Automated Filtering Big Visual Data from Drones for Enhanced Visual Analytics in Construction

MirSalar Kamari[1] and Youngjib Ham[2]

[1]Graduate Student, Dept. of Construction Science, Texas A&M Univ., 3137 TAMU, College Station, TX 77843. E-mail: kamari@tamu.edu
[2]Assistant Professor, Dept. of Construction Science, Texas A&M Univ., Francis Hall 329B, 3137 TAMU, College Station, TX 77843. E-mail: yham@tamu.edu

Abstract

Nowadays, to assess and document construction and building performance, large amount of visual data are captured and stored through camera equipped platforms such as wearable cameras, unmanned aerial/ground vehicles, and smart phones. However, due to the nonstop fashion in recording such visual data, not all of the frames in captured consecutive footages are intentionally taken, and thus not every frame is worthy of being processed for construction and building performance analysis. Since many frames will simply have non-construction related contents, before processing the visual data, the content of each recorded frame should be manually investigated depending on the association with the goal of the visual assessment. To address such challenges, this paper aims to automatically filter construction big visual data that requires no human annotations. To overcome challenges in pure discriminative approach using manually labeled images, we construct a generative model with unlabeled visual dataset, and use it to find construction-related frames in big visual dataset from jobsites. First, through composition-based snap point detection together with domain adaptation, we filter and remove most of accidently recorded frames in the footage. Then, we create discriminative classifier trained with visual data from jobsites to eliminate non-construction related images. To evaluate the reliability of the proposed method, we have obtained the ground truth based on human judgment for each photo in our testing dataset. Despite learning without any explicit labels, the proposed method shows a reasonable practical range of accuracy, which generally outperforms prior snap point detection. Through the case studies, the fidelity of the algorithm is discussed in detail. By being able to focus on selective visual data, practitioners will spend less time on browsing large amounts of visual data; rather spend more time on looking at how to leverage the visual data to facilitate decision-makings in built environments.

INTRODUCTION

With an advent of portable devices for visual sensing such as the GoPro camera mounted on a hard hat or smartphones & tablets, the as-is state of jobsites is effortlessly recorded on daily, weekly or monthly basis. Moreover, recently, there has been a substantial growth in the usage of Unmanned Aerial Vehicles (UAVs) capturing the as-is conditions of built environments (Ham et al. 2016). Given such captured visual data, construction researchers have primarily studied in construction scene understanding (e.g., equipment detection & tracking, worker action recognition

& classification, or 3D point cloud generation & segmentation) to measure different performance metrics in given built environments, and yielded promising results with significant potentials (Yang et al. 2015). For example, to predict potential risks of wind-induced cascading damages to construction projects and to infer the negative impacts on neighboring communities, we can collect large-scale visual data representing the as-is jobsite condition before extreme wind-related events occur (e.g., during hurricane warning), and recognize potential at-risk construction resources and at-risk states of equipment in jobsites (Ham et al. 2017). The outcomes of such visual recognition can directly communicate with people in jobsites, which enable to identify areas in need of protection, and further can be used for disaster simulation as an input. In addition, by taking jobsite live-stream footage, a vision-based detection and classification algorithms would be used to assess the safety of any workers through detection of their hard hat (Park et al. 2015). Such vision-based system would have potential to further advance monitoring the safety metrics on daily basis, and will empower the in-charge jobsite disciplines to constantly keep an eye on safety metric reports, and minimize possible injuries or death-tolls resulted by construction activities. On top of those, there have been many other research efforts to use multimodal visual data from built environments (e.g., thermography-driven building energy performance modeling and analysis (Ham and Golparvar-Fard 2015), occupational safety analysis (Han and Lee 2013), pavement management (Koch and Brilakis 2011), construction performance analytics (Han and Golparvar-Fard 2017), etc.).

