing for video-based construction workface assessment has not been evaluated either. For example, recruiting non-expert annotators from AMT may negatively affect the quality of the assessments. Beyond addressing technical challenges, detailed experiments are necessary to examine the potential of non-experts against an expert control group, and formulate new strategies for improving the accuracy of crowdsourcing tasks.

3 Method

To address challenges of applying crowdsourcing to video-based construction workface assessment task, this paper introduces a new framework for crowdsourcing video-based workface assessment. The prototyped platform benefits from intuitive user interfaces that enable construction workface assessment data retrieval, visualization, and cross-validation. It also exhibits several automated methods that support propagating the annotations from a subset of keyframes to the remaining frames. Several preliminary experiments are conducted on different methods for annotation and their frequency. The most appropriate video length for annotation to fine tune parameters of the automated methods is also investigated. A new taxonomy for construction activities is decoded as well. The performance of the no-experts is analysed again an expert group for detecting, tracking, and recognizing worker activities. Applying cross-validation methods to improve workface assessment accuracy is also investigated. Figure 1 shows an

overview of the crowdsourcing framework for workface assessment. In the following, various modules of the prototyped platform and the validation experiments are discussed in detail.

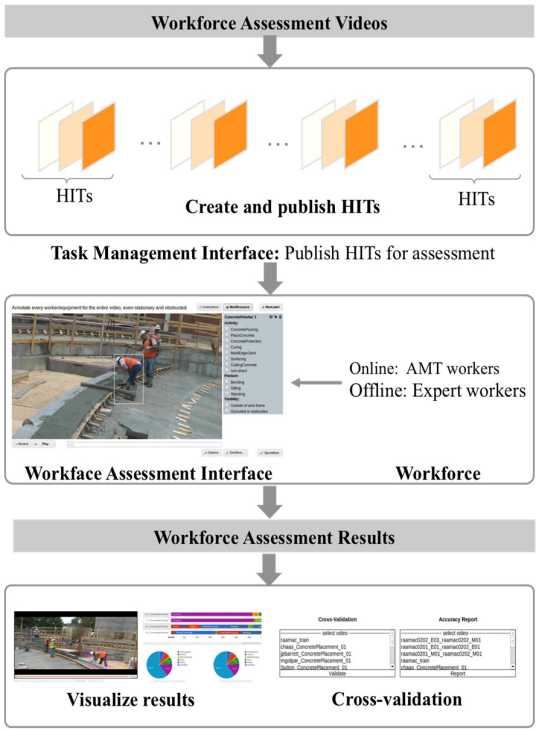

Figure 1. The workflow in the crowdsourcing workface assessment tool.

3.1 Activity Analysis User Interface

"Task Management" and "Workface Assessment" are the two main interfaces in the prototyped platform. Task Management Interface assists requesters, e.g. site engineers, or researchers, to manage publication of workface assessment tasks and retrieve their results. The workface Assessment Interface provides the annotators with access to complete video-based workface assessment tasks. As shown in Figure 1, at first, the requesters use Task Management Interface to break a video of construction operations into several Human Intelligent Tasks (HITs) and publish each shorter video clip online or offline. Annotators can then accept published HITs to generate workface assessment results using the Workface Assessment Interface. When all HITs belonging to the same posted video are completed, requesters can retrieve, visualize, and cross-validate the results. They can also generate formal assessment reports through the Task Management Interface (see Figure 2).

The Task Management interface includes: 1) "Video Links"—assisting requesters in managing access of the annotators and controlling the quality of their work, 2) "Video Upload"—supporting crowdsourcing workface assessment and collection of large-scale ground-truth dataset for both academia and industry, 3) "Cross-Validation and Accuracy"—enabling quality control on both pre and post-assessment steps to avoid using inaccurate results produced by non-expert annotators, and 4) "Video Visualization"—supporting retrieval of workface assessment information at any level of granularity and presenting workface assessment results in form of annotated videos, crew-balance charts, or pie charts. The Workface Assessment interface, as shown in Figure 3, includes: 1) assessment function—enabling "role-activity-tool-posture" annotation and 2) supporting functions for creating and customizing new resources, and adjusting "monitoring" settings.

