VISION-BASED WORKFACE ASSESSMENT USING DEPTH IMAGES FOR ACTIVITY ANALYSIS OF INTERIOR CONSTRUCTION OPERATIONS

BY

ARDALAN KHOSROWPOUR

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign 2013

Urbana, Illinois

Adviser:

Assistant Professor Mani Golparvar-Fard

ABSTRACT

Workface assessment –the process of determining the overall activity rates of onsite construction workers throughout a day– typically involves manual visual observations which are time-consuming and labor-intensive. To minimize subjectivity and the time required for conducting detailed assessments, and allowing managers to spend their time on the more important task of assessing and implementing improvements, we propose a new inexpensive vision-based method using RGB-D sensors that is applicable to interior construction operations. This is particularly a challenging task as construction activities have a large range of intra-class variability including varying sequences of body posture and time-spent on each individual activity. On the other hand, the state-of-the-art skeleton extraction algorithms from RGB-D sequences are not robust enough especially when workers interact with tools or self-occlude the camera's field-of-view. Existing vision-based methods are also rather limited as they can primarily classify "atomic" activities from RGB-D sequences involving one worker conducting a single activity.

To address these limitations, our proposed original method involves three main components: 1) an algorithm for detecting, tracking, and extracting body skeleton features from depth images; 2) A discriminative bag-of-poses activity classifier trained using multiple Support Vector Machines for classifying single visual activities from a given body skeleton sequence; and 3) a Hidden Markov model with a Kernel Density Estimation function to represent emission probabilities in form of a statistical distribution of single activity classifiers. For training and testing purposes, we also introduce a new dataset of eleven RGB-D sequences for interior drywall construction operations involving three actual construction workers conducting eight different activities in various interior locations. Our experimental results with an average accuracy of 76%

on the testing dataset show the promise of vision-based methods using RGB-D sequences for facilitating the activity analysis workface assessment.

ACKNOWLEDGMENT

I would like to express sincerest gratitude to my advisor Dr. Mani Golparvar-Fard for giving me the opportunity to be a part of this project and all the guidance and support he has given me along the way. I would never have been able to finish my thesis if it was not for his willingness to help and patience.

I would like to thank Igor Fedorov, Aleksander Holynski for their contribution to the development of atomic activity recognition model and also, Simin Liu for her contribution in data collection/labeling process. Furthermore, I would like to thank Poettker Construction for their co-operation and providing human and material resources for data collection.

This work is also partially supported by University of Illinois Department of Civil and Environmental Engineering's Innovation Grant and also "el Patrimonio Autonomo Fondo Nacional de Financiamiento para la Ciencia, la Tecnologia y la Innovacion, Francisco Jose De Caldas" under contract RC No. 0394-2012 with Universidad del Norte.

The last but not the least, I am most grateful for my family's support, guidance, and encouragement and also my beloved girlfriend, Melika, for her endless love, patience, and incredible support through the course of Master.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Several recent research studies have shown the feasibility of construction activity analysis and its positive correlation with improved direct-work rates [1-6]. According to the Construction Industry Institute (CII) [2,4], successful implementation of activity analysis involves two key steps: (1) continuous workface assessment, and (2) planning and implementing improvements. As the first step, workface assessment –the process of determining the overall activity rates of onsite construction workers throughout a day– involves an observer walking along randomly selected pre-defined routes, and characterizing the activity of each worker seen [2]. Nevertheless, visual observation at high level of confidence is constrained by the high cost associated with performing manual data collection, the risk of interfering the activities under observation, and the tendency to produce inaccurate data [7,8]. For example, obtaining 95% confidence in workface assessment considering above mentioned constraints requires a minimum of 5,100 random observations for a 10 hour working shift regardless of the worker population, activity type, or the job size [9]. Consequently to avoid the over-productiveness phenomenon caused by close surveillance and observation of workers – the Howthorne Effect— distance limit instructions for manual visual observations are proposed [9]. Initiating random routes and times, obeying standard distance limits to the workers, and instantaneous task-level data collection on entire job-site are some of the other key issues to be considered. Manual implementation of these tasks especially for several ongoing construction operations on a jobsite can significantly challenge frequent implementation of workface assessment which is the necessary step before improvement can be planned and implemented [2,4,10].

To address current limitations, a large body of research in the past few years have focused on methods that can automate the workface assessment. These methods range from application of sensors such as Ultra Wide Band (UWB) systems [11-13], Radio Frequency Identification (RFID) tags [14], and Global Positioning Systems (GPS) [15,16] to computer vision methods using video cameras [17-19]. Several existing methods that build on top of the non-visual sensors mainly track the locations of the workers. Without interpreting the activities of the workers and purely based on location information, deriving workface assessment data is challenging. For example for interior drywall construction activities, distinguishing between idle time, picking up a gypsum board, and measuring and cutting purely based on location data will be very difficult as during these activities the location of the worker would not change.

To address the limitations of location-based activity recognition, Joshua and Varghese [20-23] proposed an accelerometer-based method which has the capability of recognizing various activities based on movement of the body skeleton. Their method was tested for bricklaying operations at the task-level resolution and promising results have been reported. Using prior knowledge about activity locations on the jobsite, Cheng et al. [7] proposed an activity analysis method based on both location and body posture of the workers by integrating UWB – for location tracking– and commercially-available Physiological Status Monitors (PSM) with a wearable 3-axial thoracic accelerometer to derive body posture data. This method uses a single body posture and location to model and infer each activity. Still distinguishing between two activities have the same location and body pose for example idle time and measuring dimensions of a gypsum board would be challenging.

Our method is different from prior research, as we choose to use inexpensive RGB-D sensors (<$150) that can provide confidentiality in the data collection, and can detect and track body skeleton of up to six workers simultaneously and in real-time. Confidentiality here means that the identity of the workers remain unknown as we only track their body skeleton. To generalize the applicability of our method, we also do not assume any prior knowledge about expected activities in certain locations on the jobsite. Also rather than directly interpreting location and single body posture to derive activities as in [7], we propose histograms of body posture from RGB-D sequences to capture tabulated frequencies of a large number of key body postures for construction activities and use learning methods to train and infer these activities in a principled way. Our original method involves three main components: 1) an algorithm for detecting, tracking, and extracting body skeleton features from depth images captured using the RGB-D sensors; 2) a discriminative bag-of-poses activity classifier trained using multiple Support Vector Machines for classifying single visual activities from a given body skeleton sequence; and 3) a Hidden Markov Model (HMM) with Kernel Density Estimation (KDE) to represent emission probabilities in form of a statistical distribution of single activity classifiers. For training and testing purposes, we also introduce a new dataset of eleven RGB-D sequences for interior drywall construction operations involving three actual construction workers conducting eight different activities in various interior locations. Instead of manually collecting and analyzing workface data, the proposed method allows project managers to spend their time on correctly interpreting the results which is key to increasing productive activities in construction [6] and according to [24-27] requires more attention because conditions may differ from one project to another. In the following, we review the related work on vision-based methods.