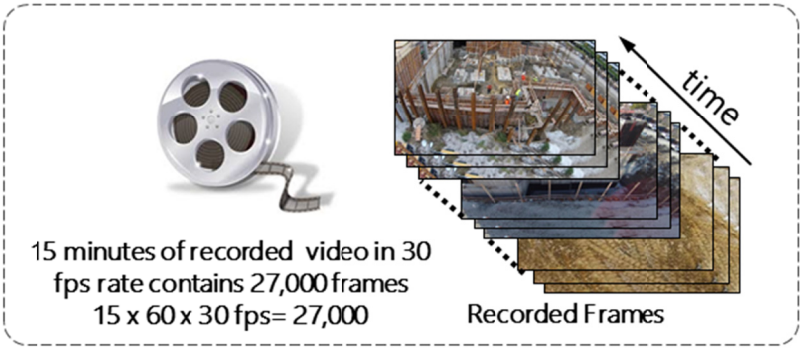

Figure 1. An example of the size of visual data needed to be filtered by site engineers

Figure 2. Examples of non-construction related images from jobsites with poor content relevance (Top), and construction images of great interest containing the as-is jobsite information (Bottom)

Despite the benefits from easily and quickly securing large amount of visual data and analyzing them to extract domain knowledge, the critical challenge is now the 'Scale' issue: how to efficiently process 'Big' visual data, large number of images or long sequence videos, beyond colleting and storing them. In practice, most vision-based platforms involve the three main phases: (1) data acquisition, (2) data classification, and (3) processing. In most cases, because of a lot of unintentionally taken frames that are not related to performance metrics and/or decision-makings, the selection of photoworthy or storytelling frames from big visual dataset should be carried out for more efficient level of assessments. For example, as can be seen in Figure 1, jobsite video at 30 fps captured for 15 minutes involves 27,000 frames. Dealing with this number of photos on daily basis for the aforementioned project-related decision-makings is not a trivial task. This poses a daily challenge for site engineers or practitioners to browse and select well-intentioned sets of construction images (Figure 2) before processing them. To avoid such time-consuming and labor-intensive tasks by practitioners, it would be easy to think of the problem in discriminative terms (i.e., training large dataset and performing classifiers). While training a discriminative classifier using manually labeled exemplars has proven successful for learning high-level image properties, it is not trivial to secure adequate and unbiased labeled visual data that people manually mark frames that likely appear intentional, which would be susceptible to bias and/or difficult to scale. To address such challenges in training a discriminative classifier based on manually labeled exemplars, there has been an effort to select optimal key frames in the recorded footage to generate point clouds (Rashidi et al. 2013). In the case of point could generations, not all frames in given video are worth being processed. This method involves high quality frame filtering, assessing overlaps between adjacent frames, measuring the baseline length, excluding blurry frames, computing distributions of features in each frame, and optimizing numbers of the selected key frames. Despite their benefits for image-based point cloud generations, there is a lack of the human-like judgment capability to assess the contents in each frame of big visual dataset and filter important storytelling images. In this sense, this prior work may not be fully

served as a benchmark to perform a robust photo documentation platform, since it is hard to assess the importance of domain-specific contents in given images.

To address challenges in pure discriminative approach (that uses manually labeled frames to train a classifier) and to select domain-specific key frames based on the importance of the contents in the frames, this paper aims to automatically filter big visual data from jobsites, enabling to find construction-related frames for a variety of project-related decision-makings. To evaluate the reliability of the proposed method, we have obtained the ground truth based on human judgment for each photo in our testing dataset. Then, we have explored the performance gain from the proposed approach for detecting photoworthy frames in construction big visual data. Finally, we have concluded a considerable improvement in precision with our two-step filtering approach.

METHODOLOGY

The goal of this paper is to automatically filter construction big visual data of great interest without manual human annotations. Figure 3 illustrates the overview. First, every single frame in the entire footage (obtained from Unmanned Aerial Vehicles (UAVs)) is given a score which assess their likelihood of being an intentionally taken photo (i.e., snap point). In this step, training with the SUN image database (Xiao et al. 2010) enables to filter and remove most of accidently recorded frames in the footage. Next, we leverage a discriminative classifier that is trained with big visual data from jobsites to accurately eliminate non-construction related images.

