Automatic Recognition of Construction Worker Activities Using Dense Trajectories

Jun Yang[a], Zhongke Shi[b] and Ziyan Wu[a]

[a]School of Mechanics, Civil Engineering and Architecture, Northwestern Polytechnical University, China
[b]School of Automation, Northwestern Polytechnical University, China
E-mail: junyang@nwpu.edu.cn, zkeshi@nwpu.edu.cn, zywu@nwpu.edu.cn

ABSTRACT

Wide spread monitoring cameras on construction sites provide large amount of information for construction management. The emerging of computer vision and machine learning technologies enables automatic recognition of construction activities from videos. As the executors of construction, the activities of construction workers have strong impact on productivity and progress. Compared to machine work, manual work is more subjective and may differ largely in operation flow and productivity from one worker to another. Hence only a handful of work study on vision based activity recognition of construction workers. Lacking of publicly available datasets is one of the main reasons that currently hinder advancement. The paper studies manual work of construction workers comprehensively, selects 11 common types of activities and establishes a new real world video dataset with 1176 instances. For activity recognition, a cutting-edge video description method, dense trajectories, has been applied. Support vector machines are integrated with a bag-of-features pipeline for activity learning and classification. Performance on multiple types of descriptors (Histograms of Oriented Gradients - HOG, Histograms of Optical Flow - HOF, Motion Boundary Histogram - MBH) and their combination has been evaluated. Experimental results show that the proposed system has achieved a state-of-art performance on the new dataset.

Keywords –

Worker; Activity recognition; Dense trajectories; Machine learning

1 Introduction

Construction is a dynamic conversion process from inputs – workers, equipment and materials to output – the built infrastructure. Monitoring site operations, especially workers and equipment's activities, is significant for productivity evaluation, progress estimation and quality control. Most modern practices lean on foremen collecting information from construction site by means of onsite observations, survey or interview. Post processing is required to analyse collected data. The entire procedure is not only labor intensive, cost sensitive and error prone, but taking away from the more important tasks of identifying opportunities for performance improvements, reviewing alternatives, and conducting what-if analysis [1]. The need for "automated operation analysis" becomes clearer and is in fact also highlighted by the U.S. National Academy of Engineering.

Nowadays, video cameras become more and more prevalent in construction sites. The recorded videos provide large amount of information for construction monitoring and management. The emerging of computer vision and machine learning technologies enables analyzing construction activities from videos automatically. In the past decade, many researchers have dedicated to this field.

One main stream method is to detect, track workers and equipment and analyze their activities by poses or trajectories combining prior knowledge. Zou and Kim [2] track the excavator by appearance and judge the idle time through its movement status. Azar et al. [3] detect and track the excavator and dump truck simultaneously to analyze the dirt loading cycle. Gong and Caldas [4] detect a concrete bucket in video streams through machine learning and estimate its travel cycles based on the prior knowledge of construction site layout. Yang et al. [5] perform similar work of monitoring concrete placement activity by tracking the crane jib through 3D pose estimation. Peddi et al. [6] track workers tying rebar through blob matching, extract skeletons for pose estimation and classify their working status into 'effective, ineffective and contributory' by poses.

Problems with the above mentioned method are apparent. In cluttered construction scenarios, or under some situations with a hand-held camera, it is difficult to detect and track construction entities stably through a long duration. Errors from previous stages (detection

and tracking) will accumulate and affect the final activity analysis adversely. To solve this problem, a recent trend is to adopt a machine learning based Bag-of-Feature (BoF) framework for activity recognition without detect or track any entities explicitly. One pioneer work from Gong et al. [7] recognizes worker and equipment activities by learning various motion patterns from spatio-temporal features of videos. Similarly, Golparvar-Fard et al. [8] also adopts BoF pipeline to recognize actions of earthmoving equipment. However, instead of using generative models, they use discriminative support vector machines. Hand-held tools are closely related with the task being executed. Hence Kim and Caldas [9] propose a novel method to improve action recognition by combining skeleton based body gesture recognition with the detection of construction objects.

