Automation in Construction 94 (2018) 360–370

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks

Xiaochun Luo[a], Heng Li[a], Dongping Cao[b,*], Yantao Yu[a], Xincong Yang[a], Ting Huang[a]

[a] Dept. of Building and Real Estate, Hong Kong Polytechnic Univ., Hung Hom, Kowloon, Hong Kong
[b] Dept. of Construction Management and Real Estate, School of Economics and Management, Tongji Univ., 1239 Siping Rd., Shanghai 200092, China

ARTICLE INFO

Keywords:
Labor productivity evaluation
Work sampling
Two-stream convolutional networks
Activity recognition

ABSTRACT

Capturing the working states of workers on foot allows managers to precisely quantify and benchmark labor productivity, which in turn enables them to evaluate productivity losses and identify causes. Work sampling is a widely used method for this task, while suffers from low efficiency as only one worker is selected for each observation. Attentional selection asymmetry can also bias its uniform object selection assumption. Existing vision-based methods are primarily oriented towards recognizing single, separated activities involving few workers or equipment. In this paper, we introduce an activity recognition method, which receives surveillance videos as input and produces diverse and continuous activity labels of individual workers in the field of view. Convolutional networks are used to recognize activities, which are encoded in spatial and temporal streams. A new fusion strategy is developed to combine the recognition results of the two streams. The experimental results show that our activity recognition method has achieved an average accuracy of 80.5%, which is comparable with the state-of-the-art of activity recognition in the computer vision community, given the severe camera motion and low resolution of site surveillance videos and the marginal inter-class difference and significant intra-class variation of workers' activities. We also demonstrate that our method can underpin the implementation of efficient and objective work sampling. The training and test datasets of the study are publicly available.

1. Introduction

Labor costs are of prime importance because, for example, they make up from 25 to 35% of construction project costs in Hong Kong [1]. Therefore, continuously capturing their working states allows us to precisely quantify and benchmark labor productivity, which in turn enables us to evaluate productivity losses and identify causes, as implemented by the conventional work sampling method [2]. However, only one worker is selected in an observation with the manual method. Furthermore, it assumes uniform object selection that each worker has the same chance of being observed as any other worker. In practice, however, the expected uniform selection could be biased due to attentional selection asymmetry [3], which guides attention in a new context and leads to a preference for stimuli that have been selected more often in the previous context. There is a need for efficient and objective work sampling.

In the past decade, vision-based methods have been the subject of much systematic investigation due to the easy accessibility and informativeness of image or video data and the non-invasiveness to activities of interest. These efforts motivate us to think that vision-based techniques have great potential to effectively and objectively capture workers' working states and time utilization information.

In the construction surveillance practice, a camera is usually mounted at a high and remote location to have a wide view, e.g., under the operator cab of a tower crane or somewhere at surrounding facilities. It reduces the costs of deployment and maintenance of surveillance, and meanwhile, highlights that video-based work sampling should be based on this kind of site surveillance. This observation excludes those algorithms based on detailed visual features, such as skeleton-based posture and activity recognition [4–9] because site surveillance videos cannot provide those detailed features. Furthermore, the methods for recognizing single, separated activities that involve few workers [10,11] or equipment [10,12] is of limited use. There is a need for a method that can recognize diverse and continuous workers' activities in site surveillance videos.

In this paper, we introduce a new vision-based activity recognition method, which takes surveillance videos as input and produces diverse and continuous activity labels of individual workers in the field of view. There are four steps in the method: 1) tracking workers with an object tracking technology, 2) extracting spatial and temporal streams with an

[*] Corresponding author.
E-mail address: dongping.cao@tongji.edu.cn (D. Cao).

https://doi.org/10.1016/j.autcon.2018.07.011
Received 8 December 2017; Received in revised form 14 June 2018; Accepted 12 July 2018
Available online 25 July 2018

optical flow estimation technology, 3) classifying the activities encoded in the two streams independently, and 4) fusing the results of the two streams to produce final activity estimates. The experimental results show that our method has achieved an average recognition accuracy of around 80.5%, which is comparable with the best results produced in the computer vision community. We also demonstrate that our method can underpin the implementation of efficient and objective work sampling for labor productivity evaluation.

2. Related work

In this section, we review the work published in the computer vision domain about human activity recognition and that in the construction domain regarding vision-based activity recognition and productivity analysis. We note that there are sensor-based solutions for worker activity recognition such as [13,14]. However, the intrusive nature of this kind of solutions can hinder their practical application. Also, the communication requirements for data collection when involving multiple workers dispersed on sites could become an unbridgeable gap. Additionally, there are solutions using depth cameras for indoor activity recognition such as [15]. As this study primarily focuses on using site surveillance videos to facilitate work sampling, we narrow our review scope to the most relevant studies, namely activity recognition based on average cameras. We refer to the reader to more comprehensive reviews on vision-based techniques for performance monitoring in [16], construction safety and health monitoring in [17], and temporary resource sensing and tracking in [18].

2.1. Human activity recognition in computer vision

Human activity recognition is an active research topic in the computer vision community with many critical applications, including human-computer interfaces, content-based video indexing, video surveillance, and robotics [19–23]. Human activities can be categorized into four types according to their complexity: gestures, actions, interactions, and group activities [20]. Gestures are the atomic components describing the meaningful motions of a person, like “stretching an arm” and “raising a leg.” Actions are single-person activities composed of multiple gestures, such as “walking,” “lifting,” and “pushing.” Interactions are activities involving two or more persons and objects. “A carpenter is cutting formwork templates” is a good example of such kind of interactions. Finally, group activities are those performed by multiple persons and objects. “A group of rebar workers is fixing rebar” and “a group of concrete workers is placing concrete of columns” are typical examples. In the construction context, human-object interactions are the focus of this paper.

Two types of approaches for activity recognition are investigated depending on activity complexity: single-layer approaches and multiple-lay approaches [20]. Due to their nature, the primary objective of the single-layered approaches has been to analyze relatively simple (and short) sequential movements of humans, such as “walking,” “lifting,” and “pushing.” They are labeled as “single-layer” because they represent and recognize human activities directly based on sequences of images. On the other hand, hierarchical approaches represent high-level human activities by describing them based on other simpler activities. Therefore, they are suitable for the analysis of complex activities.

Recent years have seen two representative tracks of the single-layer approach to action recognition regarding the methods for feature representation. The first track employs hand-crafted local features such as dense trajectories (DTs) [24] and improved dense trajectories (IDTs) [25]. IDTs extend DTs through correcting camera motion, which is estimated by matching feature points between frames using SURF [26] descriptors and DTs, resulting in the warped optical flow. The approach based on IDTs has proved to perform best on a variety of datasets like UCF-101 and HMDB-51 [25].

The second track attempts to understand human actions by incorporating spatial and temporal information extracted from RGB and optical flow images [27]. This track is pioneered by Simonyan and Zisserman [28]. They proposed two-stream Convolutional Networks (ConvNets), which consists of a spatial stream ConvNet with eight layers, taking single RGB frames as input and providing action class scores as output, and a temporal ConvNet with the same structure, while receiving multiple-frame optical flow as input and outputting action class scores. The two stream results are finally fused to provide the final class score. The two-stream structure shows comparable performance with DTs on UCF-101 and HMDB-51 [28].

Recently, the two-stream approach is extended to obtain better performance. For example, Wang et al. [29] introduce temporal segment networks (TSNs) to tackle the problem of action recognition in untrimmed videos. Zhang et al. [30] replace optical flows with motion vectors to accelerate the whole learning process, which initially suffers from the most computationally expensive step of calculating optical flows. Ma et al. [27] examined the differences and the distinguishing factors between various methods using recurrent neural networks or ConvNets on temporally-constructed feature vectors (Temporal-ConvNet). They found that both recurrent neural networks and Temporal-ConvNets on spatiotemporal feature matrices can exploit spatiotemporal dynamics to improve the overall performance. Most importantly, they found that each of these methods requires proper care to achieve state-of-the-art performance.

2.2. Construction activity recognition and productivity analysis

Publications that concentrate on atomic action recognition in construction more frequently adopt space-time methods. For example, Golparvar-Fard et al. [12] presented a method using spatiotemporal features and support vector machine (SVM) classifiers to estimate the action of earthmoving equipment (excavators and trucks). Yang et al. [11] introduced a study using DTs [24, 31] to recognize workers' actions and reported the best average accuracy of 59%, which represents the state of the art of recognizing workers' activities.