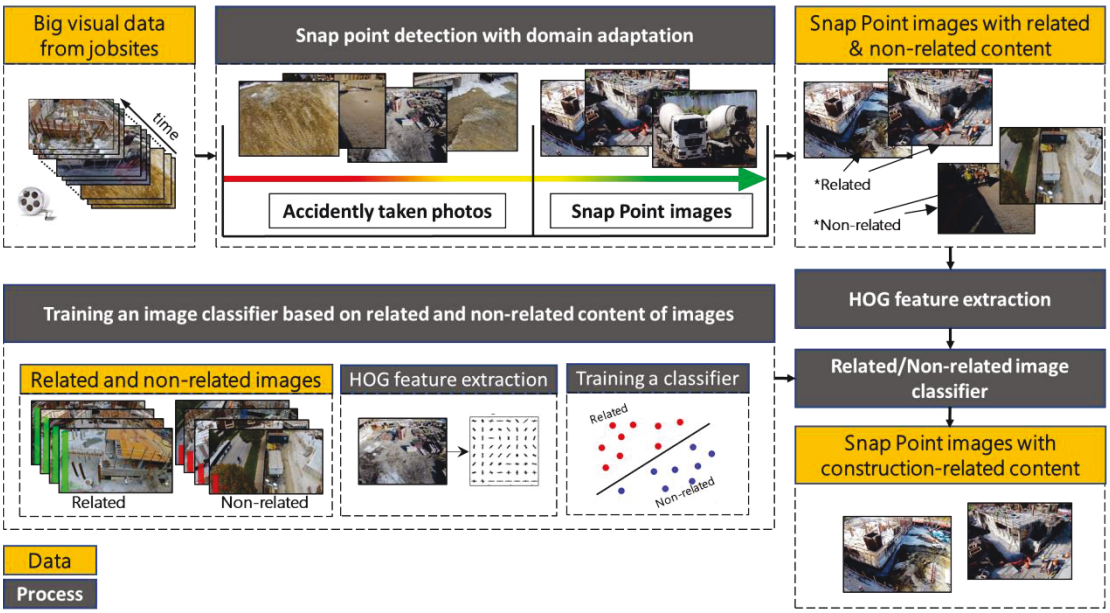

Figure 3. An overview of the data and the process proposed in this paper

Detecting Snap Points with Domain Adaptation

Detecting snap points in any footage involves numerous challenges to overcome. First, snap point images may not share the same object or article between each other. In general, their contents could be anything. This variety of objects in snap point images requires obtaining a vast variety of training dataset. This is a very labor intensive task and demands human judgment on each captured frame. Second, the height of camera from which images are taken should match between training and testing dataset. Because the videos are captured from different devices such as head-mounted GoPro cameras or flying drones, the heights where the images are taken would be different. This causes poor matches between training and testing domains. Third, the quality of images that are captured from portable devices are often low, so they may not match well with high quality photos captured intentionally on the web. Thus, obtaining training dataset for diverse scenes to match with images from different height is not trivial. In this paper, to initially remove accidentally taken photos from the captured footage, we build upon the snap point detection of (Xiong and Grauman 2014) which is trained on 130K images from the SUN database. The SUN dataset involves 130K of human taken photos for 899 categories. Even though very small portion of the SUN dataset categories are related to the construction domain, performing the snap point detection with such general dataset would eliminate large portion of unintentionally taken photos in construction big visual dataset.

Feature Extraction

First, discriminative features of images for both training and testing dataset is extracted. Then, those features extracted from different domains are studied in terms of the similarities. The extracted features could be named as Dense-SIFT, HOG, GIST, SSIM, line alignment, and motion blur. Because most intentionally taken images are aligned with horizon, we intend to extract 'line alignment feature' (Košecká and Zhang 2002) to remove those photos that are not well aligned with the horizon. In addition, the motion blur feature (Crete et al. 2007) is explored to remove blurry images, since intentionally taken photos are not typically blurry. Other sets of features are studied to see how much testing dataset agrees with the training domain. Once features are extracted, their variances are studied through the Principle Component Analysis (PCA) to derive eigenvectors that compactly capture higher variances of the features and substantially reduce the size of them. Combinations of these eigenvectors for each feature are then concatenated. The overall performance strongly depends on how many eigenvectors of each feature type are stored in the concatenation array. It is noted that more eigenvectors of features in the concatenation array does not always guarantee the higher accuracy. We assessed the precision of the method over different numbers of eigenvectors for each feature in the concatenation array and took the best values that ensured the higher precision.

