Joint Segmentation and Recognition of Worker Actions using Semi-Markov Models

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

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

Abstract –

Vision-based automated recognition of worker actions has gain lots of interest during the past few years. However, existing research all requires pre-segmented video clips, which is not applicable in the real situation. Furthermore, pre-segmented videos abandon the temporal information of action transition. A joint action segmentation and recognition method, which can segment continuous video stream while recognizing the action type for each segment, is an urgent need. In this paper, we model the worker actions with a discriminative semi-Markov model. In the model, a set of features is defined to capture both the local and global characteristics of each action cycle. Then the semi-Markov model is formulated as an optimization problem and solved by the cutting plane method for simultaneous action segmentation and recognition. Scale-Invariant Feature Transform (SIFT) is applied to detect feature points in the region of interest in every frame. Two descriptors (Histograms of Oriented Gradients – HOG, Histograms of Optical Flow – HOF), are computed in the feature points to encode the scenario and motion flow simultaneously. Finally, the Bag-of-Feature strategy is adopted for feature representation. Experimental results from real world construction videos show that the proposed method is able to segment and recognize continuous worker actions correctly, resulting in a prospecting application in automated productivity analysis.

Keywords –

Action Segmentation; Action Recognition; Worker; Semi-Markov Models

1 Introduction

Worker activities have strong impact on productivity, progress and quality of construction projects. Traditional activity analysis usually involves foremen collecting data through onsite observations, survey or interview, and analyzing data offline [1, 2]. This procedure is labor intensive, cost sensitive and can be prone to error. Lacking of real-time information is another major concern.

With the prevalence of onsite cameras and the emergence of advanced computer vision technologies, video based construction operation analysis has become a new trend in recent years [3, 4]. Either workers or equipment are detected and tracked through image sequences for further activity analysis [5-9]. Or spatial-temporal features are extracted directly from videos for action recognition [10-12]. The latter, for being robust to changing view angle and moving cameras, has attracted lots of attentions.

Gong et al. [10] adopted the Bag-of-Feaures pipeline to recognize worker and equipment activities. Spatio-temporal features were extracted, clustered and quantized for motion pattern learning and prediction. Golparvar-Fard et al. [11] also adopted similar strategy for action recognition of earthmoving equipment. Yang et al [12] established a bigger worker action dataset and achieved a state-of-art performance on the new dataset by using dense trajectories for feature description.

Though the above mentioned research has made remarkable achievement, their limitation is obvious. They are all based on pre-segmented video clips by assuming that each clip only contains a single action cycle. In real world application, analyzing a continuous video stream with repetitive actions or sequential actions is often expected. What is more, pre-segmented video clips abandon the temporal information of action transition, which is valuable for operation flow analysis.

Hence, an automated action segmentation and recognition method is needed for continuous operation analysis.

In computer vision field, traditional solution is to treat action segmentation and recognition separately. Temporal segmentation is applied to partition videos into coherent constituent parts and recognition could then be simply carried out as categorization of the action classes corresponding to the segments. Typically three types of methods can be applied for segmentation: boundary detection, sliding windows and compositional approaches [13].

However, it is often difficult to segment a video into actions purely based on low-level cues. What is more, separating segmentation from classification may result in important loss of information related to the actions. A recent trend is to solve segmentation and classification jointly [14, 15].

Hoai et al [14] proposed a discriminative temporal extension of the spatial bag-of-words model for joint segmentation and recognition of human actions. A multi-class SVM was applied for classification and segmentation inference was done with dynamic programming. Shi et al [15] presented a discriminative semi-Markov model to define features over boundary frames, segments and also neighboring segments. The inference problem of simultaneous segmentation and recognition was solved efficiently using optimization algorithms. Their method exhibited good performance on repetitive action segmentation and recognition.

Inspired by Shi et al [15], in this paper, we adopt the discriminative semi-Markov model to represent continuous worker actions. Scale-Invariant Feature Transform (SIFT) is applied for feature point detection. Two descriptors (Histograms of Oriented Gradients – HOG, Histograms of Optical Flow – HOF), are computed in the feature points to encode the scenario and motion flow simultaneously. Then the semi-Markov model is solved by the cutting plane method for simultaneous action segmentation and recognition.

The rest of the paper is organized as follows. Section 2 illustrates the algorithm in detail. Section 3 gives out the experimental results. Section 4 concludes the paper.

