Journal of Civil Engineering and Management
ISSN 1392-3730 / eISSN 1822-3605
2018 Volume 24 Issue 7: 568–580
https://doi.org/10.3846/jcem.2018.6133

ENHANCING ACTION RECOGNITION OF CONSTRUCTION WORKERS USING DATA-DRIVEN SCENE PARSING

Jun YANG

School of Automation, Northwestern Polytechnical University, Xi'an, China

Received 09 March 2018; accepted 26 August 2018

Abstract. Vision-based action recognition of construction workers has attracted increasing attention for its diverse applications. Though state-of-the-art performances have been achieved using spatial-temporal features in previous studies, considerable challenges remain in the context of cluttered and dynamic construction sites. Considering that workers actions are closely related to various construction entities, this paper proposes a novel system on enhancing action recognition using semantic information. A data-driven scene parsing method, named label transfer, is adopted to recognize construction entities in the entire scene. A probabilistic model of actions with context is established. Worker actions are first classified using dense trajectories, and then improved by construction object recognition. The experimental results on a comprehensive dataset show that the proposed system outperforms the baseline algorithm by 10.5%. The paper provides a new solution to integrate semantic information globally, other than conventional object detection, which can only depict local context. The proposed system is especially suitable for construction sites, where semantic information is rich from local objects to global surroundings. As compared to other methods using object detection to integrate context information, it is easy to implement, requiring no tedious training or parameter tuning, and is scalable to the number of recognizable objects.

Keywords: worker, action recognition, scene parsing, computer vision, context.

Introduction

Effective and timely analysis of workforce activity is essential for productivity measurement, progress evaluation, safety monitoring, and labor force training (Gouett *et al.* 2011; Gerek *et al.* 2014; Akhavian, Behzadan 2016; Han *et al.* 2014). Current efforts typically lean on visual observation and manual analysis, including an array of project-level information systems, direct observation methods, and survey/interview-based methods (Kim, Caldas 2013). This is usually a tedious and high cost task because valuable visual observations at a high confidence level usually require hours of continuous observation (CII 2010), in addition to the very considerable amount of time required for data analysis. Furthermore, workers may alter their behavior when being noticeably observed. Construction activities will then unintentionally diverge from the norm.

There is an urgent need for automated activity analysis. In the past decade, information technology has been applied in research field to collect operation data and analyze construction activity automatically. Some methods are based on tracking locations of construction entities (Navon, Goldschmidt 2010; Cho *et al.* 2014). They use various sensor systems, such as ultra-wide band (UWB) (Cheng *et al.* 2011), global positioning (GPS) (Pradhananga, Teizer 2013), or radio frequency identification (RFID) (Costin *et al.* 2012), to track workers or equipment and interpret their activities using prior knowledge of the site layout. Other methods are based on recognizing gestures of construction entities. They capture the body movements of construction workers by means of wearable accelerometers (Joshua, Varghese 2011, 2013), embedded smartphone sensors (Akhavian, Behzadan 2015, 2016) or motion capture system (Han *et al.* 2014) and then recognize activities by machine learning.

Compared to the aforementioned technologies, video cameras capture wide range information non-intrusively in a relatively low cost. With the aid of computer vision technology, both construction activities (Yang *et al.* 2015) and as-built infrastructures (Fathi *et al.* 2015) can be analyzed automatically. Vision-based action recognition is the first step of activity analysis. Recent years, several researchers have studied spatial-temporal feature based worker action recognition. The state-of-art recognition rate report-

*Corresponding author. E-mail: *junyang@nwpu.edu.cn*

ed for 11 action types was merely 59% (Yang *et al.* 2016), which is not satisfying for further analysis in real application. Video cameras record vivid content in construction sites, including workers, equipment, tools, materials, and temporary facilities or structures. All these semantic information naturally offers supplementary evidence for action recognition. However, existing action recognition methods mainly rely on analyzing workers movement without considering semantic information.

To address this issue, we propose utilizing semantic information to enhance worker action recognition. Conventional approach of obtaining semantic information is object detection (Kim, Caldas 2013; Gupta *et al.* 2009; Yao, Fei-Fei 2010a, 2010b). It detects only a small and fixed set of objects, while abandoning other semantically valuable information. Scene parsing is able to depict the entire scene non-parametrically. Compared to object detection, they require considerably less time for training and system tuning, but supply more semantic information (Liu *et al.* 2011a). So we adopt a scene parsing method to obtain semantic information from construction site. Our system comprises four modules: (1) taxonomy of construction objects is established to describe the trades-related context, (2) a scene parsing method named 'label transfer' is adopted for construction object recognition, (3) the probabilistic model depicting the relationships between objects and worker actions is learned from training data, and (4) the baseline of worker action recognition is improved by using semantic information.

Promising experimental results are achieved on a publicly available dataset, which contains 500 video clips recorded in real construction sites, covering 11 types of worker actions. Algorithm with the state-of-the-art performance of worker action recognition (Yang *et al.* 2016) was selected as the baseline. The experimental results showed that the proposed system outperforms the baseline algorithm by 10.5% on average.

The main contribution of this paper is that the idea of adopting a scene parsing method for action recognition enhancement is novel. It provides a new solution to integrate semantic information globally, other than conventional object detection, which can only depict local context. The proposed system is especially suitable to be applied in construction sites, where semantic information is rich from local objects to global surroundings.

1. Related work

During the past decade, many researchers have applied computer vision technologies to construction operation analysis (Yang *et al.* 2015; Seo *et al.* 2015; Teizer 2015). In early studies, usually workers were detected (Rezazadeh Azar, Mccabe 2012; Memarzadeh *et al.* 2013) and tracked (Peddi *et al.* 2009; Yang *et al.* 2010) or equipment was tracked (Zou, Kim 2007; Bugler *et al.* 2014; Brilakis *et al.* 2011), and then their activities were analyzed by trajectories using prior knowledge of the site layout (Rezazadeh Azar *et al.* 2012; Yang *et al.* 2014; Gong, Caldas 2011). However, trajectories are not always sufficient for construction operation analysis, in particular when trades are practiced at a fixed spot without obvious location changes. Under such circumstances, it is more important to describe the body movements of equipment or workers. Spatial-temporal descriptors effectively depict motion through the space and time domains, as well as capturing salient background information (Laptev 2005; Dollar *et al.* 2005). A recent trend is to adopt the spatial-temporal feature descriptors in a bag-of-features pipeline for action recognition without using explicit object detection or tracking (Laptev *et al.* 2008; Wang *et al.* 2013).

Gong *et al.* (2011) utilized the 3D Harris detector as the feature detector, histogram of gradient (HoG) and histogram of optical flow (HoF) as feature descriptors, and Bayesian network models as the learning method in worker and backhoe action recognition. Golparvar-Fard *et al.* (2013) focused on action recognition of earth moving equipment. They used a Gabor filter as the feature detector, HoG and HoF as descriptors, and support vector machines (SVMs) for action learning. Yang *et al.* (2016) studied worker action recognition. They established a new dataset with 1176 video clips, covering 11 types of trades. A state-of-the-art recognition rate was achieved using dense trajectory description (Wang *et al.* 2013).

Evidence garnered from cognitive science research shows that humans require semantic information, such as the context, scene, or interacting objects, to recognize actions (Biederman *et al.* 1982). In many studies in the computer vision field, attempts have been made to include semantic information in human activity recognition (Ziaeefard, Bergevin 2015; Onofri *et al.* 2016; Herath *et al.* 2017). Marszalek *et al.* (2009) exploited the context of natural dynamic scenes for human action recognition in video clips. They used movie scripts for annotation and discovered the reoccurring relation between scenes and actions. Instead of focusing on the scene in general, Ullah *et al.* (2010) proposed improving action recognition by disambiguating local space-time features and integrating additional non-local cues. They decomposed videos into region classes and augmented local features with corresponding region-class labels. Gupta *et al.* (2009) presented a Bayesian approach for gaining an understanding of human-object interactions. In their method, spatial and functional constraints were applied for coherence semantic interpretation.

Construction workers interact with tools, equipment, materials, or other workers frequently in order to complete their tasks. However, in few studies was an attempt made to understand worker activities using semantic information. A pioneering work by Kim and Caldas (2013) was the first to use tools information to improve worker action recognition. In their study, worker actions are recorded as skeleton movements by Microsoft KINECT system, and classified by a Gaussian mixture model. Three types of actions, "caulking", "hammering", and "screwing", were involved.