Domain Adaptation and Computing Similarity

Since the recorded footages are sometimes very low resolution, thus they may not match well with the high-resolution training dataset. Therefore, there should be a domain invariant feature space to connect training and testing domains together. In other words, an infinite dimensional geodesic path connecting the training and the

testing domains should be established. This will allow two different domains to be attached via common feature subspaces in the middle, and ultimately decreases the mismatches that are initiated due to the differences in the resolution and camera characteristics of the two domains. Building upon (Gong et al. 2012), the geodesic path can be expressed as following:

$$K_{GFK}(x_i, x_j) = (z_i^{\infty}, z_j^{\infty}) = \int_0^1 (\Phi(t)^T x_i)^T (\Phi(t)^T x_j) dt \quad (1)$$

where x_i, x_j represent the visual features of training and testing dataset respectively, and $z_i^{\infty}, z_j^{\infty}$ denote the infinite dimension containing all of projections of x_i, x_j along geodesic path shown by $\Phi(t)$. The parameter t is which changes from 0 to 1 to represent the distance of projections which transfers from training to testing domains respectively. After obtaining a geodesic path for two domains to match, the similarity between each other can be obtained, which is shown with K_{GFK}. The obtained similarity is a key parameter to judge if a testing image is an intentionally taken photo or not.

Snap Point Prediction

For any given testing frames, set of images from the training domain with the highest similarity to them, in other words the highest Geodesic Flow Kernel (GFK) values could be retrieved. Let's say we have retrieved *k* number of photos in training domain that are similar to the testing frame. Now, the intentionally taken confidence for the testing frame could be calculated from the summation of similarities of *k* number of retrieved images to that testing frame, as shown below:

$$S(x^e) = \sum_{j=1}^{k} K_{GFK}(x^e, x^w{}_j) \quad (2)$$

where, x^e denotes the testing frame descriptor, and x^w be the retrieved similar images from given training dataset. The parameter k is the number of retrieved similar photos from the training dataset and $S(x^e)$ represents the confidence of snap point for the testing frame. The higher values of $S(x^e)$ indicates the higher chance of a testing frame to be on purposely recorded. With all these parameters, we now can predict snap point frames for any given footages.

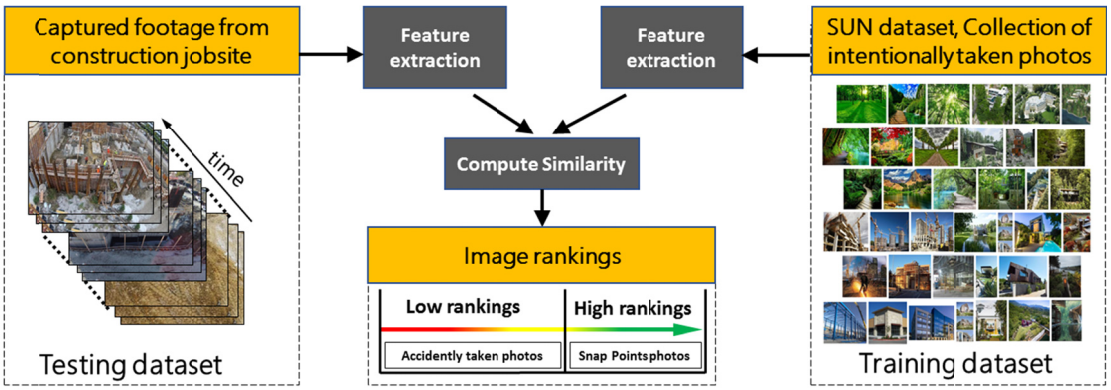

Figure 4. Snap point detection in our two-step filtering approach to screen construction big visual data of great interest without manual human annotations