Luo et al. [32] introduced a method to recognize diverse construction activities in site images with deep convolutional networks. The method recognizes construction-related objects with the Faster R-CNN [33], constructs relevance networks based on these objects, and derives the activities by matching activity patterns, which are defined by relevance relationship among those detected objects. This method focuses on separated, still images and therefore misses the temporal information encoded in consecutive frames. Recently, deep learning is also used to detect unsafe behaviors such as workers not wearing hardhats [34], not wearing safety belts [35], and misusing ladders [36]. However, these methods focus on object detection or activity recognition involving a single worker in the field of view.

Earlier literature on construction activity recognition particularly focuses on the productivity analysis of construction equipment [37–41]. Pioneering this task, Gong and Caldas [38] introduced a vision-based method to analyze the productivity of concrete pouring of a tower crane and concrete buckets. Their method breaks down construction operations into a variety of working task elements and describes how these elements unfold in planned locations and sequences. Similarly, Bügler et al. [37] proposed a rule-based interaction activity detection method for analyzing earthwork productivity of excavators and dump trucks. In their method, the activity state (i.e., static, moving, absent, or filling) is checked when an excavator and a dump truck are in proximity. Besides, Rezazadeh Azar et al. [39] proposed another rule-based method for analyzing dirt loading cycles of excavators and dump trucks. The logic reasoning checks equipment orientations for filling, and then the SVM classifier detects the earthwork states according to the distances between the base point of the excavator and four corners of the dump truck.

Yang et al. [40] presented a stochastic activity recognition method to infer two-state tower crane activities (i.e., concrete pouring and non-concrete material movement) using crane jib trajectories and site layout information. In the method, the jib angle trajectory is tracked with a 2D to 3D rigid pose tracking algorithm, and a probabilistic graph model was introduced to process the tracking results as well as recognize crane activities.

In summary, a considerable amount of literature on visual object detection [42–47] and construction activity recognition [10–12,39,41,48–50] has been published. These studies have contributed to taking a significant step forward in introducing computer vision technologies to the time-consuming tasks [16]. However, they are oriented towards recognizing single, separated activities involving few workers [10,11] or equipment [10,12]. Furthermore, the reported low recognition accuracies impede their practical applications.

There is a need for a method that can recognize diverse and continuous construction activities in site surveillance videos.

3. Research scope and challenges

This study focuses on using site surveillance videos to recognize workers' diverse and continuous activities. Dynamic and complicated construction situations impose a major challenge to vision-based activity recognition methods. Recognizing workers' activities in site surveillance videos is more challenging than recognizing those activities recorded in the publicly available datasets like UCF-101 and HMDB-51 for the following reasons. Firstly, severe camera motion could happen if a camera is mounted under the operator cab of a tower crane. Secondly, the low resolution only allows coarse visual features to extracted. Thirdly, in a video stream of labor activities, the inter-class difference could be marginal, while intra-class variation could be significant. Continuity of labor activities makes the difference between two consecutive but different activities marginal, such as the difference between “moving” and “transporting rebar”, and that between “fixing formwork” and “placing formwork.” Meanwhile, workers in an identical activity can also take various postures. For example, they can fix rebar by standing, stooping, squatting, and stretching. Fourthly, occlusion happens frequently due to ubiquitous construction components, large equipment, and temporal materials and facilities. The reported best average accuracy 59% [11] of workers' activity recognition partially proves that.

4. Method development

Motivated by the success of two-stream ConvNets in activity recognition, we lend the latest method in this line (i.e., TSNs) [29] to recognizing workers' activities in site surveillance videos. The overall flowchart of our method is summarized in Fig. 1. There are four steps. The first step is to track workers to spatiotemporally segment workers of interest. The second step is to extract spatial and temporal streams. The third step is to recognize their activities with two-stream ConvNets. The last step is to fuse the recognition results from spatial and temporal streams to obtain activity estimates.

4.1. Tracking workers

Tracking individual workers is of utmost importance for constructing the vision-based solution. In the computer vision community, there are two categories of technologies for this task: single object tracking (SOT) and multiple object tracking (MOT). The SOT approach tackles one object at a time by differentiating between the object of interest and the background, while the MOT approach follows multiple objects synchronously. MDNet [51] is a SOT algorithm based on the representations from a discriminatively trained ConvNet. It pre-trains a ConvNet to obtain a generic target representation with the shared layers, initializes the last three volatile layers with the target feature in the first frame, and updates the three layers to get the robustness and accuracy.

We use MDNet to track workers and create temporally and spatially cropped videos. Manually specifying bounding boxes of workers, in the beginning, is necessary to start the tracking process. Multiple bounding boxes can be specified in the beginning to speed up the process, and a for-loop is implemented to track workers one by one automatically. However, another problem arises with a long video clip when workers enter the field of view in the frames after we create bounding boxes. A workaround is to trim long surveillance clips into relatively short ones for tracking workers since a session of conventional work sampling is to observe an operation for a limited time. In our study, we trim long surveillance videos into one-minute clips to reduce the number of workers not tracked because they present in a frame after we specify bounding boxes.

Discretizing continuous activities into finite atomic activities is a reasonable resolution to recognizing activities in long video streams because exhaustively identifying all activities with which a worker could appear is practically impossible. The problem then reduces to what time granularity should be used. In this study, an activity unit is designed to last 3 s for two reasons. An activity unit should carry enough spatial and temporal variance across consecutive frames to make it recognizable. In other words, it should not be trimmed too short to be informatively similar with any frame of it. On the contrary, an activity unit should be temporally short to contain only one activity to maintain its atomicity. We use 3 s as the length of activity unit, which is also used in the datasets of the computer vison community, e.g., ActivityNet [52] and Google Research AVA [53]. Given this simplification, we introduce an extended bounding box, which is the minimum box covering all bounding boxes in 3 s, as illustrated in Fig. 2. As a result, an activity unit is a spatiotemporal segment created with an extended bounding box.

4.2. Extract spatial and temporal streams

Having activity units, we use FlowNet 2.0 [54] to extract spatial and temporal streams of individual workers. FlowNet [55] demonstrates that ConvNets can also be successfully used to estimate optical flow by solving it as a supervised learning task. Flownet 2.0 extends FlowNet by significantly improving estimation quality; it decreases the estimation error by more than 50% and performs on a par with state-of-the-art methods while running at interactive frame rates.

In this study, we use spatial streams and temporal streams. Spatial streams are virtually RGB images encoding static appearance at specific time points, which lack the contextual information about previous and next frames. Complementarily, temporal streams are essentially warped optical flow fields generated with FlowNet 2.0 [54], which capture the motion information of objects of interest with camera motion corrected. Specifically, temporal streams are represented with two normalized gray images, one for horizontal (x-direction) motion and another for vertical (y-direction) motion in the image coordinate. Fig. 3 shows an exemplary RGB image stack and its corresponding optical flow fields in both x and y directions.

4.3. Recognizing activities based on two streams

We use TSNs [29] to recognize workers' activities by taking the two streams as input. At the time of writing, TSNs represent the state of the art of activity recognition on two popular activity datasets with accuracies of 69.4% on HMDB-51 [56] and 94.2% on UCF-101 [57] respectively. TSNs choose the Inception with Batch Normalization (BN-Inception) [58] as the primary building block to encode the spatial and temporal streams due to its good balance between accuracy and efficiency [29]. In the seminal introduction of two-stream ConvNets for activity recognition in [28], a video is encoded with a single RGB frame, representing appearance information, and a single stack of optical flow fields, representing motion information. This design ignores the