2 Methodology

The paper aims at designing a joint action segmentation and recognition system for worker activity analysis. Some basic assumptions are as follows.

First, the system input is continuous video stream captured from construction site either by mounted surveillance cameras or hand-held cameras. Hence no static background is required. What is more, the algorithm is expected to handle various points of view. Second, cyclic worker actions, such as bricking or nailing, are the major concern. That is to say, the action should have clear starting point and ending point.

The overall workflow of the proposed system is shown in Figure 1. As can been seen, the system flow is divided into two pipelines: training and testing. First, feature points in labeled video clips are extracted by SIFT detector. Then, HoG and HoF descriptors are computed on the feature points. Feature vectors are quantized using the Bag-of-Features strategy. After that, feature functions are formed according to their definitions. Semi-Markov Model is learned through optimization. During testing, feature functions are formed using the codebook generated in the training stage. Then the trained SMM model is applied to infer the continuous action boundaries and types. More details are illustrated as follows.

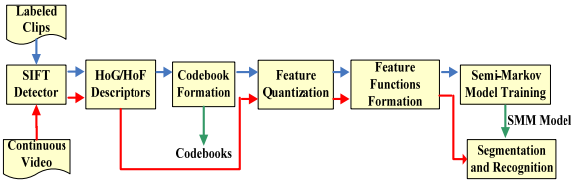


Fig 1. System Overview

2.1 Semi-Markov Model

The joint action segmentation and recognition problem is formulated as a convex optimization over a probabilistic semi-Markov model [15]. To make the paper self contained, a brief description of the model is provided as follows.

Considering a graph model, each node corresponds to a segment of video frames having the same action label, and each edge depicts the statistical dependency between adjacent segments. Let $C = \{1, \ldots c\}$ represent the action labels. Given a video sequence X of length m, assume there are l segments with boundaries $\{n_k\}_{k=0}^{m-1}$. Then action sequence label Y can be represented as $Y = \{(n_k, c_k)\}_{k=0}^{l-1}$, where each pair (n_k, c_k) denotes the starting position and the corresponding action label for the k th segment.

Model parameter is denoted as W. And $\Phi(X, Y)$ represents a feature map over the joint input-output space. Assume that the conditional probability distribution over action sequence label Y given current observation sequence is a log-linear model,

$$\log p(Y \mid X, W) = \langle W, \Phi(X, Y) \rangle - A_W(X)$$

Where $A_W(X)$ is a normalization constant to ensure

$p(Y \mid X, W)$ is a valid probability distribution. And $\Phi(X, Y)$ is defined as,

$$\Phi(X,Y) = (\sum_{i=0}^{l-1} \phi_1(X, n_i, c_i), \sum_{i=0}^{l-1} \phi_2(X, n_i, n_{i+1,} c_i), \sum_{i=0}^{l-1} \phi_3(X, n_i, n_{i+1}, c_i, c_{i+1}))$$

In which, ϕ_1 and ϕ_2 depicts the observation-label dependencies within the current action segment with ϕ_1 focusing on the segment's boundary frame and ϕ_2 describing the global characteristics of the current segment. While ϕ_3 encodes the interaction between two neighboring segments.

Given an unseen video sequence X , its action sequence can be labeled optimally by solving the following maximum likelihood decoding problem

$$Y^* = \arg\max_Y \log p(Y \mid X, W).$$

Learning is accomplished by solving a regularized optimization problem with respect to the parameter W . The goal is to make W be bounded to avoid over-fitting, while maximizing the minimum log ratio of the conditional probabilities,

$$\min_{W,\xi} \frac{\|W\|^2}{2} + \frac{\eta}{T} \sum_t \xi_t$$

$$s.t. \quad \langle W, \Delta\Phi(X_t, Y) \rangle \geq \Delta(Y_t, Y) - \xi_t \quad \forall t, Y$$

for the set of video sequences $\{t : t \in 1, \ldots, T\}$. In the equation, $\Delta\Phi(X_t, Y) := \Phi(X_t, Y_t) - \Phi(X_t - Y)$ and ξ_t is a non-negative slack variable to account for the non-separable case.

For the sake of completeness, the dual program is given as

$$\max_\alpha \sum_{t,Y} \alpha_{t,Y} \Delta(Y_t, Y) - \frac{\eta}{2} \left\| \sum_{t,Y} \alpha_{t,Y} \Delta\Phi(X_t, Y) \right\|^2$$

$$\text{s.t.} \quad \alpha_{t,Y} \in M \quad \forall t$$

Where M represents the probability simplex constraints.