Filtering Non-Construction Related Frames

Despite the initial filtering of large portions of unworthy visual data, the outcomes are most likely to still include many images without construction-related contents. This is because of the comprehensiveness of the SUN database used for training during initial filtering, not focused on the construction domain. Therefore, to automatically detect a handful of highly ranked snap point frames in construction, we need to further investigate the domain-specific content and eliminate the images that are not related to our purpose. In this regard, as a proof of concept, we leveraged 538 construction photos from UAVs with construction related content as "positive" sample, and 60 images with non-construction related content as "negative" one. Then, we extracted the HOG features and trained a Support Vector Machine (SVM) classifier based on positive and negative datasets. To put the effect of our trained classifier in place and combine the results with the obtained snap point rankings, we first explore construction related and non-construction related contents in our testing dataset through our trained classifier. After obtaining related and non-related classification for all the testing frames, we assign a very low ranking for images with non-construction related content, but for images with construction related content, we leave the snap point score as they are. This will allow us to obtain the modified rankings for our testing dataset. It is noted that images used to train the classifier were not used for testing purposes, and both groups were taken at different stages in an ongoing construction jobsite. Figure 5 represent the algorithmic scheme for ranking modification.

Input: Ranking snap point images with specific order (*SnpRank: 0<SnpRank<1*)
Related/non-related classification results with the same specific order (*ClsRank: ClsRank =0 or 1*)
Output: Modified ranking for testing dataset (*ModfdRank*)

1*ModfdRank = SnpRank*
2 for *i = 1* to *number of images*
3 if *i*th member of array *ClsRank* equals 0 (image content is non-construction related)
4 Assign *(0.01 + rand/100) to i*th member of *ModfdRank*
5 end
6 end
7 Return *ModfdRank*

Figure 5. Algorithmic scheme for integrating snap point ranking and classification results

EXPERIMENTAL RESULTS AND DISCUSSIONS

To validate the proposed method, we obtained image scores based on human judgment and compared them with the scores coming out of our method. Initially, we implemented an interface to load each image and a slider enabling to assign scores for each image based on the importance and the level of details they provide in the construction domain. The value of slider could be altered from 0 to 100. For comparison, we categorized scores within three main classes: 1- poor, 2-fair, and 3-good. In this paper, as a proof of concept, images receiving scores from 0 to 33 were considered within the "poor" category, score of images from 33 to 66 range were assigned to the "fair" category, and images getting scores more than 66 were specified in the "good" category. We asked people to consider images as higher category if large portion of them cover construction-related content. Later, we can consider images in the "poor" and "good" category to better train our classifier. To avoid biased human judgment, we have carried out the scoring tasks for multiple times, and averaged scores for each image before comparing all of them with scores obtained via our method. Figure 6 represents examples of the categorization of our testing visual data. As a proof of concept, Figure 7 shows the confusion matrix for our SVM classifier, which shows the practical reasonable level of accuracy in recognition for the purpose of filtering big visual data.

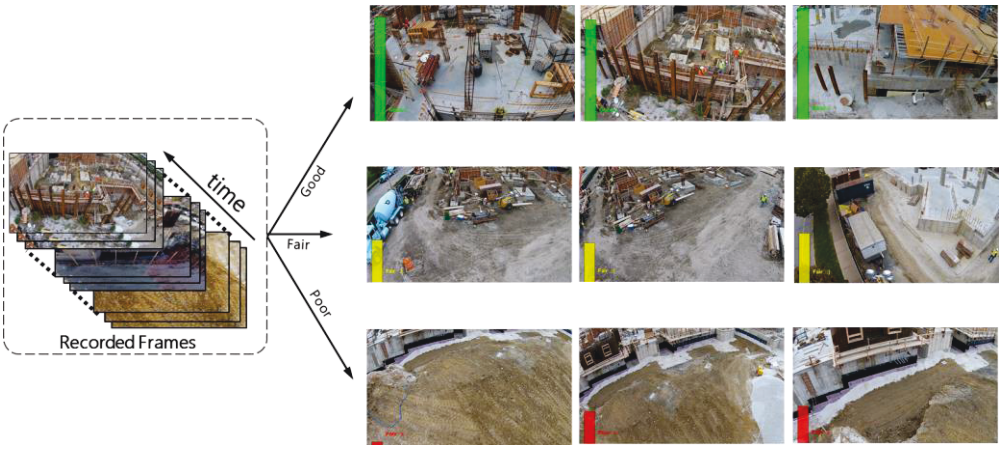

Figure 6. Examples of ground truth classification in visual dataset from jobsites