Figure 2. The Task Management interface: (top) annotated video; (middle) crew-balance chart; and bottom) detailed and CII-type activity pie charts. The annotations in the upper left corner of each box overlaid on the annotated video shows the role-activity-tool-body posture of each worker.

3.2 Taxonomy of Construction Activities

A new taxonomy is introduced to decode complex construction activities with the following format: *worker role* is conducting *CII activity category* (c_{II}) in form of a *visual activity category* (c_v) using *tool* (t_v), at *body posture* (p_v), and is *visible*, *occluded*, or *outside of the video frame*. As a starting point, "worker type" contains 20 different roles of construction worker such as Concrete Finisher, Carpenter, Electrician, Bricklayer, etc. Second layer—"CII activity categories"—describes

worker activities in form of direct and non-direct work as defined by CII. The third layer is the "visual activity categories" which provides detailed information on "activities", "tools" and "body posture" related to the direct work activities. "Tools" can play an important role here since they have distinct visual appearances and because different types of "activities" require different types of tools. An assistive interface that provides illustrative images has been designed to aid non-expert annotators in selecting different tools out of a large collection. "Posture" can also provide very useful information for differentiating different activities. The detailed representation of activity-tool-body posture in this visual activity layer is beneficial to the extraction of proper visual features and devising appropriate computer vision methods. In addition, this taxonomy describes non-direct works with worker body postures to enable a better synthesis of the work activities. Since construction videos typically exhibit sever occlusions and to avoid introducing noise in video-based activity datasets, the visibility of the workers– i.e. whether workers are occluded and/or are outside of video is also labelled. The complete taxonomy for concrete placement operations is available at http://activityanalysis.cee.illinois.edu/.

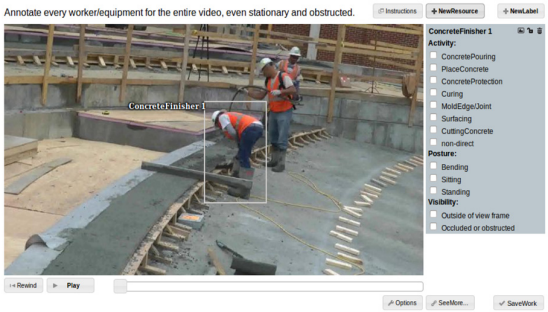

Figure 3. Workface Assessment Interface

3.3 Extrapolating Keyframe Annotations

Although crowdsourcing can reduce human efforts, time and cost for workface assessment, video annotation still needs strategies to propagate assessment results from a sparse set of keyframes. For propagating user annotations from keyframes to non-key frames, both linear and detection-based extrapolation methods are implemented. T is defined as the total number of frames, and $B = [x_{min}, x_{max}, y_{min}, y_{max}]$ as the 2D pixel coordinates of each annotation bounding box. x_{min}, y_{max} denote the coordinates of upper-left corner, and x_{max}, y_{min} denote the coordinates of the lower-right corner of the bounding box. $B_t (0 \leq t \leq T)$ is then defined as the coordinates of the bounding box at time t.

Linear extrapolation method assumes 2D constant velocity in both x and y direction; and if a point in x direction is at $B_0(x_{min})$ and then at $B_T(x_{min})$ the $B_t(x_{min})$ follows Eq. 1. Applying coordinates of the keyframe bounding boxes, B_t can be calculated as Eq. 2.