The above problem can be solved approximately using optimization techniques such as cutting plane [16] or the bundle method [17].

2.2 Feature extraction and description

A frame based feature representation is expected to encode the feature map $\Phi(X, Y)$. Considering the need for robustness to illumination, scale and view angle change, a local feature detector SIFT [18] is adopted. After feature points being detected, technically the SIFT descriptor can be applied for feature description. However, considering that action segmentation and recognition is not a task based on a single frame but usually requiring dynamic information from neighboring frames, we adopt HoG (Histograms of Oriented Gradients) [19] and HoF (Histogram of Optical Flow) [20] descriptors instead. HoG is used to encode static appearance information, while HoF is to capture the local motion information.

Descriptor are computed on a 3D video patch in the neighborhood of each detected SIFT point. The patch is partition into a grid with 3x3x2 spatio-temporal blocks. 4 bins HoG descriptors and 5-bin HoF descriptors are then computed for all blocks and concatenated into 72-element and 96-element descriptors respectively, finally forming a 168-dimensinal feature vector. For dimension reduction, the Bag-of-Features strategy is applied. A 50-dimensinal codebook is formed by clustering on all features. Then each feature vector is mapped to codebook centers, resulting in a quantized histogram.

Finally, feature functions ϕ_1, ϕ_2, ϕ_3 are computed. ϕ_1 describes the boundary frame features.

$$\phi_1(X, n_i, c_i) = \psi_1(X, n_i) \otimes c_i$$

Where $\otimes$ denotes the tensor products. ψ_1 is the concatenation of a constant 1 and the histogram vector on the boundary frame.

ϕ_2 captures features on the current segment.

$$\phi_2(X, n_i, n_{i+1}, c_i) = \psi_2(X, n_i, n_{i+1}) \otimes c_i$$

Where $\psi_2(X, n_i, n_{i+1})$ contains three components: the length of the segment, the mean and the variance of the histogram vector of the segment (from frame n_i to $n_{i+1} - 1$).

ϕ_3 is used to depict features on neighboring segments.

$$\phi_3(X, n_i, n_{i+1}, c_i, c_{i+1}) = \psi_3(X, n_i, n_{i+1}) \otimes c_i \otimes c_{i+1}$$

It is the concatenation of four components: the mean of the histogram vector from frames n_i to $n_{i+1} - 1$, and from frames n_{i+1} to $n_{i+1} + d$; the variance of the histogram vector from frames n_i to $n_{i+1} - 1$, and from frames n_{i+1} to $n_{i+1} + d$. d is the minimum duration of a segment defined by the user.

3 Experimental results

This section describes experimental results on two groups of data: the repetitive actions and the sequential actions.

The original data is selected from the worker action dataset proposed by the authors previously [12]. Four types of worker actions are selected: 'Nailing', 'TieRebar', 'Shoveling' and 'LayBrick'. These actions are all cyclic with starting and finishing points for each cycle, offering a clear sign for segmentation. For each action type, we pick 3 workers. Each one performs the action 6 times. So there are totally 72 video clips (in resolution 320*240，frame rate 25 fps).

Figure 2 displays sample frames extracted from the dataset. As can be seen, various points of view are covered in the dataset. Furthermore, different workers may execute the same action differently. These all propose difficulties for action segmentation and recognition.

For algorithm evaluation, the segmentation performance and the recognition performance are measured separately. The algorithm will assign each frame a segment label and an action label. By comparing the segmentation and recognition results to the ground truth, the overall frame-level accuracy can be calculated as the ratio between the number of agreements over the total number of frames. Confusion matrix is computed to further evaluate the recognition performance.

Fig 2. Sample frames of the dataset (From top row to bottom: nailing, kntRod, shoveling and bricking, respectively)

3.1 Repetitive actions

Video clips from the same action type and the same worker are concatenated into longer video sequences. For each action type, two workers are used for training. One worker is left out for testing. Eventually, 8 video sequences are for training, 4 for testing. Each sequence contains 6 action clips.

The overall frame-level accuracy for action segmentation is 88.3%. The results for each sequence are shown in Fig. 3, in which different color represents different action type and black boundary of each bar gives the segmentation result. In each subfigure, 'Truth' represents the ground truth and 'Ours' represents the algorithm output. It can be seen that the number of segments is correct for all sequences. The segments boundaries are not very satisfying.