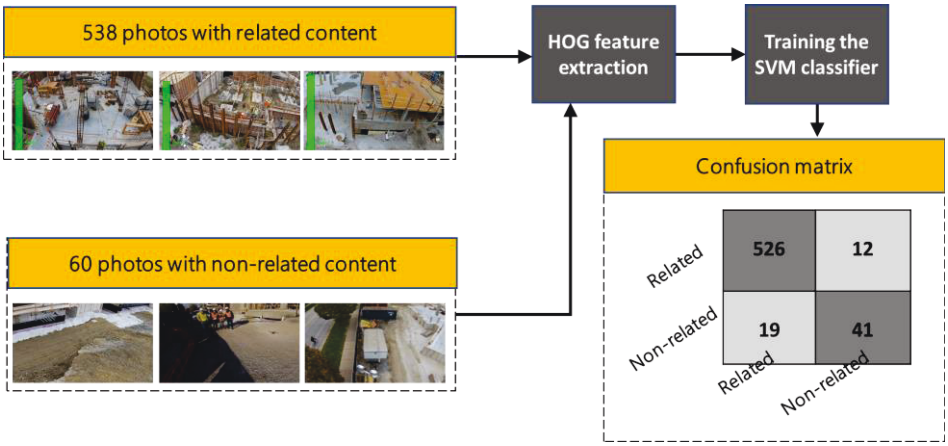

Figure 7. Training a discriminative classifier in our two-step filtering approach for construction visual dataset from jobsites

Having the ground-truth data and the scores of images from the proposed two-step filtering method in hand, we now can report the precision-recall curve to validate the reliability of our method. To report the performance gain, we benchmarked our method with the outcomes from prior snap point detection in web photo prior (Xiong and Grauman 2014). Figure 8 shows the improved performance of our method over (Xiong and Grauman 2014). The precision-recall curve indicates that with our trained classifier in place, the given big visual data from jobsites could be filtered more accurately. This is because as discussed above, we have implemented two-step classifiers to better assess the confidence of snap points that are related to the construction domain. The outcomes of the proposed method would be further refined to achieve more focused and related results.

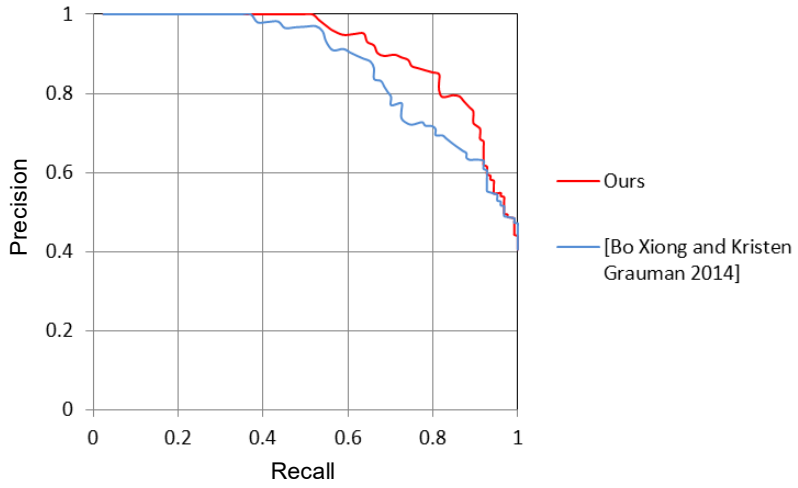

Figure 8. Performance gain from the proposed method in the precision-recall curve

CONCLUSION

To easily and quickly obtain visual data from evolving construction jobsites, buildings, landscapes and infrastructures, UAVs are widely used to constantly cruise around them. Despite their benefits, it is likely that more than half of recorded visual data are completely irrelevant or are taken from poor view point. This paper proposed a two-step method to automatically filter construction big visual data recoded from UAVs. Our method eliminates time-consuming human judgment to select construction-relevant frames from recorded big visual dataset. Experimental results show the performance gain from the proposed two-step filtering over the existing approach. Automated approach for filtering big visual data enables jobsite engineers and practitioners to obtain very dense and yet focused images they need for project management. Ultimately, by being able to focus on selective frames, practitioners can spend less time on browsing large amounts of visual data; rather spend more time on looking at performance problems potentially occurred in jobsites.