Even with remarkable achievements on machine learning based construction activity recognition, some open issues still exist. First, workers' activities have not been studied extensively. The work of Gong et al [7] only covers five categories of worker actions in formwork activity with 300 video segments, which is relatively small compared to some state-of-art human action dataset [10,11] in computer vision field. Different from equipment activity, workers' activities are more subjective and differ largely in operation flow and productivity according to different individuals. Generally, the types of activities conducted by workers vary from trades to trades. Yet some activities may share similar visual features. Intra-class difference and inter-class similarity introduce big difficulties to vision-based activity recognition. Lacking of publicly available datasets is another main reason that currently hinders advancement. Second, from the algorithm aspect, previous works adopt a spatio-temporal feature description, which is extracted from a joint 3D domain of 2D space and 1D time. However, the space domain and the time domain in videos have different characteristics naturally. More intuitive way is to treat them differently. For example, tracking interest points through video is such an option. Features from joint 3D space depict video information at a given location in space and time while tracked trajectories of given interest points capture motion information better [12]. This is significant for activity recognition.

Based on the above discussion, in this paper we adopt a cutting-edge video representation method – dense trajectories for workers activity recognition. Dense trajectories are obtained by dense sampling and tracking in the dense optical field. It has been tested in several human action datasets and achieved state-of-art performance [12]. Multiple feature descriptors and their combination have been evaluated. To test our proposed system, a new large scale data set of worker activities covering a wide range of trades, is established by capturing videos from the real construction site.

The rest of the paper is organized as follows. Section 2 describes the methodology in detail by illustrating dense trajectories and related feature descriptors, as well as the classification method. Section 3 presents the new data set. Section 4 gives out experimental results. Section 5 concludes the paper.

2 Methodology

The overall workflow of the proposed system is shown in Figure 1. As it can been seen, the system is built upon a Bag-of-Feature scheme. First, video clips are represented by visual feature descriptors. Specifically, dense trajectories are generated by dense sampling and tracking on a dense optical flow field. Then various descriptors are computed along the dense trajectories. Second, codebooks per each description channel are constructed using k-means clustering and descriptors are quantized by assigning to the nearest vocabulary word. Third, a non-linear SVM (Support Vector Machines) is adopted for classification. More details are illustrated as follows.

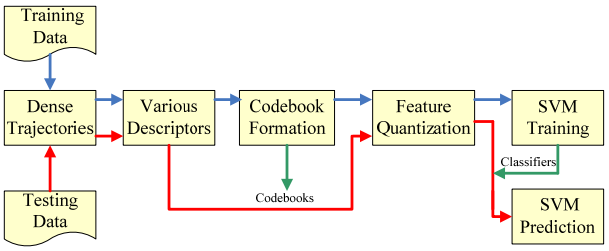

Figure 1. The system overview

2.1 Dense trajectories generation

Dense trajectories are obtained by densely sampling and tracking feature points on multiple spatial scales separately. For each spatial scale, feature points are densely and equally sampled by step of W pixels. The spatial scale increases by a factor of $1/\sqrt{2}$. Since the key point of activity recognition is to capture the motion patterns, feature points in homogeneous image areas are removed by thresholding on the eigenvalues of the auto-correlation matrix for each frame [13].

Then for each frame I_t, dense optical flow field $\omega_t = (u_t, v_t)$ are extracted, in which u_t and v_t represent the horizontal and vertical component separately. For a given point $P_t = (x_t, y_t)$, its tracked position in the next frame is smoothed by a median filter on ω_t:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{(x_t, y_t)} \quad (1)$$

where M is the median filtering kernel.

Trajectories are formed by concatenating points from subsequent frames: $(P_t, P_{t+1}, P_{t+2,} \cdots)$. To restrain tracking drifting, the length of the trajectories is limited to L frames. Static trajectories and trajectories with sudden large displacements are filtered out in post processing steps. To ensure dense coverage of trajectories, a new point is added if no tracked point is found in a $W \times W$ neighbourhood. The empirical values used in the experiments are $L = 15, W = 5$.

2.2 Descriptors

Three types of descriptors, namely HoG (Histograms of Oriented Gradients) [14], HoF (Histogram of Optical Flow) [15] and MBH (Motion Boundary Histograms) [16] are used in our system. HoG is designed to encode static appearance information. HoF is good at capturing the local motion information. MBH represents the gradient of the optical flow. Hence it will remove information from constant camera motion and keep changes in the flow field.

Descriptors are computed within a space-time volume aligned with a trajectory to encode the motion information. The volume size is $N \times N \times L$, where N is in pixels and L is the frame length of the trajectory. Considering the structure information, the volume is further divided into a smaller size $n_\sigma \times n_\sigma \times n_\tau$. In the experiments, the default values are $N = 32, n_\sigma = 2, n_\tau = 3$. With orientations quantized into 8 bins for HoG and 9 bins for HoF (one extra zero bin), the final descriptor size ends up 96 for HoG and 108 for HoF. Specifically, the MBH descriptor is split into horizontal component MBHx and vertical components MBHy, whose sizes are both 96.