$$B_t(x_{min}) = (T - t) \times \frac{B_T(x_{min}) - B_0(x_{min})}{T - 0} \quad (1)$$

$$B_t = (T - t) \times \frac{B_T - B_0}{T} \quad (2)$$

Detection-based extrapolation method treats keyframe annotations as positive samples to train machine learning classifiers. The visual feature descriptors x_i will consist of HOG and HOC, which are as shown in Eq. 3:

$$x_i = \begin{bmatrix} HOG \\ HSV \end{bmatrix} \quad (3)$$

where, HOG is computed based on [25], and HOC is a nine-dimensional feature containing 3 means and 6 covariance computed from Hue, Saturation and Value color channels. Positive samples are the user-annotated keyframes and negative samples are automatically extracted from background. Here x_i refers to both of these annotations. To learn a specific visual classifier that is able to assign high classification scores to the positive samples, the same procedure in [14] is followed and a Support Vector Machine (SVM) classifier is introduced per resource type. Each SVM classifier is trained by feeding positive/negative samples $(x_i, +1)/(x_i, -1)$ to optimize maximum margin objective function. Due to the presence of frequent occlusions and background clutter on construction sites, it is difficult to propagate non-key frame annotations at an accuracy of 100%. Therefore, the constrained tracking of [23] is applied to reduce the error in detection of workers and equipment. Constrained tracking finds the best candidate from all possible detections for each frame to constitute a path with minimum cost. This path is defined as $B_{0:T} = \{B_o, B_1, \dots, B_{T-1}, B_T\}$, where B_0 and B_T are manually generated keyframe annotations and $B_o, B_1, \dots, B_{T-1}, B_T$ are automatically generated from the trained SVM classifiers. The optimization problem is then defined as:

$$\underset{b_{1:T-1}}{\operatorname{argmin}} \sum_{t=1}^{T-1} U_t(B_t) + P(B_t, B_{t-1}) \quad (4)$$

where the unary cost $U_t(B_t)$ is defined by Eq. 5 and pairwise cost $P(B_t, B_{t-1})$ is defined by Eq. 6:

$$min(-w \cdot \phi(B_t), \alpha_1) + \alpha_2 \|B_t - B_t^{lin}\| \quad (5)$$

$$P(B_t, B_{t-1}) = \alpha_3 \|B_t - B_{t-1}\|^2 \quad (6)$$

The unary cost $U_t(B_t)$ calculates the cost of the potential detection in each frame by the score of the visual classifier and l_2 -norm of its bounding box difference between the SVM detection and the linear extrapolation. The SVM associates the most possible

prediction with the highest score to minimize the most likely cost for the detection. Here, $-w \cdot \phi(B_t)$ is used as the score of the visual classifier. Due to the presence of occlusions, some video frames may not contain ground truth detection. These frames cause false negatives with small scores to be the potential B_t. In such situations, the annotations for non-key frames will rely on the linear extrapolation method, and will replace classifier score $-w \cdot \phi(B_t)$ with a very small (zero number) α_1. The pairwise cost calculates smoothness of the detection path for each worker. Here, the position of the bounding box does not change if the camera motion is minimal. Thus, a true path should have the minimum pairwise cost among all possible candidates. This pairwise cost has been adopted as a gauge to test and select the best candidates for the path in each frame.

3.4 Annotating Multiple Workers

Annotating multiple resources in a jobsite video is common, because most construction crews involve a large number of workers. Therefore, an efficient annotation method is imperative to reduce time and guarantee quality for dense annotations. In the experiment section, three annotation methods are validated to address multiple resources annotation tasks; their performance under different conditions is analysed as well. These annotation methods include one-by-one, all-at-once, and role-at-once. One-by-one annotation method requires the annotators to annotate/update a construction resource to the full length of each video, and rewind video to start for next resource until all resources has been annotated/updated. All-at-one annotation method asks the annotators to annotated/update all resources simultaneously in the same frame, and do so until the end of the video. Role-at-once annotation method is the combination of one-by-one and all-at-once, which is to annotate/update resources with the same role simultaneously and rewind video to annotate/update the next group of resources. One-by-one annotation method needs annotators to watch video for N times (i.e., N is resource number). All-at-once annotation method requires annotators to watch the entire video prior to conducting annotation and to watch video twice in total. And, role-at-once requires annotators to watch video M time (i.e., M is number of roles).