CHAPTER 2

LITERATURE REVIEW

The advent of high-resolution video cameras, high storage databases, and availability of Internet over the past few years have transformed the ongoing construction operations' documentation methods. Today, it is common for owners and contractors to have web cameras continuously monitoring their onsite construction activities. Building on the state-of-the-art algorithms in computer vision and leveraging these existing web cameras, several recent methods have been proposed that focus on detecting construction workers and equipment [28-31], tracking their location in 2D and 3D [32,33], recognizing their activities based on their locations [34], or classifying atomic activities from videos containing a worker or equipment performing a single activity [35-37]. Teizer and Vela [18] reviewed and compared existing computer vision tracking methods based on RGB images and highlighted the challenges of workface interaction such as occlusion, visual clutter, and photometric visual variability in construction site. In addition to tracking worker location, activity recognition using video cameras has been a research area that has received attention over the past few years. To distinguish from construction activity analysis, by worker activity recognition, we mean detecting and documenting activities that are conducted by workers as part of the workface assessment process. Peddi et al. [17] proposed a method which classifies workers activity into three main categories—effective, ineffective, and contributory—based on a pose detection and tracking algorithms. Golparvar-Fard et al. [35] presented an algorithm which learns the distributions of spatio-temporal features and individual activity categories for earthmoving equipment using a multi-class Support Vector Machine (SVM) learning/inference model.

Over the last two to three years, the advent of depth sensors such as Microsoft Kinect, PrimeSense Carmine, and Time of Flight (TOF) have significantly facilitated "detection and tracking people". Because this sensor facilitates detecting, tracking body skeleton, and recognizing its pose in 3D – which has been a challenging component of video-based methods – recent studies in computer vision has focused on using depth maps for conducting activity recognition [38-42]. Similarly in the Architecture/Engineering/Construction (AEC) community, detecting and tracking skeleton and body pose estimation has quickly enabled research on vision-based methods for monitoring workers safety, health monitoring, and activity analysis [8,43-45]. Han et al. [43] studied consequent motion-analysis techniques to detect the unsafe activities of workers by transforming the motion data onto a three-dimensional space and learn a classifier to detect unsafe activities. Soumitry and Teizer [44] proposed a rule-based classification of worker activities of as ergonomic or non-ergonomic that can be beneficial for worker training, education, safety, and health. Escorcia et al. [8] proposed a discriminative learning/inference model for classifying activities conducted by single workers self-contained in short videos. As a first step, the proposed method classified single activities per RGB-D sequence containing one worker, however it did not model the variability in duration of single activities nor the frequency or sequence of their transitions from one activity to another.

Over the past few years, numerous studies in computer vision community have focused on activity recognition from various sequence lengths of RGB-D images (e.g., [38-42]). Sung et al. [40] proposed a method based on detecting and tracking body skeleton and representing them as histogram of gradient (HOG) features for generic human activity recognition from long sequences of RGB-D images. The structure of activities were learnt through hierarchical maximum entropy Markov model (MEMM) and dynamic programming. The RGB-D images used in their

experiments were mainly first-person views consisting of simple activities captured under controlled conditions; for example brushing teeth, drinking water, cooking, opening pill container, etc. Similar to the concept proposed by Yao and Fei-Fei [46], Koppula et al. [38] jointly represented human sub-activities and their object interactions based on their affordance. In a most recent work by Koppula and Saxena [47], spatio-temporal structure of activities has been modeled using conditional random field and as a result, the accuracy of activity recognition has been significantly improved compared to their previous work. These studies have all been conducted with data captured under controlled conditions with no occlusions. More precisely, for training/testing these methods, the depth sensors were placed in front of the human body which is more suitable for game/robotic applications as mostly a first-person view is needed in those cases. Unfortunately in construction activities, workface interaction and occlusions (both self-occlusions and those caused by construction objects) are the most dominant conditions of a scene. Also it is practically impossible to place the depth sensor in front of the worker. Placing the sensor behind or on the sides of the operations to document worker activities causes a larger range of intra-class variability in the body posture. Also these methods were mainly tested for simple activities where the duration of each activity was rather consistent. High intra-class variability in the duration of construction activities both among multiple workers and also for single worker spending different amount of time for repeating similar activity is yet another challenging factor in learning the structure of construction operations.

Despite the existence of various methods for tracking workers or recognizing their atomic activities, to our knowledge the problem of inferring a time-series of duration-variant activities for construction workface assessment from a sequence of depth images is not studied before. Thus, the main contribution of this article is the modeling of the worker activity recognition, recognizing

their durations, and transitions from one activity to another as an inference problem with a Viterbi algorithm and a Hidden Markov Model (HMM). We also introduce a new dataset of eleven RGB-D sequences for interior drywall construction operations involving three construction workers conducting eight different activities in various interior locations. The structure of the rest of the paper is as follows: In chapter 3, our approach to workface assessment is introduced. Next, the HMM and the multiple atomic activity classifiers used as part of the HMM to classify single activities from a given sequence of depth images are reviewed. Our new dataset and experimental setup are presented in chapter 4. The result of training and inferring construction activities from actual construction activity data collected from construction projects are presented in chapter 5. The perceived benefits and observed limitations of our method are discussed in chapter 6.

CHAPTER 3

METHOD

We propose a supervised machine learning based method for workface assessment from RGB-D sequences. For training our model, we collected ground-truthed labelled data from actual interior drywall construction operations. Our input is depth images from a Microsoft Kinect sensor, from which we extract body posture features that are fed into our learning/inference algorithm. To do so, we assume the sensor is setup on a tripod in an approximate distance of three and half meters so that the workers are within the field-of-view of the depth sensor. In our experiments, we use a PrimeSense User Tracker [48] to detect and track the body posture, recognize the human pose, and generate body skeletons based on single depth images. Using body posture features extracted from this tracker, we train a Hidden Markov Model (HMM) which captures spatio-temporal properties of construction activities and models the transition between activities over time. Figure 1 shows the setup of the sensor on a camera tripod both in lab and actual construction site settings.

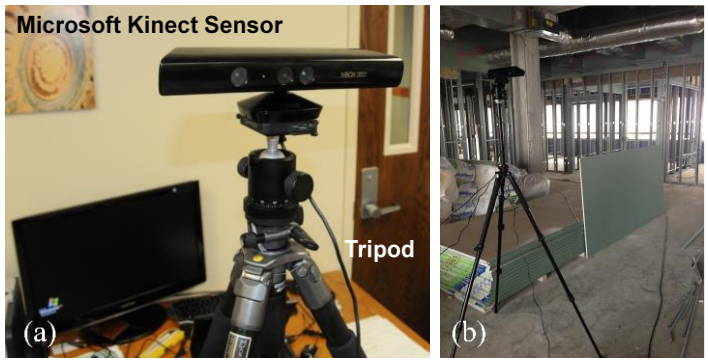


Figure 1: Microsoft Kinect sensor setup on a tripod in lab and actual construction site settings.