In addition to the impact on jobsite management in the practical aspect, selective photo log would be treated as a benchmark to train advanced image classifiers or to carry out image segmentation tasks. Furthermore, more fluent process to generate point clouds could be achieved through automated removal of non-construction related images that would confuse the algorithms to generate 3D point clouds. By coupling with autonomous UAV navigation platforms, the proposed method can also support to obtain better viewpoints to provide better snapshots taken from jobsites. This can provide the freedom to proceed or quit the task of recording visual data thanks to constantly evaluating the quality of recorded frames.

Future works involve exploring visual data obtained from head-mounted GoPro cameras in jobsites and validating the reliability of using ground-level big visual data from jobsites. We believe it can enhance the accuracy of the algorithm through enhancing training domain. In addition, in the near future, we will leverage more visual data to feed snap point detection platforms and increase the level of accuracy. All these are currently being explored as part of our ongoing research.

ACKNOWLEDGEMENT

This material is in part based upon work supported by the National Science Foundation (NSF) under CMMI Award #1635378. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

Crete, F., Dolmiere, T., Ladret, P., and Nicolas, M. (2007). "The blur effect: perception and estimation with a new no-reference perceptual blur metric." *Human vision and electronic imaging*, 64920.

Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). "Geodesic flow kernel for unsupervised domain adaptation." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2066–2073.

Ham, Y., and Golparvar-Fard, M. (2015). "Three-Dimensional Thermography-Based Method for Cost-Benefit Analysis of Energy Efficiency Building Envelope Retrofits." *Journal of Computing in Civil Engineering*, 29(4), B4014009.

Ham, Y., Han, K. K., Lin, J. J., and Golparvar-Fard, M. (2016). "Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works." *Visualization in Engineering*, 4(1), 1.

Ham, Y., Lee, S. J., and Chowdhury, A. G. (2017). "Imaging-to-Simulation Framework for Improving Disaster Preparedness of Construction Projects and Neighboring Communities." *Computing in Civil Engineering 2017*, American Society of Civil Engineers, Reston, VA, 230–237.

Han, K. K., and Golparvar-Fard, M. (2017). "Potential of big visual data and building information modeling for construction performance analytics: An exploratory study." *Automation in Construction*, 73, 184–198.

Han, S., and Lee, S. (2013). "A vision-based motion capture and recognition framework for behavior-based safety management." *Automation in Construction*, Elsevier, 35, 131–141.

Koch, C., and Brilakis, I. (2011). "Pothole detection in asphalt pavement images." *Advanced Engineering Informatics*, 25(3), 507–515.

Košecká, J., and Zhang, W. (2002). "Video Compass." Springer, Berlin, Heidelberg, 476–490.

Park, M.-W., Elsafty, N., and Zhu, Z. (2015). "Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction Workers." *Journal of Construction Engineering and Management*, 141(9), 4015024.

Rashidi, A., Dai, F., Brilakis, I., and Vela, P. (2013). "Optimized selection of key frames for monocular videogrammetric surveying of civil infrastructure." *Advanced Engineering Informatics*, Elsevier, 27(2), 270–282.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). "Sun database: Large-scale scene recognition from abbey to zoo." *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, 3485–3492.

Xiong, B., and Grauman, K. (2014). "Detecting Snap Points in Egocentric Video with a Web Photo Prior." *Computer Vision -- ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, D.

Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds., Springer International Publishing, Cham, 282–298.

Yang, J., Park, M.-W., Vela, P. A., and Golparvar-Fard, M. (2015). "Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future." *Advanced Engineering Informatics*, 29(2), 211–224.