3.5 Quality Assurance/Control for AMT

The AMT is a marketplace of tens to hundreds of thousands of annotators for solving HITs quickly and effortlessly. However, due to poor performance of the annotators, quick assessments may lead to erroneous results. To lower the risk of obtaining poor quality results, pre and post-assessment quality control steps are defined as follows: 1) pre-assessment—adding a short testing video—for which the ground truth is previously generated—to the start of each HIT, and 2) post-assessment—repeated-labelling a video for multiple times to deal with noisy data for quality improvement purposes. Although repeated-labelling improves the quality of workface assessment, the unnecessary repeated-labelling will cost extra money and time. To save annotation cost and time, it is necessary to find an optimal repeat time. Thus, cross-validation is conducted to examine how the accuracy changes based on different repeat times:

3.6 Designing Micro-tasks from Site Videos

A video of a single construction operation is typically several hours long. This makes assessing it much more difficult than completing a typical short-length micro-task (HITs) on the AMT. To make crowdsourcing feasible, an entire video is broken down into several shorter HITs. For effectiveness, 1) the length of a HIT, and 2) annotation frequency are considered. To find the optimal parameters for crowdsourcing video-based workface assessment, several experiments are conducted using different length of a HIT and annotation frequency. The experiments are discussed in the following section.

4 Experiments

4.1 Setup and Performance Measures

Three separate experiments are conducted with the annotators from the controlled group of construction experts to investigate the impact of the different annotation methods, video lengths, and annotation frequencies on the accuracy of the workface assessment results. To validate the hypothesis that crowdsourcing video-based construction workface assessment through the AMT marketplace is a reliable approach, we conducted two experiments to 1) compare the performance of non-expert annotators with the controlled group of construction experts; and 2) test the performance of the post-assessment quality control procedure and explore the best repeated-labelling times for desirable level of accuracy by experimenting cross-validation with different randomly selected folds from both expert and non-expert annotation results. To compare experiments and choose the parameters that achieve optimal performance, two validation measures were chosen: 1) the annotation time spent to complete each experiment and 2) the accuracy of the workface assessment results.

4.2 Results and Discussions

The experiments for selecting the best annotation method include leveraging one-by-one, all-at-once, and role-at-once methods to annotate “Easy”, “Normal”, and

"Hard" videos. To examine the relationship between video length and the annotation time and accuracy of workface assessment, experiments are conducted using different video lengths of 10, 30, and 60 seconds. To explore the trade-off between annotation frequency and annotation time and accuracy, three fixed annotation frequencies of 3-time, 5-time, and 9-time per minute are experimented. The experimental results of annotation time and accuracy for each experiment are reported in Tables 1 and 2 respectively.

Table 1. Total annotation time on each experiment (s).

	Videos with different levels of difficulty		
	Easy	Normal	Hard
	Time for each annotation method		
AM01	8,380	20,841	18,232
AM02	9,034	22,801	12,495
AM03	10,943	24,459	18,298
	Time for each video length		
10s	8,380	20,841	18,232
30s	8,349	21,784	10,227
60s	5,050	12,520	7,812
	Time for each annotation Frequency		
9	4,586	8,229	4,975
5	2,572	5,668	3,630
3	1,636	3,451	3,573

Table 2. Average accuracy for each experiment (%).