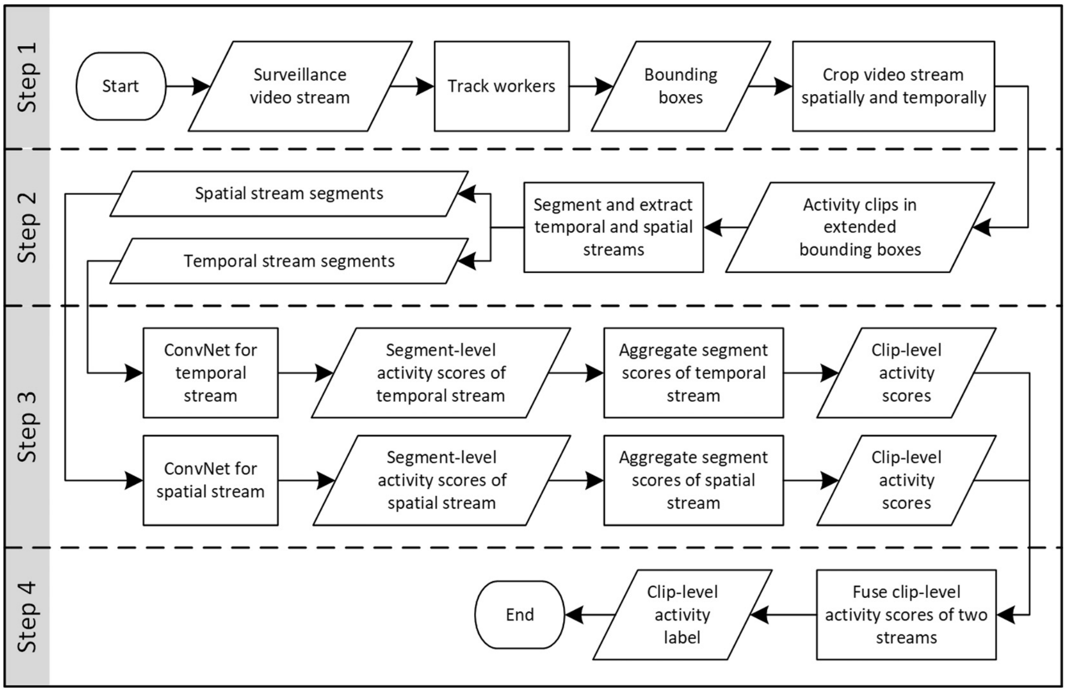

Step 1
Start
Surveillance video stream
Track workers
Bounding boxes
Crop video stream spatially and temporally
Step 2
Spatial stream segments
Temporal stream segments
Segment and extract temporal and spatial streams
Activity clips in extended bounding boxes
Step 3
ConvNet for temporal stream
Segment-level activity scores of temporal stream
Aggregate segment scores of temporal stream
Clip-level activity scores
ConvNet for spatial stream
Segment-level activity scores of spatial stream
Aggregate segment scores of spatial stream
Clip-level activity scores
Step 4
End
Clip-level activity label
Fuse clip-level activity scores of two streams

Fig. 1. Overall flowchart.

information captured by long-range temporal structures. To address this problem, TSNs operate on a sequence of short snippets sparsely sampled from an entire video [29]. In TSNs, each snippet in this sequence will produce its preliminary prediction of the activity classes, and then a consensus among the snippets will be derived as the video-level prediction. Therefore, we conjecture that TSNs are suitable to tackle the complex workers' construction activities, which consist of diverse atomic motions during a relatively long period.

4.4. Fusing results of two streams

Fusing the estimation scores of the spatial and temporal networks is critical to taking advantage of their proprietary prediction capacity. The fusion strategy can be implemented with video-level predication scores of all activity classes. Take a video V as an example to explain the fusion strategy. The video level activity scores from the spatial stream network are a vector $S = (s_1, s_2, \ldots, s_N)$, where N is the number of activity classes, which is 16 in our study. Similarly, the activity scores from the temporal stream network are expressed with $T = (t_1, t_2, \ldots, t_N)$. In [29], the weighted average is adopted and evaluated; the weighted result of two streams is then $w_s S + w_t T$, where w_s and w_t are two scalars weighting the results of the spatial and temporal stream networks respectively. The fused video-level activity prediction is the activity class with the maximum score. The confidence of each activity prediction is derived with the widely used softmax function; for example, a video has the probability of $e^{z_i}/\Sigma_{j=1}{}^{N} e^{z_j}$ to be activity class i, where z_j is equal to $w_s s_j + w_t t_j$.

However, due to marginal inter-class difference and significant intra-class variation, it is challenging for human experts to distinguish those activities in short clips that are spatially cropped to contain the worker of interest. Furthermore, construction sites are generally congested, and workers' working areas and paths are frequently overlapped. The cropped videos contain, with a high probability, other workers irrelevant to the activity of interest. In combination with low resolution and severe camera motion, they make construction activity recognition in site surveillance videos relatively challenging. The segments of a short clip could be classified as different activities and allocated with various confidence values. Therefore, excluding those segments with low confidence levels is conjectured to be able to improve video-level recognition accuracy.

Based on this observation, we propose a new method to fuse the two streams. It is different from the weighted average method used in [29], where S and T are the mean of M segments' results, namely $S = \frac{1}{M}\sum_{k=1}^{M} S_k$ and $T = \frac{1}{M}\sum_{k=1}^{M} T_k$. In our method, S and T are respectively the sum of the top K_s spatial stream segments and that of top K_t temporal stream segments. Eqs. (1) and (2) formalize their calculation,

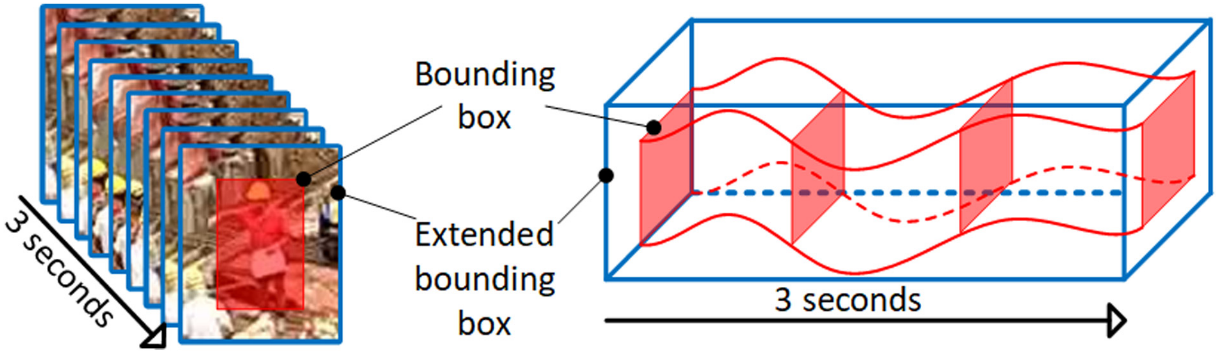


Fig. 2. Illustration of activity units.

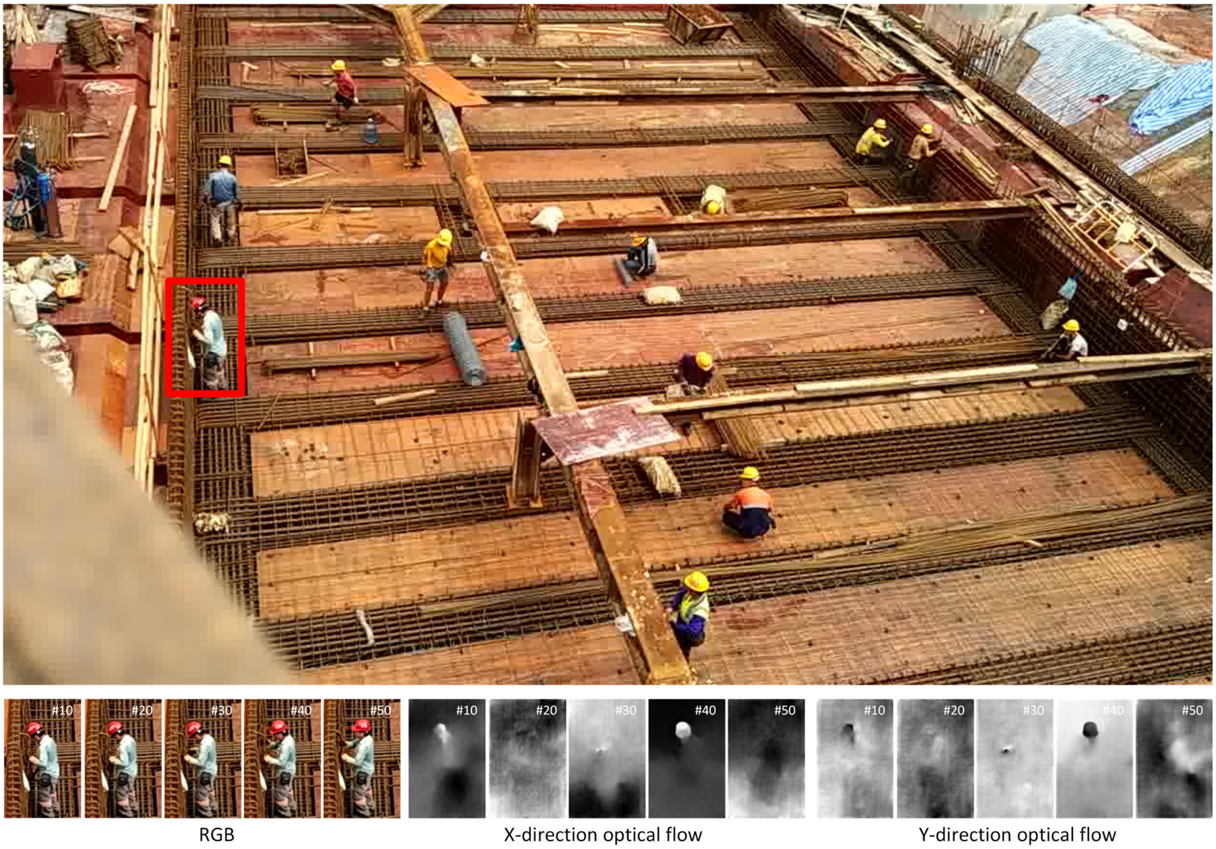


Fig. 3. Snapshot of surveillance videos with RGB stack and optical flow.