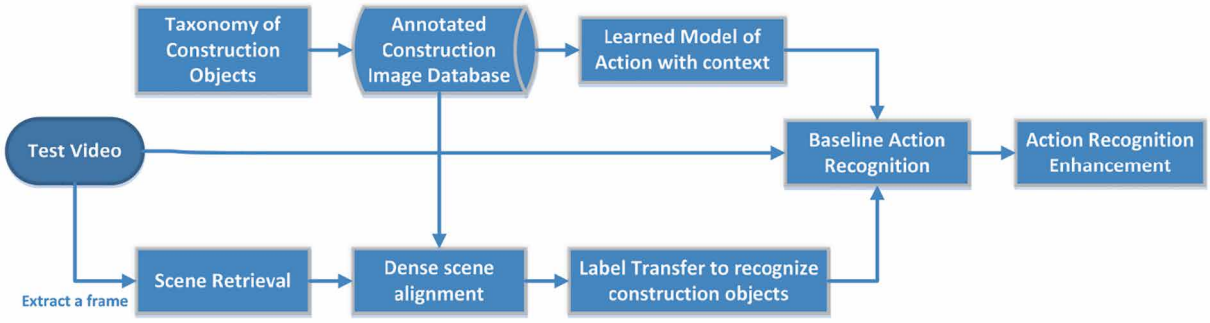

Figure 1. System overview

Then, handheld tools were recognized and included in a probabilistic model to improve action recognition. The study was conducted in a controlled indoor environment. Only handheld tools were considered as context, and other semantic information was not utilized. Construction sites are usually cluttered and dynamic and thus handheld tools may be invisible as a result of occlusion. Luo *et al.* (2018) proposed a novel scheme to interpret interaction activities from still site images using semantic information. Construction-related objects were detected using convolutional neural networks. Semantic relevance and spatial relevance were established to describe relevance between two objects. Then activities were recognized through predefined activity patterns. Their method has the advantage to interpret multiple activities simultaneously from wide-range surveillance images. It can be applied as guidance to further detailed action analysis.

Construction sites evolve over time and involve a large number of entities (Teizer 2015). Kim *et al.* (2016) proposed a novel scheme to recognize construction objects in the entire image using a data-driven scene parsing method. The system was nonparametric and scalable to the number of recognizable objects. An average pixel-wise recognition rate of 81.48% was achieved for real construction site images. Inspired by Kim *et al.* (2016) and Kim and Caldas (2013), we propose a system using context information obtained by data-driven scene parsing to enhance action recognition of construction workers.

The differences between the proposed system and the two closely related studies are as follows. Kim *et al.* (2016) introduced an existing scene parsing method 'Label Transfer' to construction objects recognition. We adopted the same scene parsing method to obtain semantic information from the entire construction site. Based on our own taxonomy, the semantic information was integrated to enhance worker action recognition. Kim and Caldas (2013) recognized three types of hand-held tools by conventional object detection and applied the tools information to improve skeleton-based worker action recognition. Our scene parsing based system can recognize more types of construction objects easily and obtain semantic information globally. Furthermore, compared to KINECT sensor, video cameras has no strict constrains on indoor or outdoor environment.

2. Methodology

The overall workflow of the proposed system is shown in Figure 1. As can been seen, it comprises three pipelines. The first pipeline is to build taxonomy of construction objects (will be introduced in Section 2.3), annotate the image database (Section 3.1), and learn the model of construction actions with context (Section 2.3). The second pipeline is data-driven scene parsing (Section 2.1), which involves three modules: scene retrieval, dense scene alignment, and label transfer-based object recognition. The third pipeline is to apply the object recognition results to the baseline algorithm (Section 2.2) for action recognition enhancement (Section 2.3). Among these procedures, data-driven scene parsing adopts an existing algorithm label transfer (Liu *et al.* 2011a). And the baseline of worker action recognition was originally developed by Yang *et al.* (2016).

Note that action recognition is conducted on videos while scene parsing is based on images. Therefore, given a test video, its first frame is extracted for scene parsing.

2.1. Scene parsing-based construction object recognition

Scene parsing is to segment and parse an image into different image regions associated with semantic categories, such as worker, scaffold, rebar, and hammer. In our system, a nonparametric scene parsing method named label transfer (Liu *et al.* 2011a) is adopted for construction object recognition. An image dataset is required for scene parsing. All images in the dataset should be annotated manually with object category labels. The first step of the algorithm is called 'scene retrieval', which is to match an input image (usually called a query) with similar images (usually called neighbors) in the database. Then the second step 'dense scene alignment' is to establish dense scene correspondence between the query image and each of the retrieved nearest neighbors. Lastly, 'label transfer' is to mapping the annotations from the nearest neighbors to the query image according to the estimated dense correspondence. Details are described as follows.

Scene Retrieval. Given a query image, scene retrieval is to find a set of nearest neighbors that share similar scene configuration with the query. *K*-NN model and

ϵ-NN model are two commonly used models to find the nearest neighbors. *K*-NN model takes the *K* closest images to the query while ϵ-NN model finds all images within $(1+\epsilon)$ times the minimum distance from the query. In our system, these two models are generalized as $<K, \epsilon>$ -NN model, defined as:

$$\mathcal{N}(x) = \{y_i \mid \text{dist}(x, y_i) \le (1+\epsilon)\text{dist}(x, y_i), y_i = \text{argmin dist}(x, y_i), i \le K\}, \quad (1)$$

where is the query image, y_i represents one of the nearest neighbors, and $dist(.,.)$ is the distance function. The Euclidean distance of the GIST descriptor (Oliva, Torralba 2001) was used in this study.

Dense Scene Alignment. In order to transfer existing annotations to a query image, dense correspondence needs to be established between the query image and its nearest neighbors. SIFT flow is used to find correspondence by matching local SIFT descriptors (Liu *et al.* 2011b). The energy function of SIFT flow is defined as:

$$E(\mathbf{w}) = \sum_{\mathbf{p}} \min(\| s_1(\mathbf{p}) - s_2(\mathbf{p}+\mathbf{w}(\mathbf{p})) \|_1, t) + \quad (2)$$

$$\sum_{\mathbf{p}} \eta(|u(\mathbf{p})| + |v(\mathbf{p})|) + \quad (3)$$

$$\sum_{(\mathbf{p},\mathbf{q})\in\varepsilon} \min(\lambda |u(\mathbf{p}) - u(\mathbf{q})|, d) + \min(\lambda |v(\mathbf{p}) - v(\mathbf{q})|, d), \quad (4)$$

where $\mathbf{p} = (x, y)$ is the pixel coordinate, $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector at point $\mathbf{p}$, s_1 and s_2 represent the local SIFT descriptor for two images, and ε contains all nearest neighbors. The three terms in the energy function have different control objectives: the data term (Eqn (2)) constrains the SIFT descriptor to be matched along the flow vector, the small displacement term (Eqn (3)) ensures the flow vectors are as small as possible, and the spatial regularization term (Eqn (4)) constrains the flow vector of adjacent pixels to be similar. By minimizing the energy, the top M re-ranked votes are retrieved from the $<K, \epsilon>$ -nearest neighbor ($M \le K$). This set contains the candidates for label transfer to the query image.