	Com.	B. B.	Role	Activity	Posture	Tool
	Accuracy of each annotation method					
AM1	0.95	0.98	0.83	0.80	0.97	0.84
AM2	0.95	0.98	0.88	0.71	0.97	0.83
AM3	0.97	0.98	0.99	0.80	0.96	0.82
	Accuracy of each video length					
10s	0.95	0.98	0.83	0.80	0.97	0.84
30s	0.90	0.95	0.88	0.81	0.97	0.83
60s	0.90	0.96	0.90	0.82	0.97	0.74
	Accuracy of each annotation frequency					
9	0.94	0.90	0.99	0.78	0.93	0.83
5	0.87	0.85	0.95	0.74	0.94	0.86
3	0.97	0.83	0.77	0.66	0.95	0.85

* Com. = Completeness; B.B. = Bounding Box

To validate the reliability of crowdsourcing on the AMT marketplace, experiments are conducted to compare the annotation time and accuracy of a large pool of non-experts (10+) against a controlled group of construction experts (5+). Figure 4 shows the difference in the annotation time between the non-expert and expert annotators. The difference in annotation time between the non-expert and expert control groups are compared and the linear regression is used to interpolate all observation points. The average difference in accuracies between these groups is shown in Table 3.

Table 3. The difference in accuracies between the expert and not-expert annotators.

The accuracy of a expert is … than a non-expert	Completeness	Bounding Box
	+0.02	+0.01
	Role	Activity
	+0.02	+0.01
	Posture	Tool
	+0.02	+0.01

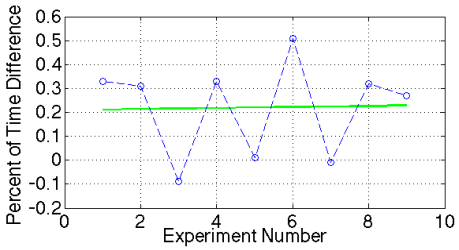

Figure 4. Annotation time difference between expert and non-expert annotators.

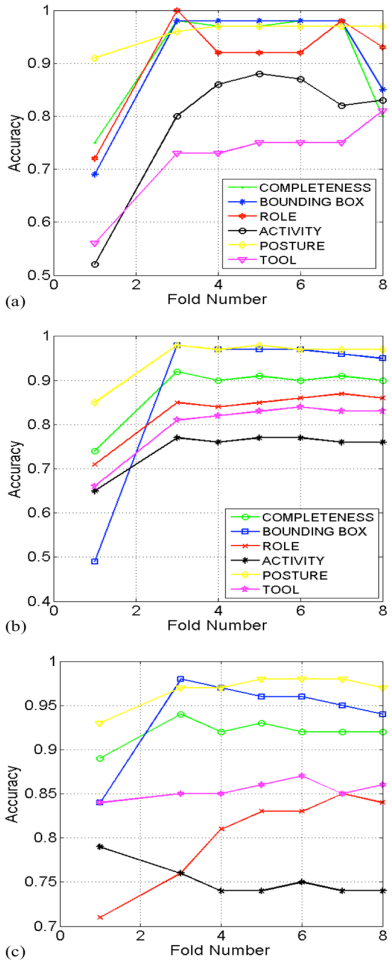

Figure 5. Cross-Validation results for (a) Easy, (b) Normal, and (c) Hard Videos.

To explore the best repeated-labelling times and guarantee satisfactory workface assessment results, different fold experiments were conducted, including 1-fold (i.e., original annotation), 3-fold to 8-fold cross-validations with one step increments. Figure 5 shows the accuracies of different fold cross-validations for each video category.

The experimental results validate the hypothesis that crowdsourcing construction activity analysis from jobsite videos on the AMT, a marketplace with non-expert annotators, is a reliable approach for conducting activity analysis. Particularly, it is shown that expert annotators are, on average, 22% faster than non-expert annotators in term of their annotation time. However, the accuracy of annotation among the non-experts is within 3% of the accuracy of the expert groups. To fine-tune the platform, the impact of different annotation methods, different HITs video lengths, and the frequency of requiring annotations were examined. Based on these experimental results, the following patterns are concluded:

1. The one-by-one annotation method works best with videos that have a small number of construction workers and also high frequency of changes in activities; whereas the all-at-once annotation method works best with videos that have high number of construction workers and with low frequency of changes in work activities;