where $S_k^{'}$ and $T_k^{'}$ are sorted score vectors of spatial and temporal segments in descending order according to their highest score element. Note that there is no explicit weighting in our method; K_s and K_t implicitly allocate the weights to two streams. Then, the fused score $S + T$ is fed to the softmax function to calculate the confidence degree $p_{i\in(1,2,...,N)}$ of each activity label, into which a video could fall. Eq. (3) summarizes the fusion operation, where $P = (p_1, p_2, ..., p_N)$. Finally, it is deduced that a video contains the activity $\mathscr{A}$ with the highest confidence degree.

$$S = \sum_{k=1}^{K_s} S_k' \tag{1}$$

$$T = \sum_{k=1}^{K_t} T_k' \tag{2}$$

$$P = softmax(S + T) \tag{3}$$

$$\mathscr{A} = argmax_{i\in(1,2,...,N)}(P) \tag{4}$$

5. Experiment and results

In this section, we report the experimental results concerning construction activity recognition. Specifically, we introduce the training and test sets, the training process, and the activity recognition results. We compare the performance of our new fusion strategy with the strategy adopted in [29]. Besides, we demonstrate that our method can be used to implement efficient and objective work sampling through a simple case.

5.1. Training and test sets

The surveillance videos were filmed with a resolution of 1280 × 720 pixels at 30 frames per second from the construction site of an office building project in Hong Kong using a pan–tilt–zoom camera. It was mounted on a scaffolding system with mild motion at the height of around 15 m to the working floor and from a distance close to the boundary of the site. The camera was configured according to the typical middle-field filming settings, in which facial features may not be visible due to the poor resolution, while the lines that delimit the head and shoulders of an individual are still informative cues to identify people in images. Fig. 3 also shows an image sampled from one of the videos and illustrates the site surveillance. The video filming spanned over 41 days between April 2016 and May 2016 and produced 76 video clips ranging from 1 min to 15 min.

Recognizing construction activities of workers on foot is closely relevant to work sampling, which is a widely used productivity evaluation approach [2]. It classifies the workers' activities as one of three modes: productive, semi-productive, and non-productive. An observer is employed to record all observations by entering checkmarks under the appropriate mode of the activities observed. Finally, all checkmarks are added up to calculate the activity percentage under each mode, which represents the productivity. Our long-term goal is to implement automatic work sampling, while we focus on rebar workers and formwork workers by covering their continuous activities at this initial stage and develop an activity taxonomy consisting of 16 classes of activities, referencing the three modes, as shown in Table 1. The semi-productive and non-productive modes considered in our study differentiate ours from the previous studies that only focus on the activities under the productive mode. Our taxonomy is envisaged to cover all these modes and therefore has a high potential to be used in practice.

As stated in Section Method Development, we use MDNet [51] to track workers one by one to produce bounding box sequences, which allow us to spatiotemporally segment videos into activity unit clips. These clips are 3-s long and primarily contain the worker of interest

Table 1
Activity taxonomy.

Trade	No.	Activity	Mode[a]	Specification and examples
Common	1	Measuring	1	Measuring formwork, purlins, and rebar with a tape.
	2	Moving	2	Moving with hands empty.
	3	Preparing	2	Preparing auxiliary materials or setting up equipment for subsequent tasks, e.g., rebar workers disentangling steel binding wire and formwork workers relocating handheld saws.
	4	Resting	3	Standing still, standing and drinking water, or standing and wiping perspiration.
Formwork workers	5	Fixing formwork	1	Fixing formwork and purlins by hammering nails.
	6	Machining formwork	1	Cutting formwork and purlins with a saw platform or a handheld saw.
	7	Placing formwork	1	Arranging formwork and purlins before fixing or settling them down after transporting.
	8	Taking formwork	2	Picking up formwork and purlins before placing or transporting.
	9	Transporting formwork	2	Transporting formwork and purlins from a location to another in a manual way such as carrying and shouldering.
Rebar workers	10	Connecting rebar	1	Connecting rebar with sleeve joints.
	11	Fixing rebar	1	Binding rebar with steel binding wire.
	12	Machining rebar	1	Cutting and bending rebar with a bar cutter/bender.
	13	Placing rebar	1	Arranging rebar before fixing or settling it down after transporting.
	14	Taking rebar	2	Picking up rebar before placing or transporting.
	15	Transporting rebar	2	Transporting rebar from a location to another in a manual way such as carrying and shouldering.
	16	Welding rebar	1	Connecting rebar via arc welding.

[a] Mode 1, productive; mode 2, semi-productive; mode 3, non-productive.

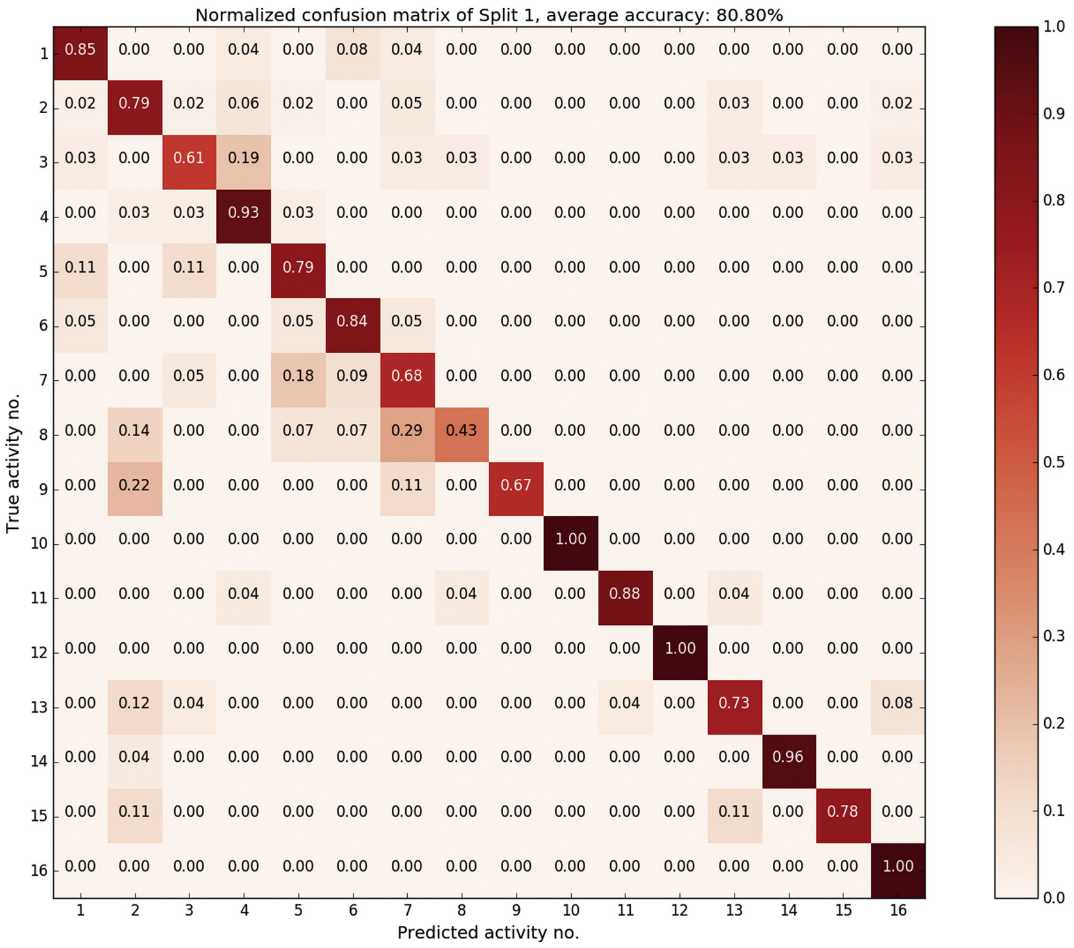


Fig. 4. Confusion matrix of recognition results based on Split 1.

with an extended bounding box. Note that other irrelevant workers frequently appear in atomic activity clips due to site congestion and that we reserve these 'disturbances' as they are part of the actual situations. Having these atomic activity clips, we manually label them with activity classes by categorizing them into different class folds, as summarized in Table 1.