Label Transfer. Now, the scene parsing problem can be formulated as the label transfer problem from the matching candidates to the query image. Let I be the query image with its SIFT image and s be the candidate set, in which s_i, c_i, and $\mathbf{w}_i$ are the SIFT image, annotation, and SIFT flow field (from s to s_i) of the i th candidate, respectively. A probabilistic Markov random field model is built to parse image I. As shown in Eqn (5), the posterior probability contains three components: likelihood, prior, and spatial smoothness. The pixels in the query image are labeled by minimizing

$$-\log P(c \mid I, s, \{s_i, c_i, \mathbf{w}_i\}) = \sum_{\mathbf{p}} \psi(c(\mathbf{p}); s, s_{i'}) + \alpha \sum_{\mathbf{p}} \lambda(c(\mathbf{p})) + \beta \sum_{\{\mathbf{p},\mathbf{q}\}\in\varepsilon} \phi(c(\mathbf{p}), c(\mathbf{q}); I) + \log Z. \quad (5)$$

The likelihood term is defined as:

$$\psi(c(\mathbf{p}) = l) = \begin{cases} \min_{i\in\Omega_{\mathbf{p},l}} \| s(\mathbf{p}) - s_i(\mathbf{p}+\mathbf{w}(\mathbf{p})) \|, & \Omega_{\mathbf{p},l} \ne 0, \\ \tau, & \Omega_{\mathbf{p},l} = 0, \end{cases} \quad (6)$$

where $\Omega_{\mathbf{p},l} = \{i; c_i(\mathbf{p}+\mathbf{w}(\mathbf{p})) = l\}, l = 1, \ldots, L$ is the index set of the candidate images, the label of which is l after being warped to pixel $\mathbf{p}$. τ is the value of the maximum difference of the SIFT feature: $\tau = \max_{s_1, s_2, \mathbf{p}} \| s_1(\mathbf{p}) - s_2(\mathbf{p}) \|$.

The prior term $\lambda(c(\mathbf{p}) = l)$ is the prior probability of object type *l* appearing at pixel **p**. It is estimated by counting the occurrences of object type *l* at pixel **p** during training:

$$\lambda(c(\mathbf{p}) = l) = -\log \text{hist}_l(\mathbf{p}), \quad (7)$$

where $\text{hist}_l(\mathbf{p})$ is the spatial histogram of object type *l*.

The smoothness term sets the neighboring pixels to have the same label when no other information is available. The chance of neighboring pixels having different labels is proportional to the luminance of the image edges:

$$\phi(c(\mathbf{p}), c(\mathbf{q})) = \delta[c(\mathbf{p}) \ne c(\mathbf{q})] \left(\frac{\xi + e^{-\gamma \|I(\mathbf{p}) - I(\mathbf{q})\|^2}}{\xi + 1}\right), \quad (8)$$

where $\gamma = (2 < \| I(\mathbf{p}) - I(\mathbf{q}) \|^2 >)^{-1}$.

2.2. Baseline algorithm for action recognition

For worker action recognition, a cutting-edge video description method, dense trajectories, was applied in a bag-of-features pipeline (Wang *et al.* 2013). Given an input video, each frame is densely sampled to obtain a set of points. These points are tracked based on displacement information from a dense optical flow field. Displacement record of a point in the time domain is a trajectory. Feature descriptors (HoG, HoF, or motion boundary histograms (MBH)) are computed along trajectories. The spatio and temporal feature descriptions of all points are concatenated to generate the video description, so called dense trajectories. Then, codebooks are generated by K-means clustering and descriptors are assigned to their nearest vocabulary word for quantization. Non-linear SVMs are trained for action recognition. We refer the readers to Yang *et al.* (2016) for more details.

2.3. Enhancing action recognition by scene parsing-based object recognition

Taxonomy of construction objects. Taxonomy of construction objects is a category structure to define interested construction objects. Reasonable taxonomy is important for effectively understanding a construction scene. Three rules are followed in the establishment of taxonomy. First, only construction entities are considered. There are five main categories: personnel, equipment, tools, materials, and temporary facilities/structures. The natural environment (e.g., trees) is not included. Second, action recognition enhancement is accomplished by assuming that

the occurrences of certain types of objects are indicators of related actions. Hence, the granularity of objects in taxonomy is rendered coarse, only focusing on the objects' existence without considering their quantities. For example, we annotate "bricks" instead of single pieces of "brick". The reader can refer to a list of frequently used temporary construction resources in Teizer (2015). The proposed taxonomy is shown in Figure 2. As can been seen, there are attributes under each category, totally 54 types of objects.

Enhancing Action Recognition with Semantic Information. Now, we are ready to improve action recognition using semantic information obtained from scene parsing. Given an input video x, let $g_a(x)$ and $g_s(x)$ be the score vectors of action classification and object recognition, respectively. The model of action recognition with context is defined as (Marszalek *et al.* 2009):

$$g_{a'}(x) = g_a(x) + \tau w g_s(x), \tag{9}$$

where $g_{a'}(x)$ is the new score vector for all action classes, $g_s(x)$ is the score vector for all object classes. τ is a weighting parameter, decided by experiments. w is the conditional probability matrix encoding the occurrence probability of a certain type of object given an action, which means $w = p(\text{Object} \mid \text{Action})$. It can be estimated from the annotated training data. Finally, the action type is determined by the maximum value of $g_{a'}(x)$.

Note that, although we selected specific algorithms for baseline action recognition in this study, technically any algorithm that fits can be plugged into Eqn (9) for context-enhanced action recognition.

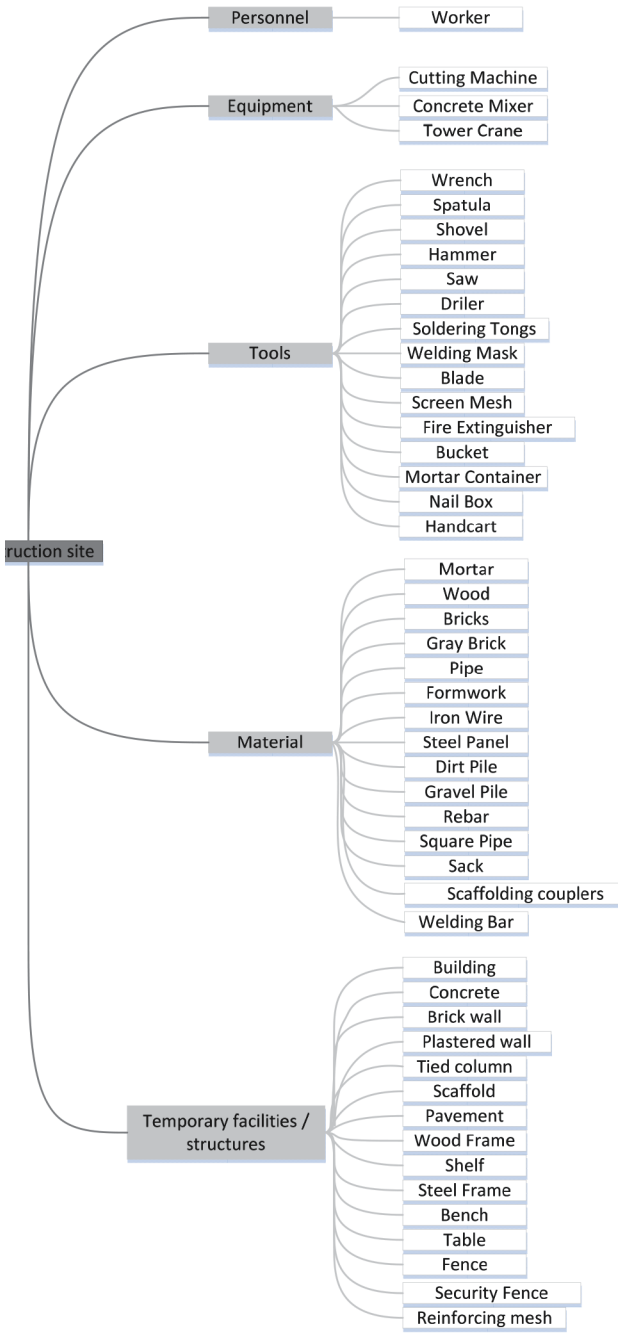

Figure 2. Proposed taxonomy of construction objects

3. Experimental results

3.1. Data preparation and experimental setup

A publicly available worker action dataset (Yang *et al.* 2016) was used in our experiments. The original dataset involves 11 types of worker actions: "LayBrick", "Transporting", "CutPlate", "Drilling", "TieRebar", "Nailing", "Plastering", "Shoveling", "Bolting", "Welding", and "Sawing". The action type "Transporting" was excluded from the experiments since it is not specifically related to semantic information. Meanwhile, 50 video clips covering different workers and view angles are extracted from each action type in order to obtain an equally distributed dataset. The final dataset with 500 video clips was divided into halves randomly for training and testing.

For scene parsing, the first frames of all videos were annotated according to the proposed taxonomy using the software LabelMe (Russell *et al.* 2008) for training and evaluation. The open source code of label transfer scene parsing is available at Liu *et al.* (2011a).