3.1. Body Pose Features

PrimeSense User Tracker estimates human pose and extracts the body skeleton from a single depth image through model fitting using dense correspondences between depth data and an articulated human model as a local optimization method [48]. Detected skeletons are tracked in real-time for up to six workers in the field-of-view as long as the workers have not left the scene for more than 10 seconds. Here, the skeleton is described by the length of the links and the joint angles. Specifically, we capture 3D Euclidean coordinates and the orientation of each joint with respect to the standard "T" skeleton pose shown in Figure 2.

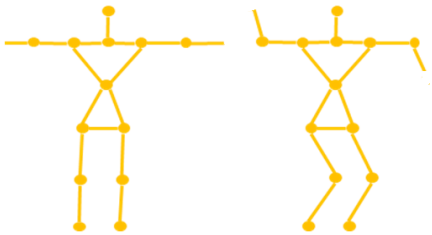
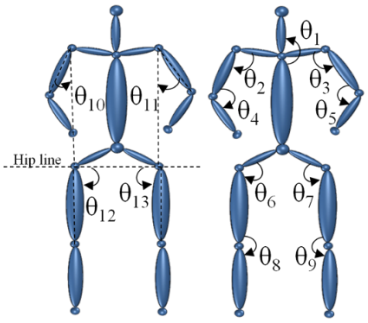


Figure 2: Left: the standard skeleton used for our model; Right: An example of the extracted body skeleton. In our model, the Euclidean coordinates and the orientations are calculated in reference to the standard skeleton.

From the 15 detected joints for the body skeleton, we track only thirteen joints removing the ones that represent the worker's hands. Our initial experiments show that the poor detectability of these joints creates noise in the dataset. Also these nodes do not play a critical role in defining distinct body postures for each construction activity. For each joint, an orientation matrix of 3×3 is documented. Instead of using the location and absolute orientation, we choose to use a single relative adjacent orientation per body joint. This reduces the size of the body pose descriptor to an overall 13-dimensional feature vector. Figure 3 shows these skeleton features.



Features	
θ_1	Angle between neck-head and neck-torso segments.
θ_2 ,θ_3	Angle between shoulder-neck and shoulder-elbow segments.
θ_4 ,θ_5	Angle between elbow-shoulder and elbow-hand segments.
θ_6 ,θ_7	Angle between hip-torso and hip-knee segments
θ_8 ,θ_9	Angle between knee-hip and knee-foot segments.
θ_{10} ,θ_{11}	Angle between arm and shoulder-hip line.
θ_{12} ,θ_{13}	Angle between thigh and hip line.

Figure 3: Body skeleton features which is extracted from the depth-images; the input to our proposed worker activity analysis algorithm.

Figure 4 shows the process of extracting the body skeleton from the depth-image which is fully automatic. Our representation of the body skeleton makes it invariant to the field-of-view, view angle, and anthropomorphic differences.

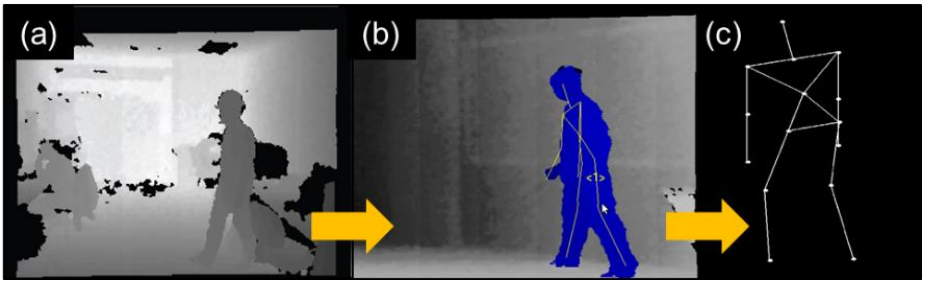


Figure 4: Extracting the worker body skeleton using a thirteen-joint body skeleton feature vector.

3.2. Model Formulation

Construction operations are complex and dynamic; nonetheless they could be represented in form of crew-balance charts; i.e., structured sequence of individual worker activities. Thus the learning algorithm should model the construction operations using body pose features considering their intra-class variability in worker activities. For example, drywall construction comprises a series of activities such as “picking up”, “holding”, “measuring”, “cutting” and “breaking” a

gypsum board to “walking”, “idling”, etc. From now on, we will call these sub-activities as “atomic activities”. Also there is no pre-determined sequence to these activities; i.e., there is no guarantee that a worker will “cut” a board right after “measurement” atomic activity. Furthermore the duration of these atomic activities vary among several workers. Even a single worker would spend different amount of time repeating the same atomic activity. These implies the following: (1) construction operations have a graph structure with their atomic activities and also (2) the order of their appearance and the duration of each cannot be exactly predicted in advance. Therefore for each operation C, we will represent a group of atomic activities ($c_i \in C$), and allow our method determine these activities, their duration, and also their sequence during modeling/inference.

3.2.1. Hidden Markov Model

There has been a great interest in modifying dynamic Bayesian models in order to ease the learning procedure or increase the model’s complexity [50-56,38] however, to model and infer a time-series of duration-variable atomic activities for a single detected worker from a sequence of RGB-D image, we construct a standard Hidden Markov Model (HMM). HMMs are dynamic Bayesian networks characterized by three main probabilities: prior probability, transition probability, and emission probability [57,58]. Assuming $X = \{x_1, x_2, \dots, x_T\}$ is a set of T states (hidden or unobserved), and $O = \{o_1, o_2, \dots, o_T\}$ is a set of K possible outcomes for T states (also known as emissions or observations) where the distribution of o_t only depends on x_t, and x_t only depends on x_{t-1}, we denote HMM as a 3-tuple $\lambda = \{A, B, \Pi\}$ where $\Pi = \{\pi_i\}$ is the vector of initial state probabilities with occurrence rate of each activity, and $A = \{a_{ij}\}$ is the state transition probability matrix that stores the probabilities of transitioning from state x_i to state x_j. The matrix

$B = \{b_i\ (k)\}$ stores the emission probabilities (e.g., the probability of observing outcome k from state $x_t = i$).

$$\Pi = \{\pi_i\} = P(x_1 = i) \quad for\ \ 1 \le i \le T \tag{1}$$

$$A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\ \}\ for\ 1 \le i, j \le T \tag{2}$$

$$B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\} \quad for\ 1 \le i \le T\ and\ \ for\ 1 \le k \le K \tag{3}$$

Since in our approach the knowledge of the ground-truth values of the hidden state variables are known at the learning stage, we can build discriminative classifiers per atomic activity and train the emission probabilities using the scores of these discriminative classifiers in a supervised fashion (discussed in section 3.2.2).