In the study, we adopt the three-fold cross-validation strategy, which is also used in [29]. In each split, the test set consists of one-third of clips and the training set is composed of the rest. To have relatively uniform representation in a split, we firstly sort all clips according to date, time, and worker numbers. Then we select the ith clip out of every three into the test set, where $i \in \{1, 2, 3\}$ is the number of the split.

5.2. Settings for network training

We follow the recommendations by [29] about the number of

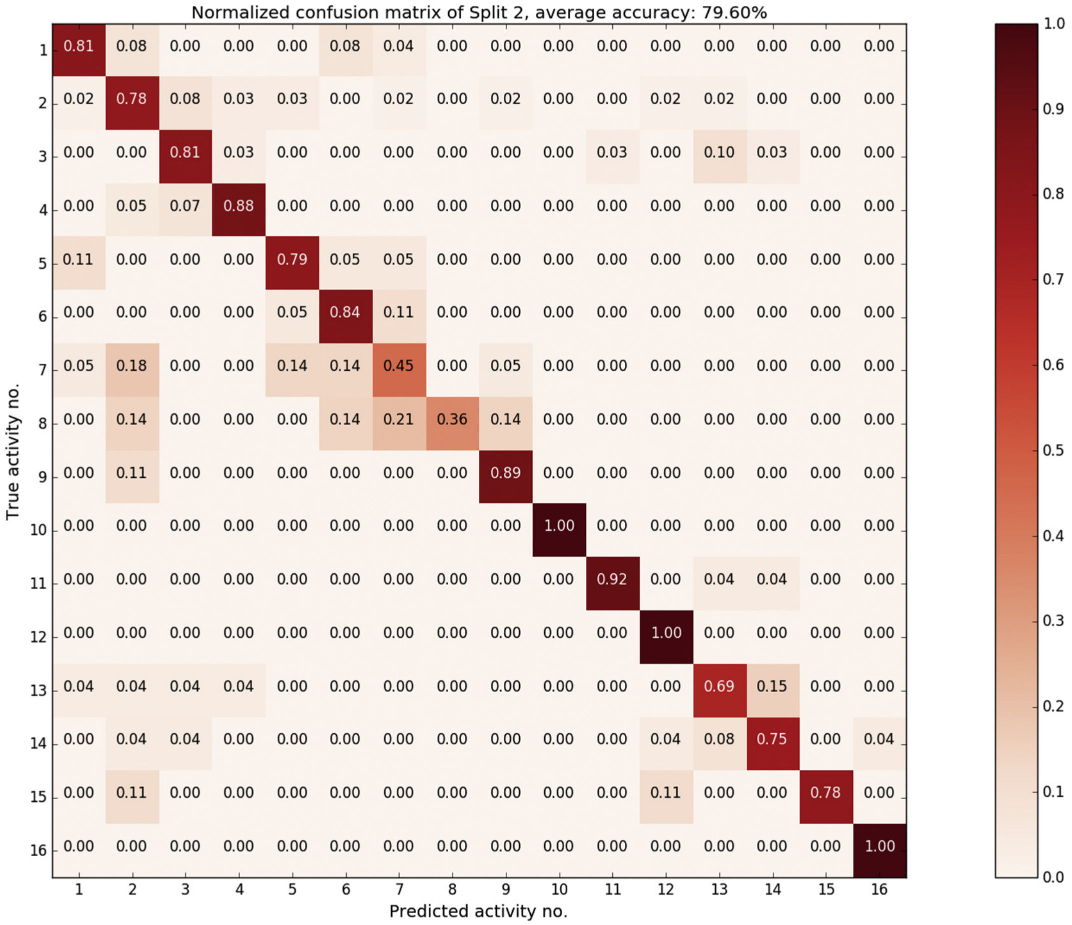


Fig. 5. Confusion matrix of recognition results based on Split 2.

segments. It is set as three at the training stage, which means that an activity unit video is divided into three segments. For spatial networks, an RGB frame is randomly sampled in a segment, while for temporal networks, five consecutive x-direction optical flow images and five corresponding y-direction optical flow images are randomly selected and sequentially stacked as the input. All RGB frames and optical flow images are reshaped to 224 × 224 pixels at the data layer of the networks.

Our network weights are initialized with pre-trained models from ImageNet [59]. We use the multiple step stochastic gradient descent policy to learn the parameters of both spatial networks and temporal networks. For spatial networks, the learning rate is initialized as 0.001, which reduces to its 10% every 5000 iterations. The whole training procedure stops after a total of 15,000 iterations. For temporal networks, the initial learning rate is 0.005, and we iteratively decrease it to its 10% every 5000 iterations. Similarly, the maximum iteration number is set to 15,000. We use a Nvidia Geforece GTX 1080 GPU to train our networks, which were implemented with a modified version of Caffe [29]. The average training time on the three training splits is around 141 min for temporal networks and 104 min for spatial networks.

5.3. Settings for network testing

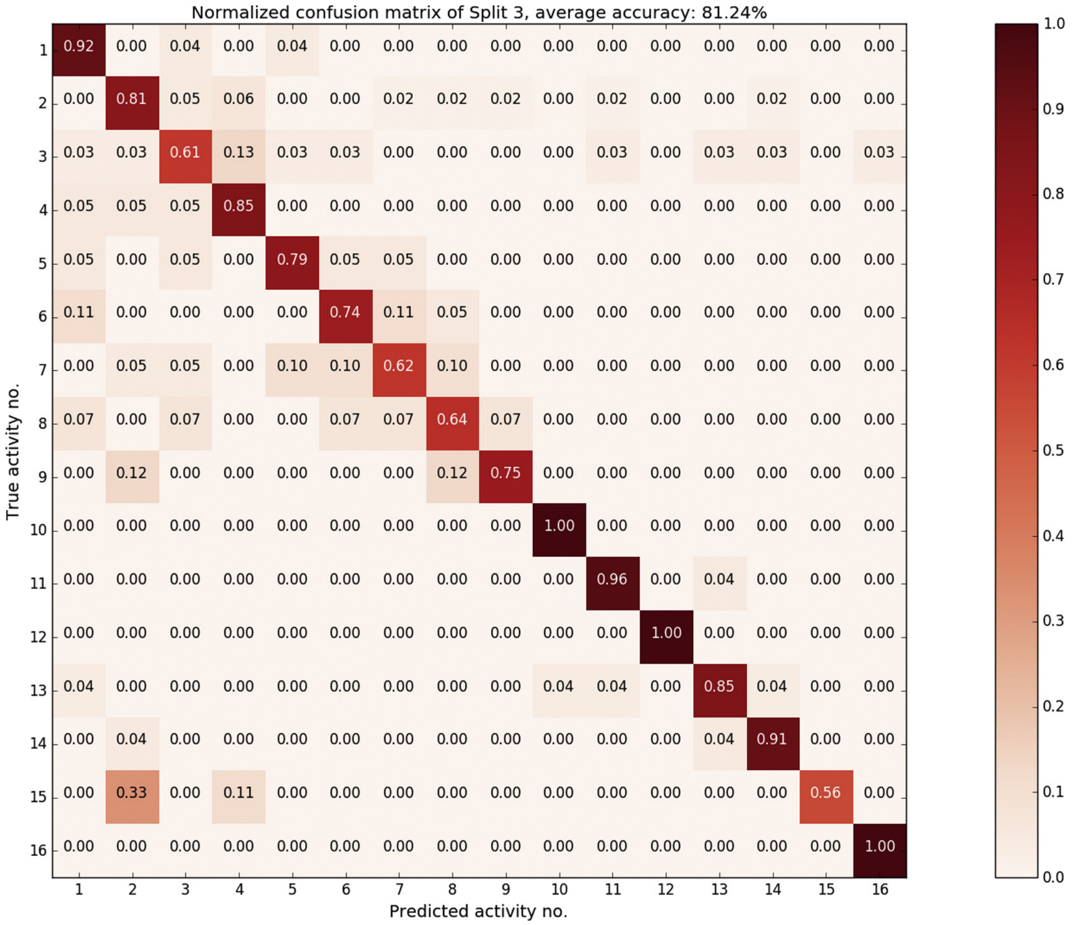
We use the same settings regarding segmentation and augmentation at the test stage in [29]. For testing the spatial and temporal networks, the number of segments was set as 25. In addition, all RGB frames and optical flow images are reshaped to 340 × 256 pixels before feeding to the ConvNet. Sliding a 224 × 224 pixels window on these 340 × 256 pixels images generates ten augmented samples. Therefore, there are ten recognition results in either a spatial segment (i.e., an RGB frame with ten augmented frames) or a temporal segment (i.e., a stack of optical flow images with ten augmented stacks). The average aggregation strategy was adopted to take advantage of the augmentation; for a segment, ten classification score vectors based on an augmented input are averaged class-wisely to generate the final score vector of that segment.