Two sample images from each action type, together with their annotations, are displayed in Figure 3. The occurrence frequency of all the objects is shown in Figure 4.

3.2. Evaluation metrics

To evaluate the scene parsing performance, two metrics were used: average pixel-wise recognition rate $\bar{r}$ and per-class average rate r_l (Liu *et al.* 2011a):

$$\bar{r} = \frac{1}{\sum_i m_i} \sum_i \sum_{\mathbf{p} \in \Lambda_i} \mathbf{1}(o(\mathbf{p}) = a(\mathbf{p}), a(\mathbf{p}) > 0), \tag{10}$$

where, for pixel $\mathbf{p}$ in image i, $a(\mathbf{p})$ is the annotated ground truth, $o(\mathbf{p})$ is the output of label transfer, Λ_i represents the image lattice for test image i, and $m_i = \sum_{\mathbf{p} \in \Lambda_i} \mathbf{1}(a(p) > 0)$ represents the number of all labelled pixels in image i (note that some pixels may be unlabeled). The per-class average rate is estimated as:

$$r_l = \frac{\sum_i \sum_{\mathbf{p}\in\Lambda_i} \mathbf{1}(o(\mathbf{p}) = a(\mathbf{p}), a(\mathbf{p}) = l)}{\sum_i \sum_{\mathbf{p}\in\Lambda_i} \mathbf{1}(a(\mathbf{p}) = l)}, \quad l = 1,\ldots,L. \qquad (11)$$

For action recognition evaluation, average precision (AP) was adopted, which approximates the area under a recall-precision curve (Everingham *et al.* 2008). For overall system performance, the average AP (AAP) was computed.

3.3. Results for scene parsing-based object recognition

We selected one test image from each action type to show the results of label transfer in Figure 5. Different objects are labeled in different colors according to the legend on the right. The query image from the test set is displayed in Figure 5(a). The best match from Figure 5(a)'s nearest neighbors together with its corresponding annotation is shown in Figures 5(b) and 5(c), respectively. Figure 5(d)

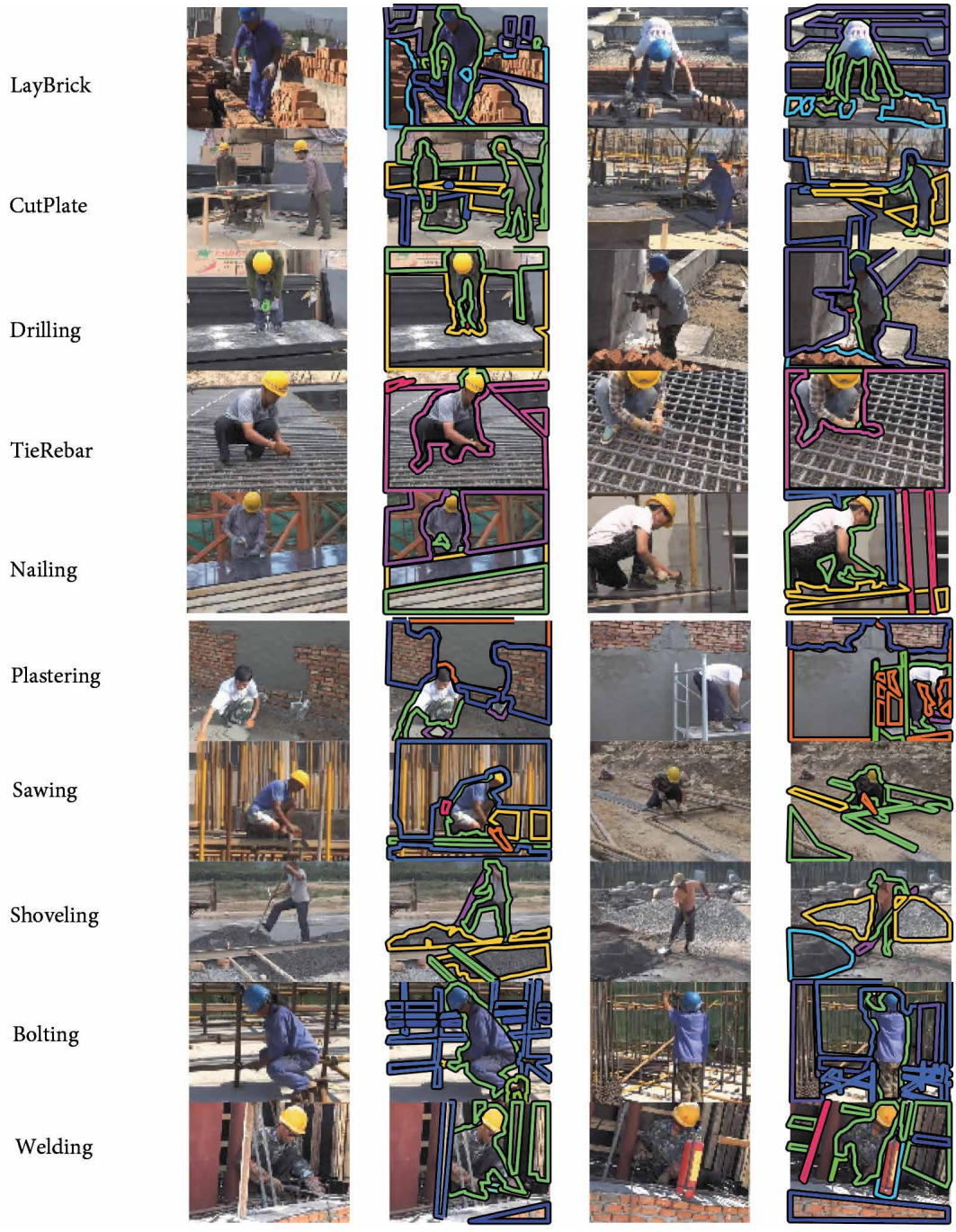

Figure 3. Sample images and their annotations in the dataset

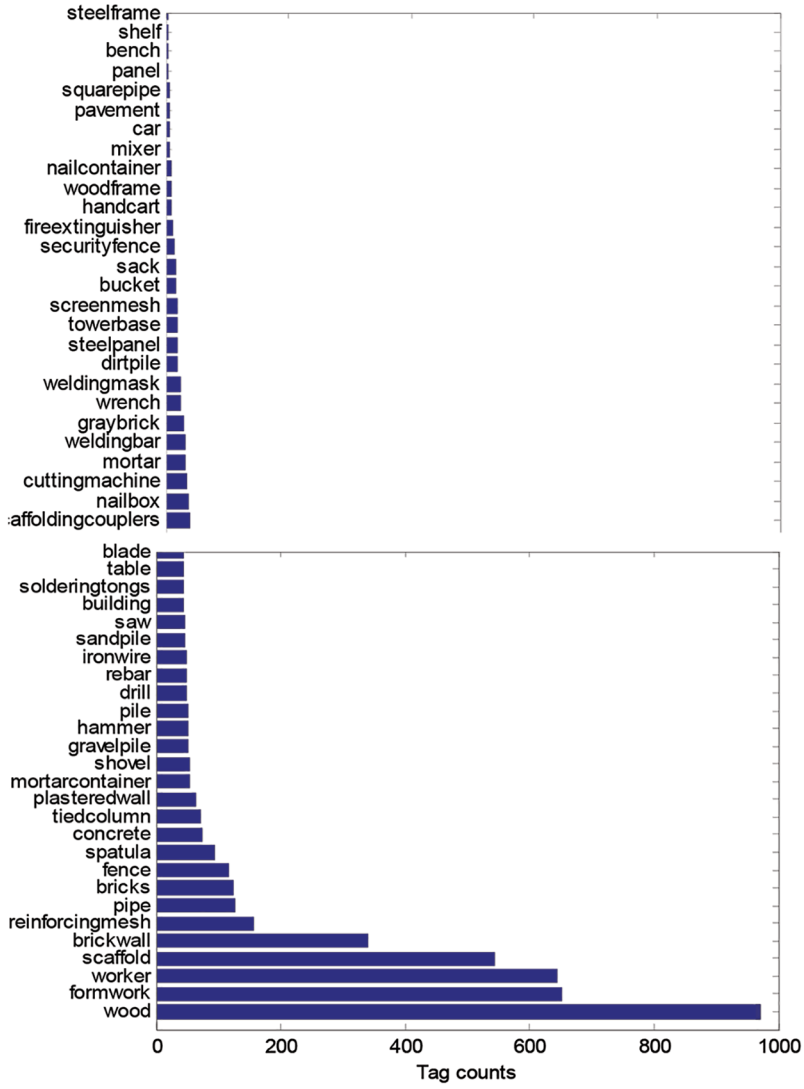

Figure 4. Occurrence frequency of all objects in ascending order

shows the warped image, where the matched SIFT flow field is applied to warp the RGB image of the nearest neighbor to the query. Figures 5(e), and 5(f) are the object recognition results and the ground truth produced by user annotation, respectively. The dark gray pixels represent "unlabeled" pixels. The system does not generate an unlabeled output, but attempts to produce a prediction for as many objects as possible.