As shown in Figure 5, our HMM is as follows: The state space X consists of all depth frames captured from the depth sensor each of which corresponds to one atomic activity. Specifically, we assume each depth frame (t) can be labeled with an atomic activity (o_t) using our discriminative classifiers based on a given sequences of depth images right before and after the particular frame of interest:$[t - \delta, t + \delta]$ where δ is a small time step.

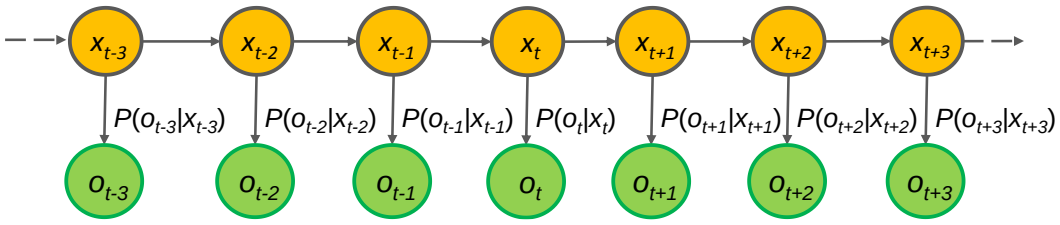


Figure 5: The HMM graphical representation where $X = \{x_1, x_2, \dots, x_T\}$ is a set of T states (hidden or unobserved), and $O = \{o_1, o_2, \dots, o_T\}$ is a set of K possible outcomes for T states (also known as emissions or observations) and the distribution of o_t only depends on x_t.

To construct the HMM, the K possible outcomes of observations $O = \{o_1, o_2, \dots, o_T\}$ and their transition from one to another must also be defined and learned through the ground-truth labelled data. In our model, as shown in Figure 6, the set of K possible outcomes at each depth frame is simply the set of worker atomic activities which are permitted for a given construction operation (e.g., Interior drywall operation). The taxonomy of visual activities for interior drywall operation is shown in Table 1. The final outcome of inferring construction activities using HMM will be a crew-balance chart – time-series of construction activities per depth frame considering that RGB-D sensors typically document a scene at the rate of 30 frames per second.

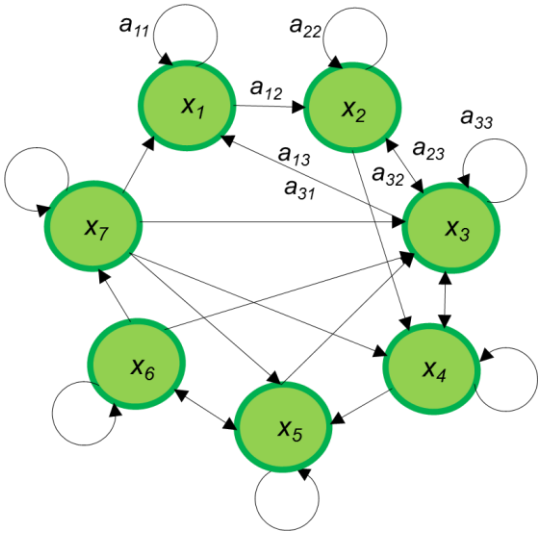


Figure 6: The state diagram for interior drywall activities shown in Table 1.

Table 1: Taxonomy of atomic worker activities within interior drywall construction operation.

Operation	Activity ID	Visual Activity
Interior Drywall Construction Operation	1	Picking up a "board"
	2	Holding a "board"
	3	Walking
	4	Putting Down a "board"
	5	Idling
	6	Measuring and Cutting a "board"
	7	Breaking a "board"

In HMM, the observations $O = \{o_1, o_2, \dots, o_T\}$ are typically drawn from continuous variable, and thus, emission probabilities can be modeled by a Probability Density Functions (PDF). The common practice for modeling emission probabilities is to use mixture of Gaussians, where B will be fully described by the weight, mean, and variance of all the Gaussian components [58]. Equation 4 shows the probability density function of a random variable, $P(\theta)$, as:

$$p(\theta) = \sum_{\tau=1}^{\Gamma} \alpha_\tau G(\tau, \mu_\tau, \sigma_\tau^2) \tag{4}$$

where $G(.)$ is the Gaussian function and Γ is the number of Gaussian components. Nonetheless, several research works [58] show that mixture of Gaussians has limitations 1) when they are used to model a PDF with more number of modes than the components, and 2) when PDF has uniform regions. To address these limitations, we use Kernel Density Estimation (KDE). KDE is a data driven, non-parametric approach to probability mass function calculation which is used to model the PDF with a Gaussian Kernel which is described in the following:

3.2.2. Kernel Density Estimation Function

Given a set of possible scores $S_a^i = \{s_{1,a}^i, s_{2,a}^i, \dots, s_{m,a}^i\}$ from a discriminative classifier for a specific atomic activity a_i over a dataset with m instances, emission probability is the likelihood

of a new predicted score s'_i belonging to a certain category S_n with respect to estimated probability density function $P(S_n)$. Equation 5 represents a KDE function with a Gaussian kernel:

$$P_{KDE}(S_n) = \frac{1}{N}\sum_{i=1}^{m} G(s_n; s_{i,n}, \sigma^2) \tag{5}$$

In a Gaussian KDE model, each component is located on a sample and the only parameter which is required to be estimated is the variance σ^2. Figure 7 illustrates a one dimensional Gaussian KDE for all seven atomic activities considered for interior drywall construction operations.

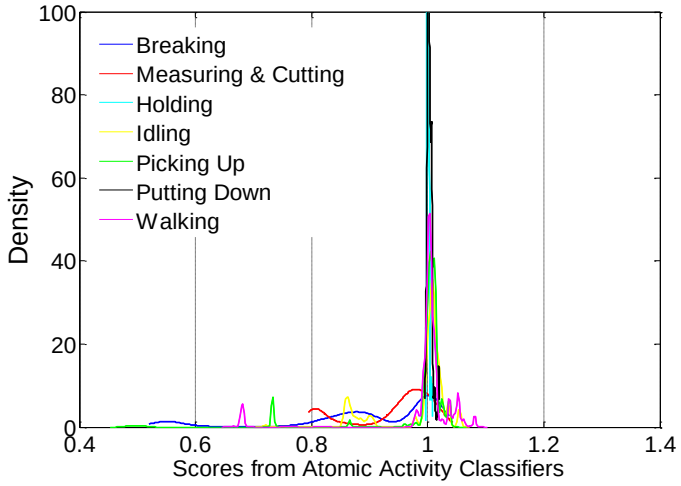


Figure 7: Gaussian kernel density estimation for all seven atomic activity categories in interior drywall construction operation.