2. Increasing a HIT video length can reduce the annotation time. For example, the 60-second long videos save 47% and 37% annotation time compared to 10 and 30-second long videos. The accuracy of workface assessment results also slightly improves with an increase in the HIT video length;

3. Manual annotation of a sparse set of keyframe is reliable for achieving complete frame-by-frame annotations. At the extreme case, the 3-time per minute annotation frequency reduces the average annotation time by 57% while dropping the accuracy of workface assessment only by 7%. The prominent decrease in annotation time by 3-times per minute could also result in cost savings, because AMT charges "requester" based on the time "annotator" devotes to each HIT;

4. A 3-fold cross-validation provides the best accuracy-cost trade-off for workface assessment. Increasing the fold (beyond 3-folds) does not increase accuracy. Also quality control steps are important to guarantee that reliability of the assessment results. The repeated labelling can also improve the accuracy of the workface assessment. The experiments suggests that the optimal performance can be achieved with a 3-fold—i.e., hiring three AMT annotators per HIT.

5 Conclusion and Future Work

This paper presents a novel method that supports crowdsourcing construction activity analysis from jobsite video streams. The proposed method leverages human intelligence recruited from massive crowdsourcing marketplace—AMT, together with automated vision-based detection/tracking algorithms to derive timely and reliable construction workface assessment result from different challenging conditions such as sever occlusion, background clutter, and camera motions. The experimental result with average accuracy of 85% in workface assessment tasks shows the promise of proposed method. The comparisons conducted between non-experts and construction validates the hypothesis that crowdsourcing video-based construction activity analysis through AMT non-experts could achieve similar (or even the same) accuracy as conducting activity analysis by construction experts.

To improve the prototyped platform, future work can focus on the following: 1) the design of more robust detection/tracking algorithm that can work well with sparse human input to effectively generate accurate non-keyframe annotations; and 2) the design of quality control method that does not require repeated labelling, to reduce requesters' cost and avoid erroneous data to prevail at the voting stage. As part of this preliminary study, a compositional structure taxonomy for construction activities was also created that models the interactions between body posture, activities, and tools. This representation can improve detection/tracking by enhancing the propagation of manual annotations to non-key frames. Also, studies that focus on using Hidden Markov Model to automatically infer construction activities from long sequences of jobsite videos could be beneficial to detection/tracking and quality control steps. Learning a set of transition and emission probabilities between each pair of construction activities from a crowd-sourcing platform can improve inference on the categories of subsequence activity types for each frame, and in turn improve the quality control process. Finally, to comprehensively validate this new method for construction video-based analysis, a set of detailed crowdsourcing market investigations and experiments should be conducted, not only to test the technical parameters, but also to build a workflow to test the cost associated with crowd sourcing, the time span between publishing and retrieval tasks, and potential risks of affecting worker privacy by outsourcing construction video annotations containing construction workers to the crowd. The prototyped platform is publicly accessible online at the following link: http://activityanalysis.cee.illinois.edu/.

Acknowledgement

We would like to thank our industry partners for their support with data collection. We thank Prof. Carl Haas for his constructive feedbacks during the development

the workface assessment platform. This work was financially supported by University of Illinois Dept. of Civil and Environmental Eng.'s Innovation Grant. The views and opinions expressed in this paper are those of the authors and do not represent the views of the individuals or entities mentioned above.

References

[1] Construction Industry Institute (CII), Guide to activity analysis, University of Texas, Austin, TX, 2010.

[2] A. Khosrowpour, J.C. Niebles, M. Golparvar-Fard, Vision-based workface assessment using depth images for activity analysis of interior construction operations, Autom. Constr. 48 (2014) 74–87. doi:http://dx.doi.org/10.1016/j.autcon.2014.08.003.

[3] T. Cheng, M. Venugopal, J. Teizer, P.A. Vela, Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments, Autom. Constr. 20 (2011) 1173–1184.

[4] A. Giretti, A. Carbonari, B. Naticchia, M. DeGrassi, Design and first development of an automated real-time safety management system for construction sites, J. Civ. Eng. Manag. 15 (2009) 325–336.