2.3 Learning activity patterns using support vector machine

To learn and predict worker activities, a non-linear support vector machine with RBF-χ^2 kernel is adopted. Various descriptors can be combined using the following approach [15]:

$$K(H_i, H_j) = \exp(-\sum_{c \in C} \frac{1}{A_c} D_c(H_i, H_j)) \quad (2)$$

where $D_c(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{v} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}$ (3)

v is the vocabulary size. A_c is the mean value of the distances between all training samples for channel c . For multi-class classification, SVM classifier is trained using a one-against-rest strategy for each activity type. During testing, classifier with the highest confidential value will dominate the predicted activity type.

3 Dataset

To test the proposed system, a new comprehensive dataset of worker activities is established. Some unique characteristics must be considered while building up the dataset. First, there are many types of trades, such as carpenter, ironworker, etc. They all have their own workflows, related tools and working scenarios. Second, for a single trade, worker activity differs largely from one individual to another due to their personal skills and habits. Third, even for the same individual, his/her repetitive activity may have slight differences from cycle to cycle due to ever changing onsite situations.

We recorded videos from four real construction sites with a hand-held camera for two months. During capturing, different weathers, illuminations, points of views, scales and occlusions are covered. The gender of construction workers, as well as their skill levels, is also considered. Unlike some human action dataset, which is performed by actors under certain instructions [17], our videos are totally nature from the real construction work without guiding the workers.

Collected videos are segmented into short clips and annotated manually. The final data set contains 1176 video clips, covering a wide range of trades, namely carpenter, ironworker, mason, plasterer, steel fixer and welder. Totally 11 types of activities are abstracted from different trades. All video clips are in resolution of 320 by 240 with the frame rate of 30 f/s. More detailed information of the proposed data set is summarized in Table 1. As can be seen in Table 2, compared to our two closely related references [7, 8], the proposed data set has the most video clips and the biggest number of activity categories. Snapshots of video frames from different activities are shown in Figure 2.

Table 1 Summary information of the dataset

Activity Category	Num of Clips	Num of Workers
LayBrick	190	18
Transporting	54	25
CutPlate	53	7
Drilling	58	5
TieRebar	157	10
Nailing	132	17
Plastering	168	12
Shoveling	185	22
Bolting	79	18
Welding	50	4
Sawing	50	8

Table 2 Comparison with other data sets

References	Data set	Num of Clips	Num of Class
Gong et al [7]	Back-hole	150	3
	Worker	300	5
Golparvar-Fard et al [8]	Exca-vator	627	4
	Truck	233	3
Ours	Worker	1176	11

4 Experimental results

In the experiment, dense trajectories with three types of descriptors are computed using Wang's implementation [12] with parameters set as aforementioned. Then, data set is divided into two equal parts randomly. The first half is for training and the rest is for testing. A subset of 100,000 randomly selected training features is clustered using k-means to generate codebooks for each description channel. The number of visual words per descriptor is fixed to 4000, which is shown to perform well on a wide range of datasets [12]. Then all features are quantized by assigning to the nearest vocabulary word using Euclidean distance. SVM classifiers with RBF χ^2 kernel are trained using quadratic programming. The performance of our proposed system has been tested on every individual descriptor (HoG, HoF, MBH), and also the combination of all descriptors using the multi-channel approach. Notice the performance of MBH is the combination of MBHx and MBHy using multi-channel approach.

To evaluate the performance results, the confusion matrix and average per class accuracy have been adopted. Defining a confusion matrix C, $C(i, j)$ is a percent count of observations known to be type i but predicted as type j. Each column of the matrix represents the instances in a predicted activity class, while each row represents the instances in an actual activity class. The average per class accuracy is defined as follows:

$$ACC = \sum_{i=j=1}^{N} C(i, j) / N \qquad (4)$$

where N is the number of activity categories.

The confusion matrices for each type of descriptor and their combination are given in Figure 3. Generally speaking, the top three categories with high accuracy are 'LayBrick', 'TieRebar' and 'Transporting'. And the bottom three with poor performance are 'Drilling', 'Bolting' and 'Sawing'. One reason is that the former three categories contain obvious movement and relatively standard workflow. For example, a common 'LayBrick' flow is: 'pick up a brick', 'get mortar with a trowel', 'smear the mortar', 'place the brick', 'knock the brick with the trowel to fasten'. The latter three categories do not have either large movement or consistent workflow. For example, when drilling, the worker's body nearly hold still, only with the bit spinning rapidly, which is really difficult to capture in a 30 f/s video.