Here, the KDE is used to model the classification scores obtained from the discriminative classifier of each atomic activity. Because the duration for each atomic activity varies in a given long sequence of RGB-D images, we need to model the score of our atomic activity classifiers for a continuous range of durations. To simplify this, for each atomic activity (c_i), we model a normal distribution for its duration $< \mu_k, \sigma_k >$. To consider a 95% rate in confidence in modeling activity durations, we choose seven to nineteen uniformly distributed time duration steps within

$[\mu_k - 2\sigma_k, \mu_k + 2\sigma_k]$ range and for each we learn the scores from the atomic activity classifier (see Figure 8). Our initial experiments showed seventeen time durations would yield the best performance and thus we decided to choose that for our experiments. In the following section, we introduce these atomic activity classifiers.

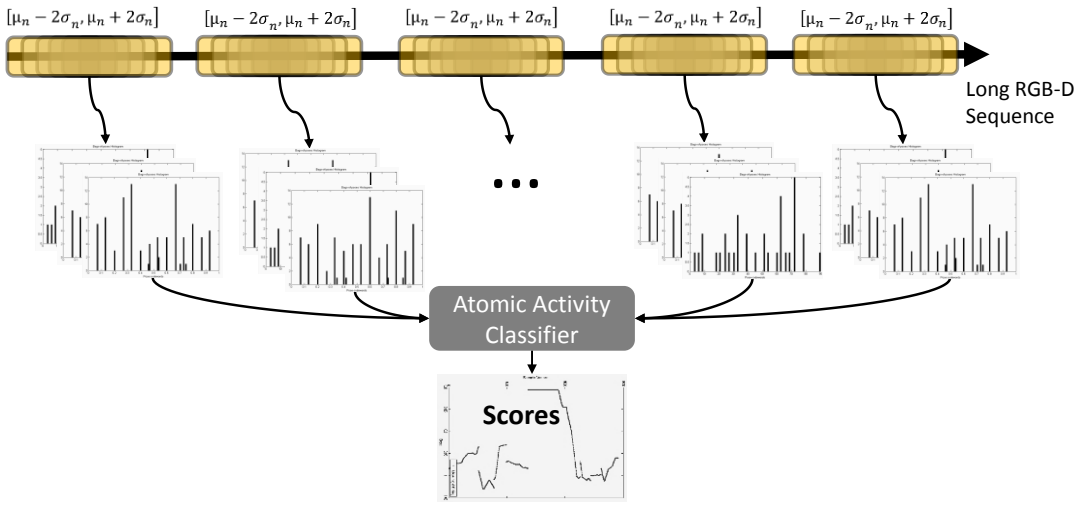


Figure 8: During the training stage, we select 7-19 time steps per frame and train the KDE for emission probabilities using scores documented for all inferences made with the multi-class atomic activity classifiers.

3.3. Atomic Activity Classification

We train a discriminative learning model for inferring and classifying atomic construction activity from a given sequence of RGB-D images $[t - \delta, t + \delta]$ at frame (t). To statistically model the sequential pattern of body skeleton features which provides a well-descriptive set of features for activity recognition and infer best activity for new sequences of body posture, we use a Bag-of-Pose model. Our proposed Bag-of-Pose model is shown in Figure 9 and is as follows:

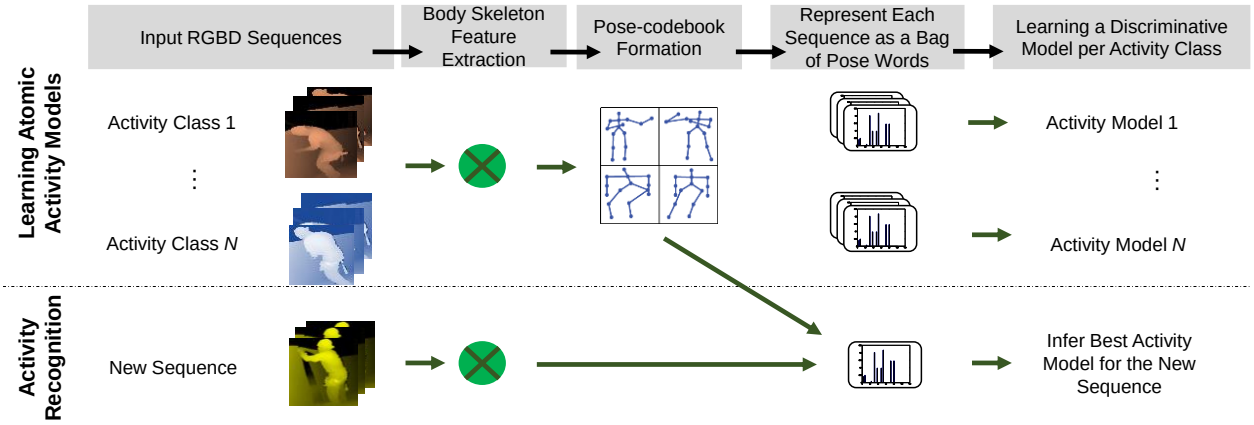


Figure 9: Our Bag-of-Pose model (both learning and inference) for multi-class atomic activity recognition from input RGBD sequence containing one worker conducting a single activity.

3.3.1. Bag of Pose Codewords for Representing Each RGB-D Sequence

At the learning stage using the ground-truthed data, we generate a collection of short RGB-D sequence with varying durations for each atomic activity class, each sequence containing one worker conducting a single atomic activity. Using the k-means clustering algorithm and the Euclidean distance as the clustering metric, the body skeleton feature vectors of the entire training dataset are clustered into a set of pose code words. Here, a similarity measurement between each frame's extracted pose and all codebook elements was done using various distance functions (L_1—distance, L_2 —distance, χ^2 significance). We chose L_2—distance function to assign the most likely pose codeword to each individual frame. The result of this process is a codebook that associates a unique cluster membership with each detected body skeleton pose. Hence, each RGB-D sequence is represented as a statistical distribution of body skeleton postures belonging to different key pose code words. For example, in "measuring" and "cutting" atomic activities there are multiple possible bending and standing poses in which a worker would be able to cut and measure a gypsum board. Rather than capturing a single pose for each of these atomic activities or

using all possible poses, we find the frequency of the most dominant body postures –from the pose codebook– that could represent these activities. Figure 10 illustrates the codebook formation process.