The recognition results are all correct for 'nailing' and 'bricking'. One segment in 'knitRod' is misclassified as 'nailing'. And all segments in 'shoveling' is misjudges as 'bricking'. The confusion matrix is shown in Fig. 4. The average recognition accuracy is 72.0%.

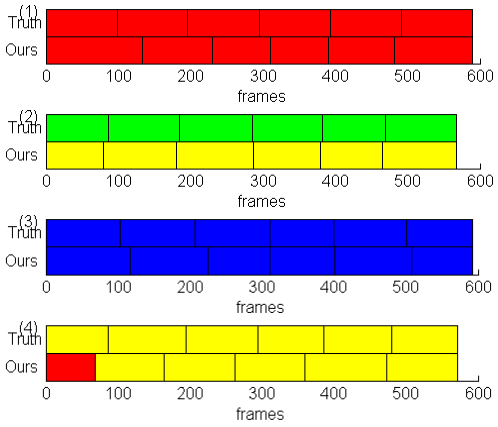


Fig.3 Automatic segmentation-recognition versus human labeled ground truth for repetitive actions. The segments are color coded; red, green, blue and yellow correspond to nailing, shoveling, bricking and knitRod classes, respectively. This figure is best seen in color.

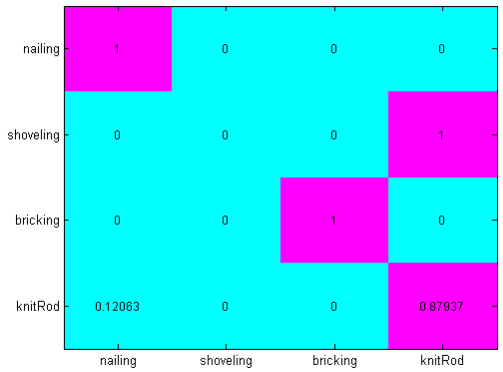
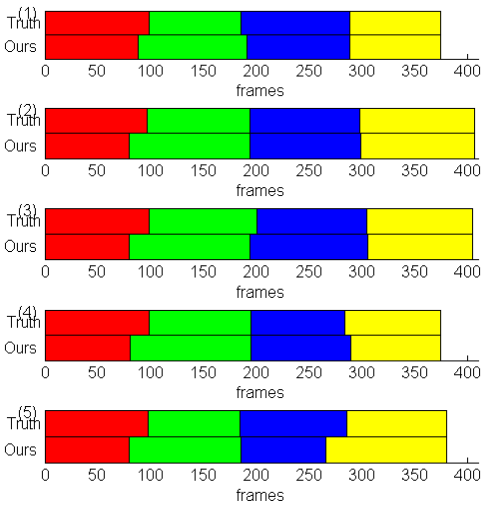


Fig. 4 Repetitive action - confusion matrix for action recognition in frame-level.

3.2 Sequential actions (Synthetic Dataset)

To simulate workers doing different task sequentially, video clips from different action types are concatenated into longer video sequences in the order of 'Nailing', 'Shoveling', 'LayBrick' and 'TieRebar', ending up 18 video sequences. 12 of them are training and the rest 6 are for testing.

The overall frame-level accuracy for action segmentation is 93.7%. The results for each sequence are shown in Fig. 5. It can be seen that the number of segments is correct for all sequences. The segments boundaries are close but not very accurate. Compared to the repetitive action result, the segmentation is slightly better, which is foreseeable because boundaries between two different actions are apparently easier to distinguish.



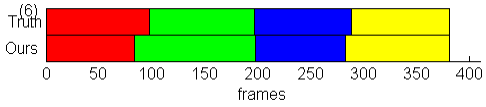


Fig. 5 Automatic segmentation-recognition versus human labeled ground truth for sequential actions. The segments are color coded; red, green, blue and yellow correspond to nailing, shoveling, bricking and knitRod classes, respectively. This figure is best seen in color.

All action type is recognized correctly per segments. The frame level confusion matrix for recognition is depicted in the confusion matrix in Fig. 6. The average accuracy is 93.7%.