In the label transfer procedure, there are five main parameters, the spatial smoothness coefficient λ, number of nearest neighbors K, number of final candidates M, prior weight α, and spatial weight β, which affect the performance of the algorithm. The spatial smoothness coefficient λ was fixed to 0.7 in the experiment, which is the optimal setting according to the study in Liu *et al.* (2011a). It was shown in Kim *et al.* (2016) that different settings of prior weight α and spatial weight β would have an effect of a magnitude of 0.001. We simply adopted the optimal configuration in Kim *et al.* (2016) with $\alpha = 0{:}06$ and $\beta = 20$. M and K are closely related to the scale of the data. Therefore, we chose different combinations of M, K and plotted the per-pixel recognition rate as a function of K for a variety of M s, as shown in Figure 6. Overall, the recognition rate increases with an increase in K. The best performance, 84.6%, is achieved when $K = 20$ and $M = 7$. Figure 7 shows the per class recognition rate under the optimal setting. As can be seen, the top five objects according to the recognition results are panel, mixer, reinforcing mesh, plastered wall, and gravel pile. The bottom five objects are shelf, hammer, wrench, saw, and scaffolding couplers. The tag count information in Figure 4 shows that the per class recognition rate is not necessarily related to their tag counts, or in other words, how frequently the objects appear in the dataset. It can also be easily noticed that the pixel area of the object has a strong correlation with its recognition rate. Usually, the smaller the object, the lower is its recognition rate. One possible explanation is that the chance that a small object is inaccurately labeled is large (Kim *et al.* 2016). Pixels not belonging to the small object lead to recognition errors. The second reason is

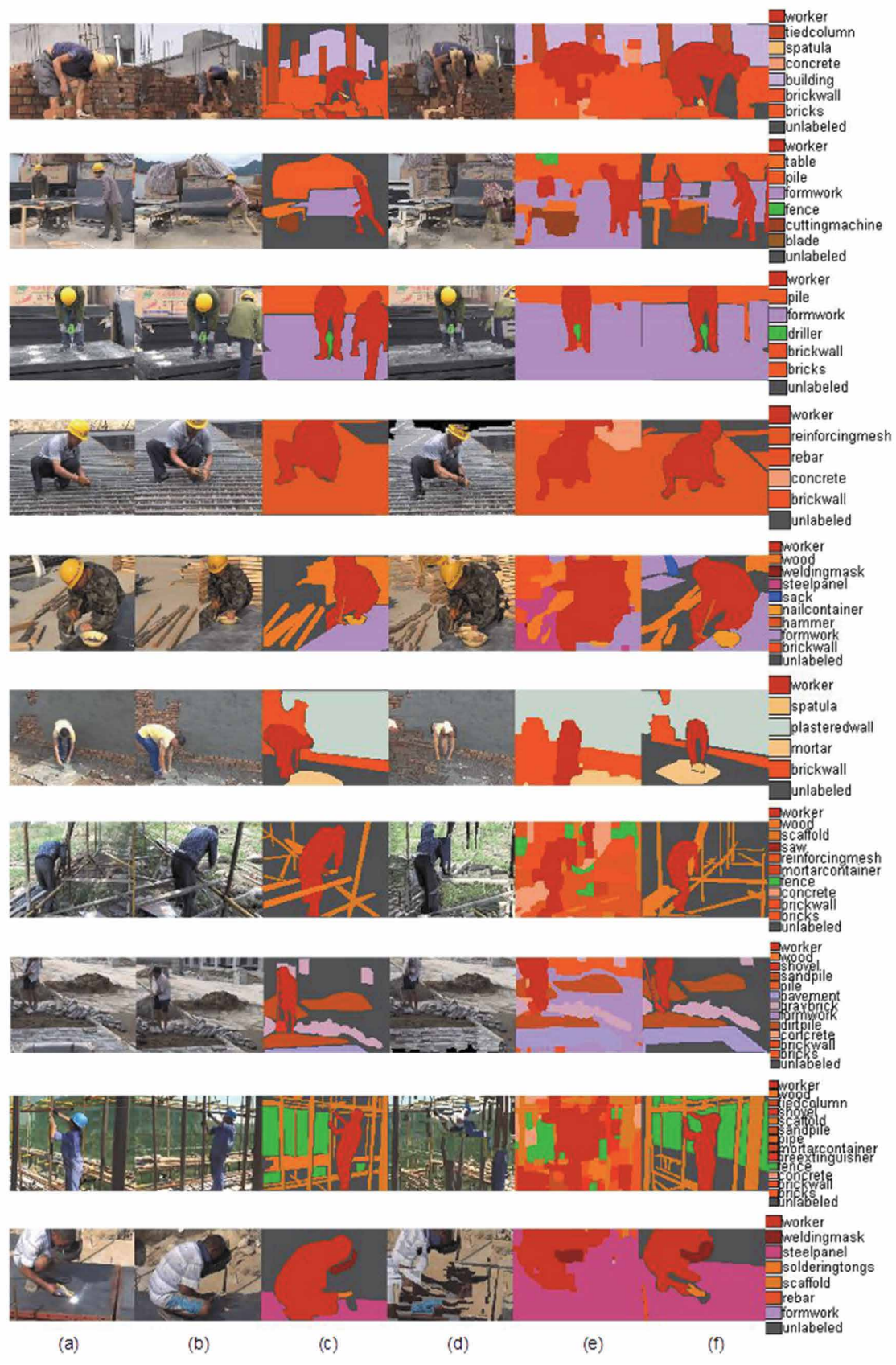

Figure 5. Scene parsing results: (a) query image, (b) the best match from nearest neighbors, (c) the annotation of the best match, (d) the warped version of (b) according to the SIFT flow field, (e) the object recognition result, and (f) the ground truth annotation of (a)

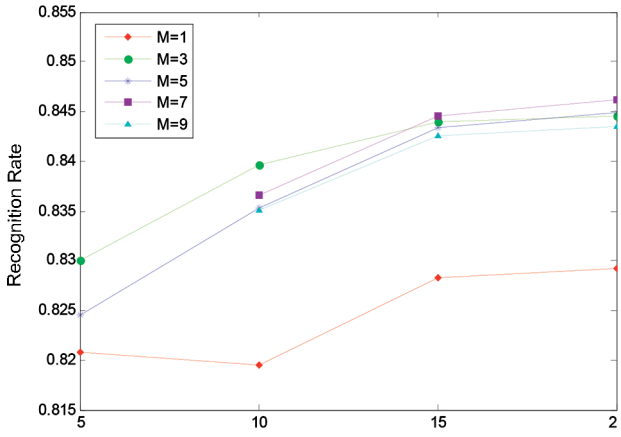

Figure 6. Per pixel recognition rate as a function of the number of nearest neighbors K and the number of voting candidates M

that the label transfer algorithm allows only one labeling for each pixel. Therefore, small objects tend to be overwhelmed by the labeling of larger objects (Liu *et al.* 2011a).

3.4. Results for action recognition enhancement

The occurrence probability of certain types of objects given an action is estimated from the annotated training data. In Figure 8, the probabilities of the top five related objects for each action type are displayed. Note that objects with probabilities equaling 1 are not ordered. Naturally, "worker" is the prerequisite for all action types. The handheld tools (drill, hammer, shovel, saw, etc.), materials (bricks, mortar, nail box, iron wire, etc.), materials (formwork, brick wall, scaffold, sand pile, etc.), and outcomes (reinforcing mesh, plastered wall, etc.) are the top categories with related corresponding actions. The estimated probability model agrees with common knowledge about construction. Meanwhile, it does not exclude any background objects or supported objects, which may have a minor relation with worker actions.