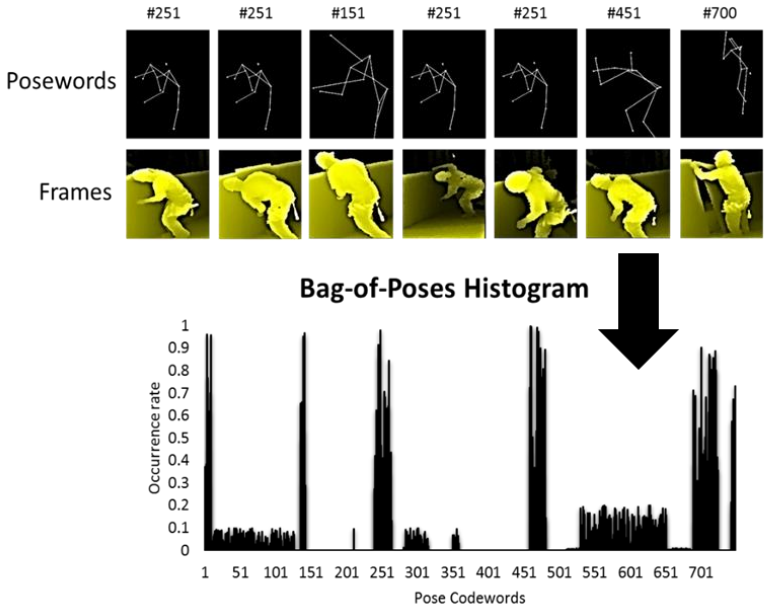


Figure 10: An example of a pose codebook generated for a sequence of body skeleton poses representing an atomic activity of a worker.

A total of 750 cluster centers are considered for the best action recognition performance. A grid-based search with various number of cluster centers from 50 to 800 with increment size of 50 was executed to find the optimum number of cluster centers which can creates the best representative pose codebook for our dataset. The effect of the codebook size (the number of pose code words) on the accuracy of our atomic activity classifiers is explored in chapter 5.

3.3.2. Learning and Inferring Atomic Activities

To learn a specific model for each atomic activity category, a multi-class one-vs-all Support Vector Machine (SVM) classifier is introduced. The SVM is a discriminative machine

learning algorithm which is based on the structural risk minimization induction principle [62]. Thus, for every atomic activity class $c_n \in C$, we train a binary SVM classifiers using given training data $\{p_i, q_i\}$ where $p_i \in P_{train}$ and possible labels are $q_i = \{+1, -1\}$. The presence of noise and occlusions, which is typical in construction site video streams, produces outliers in the SVM classifiers. Thus, we introduce slack variables χ_i which fits the desired hyperplanes with a soft margin approach. Consequently the SVM optimization problem can be written as:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to: } y_i\left(w.x_i + b\right) \geq 1 - \xi_i \quad \text{for } i = 1,\ldots,N \\ &\qquad\qquad \xi_i \geq 0 \text{ for } i = 1,\ldots,N \end{aligned} \tag{6}$$

In Equation 6, C represents a penalty constant which controls the trade-off between the margin size and the training error. We split our dataset into a disjoint training and testing ($P_{train} \cap P_{test} = \emptyset$) sets with 80% and 20% shares respectively. During the testing phase, we predict the scores $\{s_{i1}, s_{i2}, \ldots, s_{in}\}$ for each testing example $p'_i \in P_{test}$ with respect to trained hypothesis for all n existing actions $H = \{h_1, h_2, \ldots, h_n\}$. Consequently, to predict the most proper class $q'_i = \{1,2, \ldots, n\}$ for each testing example p'_i we look on the maximum score obtained from all classifiers $q'_i = Max\{s_{i1}, s_{i2}, \ldots, s_{in}\}$. Since no prior knowledge exists on separability condition of our dataset, implementing a kernel to map the data into a higher dimensional feature space can improve the classification performance. Thus, linear Gaussian Radial Basis Function (RBF), histograms intersection, and χ^2 kernels are pre-computed to evaluate the classification performance by mapping our dataset into different high dimensional spaces (see Table 2). Barla et al. [63] states that for histogram shaped data, the histogram intersection kernel usually outperforms RBF and Polynomial kernels. The histogram intersection kernel penalizes error of the sort where

one histogram has a zero in a bin where the other histogram is non-zero. This is an ideal behavior for our model as if a given pose codebook histogram has zero poses in a particular bin, it is unlikely to correspond to an atomic activity which should contain a non-zero entry in that bin. Regardless, we validates our choice of SVM kernel on the accuracy of atomic activity classifier in chapter 5.

Table 2: Various kernel functions for support vector machine classifier.

Kernel	Function
Histograms Intersection	$\sum_{i=1}^{m} \min\{d_i, d_j\}$
Gaussian Radial Basis Function	$e^{\left(-\frac{\|d_i - d_j\|^2}{2\sigma^2}\right)}$
χ^2	$1 - \sum_{i=1}^{m} \frac{2(d_i - d_j)^2}{d_i + d_j}$

3.4. Learning and Inferring HMM with KDE of Atomic Activity Classification Scores

The actual observation sequence (i.e., time-series of atomic activities) $X = \{x_1, x_2, \dots, x_T\}$ is a manifestation of some state sequence $O = \{o_1, o_2, \dots, o_T\}$ through the emission probability density function B. To infer the HMM model, the Viterbi method [49,59] which is a recursive dynamic programming algorithm is used to find the optimal and most likely sequence of the states. To find the best sequence $X = \{X_1, X_2, \dots, X_T\}$ for the given random observation $O = \{O_1, O_2, \dots, O_T\}$ we have [60,61]:

$$\delta_t(i) = \max_{X_1, X_2, \dots, X_{t-1}} P[x_1\, x_2 \dots\, x_t(= i), o_1 o_2 \dots o_t | (\pi,\, A,\, B)] \tag{7}$$

where $\delta_t(i)$ is the highest score for a single path at time t based on previous $t-1$ observations and the states sequence ends in state X_i. Induction in our formulation is:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \; . b_j(o_{t+1}) \tag{8}$$

The state and transition probabilities needed for this HMM model can be easily trained using ground-truthed data. For training the emission probability density function, we use our atomic activity recognition classifier on the short sequence data with ground-truth labels. For each activity class, we modeled a normal distribution for the durations $< \mu_k, \sigma_k >$ and stored classification scores for the range of possible durations for each atomic activity $[\mu_n - 2\sigma_n, \mu_n + 2\sigma_n]$ which guarantees a 95% rate of confidence. At the inference stage, for each frame, we breaking down the 4σ range to a set of various durations per action into seventeen steps. Among all steps, we only store the maximum classification score and use that as our observation variable. This strategy allows our model to cope with variable-duration activities.

CHAPTER 4

EXPERIMENTAL SETUP

4.1. Data Collection

Due to the lack of databases for training visual atomic activities of interior drywall construction operation, before testing our algorithm, it was necessary to create a comprehensive benchmarking RGB-D sequence dataset. The target operation consists of 7 activities: picking up, holding, walking, putting down, measuring and cutting, breaking the gypsum board, and idling. A Microsoft Kinect sensor is used to record data on an actual construction site in Central Illinois which is currently under construction and also in Newmark Civil Engineering Laboratory where same activities conducted by professional workers from the actual jobsite were replicates in the lab. The depth sensor was set with various angles and views with respect to gypsum board and worker position. Since in drywall operations, workers face the materials, there is always a minimal chance to capture data from a front view. Therefore, the only remaining alternatives which could provide a semi-clear view and informative data was to locate the sensor in various locations within a range of 45° from the sides as illustrated in Figure 11.