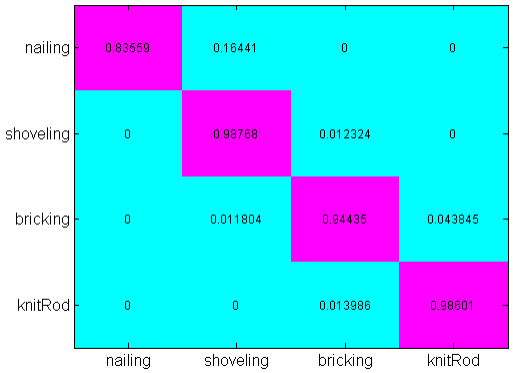


Fig. 6 Sequential action - confusion matrix for action recognition in frame-level.

3.3 Comments

From the two groups of experiments, we can see that the Semi-Markov Models method can segment continuous long construction videos and recognize worker action type for each segment. The action recognition rate is comparable to the state-of-art on pre-segmented videos [12]. The number of segmented are all correct in all testing sequences, which is important when applying the method to productivity analysis. However, the segmentation boundaries are not very accurate. One possible explanation is: worker actions are recorded from multiple points of view and different worker usually have different poses and various action length. These all introduce difficulties in segmentation.

Comparing the two groups of experiments, it can be concluded that the Semi-Markov Model can model the sequential action better than repetitive actions.

4 Conclusions

In this paper, we applied a Semi-Markov Model

method for joint action segmentation and recognition in construction videos. Experimental results showed that the method can recognize worker action type at a state-of-art accuracy while segmenting long videos into individual segments correctly in numbers. Joint action segmentation and recognition can be used for further productivity analysis and may supply useful information for workflow analysis. Future study may seek to test the methods on bigger datasets, improve the feature description method for better segmentation result.

References

[1] C. I. I. (CII) (Ed.), IR252.2a – Guide to Activity Analysis, Construction Industry Institute, Austin, TX, USA, 2010.

[2] M. C. Gouett, C. T. Haas, P. M. Goodrum, C. H. Caldas, Activity analysis for direct-work rate improvement in construction, Journal of Construction Engineering and Management 137 (2011) 1117–1124.

[3] 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, Advanced Engineering Informatics 29 (2015) 211–224.

[4] J. Teizer, Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites, Advanced Engineering Informatics 29 (2015) 225–238.

[5] M.-W. Park, C. Koch, I. Brilakis, Three-dimensional tracking of construction resources using an on-site camera system, Journal of Computing in Civil Engineering 26 (2012) 541–549.

[6] E. Rezazadeh Azar, S. Dickinson, B. McCabe, Server-customer interaction tracker: computer vision–based system to estimate dirt-loading cycles, Journal of Construction Engineering and Management 139 (2012) 785–794.

[7] J. Yang, P. Vela, J. Teizer, Z. Shi, Vision-based tower crane tracking for understanding construction activity, Journal of Computing in Civil Engineering 28 (2012) 103–112.

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

[9] M. Bugler, G. Ogunmakin, J. Teizer, P. A. Vela, A. Borrmann, 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), Cardiff, Wales.

[10] J. Gong, C. H. Caldas, C. Gordon, Learning and classifying actions of construction workers and equipment using bag-ofvideo-feature-words and bayesian network models, Advanced Engineering Informatics 25 (2011) 771–782.

[11] M. Golparvar-Fard, A. Heydarian, J. C. Niebles, Vision-based action recognition of earthmoving equipment using spatiotemporal features and support vector machine classifiers, Advanced Engineering Informatics 27 (2013) 652–663.

[12] J. Yang, Z. Shi, Z. Wu, Automatic recognition of construction worker activities using dense trajectories, in: Automation and Robotics in Construction and Mining, 32nd International Symposium on, pp. 75–81.

[13] Weinland D., Ronfard R. and Boyer E. A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition. Computer Vision and Image Understanding, 2011, 115(2): 224-241.

[14] Hoai M., Lan Z. and De la Torre F. Joint Segmentation and Classficiation of Human Actions in Video. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp 3265-3272.

[15] Shi Q., Cheng L., Wang L. And Smola A. Human Action Segmentation and Recognition using Discriminative Semi-Markov Model. International Journal of Computer Vision, 2011, 93: 22-32.

[16] Tsochantaridis I., Joachims T., Hofmann T., and Altun Y. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 6, 1453–1484.

[17] Smola A., Vishwanathan S., and Le Q. Bundle methods for machine learning. Advances in neural information processing systems, 2007, pp. 1377-1384.

[18] Lowe, D. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2), 91–110.

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

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