The average per class accuracy is 39%, 49%, 57% and 45% for descriptor HoG, HoF, MBH and combination of them separately. MBH gives out the best performance, which is mainly due to its ability to suppress camera motion. However, the combination of all descriptors does not have the best performance. This is unexpected since in [12] the combination of multiple descriptors achieved the best performance in all tested nine datasets. One possible explanation is that the activities in our construction data set are not as consistent as those in other data sets. Some activities are coarse-grained, such as 'Transporting', 'CutPlate' and 'Shoveling', who mainly involve body movement. Some are fine-grained, such as 'TieRebar' and 'Bolting', where hand movement is dominant. A few categories are somewhere in between, such as 'LayBrick' and 'Nailing', where both coarse body movement and fine hand movement are required. A naive combination of all descriptors may affect the discrimination ability adversely.

5 Conclusions

In this paper, we established a new video data set of construction worker with 1176 clips, divided into 11 categories of common activities. A cutting-edge video representation method - dense trajectories was adopted for activity recognition based on a bag-of-feature framework. Three types of descriptors, namely HoG, HoF and MBH, were computed along dense trajectories. The multi-class SVM with non-linear RBF χ^2 kernel was applied for training and classification. Performance was evaluated on all three descriptors and their combination. Experimental results showed that the MBH descriptor achieved the best accuracy of 57%. Future work may seek to combining semantic information to improve activity recognition and simultaneously segmenting and recognizing worker activities from continuous video streams.


Lay-Brick
Transp orting
Cut-Plate
Drill ing
Tie-Rebar
Nailing
Plaster-ing
Shovel ing
Bolting
Weld-ing