5.4. Performance metrics

We adopt *accuracy* and *confusion matrixes* as the metrics to evaluate the performance of our method. Accuracy simply measures how often the TSNs make the correct prediction and is a widely accepted measure for activity recognition in the computer vision community. Specifically, per-class accuracy is the ratio between the number of correct predictions concerning a specific activity class and the total number of predictions (i.e., the number of the atomic activity units of that class). In addition, average accuracy is the average of all per-class accuracies (i.e., the ratio between the sum of each per-class accuracy and the number of activity classes). We do not use precision and recall as the metrics of activity recognition performance. As in our study, activity recognition is conducted based on existing test sets, which consists of individual atomic activity units. For an activity unit, it can either be correctly labeled or wrongly labeled. In other words, there are only true positives and false positives, but no false negatives or true negatives.

5.5. Results of individual networks

We evaluate the action recognition results of spatial networks and temporal networks individually on the test sets in the three splits.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.92	0.00	0.04	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.81	0.05	0.06	0.00	0.00	0.02	0.02	0.02	0.00	0.02	0.00	0.00	0.02	0.00	0.00
3	0.03	0.03	0.61	0.13	0.03	0.03	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.03	0.00	0.03
4	0.05	0.05	0.05	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.05	0.00	0.05	0.00	0.79	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.11	0.00	0.00	0.00	0.00	0.74	0.11	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.05	0.05	0.00	0.10	0.10	0.62	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.07	0.00	0.07	0.00	0.00	0.07	0.07	0.64	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.12	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.04	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
13	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.85	0.04	0.00	0.00
14	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.91	0.00	0.00
15	0.00	0.33	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 6. Confusion matrix of recognition results based on Split 3.

Table 2
Statistics of our dataset.

Activity	Clip count	Avg. width	Avg. height	Width std.	Height std.
Measuring	77	96.2	145.7	24.1	36.2
Moving	189	158.6	171.8	52.1	37.8
Preparing	93	100.6	150.9	18.9	23.5
Resting	119	91.4	144.4	16.0	24.4
Fixing formwork	57	95.7	153.5	14.7	20.5
Machine formwork	57	104.0	131.9	15.8	16.2
Placing formwork	65	108.6	154.1	21.0	34.6
Taking formwork	42	112.5	150.5	23.6	25.3
Transporting formwork	26	164.2	167.2	52.9	26.1
Connecting rebar	23	90.0	121.6	11.8	12.6
Fixing rebar	71	89.4	151.2	15.2	37.4
Machining rebar	33	83.3	120.4	7.5	5.5
Placing rebar	78	93.8	141.9	13.7	27.0
Taking rebar	71	91.4	140.5	19.9	21.1
Transporting rebar	27	138.1	157.7	46.0	39.2
Welding rebar	27	97.1	146.5	10.5	11.0

Table 3 summarizes their performance in terms of per-class accuracies and average accuracies. We can see that the spatial ConvNet has achieved a distinct advantage over the temporal ConvNet. Our results are significantly different from those reported in [29], where the temporal ConvNet performs slightly better than the spatial ConvNet. The low resolution and camera motion in our data sets probably contributes to the major difference.

5.6. Results of fused networks

Based on the above results, we conjecture that we need a new fusion strategy, as explained in Section 4.4. In [29], the weighted average strategy at the stream level and the mean aggregation strategy at the segment level were adopted and generated the best results on HMDB-51 and UCF-101. Here, we evaluate and compare our strategy and that proposed in [29] on our test sets. In our experiment, the weights w_s and w_t use four groups of values, including 1 and 1, 1 and 0.75, 1 and 0.5, and 1 and 0.25. We adopt an exhaustive strategy to identify the values of K_s and K_t, which evaluated 625 (i.e., 25 × 25) combinations according to the resulted average accuracies out of the three splits. We find that K_s and K_t set to 24 and 9 respectively produce the best results. The experimental results show that our fusion strategy produces the best average accuracies.

5.7. Comparison with the state-of-the-art

We preliminarily compare our method with the method based on DTs reported in [11]. Note that the authors of [11] compare their method with an early method in [10], which is based on bag-of-visual-words, and prove the performance advantage of their method over the latter by the average accuracy of 59% versus 35%. As illustrated in Figs. 4–6, and summarized in Table 3, the average accuracy of our method is 80.5% (i.e., the average of 80.8%, 79.6%, and 81.2% out of the three splits), supporting us to conclude that our method has outperformed the state-of-the-art in the construction community. Although the comparison based on two different test sets could undermine the credibility of the conclusion, we believe that our datasets are more challenging due to the video quality described in Table 2.

Table 3
Comparison of per-class accuracies and average accuracies with different fusion strategies.

Split	Fusion[a]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Avg.
1	RGB	0.808	0.619	0.645	0.925	0.737	0.789	0.682	0.429	0.667	0.875	0.875	1.000	0.692	0.958	0.778	1.000	0.780
	OF	0.654	0.810	0.484	0.850	0.684	0.526	0.455	0.429	0.556	0.875	0.667	0.727	0.577	0.667	0.556	1.000	0.657
	1 + 1	0.808	0.762	0.613	0.950	0.737	0.842	0.636	0.500	0.556	0.875	0.875	1.000	0.692	0.958	0.778	1.000	0.786
	1 + 0.75	0.808	0.683	0.613	0.925	0.737	0.842	0.682	0.429	0.667	0.875	0.875	1.000	0.692	0.958	0.778	1.000	0.785
	1 + 0.5	0.808	0.635	0.613	0.925	0.737	0.842	0.682	0.429	0.667	0.875	0.875	1.000	0.692	0.958	0.778	1.000	0.782
	1 + 0.25	0.808	0.619	0.645	0.925	0.737	0.789	0.682	0.429	0.667	0.875	0.875	1.000	0.692	0.958	0.778	1.000	0.780
	Ours	0.846	0.794	0.613	0.925	0.789	0.842	0.682	0.429	0.667	1.000	0.875	1.000	0.731	0.958	0.778	1.000	**0.808**
2	RGB	0.808	0.730	0.806	0.800	0.737	0.842	0.545	0.286	0.667	1.000	0.917	1.000	0.654	0.708	0.778	1.000	0.767
	OF	0.615	0.762	0.323	0.775	0.579	0.579	0.455	0.286	0.333	0.500	0.833	0.909	0.692	0.667	0.222	1.000	0.596
	1 + 1	0.808	0.794	0.774	0.850	0.789	0.842	0.409	0.286	0.667	1.000	0.917	1.000	0.692	0.750	0.667	1.000	0.765
	1 + 0.75	0.808	0.746	0.806	0.825	0.737	0.842	0.409	0.286	0.667	1.000	0.917	1.000	0.692	0.750	0.778	1.000	0.766
	1 + 0.5	0.808	0.746	0.806	0.800	0.737	0.842	0.455	0.286	0.667	1.000	0.917	1.000	0.654	0.750	0.778	1.000	0.765
	1 + 0.25	0.808	0.730	0.806	0.800	0.737	0.842	0.500	0.286	0.667	1.000	0.917	1.000	0.654	0.708	0.778	1.000	0.765
	Ours	0.808	0.778	0.806	0.875	0.789	0.842	0.455	0.357	0.889	1.000	0.917	1.000	0.692	0.750	0.778	1.000	**0.796**
3	RGB	0.920	0.730	0.613	0.821	0.789	0.737	0.667	0.571	0.625	1.000	0.957	1.000	0.808	0.826	0.556	1.000	0.789
	OF	0.720	0.794	0.419	0.718	0.474	0.632	0.381	0.357	0.375	0.714	0.913	0.818	0.500	0.522	0.556	1.000	0.618
	1 + 1	0.920	0.810	0.548	0.846	0.789	0.684	0.571	0.429	0.625	1.000	0.957	1.000	0.846	0.870	0.556	1.000	0.778
	1 + 0.75	0.920	0.730	0.613	0.821	0.789	0.684	0.571	0.571	0.625	1.000	0.957	1.000	0.808	0.913	0.556	1.000	0.785
	1 + 0.5	0.920	0.730	0.613	0.821	0.789	0.737	0.571	0.571	0.625	1.000	0.957	1.000	0.808	0.870	0.556	1.000	0.785
	1 + 0.25	0.920	0.730	0.613	0.821	0.789	0.737	0.619	0.571	0.625	1.000	0.957	1.000	0.808	0.870	0.556	1.000	0.788
	Ours	0.920	0.810	0.613	0.846	0.789	0.737	0.619	0.643	0.750	1.000	0.957	1.000	0.846	0.913	0.556	1.000	**0.812**

[a] RGB, spatial stream; OF, optical flow representing temporal stream; $i + j$, w_s valued as i and w_t valued as j.