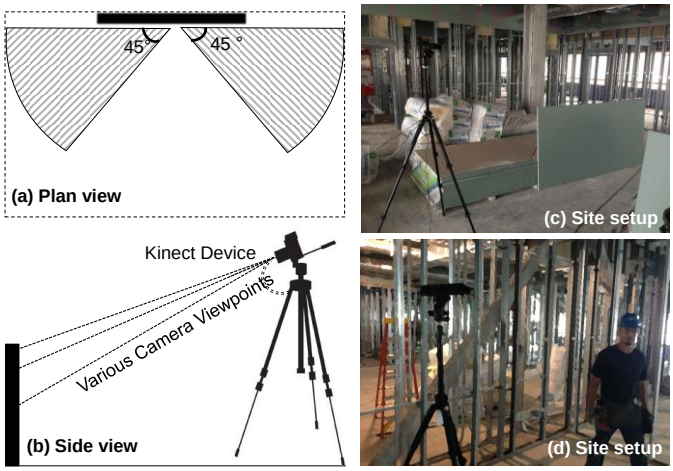


Figure 11: The Sensor setup and the location of the data collection on the actual site.

To ensure anthropomorphic invariance in training the atomic action classifier, we collected more than 300,000 depth images (>15min) from three workers with different body shapes and heights. Moreover, eleven long sequence videos were also collected and used for training and testing. All these 300,000 frames have been manually labeled and cross validated to maintain an accurate and unprecedented dataset. The most important segment of labelling process is labelling the transitional frames between different actions; hence, we addressed this challenge by defining a certain visual taxonomy for transitions to facilitate the labeling process. The key component in labeling is consistency in transitions; as long as we stay consistent, our algorithm would find the most probable sequence and correctly predict the transitions. The RGB-D sequence dataset is made public at: http://raamac.cee.illinois.edu/activityanalysis.

4.2. Performance Evaluation Measures

To quantify and benchmark the performance of the action recognition algorithm, we plot the Precision-Recall, ROC curves and study the Confusion Matrix. These metrics are extensively used in the Computer Vision and Information Retrieval communities as set-based measures. In the context of worker activity recognition, we define each as follows:

4.2.1. Precision-Recall Curve

To compare the overall average performance of the variations of the proposed activity recognition algorithms over a particular datasets, individual activity class precision values are interpolated to a set of standard recall levels (0 to 1 in variable increments of < 1). Here, precision is the fraction of retrieved activity instances that is classified as positive and is truly positive, while

recall is the fraction of positive examples that are correctly labeled [64]. Thus, precision and recall are calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

where in TP is the number of True Positives, FN is the number of False Negatives and FP is the number of False Positives. For instance, if a breaking atomic activity sequence is correctly recognized under the breaking class, it will be a TP; if a holding atomic activity sequence is incorrectly recognized as breaking, it will be a FP for the breaking class. When a breaking video is not recognized under the breaking class, then the instance is a FN. The particular rule used to interpolate precision at recall level i is to use the maximum precision obtained from the action class for any recall level great than or equal to i. For each recall level, the precision is calculated, and then the values are connected and plotted in form of a curve.

4.2.2. Confusion Matrix

The performance of the activity classifiers are analyzed using confusion matrix. This matrix returns the average accuracy per activity class using the following formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

A confusion matrix shows for each pair of atomic activity classes <c_1, c_2>, how many frames of activity c_1 were incorrectly assigned to c_2. Each column of the confusion matrices represents the

predicted atomic activity class and each row represents the actual atomic activity class. The detected TPs and FPs are compared and the percentage of the correctly predicted classes with respect to the actual atomic activity class is created and represented in each row.

CHAPTER 5

RESULTS AND DISCUSSIONS

To evaluate the performance of our algorithm, we separately conducted numerous experiments both on atomic activity recognition and HMM to find the optimum parameters and highest possible accuracies. For training/inferring atomic activities using SVM and using HMM, we used the LIBSVM [62] and HMM/KDE [65,66] open source packages. The results are reported based on 5-fold cross validated evaluations and various validation metrics mentioned above are used to discuss the algorithm performance in addition to address the advantages, limitations and potential improvement opportunities for the method:

5.1. Performance of the Atomic Activity Recognition Classifiers

There are two main parameters in atomic action recognition which should be optimized to obtain the best result; number of cluster centers in pose codebook generation and the kernel used in SVM to maintain a linearly separable dataset. In order to find the best cluster center number, a grid-based search is executed on a wide range of values from 200 centers to 800 centers with a 50 centers increment $[200:50:800]$. Cluster centers are the pose codewords that represent the entire poses in our dataset; the more the number of cluster centers, the more expressive our codebook will be however, after exceeding a certain number of clusters, the algorithm become bias to the target function and the accuracy does not increase anymore. Figure 12 illustrates obtained accuracies for various number of pose code words using histogram intersection kernel for the one-vs.-all SVM classifiers.

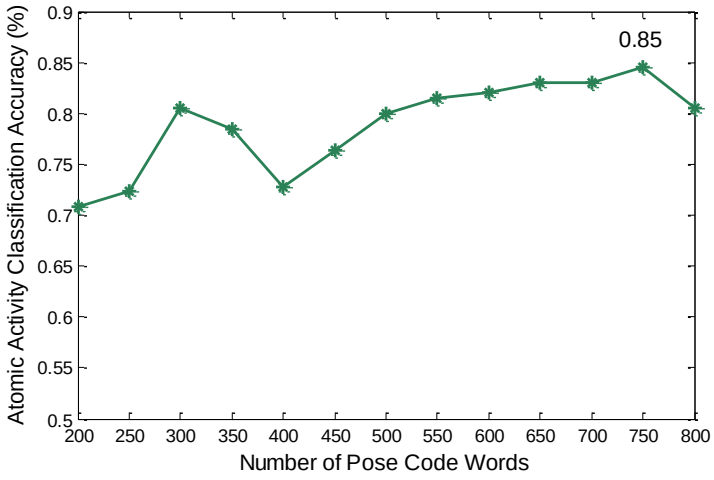


Figure 12: The atomic activity classification accuracy vs. number of pose code words. The performance was tested for one-vs.-all classifies with histogram intersection kernel.

As depicted in Figure 12, two peaks can be identified in the graph at 300 and 750 clusters corresponding to respective accuracies of 80.51% and 84.6%. Considering the first peak as a local maxima and the second one as the global maxima; there is a considerable tradeoff in terms of accuracy versus computational cost. The greater the number of clusters, the more expensive computational cost will be, and for roughly 4% increase in the accuracy, the computational time doubles. As a proof-of-concept for applicability of vision-based algorithms for activity analysis and considering that the accuracy is the first priority, we chose 750 cluster centers.