As described in Section 3.3.2, the semantic information gained from label transfer was applied to enhance action recognition. The baseline action recognition algorithm in Yang *et al.* (2016) was adopted. The parameters were set according to the best reported performance; MBH was used as the feature descriptor and the codebook size was equal to 500. The average precision per action type of the proposed system is shown in Figure 9. For comparison, the performance of the baseline algorithm, together with that of action prediction by object recognition, are also displayed. It can be seen clearly that the proposed system improved the performance of the baseline algorithm. The AAPs of the baseline algorithm, action prediction by context, and the proposed system are 69.2%, 47.9%, and 79.7%, respectively. The average gain of the proposed system as compared to the baseline is 10.5%, with a maximum gain of 20% and minimum gain of 2%. The most improved action type is "Drilling". While drilling, a worker shows no obvious body movement, and the fast spinning of the drill is difficult to capture by video descriptions. Using context information, such as the recognition of "Drill", action recognition is highly enhanced. The action type "Bolting" has the smallest gain. By closely examining the details, we discovered that the background of this action type was cluttered and lacked continuity because of the presence of scaffolding. Therefore, when the image was labeled, many pixels remained "unlabeled". During label transfer, the algorithm predicted unlabeled pixels for some unrelated objects for this action type. A poorly recognized context cannot facilitate action recognition. By tuning the parameter τ in Eqn (9), we can easily control the effect of context information on action recognition. According to our current experience, the best performance is achieved with τ equal to 0.7.

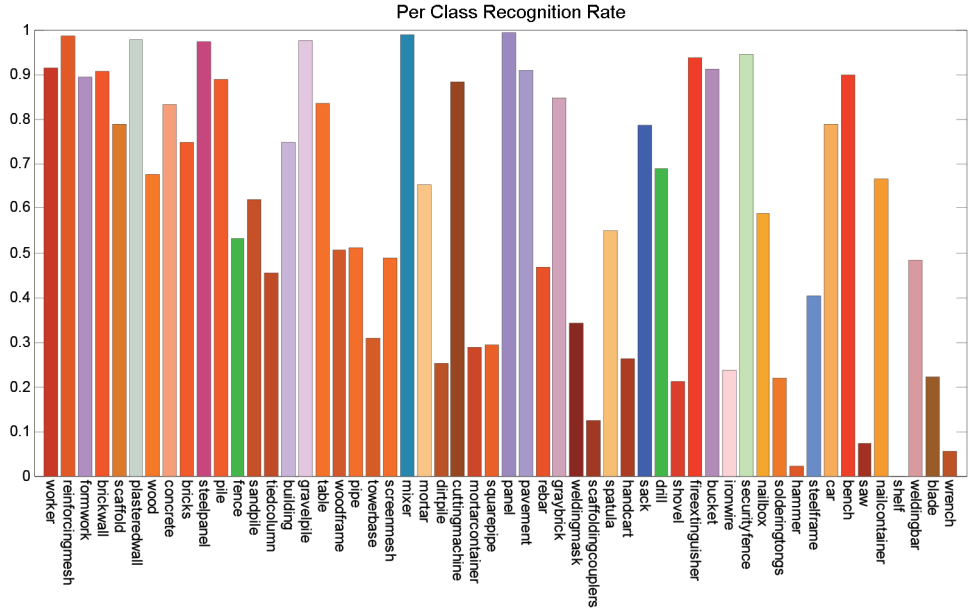

Figure 7. Per class recognition rate under the optimal parameter settings

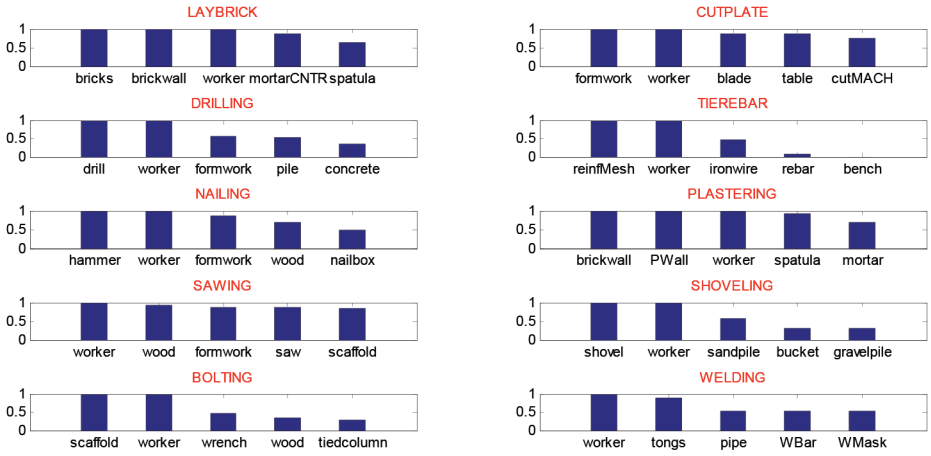

Figure 8. Probabilities of top five related objects for each action type. Note that because of space limitation, some of the objects' names are abbreviated: "mortarCNTR", "cutMACH", "reinfMesh", "PWall", "tongs", "WBar", and "WMask" represent "mortarcontainer", "cuttingmachine", "reinforcingmesh", "plasteredwall", "solderingtongs", "weldingbar", and "weldingmask", respectively

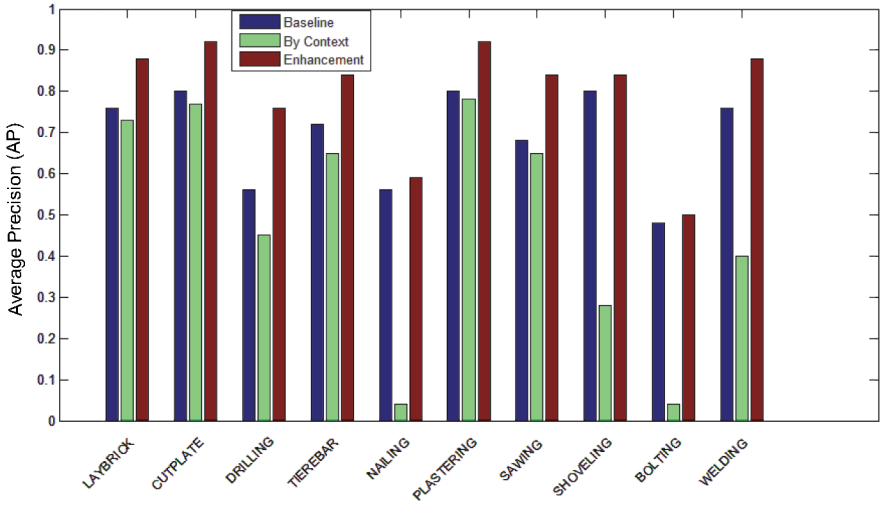

Figure 9. Comparison of the average precision of the proposed system for each action type with that of the baseline algorithm and context-based action prediction

4. Discussion

Experimental results show that worker actions recognition has been improved significantly with the integration of scene parsing based object recognition. Construction operations are essentially interactions between all the construction entities (workers, equipment, materials, tools and temporary facilities or structures). It is beneficial to combine context information for activity analysis. However, conventional object detection is computationally expensive. The proposed system is able to capture semantic information from the whole construction site conveniently. Another advantage of the system is its openness. To support the recognition of more action types, the user simply needs to add more samples from the new categories, annotate related objects, and re-train the action-context model. The tedious classifier training task required by conventional learning-based systems is not needed.

However, the limitation of the proposed system also originates in the scene parsing module. The success of scene parsing-based object recognition strongly depends on good matches between the query image and similar training images. This indicates that the system may fail when applied to an unseen construction scene. One possible solution is to enlarge the general database by annotating more construction site images (e.g., with the help of a crowd sourcing platform). When applied to new construction scenarios, it is also suggested to prepare a few annotated images of targeting scenes, which will not take many efforts using the software LabelMe and our proposed taxonomy.