Table 4
Average computation time.

Computation item[a]	Average computation time
Object tracking with MDNet	2.53 frames per second
Stream creation with FlowNet 2.0	11.30 s per activity unit
Activity recognition with the spatial ConvNet	0.87 s per activity unit
Activity recognition with the temporal ConvNet	0.79 s per activity unit

[a] The computation time is recorded using a PC with the configuration profile of a NVidia Geforce GTX 1080 GPU, an Intel Xeon CPU E5-2630, and a memory of 32 GB.

Table 5
Work sampling sheet.

Observation no.	Productive	Semi-productive	Non-productive
1	5	5	1
2	6	5	0
3	5	5	1
4	5	5	1
5	4	6	1
6	4	6	2
7	4	7	1
Total	33	39	7
Percentage (%)	41.77	49.36	8.86

5.8. Computation time

Table 4 records the computation time of worker tracking with MDNet, stream creation with FlowNet 2.0, and activity recognition with TSNs, using the same computer with a configuration profile of an NVidia Geforce GTX 1080 GPU, an Intel Xeon CPU E5-2630, and a memory of 32 GB. We view the computation time as a metric only for the reader's reference, rather than a performance metric for comparison because the computer configuration profile has a significant effect on it.

5.9. An exemplary case

To demonstrate how our activity recognition method can be embedded into the current practice, we implemented a prototype system that censuses workers and their working states continuously and automatically taking site surveillance videos as input. The work sampling system consists of three modules: 1) object tracker, which is implemented based on MDNet [60], recognizes and tracks workers individually; 2) activity recognizer, which implements the activity recognition method that we propose in this study; and 3) work analyzer, which censuses workers and their working states based on the output of the activity recognizer.

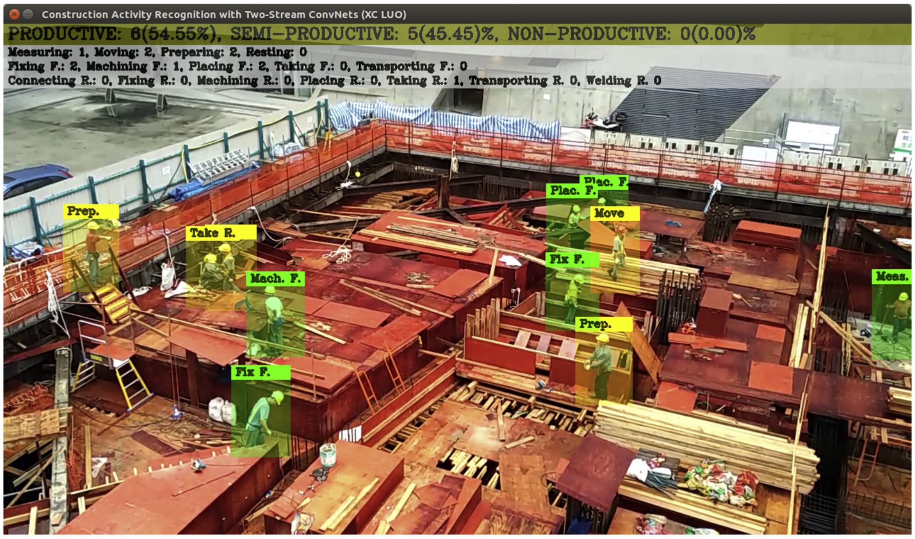
Fig. 7 illustrates the interface of the prototype system, in which the snapshot shows a 21-second site surveillance clip of formwork construction. Following the 3-second segment rule, we collected seven observation results, as summarized in Table 5. Unlike the traditional manual method, with which only one worker is randomly selected and observed in an observation, our method can observe all workers that can be detected and tracked by the object tracker at a time and therefore works more efficiently. Also, the traditional manual sampling method requires that each worker has the same chance of being observed as any other worker. While the expected uniform object selection could be biased due to attentional selection asymmetry [3], which guides attention in a new context and leads to a preference for stimuli that have been selected more often in the previous context. Our method produces objective and comprehensive observations.

6. Discussion

This section discusses our contribution to the body of knowledge as well as the research limitations.

6.1. Knowledge contribution

We contribute to the body of knowledge from three perspectives. Firstly, we integrate the recent advances in computer vision and deep learning to recognize diverse and continuous activities in middle-field site surveillance videos, which impose major challenges including the severe camera motion and low resolution and the diversity and consecutiveness of activities. To implement the combination, we propose a method to create activity units through spatiotemporal segmentation based on tracking results in 3 s (i.e., bounding box sequences), which subsequently allow us to create spatial streams and temporal streams using FlowNet 2.0. The experimental results demonstrate that combining the latest techniques of object tracking, optical flow estimation, and activity recognition to implement an efficient and objective work sampling method is technically feasible. This combination achieves the comparable performance of activity recognition with the original



```
Construction Activity Recognition with Two-Stream ConvNets (XC LUO)
PRODUCTIVE: 6(54.55%), SEMI−PRODUCTIVE: 5(45.45)%, NON−PRODUCTIVE: 0(0.00)%
Measuring: 1, Moving: 2, Preparing: 2, Resting: 0
Fixing F.: 2, Machining F.: 1, Placing F.: 2, Taking F.: 0, Transporting F.: 0
Connecting R.: 0, Fixing R.: 0, Machining R.: 0, Placing R.: 0, Taking R.: 1, Transporting R. 0, Welding R. 0
```

Fig. 7. Interface of the work sampling system.

method, given these major challenges.

Secondly, we propose a new strategy for fusing the results of the two streams. It acknowledges the characteristics of streams generated from site surveillance systems and results in better activity recognition performance. There are two strategies for combining the results from different streams in the original work, i.e., the weighted average strategy at the stream level and the mean aggregation strategy at the segment level. We take the linear combination of the results in the first K_S vectors of the spatial stream and K_T vectors of the temporal stream, to unleash the complementary discrimination capability. Our datasets are different from the frequently used activity datasets in the computer vision community, e.g., UFC-101 and HMDB-51, where the performance of spatial network and that of temporal network are close [29]. While in our case, due to the camera motion and poor resolution, the performance of spatial network significantly outperforms that of temporal network. It excludes the weighted average strategy and the mean aggregation strategy, as they equally treat the results of the segments in each stream. The linear combination of top K_S vectors of the spatial stream and K_T vectors of the spatial stream well acknowledge the performance difference among the segments in each stream. The test results show that our strategy outperforms the original strategies.

Thirdly, we develop an activity dataset based on real-life project surveillance videos and make them publicly available. Objectively evaluating and comparing technical performance is critical to research on proposing original techniques. This becomes prominent to development of vision-based techniques. In the computer vision community, there are various publicly available datasets, e.g., PASCAL [61] and MS COCO [62] for object detection and classification, UCF-101 [57] and HMDB-51 [56] for activity recognition. However, it is also a chronic problem in the construction community that there is lack of well-known datasets. One reason could be that creating a dataset for general objectives is usually expensive [63]. Another reason could be that datasets created in our community generally serve specific research objectives, which may limit their users. The dataset with a taxonomy of 16 activities of form workers and rebar workers has great potential to server other studies in construction.

6.2. Research limitations

This study has following two limitations. Firstly, there could be activities that cannot be categorized into any activity in the taxonomy. As explained in Section 4.1, we select 3 s as the time length of activity units to discretize continuous activities into finite atomic activities. Based on the time length of activity units, we carefully identify the activities that could appear with formwork workers and rebar workers. We believe that the taxonomy of 16 activities in this study covers the frequently observable activities and therefore adequately serves the testing and demonstration objectives. Secondly, manually specifying bounding boxes of workers, in the beginning, is necessary to start the tracking process with MDNet. The cold start undermines the efficiency of our method. Technically, detection-and-tracking MOT methods could be a solution to the cold-start problem, while inaccurate detection and frequent identify switching are still the major problems to be addressed.

7. Conclusion

In this paper, we introduced a new labor activity recognition method, which receives surveillance videos as input and recognizes diverse and continuous activities of each worker in the field of view. The method integrates the recent advances in computer vision and deep learning, including single object tracking with MDNet [60], optical flow estimation with FlowNet 2.0 [54,55], and activity recognition with TSNs [29]. To leverage complementary prediction capacities of two streams, we proposed a new fusion strategy, which considers the 24 best recognition scores in spatial streams and the nine best scores in temporal streams. The experimental results show that our activity recognition method has achieved an average accuracy of 80.5%, which is comparable with the state-of-the-art of activity recognition in the computer vision community, given the severe camera motion and low resolution of site surveillance videos and the marginal inter-class difference and significant intra-class variation of workers' activities. We also demonstrated that the method can underpin the implementation of efficient and objective work sampling in labor productivity evaluation and that it possesses great potential for use in practice.