[5] A. Costin, N. Pradhananga, J. Teizer, Leveraging passive RFID technology for construction resource field mobility and status monitoring in a high-rise renovation project, Autom. Constr. 24 (2012) 1–15.

[6] D. Zhai, P. Goodrum, C. Haas, C. Caldas, Relationship between Automation and Integration of Construction Information Systems and Labor Productivity, J. Constr. Eng. Manag. 135 (2009) 746–753.

[7] N. Pradhananga, J. Teizer, Automatic spatio-temporal analysis of construction site equipment operations using GPS data, Autom. Constr. 29 (2013) 107–122. doi:10.1016/j.autcon.2012.09.004.

[8] A. Peddi, L. Huan, Y. Bai, S. Kim, Development of human pose analyzing algorithms for the determination of construction productivity in real-time, Constr. Res. Congr. ASCE, (2009) 11–20.

[9] J. Teizer, P.A. Vela, Personnel tracking on construction sites using video cameras, Adv. Eng. Informatics. 23 (2009) 452–462.

[10] E. Rezazadeh Azar, B. McCabe, Automated Visual Recognition of Dump Trucks in Construction Videos, J. Comput. Civ. Eng. 26 (2012) 769–781.

[11] A. Khosrowpour, I. Fedorov, A. Holynski, and Juan Carlos Niebles, M. Golparvar-Fard, Automated Worker Activity Analysis in Indoor Environments for Direct-Work Rate Improvement from Long Sequences of RGB-D Images, in: Constr. Res. Congr. 2014, n.d.: pp. 729–738.

[12] I. Brilakis, M.-W. Park, G. Jog, Automated vision tracking of project related entities, Adv. Eng. Informatics. 25 (2011) 713–724.

[13] M.W. Park, I. Brilakis, Construction worker detection in video frames for initializing vision trackers, Autom. Constr. 28 (2012) 15–25.

[14] M. Memarzadeh, M. Golparvar-Fard, J.C. Niebles, Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors, Autom. Constr. 32 (2013) 24–37.

[15] J. Gong, C.H. Caldas, An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations, Autom. Constr. 20 (2011) 1211–1226.

[16] M. Golparvar-Fard, A. Heydarian, J.C. Niebles, Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers, Adv. Eng. Informatics. 27 (2013) 652–663.

[17] S. Chi, C.H. Caldas, Automated Object Identification Using Optical Video Cameras on Construction Sites, Comput. Civ. Infrastruct. Eng. 26 (2011) 368–380. doi:10.1111/j.1467-8667.2010.00690.x.

[18] J.Y. Kim, C.H. Caldas, Vision-based action recognition in the internal construction site using interactions between worker actions and construction objects, 2013 Proc. 30th ISARC. (2013) 661–668.

[19] J. Yang, O. Arif, P.A. Vela, J. Teizer, Z. Shi, Tracking multiple workers on construction sites using video cameras, Adv. Eng. Informatics. 24 (2010) 428–434. doi:10.1016/j.aei.2010.06.008.

[20] V. Escorcia, M.A. Dávila, M. Golparvar-Fard, J.C. Niebles, Automated Vision-based Recognition of Construction Worker Actions for Building Interior Construction Operations Using RGBD Cameras, in: Proc. Constr. Res. Congr., 2012.

[21] A. Sorokin, D. Forsyth, Utility data annotation with Amazon Mechanical Turk, in: 2008.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR, 2009.

[23] C. Vondrick, D. Patterson, D. Ramanan, Efficiently scaling up crowdsourced video annotation, Int. J. Comput. Vis. 101 (2013) 184–204.

[24] F.C. Heilbron, J.C. Niebles, Collecting and Annotating Human Activities in Web Videos, in: Proc. Int. Conf. Multimed. Retr., 2014: p. 377.

[25] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in: Comput. Vis. Pattern Recognit., 2005.