Second, the effect of "unlabeled" pixels on action recognition is negative if their object types are wrongly predicted. One solution could be to add a confidence value to prediction. If the value is smaller than a certain threshold, the prediction result will be excluded from action recognition.

Lastly, the recognition of small objects is not satisfactory in the current scheme, which may affect the system's performance given that some handheld tools are usually in small scale. One solution is to increase the accuracy of small object labeling using a segmentation method, e.g., Grabcut (Tang *et al.* 2013). Another solution is to improve the label transfer algorithm. By introducing spatial constraints between various construction entities, small objects, such as handheld tools, can be constrained to locations close to workers. Recognition will then be improved. We leave the above-mentioned issues as future work.

Conclusions

We presented a novel system that uses semantic information to enhance worker action recognition. A non-parametric data-driven scene parsing method was adopted to recognize construction objects. The model of action recognition with context is learned from the training data. Action recognition is then improved by using the recognized construction objects. Promising results were achieved on a comprehensive dataset. The proposed system outperformed the baseline algorithm by 10.5% on average, with a maximum performance gain of 20% per action type. As shown in our taxonomy, the proposed system can obtain semantic information globally. It is beneficial for construction action recognition since semantic information is naturally rich in construction sites and the relationship between semantic information and construction operations can be clearly defined. The key point of 'label transfer' is to transfer object labels to query images from existing annotated images. Hence it is easy to change the number of recognizable objects as long as new objects are annotated in the database. Compared to conventional object detection, which needs to train individual classifier for each type of object, the scene parsing method is easy to implement, requiring less tedious training or parameter tuning, and is scalable to the number of recognizable objects (Liu *et al.* 2011a). The limitation of the proposed system is that it relies on good matches between the query image and the database. So it may fail in an unseen construction scene. It is suggested that in real world application a set of completely labeled images are prepared for a particular job site.

The proposed system proves the success of context based activity analysis and can be easily applied to real construction management. Notice that in real world application, video cameras for general monitoring purpose are usually mounted statically with a wide range of view, recording multiple activities simultaneously. To analyze certain type of worker actions as described in this paper (requiring a relatively high resolution of worker movements and related objects), a possible solution is to set up multiple temporary cameras to monitor different types of workers according to the site layout. Hand-held video cameras can also be used complementarily by foremen.

Action recognition forms a foundation for management applications. To measure productivity, continuous worker activities needs to be observed and measured, which will require action segmentation for long video analysis. With prior knowledge of construction operations, progress can be evaluated based on productivity. Though regular actions types are considered in this paper, safety monitoring can be conducted by recognizing unsafe actions from safe actions.

Recent years, deep learning has been a fast growing direction in computer vision. Several studies in construction field have adopted deep learning for object detection and so derived activity interpretation (Luo *et al.* 2018; Fang *et al.* 2018). Though deep learning has achieved the state-of-art performance on object detection (Krizhevsky *et al.* 2017), its performance heavily relies on tedious training and elaborate system tuning. The proposed scene parsing based construction object recognition is nonparametric, easy to implement with few parameters. And it is scalable, which means it can easily adjust the number of object categories without cumbersome re-adjustment of the system.

From the aspect of action recognition, Ding *et al.* (2018) integrated convolution neural networks and long short-term memory to detect unsafe worker behavior. Experimental results on a relatively small dataset (200 video clips, 4 action types) showed that deep learning method outperformed feature descriptors (HoG, HoF, MBH) based method. However, a recent survey on action recognition (Herath *et al.* 2017) discovered that both descriptors based method and deep learning performed equally well on a widely used dataset HMDB-51 (7000 video clips, 51 action types). Hence comparison study of descriptors based and deep learning based action recognition still needs to be performed on larger scale construction datasets, which should be one of the future directions.

Funding

This work was supported by the Natural Science Foundation of Shaanxi Province, China under Grant (No. 2017JM5061).

References

Akhavian, R.; Behzadan, A. H. 2015. Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers, *Advanced Engineering Informatics* 29(4): 867–877. https://doi.org/10.1016/j.aei.2015.03.001

Akhavian, R.; Behzadan, A. H. 2016. Smartphone-based construction workers' activity recognition and classification, *Automation in Construction* 71: 198–209. https://doi.org/10.1016/j.autcon.2016.08.015

Biederman, I.; Mezzanotte, R. J.; Rabinowitz, J. C. 1982. Scene perception: Detecting and judging objects undergoing relational violations, *Cognitive Psychology* 14(2): 143–177. https://doi.org/10.1016/0010-0285(82)90007-x

Brilakis, I.; Park, M.; Jog, G. M. 2011. Automated vision tracking of project related entities, *Advanced Engineering Informatics* 25(4): 713–724. https://doi.org/10.1016/j.aei.2011.01.003

Bugler, M.; Ogunmakin, G.; Teizer, J.; Vela, P. A.; Borrmann, A. 2014. A comprehensive methodology for vision-based progress and activity estimation of excavation processes for productivity assessment, in *Proceedings of the 21st International Workshop: Intelligent Computing in Engineering (EG-ICE)*, 2014, Cardiff, Wales.

Cheng, T.; Venugopal, M.; Teizer, J.; Vela, P. A. 2011. Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments, *Automation in Construction* 20(8): 1173–1184. https://doi.org/10.1016/j.autcon.2011.05.001

Cho, D.; Cho, H.; Kim, D. 2014. Automatic data processing system for integrated cost and schedule control of excavation works in NATM tunnels, *Journal of Civil Engineering and Management* 20(1): 132–141. https://doi.org/10.3846/13923730.2013.801907

CII. (Ed). 2010. *IR252.2a – Guide to activity analysis*. Construction Industry Institute, Austin, TX, USA [online], [cited 02 March 2018]. Available from Internet: https://www.construction-institute.org/resources/knowledgebase/knowledge-areas/general-cii-information/topics/rt-252/pubs/ir252-2a

Costin, A. M.; Pradhananga, N.; Teizer, J. 2012. Leveraging passive RFID technology for construction resource field mobility and status monitoring in a high-rise renovation project, *Automation in Construction* 24: 1–15. https://doi.org/10.1016/j.autcon.2012.02.015

Dollar, P.; Rabaud, V.; Cottrell, G.; Belongie, S. 2005. Behavior recognition via sparse spatio-temporal features, in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, IEEE, 65–72. https://doi.org/10.1109/vspets.2005.1570899

Ding, L.; Fang, W.; Luo, H.; Love, P. E. D.; Zhong, B.; Ouyang, X. 2018. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory, *Automation in Construction* 86: 118–124. https://doi.org/10.1016/j.autcon.2017.11.002

Everingham, M.; van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. 2008. Overview and results of the classification challenge, in *The PASCAL VOC08 Challenge Workshop, in conj. with ECCV*.

Fang, Q.; Li, H.; Luo, X.; Ding, L.; Rose, T. M.; An, W.; Yu, Y. 2018. A deep learning-based method for detecting non-certified work on construction sites, *Advanced Engineering Informatics* 35: 56–68. https://doi.org/10.1016/j.aei.2018.01.001

Fathi, H.; Dai, F.; Lourakis, M. 2015. Automated as-built 3D reconstruction of civil infrastructure using computer vision: achievements, opportunities, and challenges, *Advanced Engineering Informatics* 29: 149–161. https://doi.org/10.1016/j.aei.2015.01.012

Gerek, I. H.; Erdis, E.; Mistikoglu, G.; Usmen, M. 2014. Modelling masonry crew productivity using two artificial neural network techniques, *Journal of Civil Engineering and Management* 21(2): 231–238. https://doi.org/10.3846/13923730.2013.802741

Golparvar-Fard, M.; Heydarian, A.; Niebles, J. C. 2013. Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers, *Advanced Engineering Informatics* 27(4): 652–663. https://doi.org/10.1016/j.aei.2013.09.001

Gong, J.; Caldas, C. H. 2011. An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations, *Automation in Construction* 20(8): 1211–1226. https://doi.org/10.1016/j.autcon.2011.05.005

Gong, J.; Caldas, C. H.; Gordon, C. 2011. Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and Bayesian network models, *Advanced Engineering Informatics* 25(4): 771–782. https://doi.org/10.1061/41182(416)34