Acknowledgements

The authors acknowledge the support of the Innovation and Technology Commission (ITC) of Hong Kong (ITP/036/12LP) and the Research Grants Council of Hong Kong (PolyU 152093/14E).

Appendix A. Supplementary material

Source code and datasets associated with this article can be found, in the online version, at https://github.com/eric4note/worksampling.

References

[1] S. Rowlinson, Cost Escalation in the Hong Kong Construction Industry Report, Hong Kong Construction Association (HKCA), Hong Kong SAR, 2014.
[2] S.P. Dozzi, S.M. Abourizk, Productivity in Construction, Institute for Research in Construction, National Research Council, Ottawa, ON, Canada, 1993.
[3] T.S. Braver, Motivation and Cognitive Control, Routledge, 2015.
[4] Y. Du, Y. Fu, L. Wang, Skeleton based action recognition with convolutional neural network, 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015, pp. 579–583.
[5] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2015) 1110–1118.
[6] F. Han, B. Reily, W. Hoff, H. Zhang, Space-time representation of people based on 3D skeletal data: a review, Comput. Vis. Image Underst. 158 (2017) 85–105.
[7] C. Li, Q. Zhong, D. Xie, S. Pu, Skeleton-based action recognition with convolutional neural networks, arXiv Preprint, 2017 arXiv:1704.07595.
[8] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a lie group, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2014) 588–595.
[9] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, AAAI 2 (2016) 8.
[10] J. Gong, C.H. Caldas, C. Gordon, Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models, Adv. Eng. Inform. 25 (4) (2011) 771–782.
[11] J. Yang, Z. Shi, Z. Wu, Vision-based action recognition of construction workers using dense trajectories, Adv. Eng. Inform. 30 (3) (2016) 327–336.
[12] 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. Inform. 27 (4) (2013) 652–663.
[13] R. Akhavian, A.H. Behzadan, Smartphone-based construction workers' activity recognition and classification, Autom. Constr. 71 (Part 2) (2016) 198–209.
[14] T. Cheng, J. Teizer, G.C. Migliaccio, U.C. Gatti, Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data, Autom. Constr. 29 (2013) 24–39.
[15] 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.
[16] J. Yang, M.-W. Park, P.A. Vela, M. Golparvar-Fard, Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future, Adv. Eng. Inform. 29 (2) (2015) 211–224.
[17] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inform. 29 (2) (2015) 239–251.
[18] J. Teizer, Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites, Adv. Eng. Inform. 29 (2) (2015) 225–238.
[19] J.K. Aggarwal, Q. Cai, Human motion analysis, Comput. Vis. Image Underst. 73 (3) (1999) 428–440.
[20] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput. Surv. 43 (3) (2011) 1–43.
[21] S.R. Egnor, K. Branson, Computational analysis of behavior, Annu. Rev. Neurosci. 39 (0) (2016) 217–236.
[22] M. Vrigkas, C. Nikou, I.A. Kakadiaris, A review of human activity recognition methods, Frontiers in Robotics and AI, 2 2015, p. 28.
[23] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comput. Vis. Image Underst. 115 (2) (2011) 224–241.
[24] H. Wang, C. Schmid, Action recognition by dense trajectories, 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3169–3176.
[25] H. Wang, C. Schmid, Action recognition with improved trajectories, Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.
[26] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, European Conference on Computer Vision, Springer, 2006, pp. 404–417.
[27] C.-Y. Ma, M.-H. Chen, Z. Kira, G. Alregib, TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition, arXiv Preprint, 2017 arXiv:1703.10667.
[28] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Proces. Syst. (2014) 568–576.
[29] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: towards good practices for deep action recognition, European Conference on Computer Vision, Springer, 2016, pp. 20–36.
[30] B. Zhang, L. Wang, Z. Wang, Y. Qiao, H. Wang, Real-time action recognition with enhanced motion vector CNNs, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 2718–2726.
[31] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (1) (2013) 60–79.
[32] X. Luo, H. Li, D. Cao, F. Dai, J. Seo, S. Lee, Recognizing diverse construction activities in site images via relevance networks of construction related objects detected by convolutional neural networks, J. Comput. Civ. Eng. (in press).
[33] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Adv. Neural Inf. Proces. Syst. (2015) 91–99.
[34] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, W. An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (Supplement C) (2018) 1–9.
[35] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, C. Li, Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment, Autom. Constr. 93 (2018) 148–164.
[36] L. Ding, W. Fang, H. Luo, P.E.D. Love, B. Zhong, X. Ouyang, A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory, Autom. Constr. 86 (Supplement C) (2018) 118–124.
[37] M. Bügler, A. Borrmann, G. Ogunmakin, P.A. Vela, J. Teizer, Fusion of photogrammetry and video analysis for productivity assessment of earthwork processes, Comput. Aided Civ. Inf. Eng. 32 (2) (2017) 107–123.
[38] J. Gong, C.H. Caldas, Computer vision-based video interpretation model for automated productivity analysis of construction operations, J. Comput. Civ. Eng. 24 (3) (2010) 252–263.
[39] E. Rezazadeh Azar, S. Dickinson, B. McCabe, Server-customer interaction tracker: computer vision-based system to estimate dirt-loading cycles, J. Constr. Eng. Manag. 139 (7) (2013) 785–794.
[40] J. Yang, P. Vela, J. Teizer, Z. Shi, Vision-based tower crane tracking for understanding construction activity, J. Comput. Civ. Eng. 28 (1) (2014) 103–112.
[41] J. Zou, H. Kim, Using hue, saturation, and value color space for hydraulic excavator idle time analysis, J. Comput. Civ. Eng. 21 (4) (2007) 238–246.
[42] S. Chi, C.H. Caldas, Automated object identification using optical video cameras on construction sites, Comput. Aided Civ. Inf. Eng. 26 (5) (2011) 368–380.
[43] S. Du, M. Shehata, W. Badawy, Hard hat detection in video sequences based on face features, motion and color information, ICCRD2011 — 2011 3rd International Conference on Computer Research and Development, Vol. 4 IEEE, Shanghai, China, 2011, pp. 25–29.
[44] 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.
[45] E. Rezazadeh Azar, B. McCabe, Automated visual recognition of dump trucks in construction videos, J. Comput. Civ. Eng. 26 (6) (2012) 769–781.
[46] E. Rezazadeh Azar, B. McCabe, Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos, Autom. Constr. 24 (2012) (2012) 194–202.
[47] Z. Zhu, I.J. Ndiour, I. Brilakis, P.A. Vela, Improvements to concrete column detection in live video, Proceedings of 27th International Symposium on Automation and Robotics in Construction (ISARC 2010), 2010, pp. 25–27.
[48] S. Han, S. Lee, A vision-based motion capture and recognition framework for behavior-based safety management, Autom. Constr. 35 (2013) (2013) 131–141.
[49] S. Han, S. Lee, F. Peña-Mora, Vision-based detection of unsafe actions of a construction worker: case study of ladder climbing, J. Comput. Civ. Eng. 27 (6) (2013) 635–644.
[50] S.J. Ray, J. Teizer, Real-time construction worker posture analysis for ergonomics training, Adv. Eng. Inform. 26 (2) (2012) 439–455.
[51] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, arXiv Preprint, 2015 arXiv:1502.06796.
[52] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: a large-scale video benchmark for human activity understanding, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2015) 961–970.
[53] C. Gu, C. Sun, D.A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, AVA: a video dataset of spatio-temporally localized atomic visual actions, arXiv Preprint, 2017 arXiv:1705.08421.
[54] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: evolution of optical flow estimation with deep networks, arXiv Preprint, 2016 arXiv:1612.01925.
[55] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, Flownet: learning optical flow with convolutional networks, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[56] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 2556–2563.
[57] K. Soomro, A.R. Zamir, M. Shah, UCF101: a dataset of 101 human actions classes from videos in the wild, arXiv Preprint, 2012 arXiv:1212.0402.
[58] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, arXiv Preprint, 2015 arXiv:1502.03167.
[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition, 2009, IEEE, Miami, FL, USA, 2009, pp. 248–255.
[60] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, arXiv Preprint, 2015 arXiv:1510.07945.
[61] Pascal VOC, The PASCAL Visual Object Classes, (2012).
[62] ImageNet, Microsoft COCO, ImageNet and MS COCO Visual Recognition Challenges Joint Workshop, 2015.
[63] K. Liu, M. Golparvar-Fard, Crowdsourcing construction activity analysis from jobsite video streams, J. Constr. Eng. Manag. 141 (11) (2015) 04015035.