Figure 2. Snapshots of all activity categories in our data set

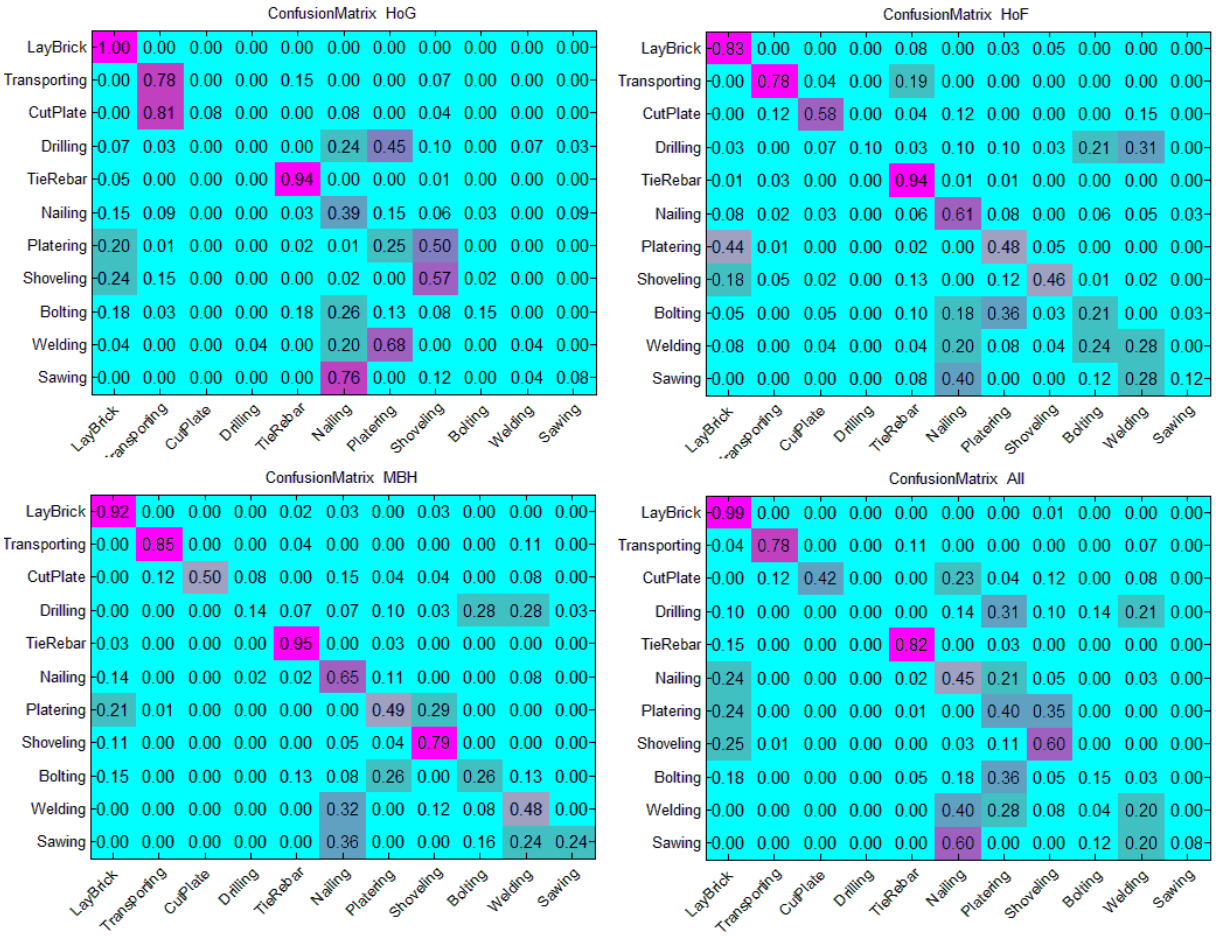

Figure 3. Confusion matrix of activity recognition using different descriptors

Acknowledgement

The work is supported by National Natural Science Foundation of China No.51208425. The authors would like to thank Mingyu Lai for data collection and Heng Wang for discussion on the algorithm.

References

[1] Yang J., Park M., Vela P. A. and Golparvar-Fard M. Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future. Advanced engineering informatics, 29(2): 211-224, 2015.

[2] Zou J. and Kim H. Using hue, saturation, and value color space for hydraulic excavator idle time analysis, Journal of Computing in Civil Engineering, 21: 238-246, 2007.

[3] Rezazadeh Azar E., Dickinson S. And McCabe B. Server-customer interaction tracker: Computer vision-based system to estimate dirt-loading cycles. Journal of Construction Engineering and Management 139: 785-794, 2013.

[4] Gong J. and Caldas C. H. An intelligent video computing method for automated productivity analysis of cyclic construction operations. In Proceedings of ASCE International Workshop on

Computing in Civil Engineering, pages. 64–73, Austin, USA, 2009.

[5] Yang J., Vela P. A., Teizer J. and Shi Z. Vision-based tower crane tracking for understanding construction activity. Journal of Computing in Civil Engineering, 28: 03–112, 2014.

[6] Kim S., Huan L., Peddi A. And Bai Y. Development of human pose analyzing algorithms for the determination of construction productivity in real-time. In Proceedings of Construction Research Congress, pages 11–20, Seattle, USA, 2009.

[7] Gong J., Caldas C. H. and Gordon C. Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and Bayesian network models. Advanced Engineering Informatics 25: 771–782, 2011.

[8] Golparvar-Fard M, Heydarian A. and Niebles J. C. Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. Advanced Engineering Informatics, 27: 652-663. 2013.

[9] Kim J. Y. and Caldas C. H. Vision based action recognition in the internal construction site using interactions between worker actions and construction objects. In Proceedings of the International Symposium on Automation and Robotics in Construction and Mining, Canada, 2013.

[10] Reddy K. K. and Shah M. Recognizing 50 human action categories of web videos. Machine Vision and Applications, 24(5): 971-981, 2013.

[11] Huehne H., Jhuang H., Garrote E., Poggio T. and Serre T. HMDB: A large video databased for human motion recognition. In Proceedings of IEEE International Conference on Computer Vision, pages 2556-2563, Barcelona, Spain, 2011.

[12] Wang H., Klaser A., Schimid C. and Liu C. Action recognition by dense trajectories. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 3169-3176, Colorado Springs, USA, 2011.

[13] Shi J. and Tomasi C. Good features to track. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, Seattle, USA, 1994.

[14] Dalal N. and Triggs B. Histograms of oriented gradients for human detection. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol 1, pages 886-893, San Diego, USA, 2005.

[15] Laptev I., Marszelek M., Schmid C. And Rozenfeld B. Learning realistic human actions from movies. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 1-8, Anchorage, USA, 2008.

[16] Dalal N., Trigg B. and Schmid C. Human detection using oriented histograms of flow and appearance. In Proceedings of European Conference on Computer Vision, Vol 2, pages 428-441, Graz, Austria, 2006.

[17] Tran D. and Sorokin A. Human activity recognition with metric learning. In Proceeding of European Conference on Computer Vision, pages. 548-561, Marseille, France, 2008.