Gouett, M. C.; Haas, C. T.; Goodrum, P. M.; Caldas, C. H. 2011. Activity analysis for direct-work rate improvement in construction, *Journal of Construction Engineering and Management* 137(12): 1117–1124. https://doi.org/10.1061/(asce)co.1943-7862.0000375

Gupta, A. K.; Kembhavi, A.; Davis, L. S. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(10): 1775–1789. https://doi.org/10.1109/tpami.2009.83

Han, S.; Lee, S.; Pena-Mora, F. 2014. Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction, *Journal of Computing in Civil Engineering* 28(5): A4014005. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000339

Herath, S.; Harandi, M. T.; Porikli, F. 2017. Going deeper into action recognition: A survey, *Image and Vision Computing* 60: 4–21. https://doi.org/10.1016/j.imavis.2017.01.010

Joshua, L.; Varghese, K. 2011. Accelerometer-based activity recognition in construction, *Journal of Computing in Civil Engineering* 25(5): 370–379. https://doi.org/10.1061/(asce)cp.1943-5487.0000097

Joshua, L.; Varghese, K. 2013. Selection of accelerometer location on bricklayers using decision trees, *Computer-Aided Civil and Infrastructure Engineering* 28(5): 372–388. https://doi.org/10.1111/mice.12002

Kim, H.; Kim, K.; Kim, H. 2016. Data-driven scene parsing method for recognizing construction site objects in the whole image, *Automation in Construction* 71: 271–282. https://doi.org/10.1016/j.autcon.2016.08.018

Kim, J. Y.; Caldas, C. H. 2013. Vision-based action recognition in the internal construction site using interactions between worker actions and construction objects, in *International Symposium on Automation and Robotics in Construction and Mining*, 661–668. https://doi.org/10.22260/isarc2013/0072

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. 2017. ImageNet classification with deep convolutional neural networks, *Communications of the ACM* 60(6): 84–90. https://doi.org/10.1145/3065386

Laptev, I. 2005. On space-time interest points, *International Journal of Computer Vision* 64(2/3): 107–123. https://doi.org/10.1109/iccv.2003.1238378

Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. 2008. Learning realistic human actions from movies, in *International Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/cvpr.2008.4587756

Liu, C.; Yuen, J.; Torralba, A. 2011a. Nonparametric scene parsing via label transfer, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(12): 2368–2382. https://doi.org/10.1109/tpami.2011.131

Liu, C.; Yuen, J.; Torralba, A. 2011b. SIFT flow: Dense correspondence across scenes and its applications, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(5): 978–994. https://doi.org/10.1109/tpami.2010.147

Luo, X.; Li, H.; Cao, D.; Dai, F.; Seo, J.; Lee, S. 2018. Recognizing diverse construction activities in site images via relevance networks of construction related objects detected by convolutional neural networks, *Journal of Computing in Civil Engineering* 32(3): 04018012. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000756

Marszalek, M.; Laptev, I.; Schmid, C. 2009. Actions in context, in *International Conference on Computer Vision and Pattern Recognition*, 2929–2936. https://doi.org/10.1109/CVPR.2009.5206557

Memarzadeh, M.; Golparvarfard, M.; Niebles, J. C. 2013. Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors, *Automation in Construction* 32: 24–37. https://doi.org/10.1016/j.autcon.2012.12.002

Navon, R.; Goldschmidt, E. 2010. Examination of worker – location measurement methods as a research tool for automated labor control, *Journal of Civil Engineering and Management* 16(2): 249–256. https://doi.org/10.3846/jcem.2010.29

Oliva, A.; Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision* 42(3): 145–175. https://doi.org/10.1023/A:1011139631724

Onofri, L.; Soda, P.; Pechenizkiy, M.; Iannello, G. 2016. A survey on using domain and contextual knowledge for human activity recognition in video streams, *Expert Systems with Applications* 63: 97–111. https://doi.org/10.1016/j.eswa.2016.06.011

Peddi, A.; Huan, L.; Bai, Y.; Kim, S. 2009. Development of human pose analyzing algorithms for the determination of construction productivity in real-time, in *Construction Research Congress*, 2009, ASCE, Seattle, WA, USA, 1: 1–20. https://doi.org/10.1061/41020(339)2

Pradhananga, N.; Teizer, J. 2013. Automatic spatiotemporal analysis of construction site equipment operations using GPS data, *Automation in Construction* 29: 107–122. https://doi.org/10.1016/j.autcon.2012.09.004

Rezazadeh Azar, E.; Mccabe, B. 2012. Part based model and spatial temporal reasoning to recognize hydraulic excavators in construction images and videos, *Automation in Construction* 24: 194–202. https://doi.org/10.1016/j.autcon.2012.03.003

Rezazadeh Azar, E.; Dickinson, S.; McCabe, B. 2012. Server-customer interaction tracker: computer vision-based system to estimate dirt-loading cycles, *Journal of Construction Engineering and Management* 139(7): 785–794. https://doi.org/10.1061/(asce)co.1943-7862.0000652

Russell, B. C.; Torralba, A.; Murphy, K.; Freeman, W. T. 2008. LabelMe: A database and web-based tool for image annotation, *International Journal of Computer Vision* 77: 157–173. https://doi.org/10.1007/s11263-007-0090-8

Seo, J.; Han, S.; Lee, S.; Kim, H. 2015. Computer vision techniques for construction safety and health monitoring, *Advanced Engineering Informatics* 29(2): 239–251. https://doi.org/10.1016/j.aei.2015.02.001

Tang, M.; Gorelick, L.; Veksler, O.; Boykov, Y. 2013. Grabcut in one cut, in *14th IEEE International Conference on Computer Vision*, 1769–1776. https://doi.org/10.1109/iccv.2013.222

Teizer, J. 2015. Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites, *Advanced Engineering Informatics* 29(2): 225–238. https://doi.org/10.1016/j.aei.2015.03.006

Ullah, M. M.; Parizi, S. N.; Laptev, I. 2010. Improving bag-of-features action recognition with non-local cues, in *Proceedings of the British Machine Vision Conference*, September 2010. BMVA Press, 95.1–95.11. https://doi.org/10.5244/c.24.95

Wang, H.; Klaser, A.; Schmid, C.; Liu, C. L. 2013. Dense trajectories and motion boundary descriptors for action recognition, *International Journal of Computer Vision* 103(1): 60–79. https://doi.org/10.1007/s11263-012-0594-8

Yang, J.; Arif, O.; Vela, P. A.; Teizer, J.; Shi, Z. 2010. Tracking multiple workers on construction sites using video cameras, *Advanced Engineering Informatics* 24(4): 428–434. https://doi.org/10.1016/j.aei.2010.06.008

Yang, J.; Vela, P.; Teizer, J.; Shi, Z. 2014. Vision-based tower crane tracking for understanding construction activity, *Journal of Computing in Civil Engineering* 28(1): 103–112. https://doi.org/10.1061/41182(416)32

Yang, J.; Park, M. W.; Vela, P. A.; Golparvar-Fard, M. 2015. Construction performance monitoring via still images, timelapse photos, and video streams: Now, tomorrow, and the future, *Advanced Engineering Informatics* 29: 211–224. https://doi.org/10.1016/j.aei.2015.01.011

Yang, J.; Shi, Z.; Wu, Z. 2016. Vision-based action recognition of construction workers using dense trajectories, *Advanced Engineering Informatics* 30(3): 327–336. https://doi.org/10.1016/j.aei.2016.04.009

Yao, B.; Fei-Fei, L. 2010a. Grouplet: A structured image representation for recognizing human and object interactions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9–16. https://doi.org/10.1109/cvpr.2010.5540234

Yao, B.; Fei-Fei, L. 2010b. Modeling mutual context of object and human pose in human-object interaction activities, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 17–24. https://doi.org/10.1109/cvpr.2010.5540235

Ziaeefard, M.; Bergevin, R. 2015. Semantic human activity recognition: A literature review, *Pattern Recognition* 48(8): 2329–2345. https://doi.org/10.1016/j.patcog.2015.03.006

Zou, J.; Kim, H. 2007. Using hue, saturation, and value color space for hydraulic excavator idle time analysis, *Journal of Computing in Civil Engineering* 21(4): 238–246. https://doi.org/10.1061/(asce)0887-3801(2007)21:4(238)