The next parameter to be set is SVM's optimum kernel function transforming our feature space to a linearly separable space. To do so, four kernel functions including, linear, Gaussian RBF, χ^2, and histogram intersection were tested with 750 cluster centers. Obtained accuracies for abovementioned kernel functions are listed in Table 3. Based on our evaluation, the histogram intersection kernel had the best performance.

Table 3: One-vs.-All Atomic Activity Classification Accuracies.

Kernel	Classification Accuracy
Histogram Intersection	% 84.6
χ^2	% 84.1
Gaussian RBF	% 81.5
Linear	% 78.46

Based on selected parameters in previous steps (histogram intersection kernel function and 750 cluster centers), we trained and tested a SVM classifier on mutually exclusive training and testing datasets. Figure 13 shows the results in precision-recall and ROC curve format and Figure 14 shows the average accuracies in the confusion matrix. As shown in Figure 14, accuracies are evenly distributed among all actions and are above 71% for all categories.

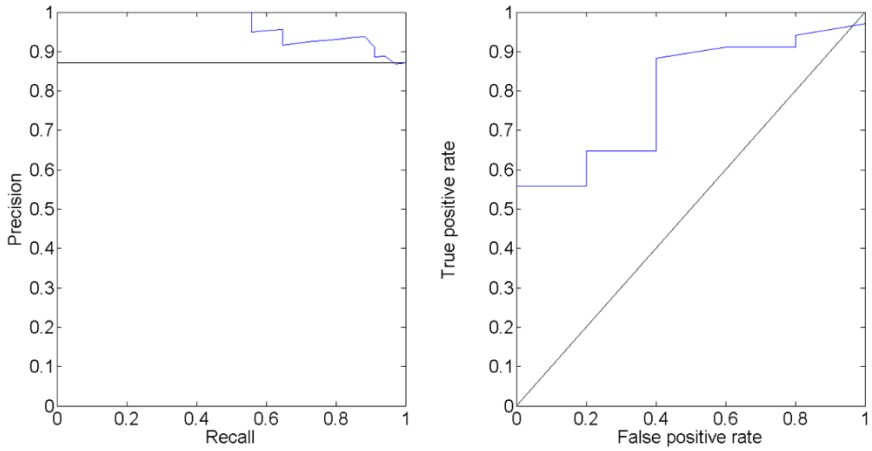


Figure 13: Precision-Recall and ROC curve for atomic action recognition with 750 cluster centers and histogram intersection kernel.

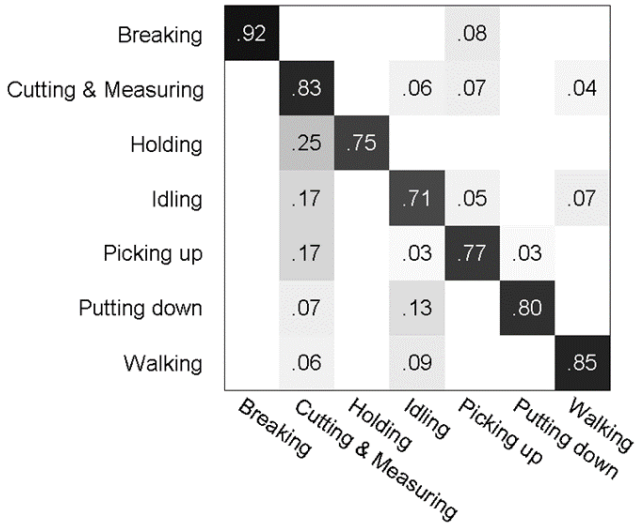


Figure 14: Confusion matrix illustrating average atomic activity recognition performance with 750 pose code words and SVM - histogram intersection kernel.

Two important patterns should be considered according to the confusion matrix, first, the lowest accuracies belong to "Holding" and "Idling" categories and second, the highest percentage of confusions are between various actions and "Cutting & Measuring" category. One of the probable roots which results in a low accuracy for "Idling" category seems to be the variety of possible poses. For example, there are idling activities in our dataset in which workers bend and use their knees as a support which shapes a quite similar pose to cutting and measuring's. "Holding" classifier also demonstrates a lower performance compared to other categories; according to a thorough investigation by authors, after reviewing all labeled videos complemented with simultaneous skeleton stream, the most contributing reasons is an inaccurate skeleton extraction due to a false depth map segmentation. As mentioned before, PrimeSense User Tracker package [48] relies on a segmentation algorithm to find human body boundary, however, proximity of a worker to any large scale object such as gypsum boards makes a significant error in segmentation process and consequently skeleton prediction. Figure 15 illustrates few examples of false calibration, segmentation, and skeleton detection. Furthermore, it is essential to point out the

reason why “Cutting” and “Measuring” activities have being categorized as a single class. The main reasoning behind this combined approach has roots in extreme similarity of the worker body poses during these activities, in addition to the weak performance of the skeleton extraction algorithm which is due to high proximity of workers to the gypsum boards. The body posture similarity causes confusion and increases the rate of misclassification at atomic activity recognition. This in turn initiates a deviation from the target function in HMM algorithm as well.

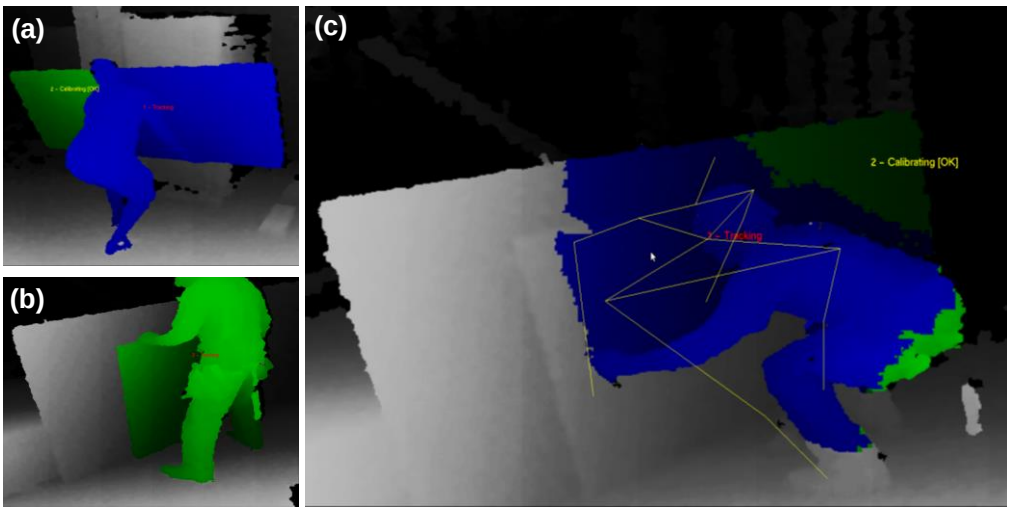


Figure 15: False calibration, segmentation, and skeleton detection due to proximity of workers and materials.

Figure 16 depicts the overall performance of atomic activity classifier with the number of score intervals that each interval contains equal samples.

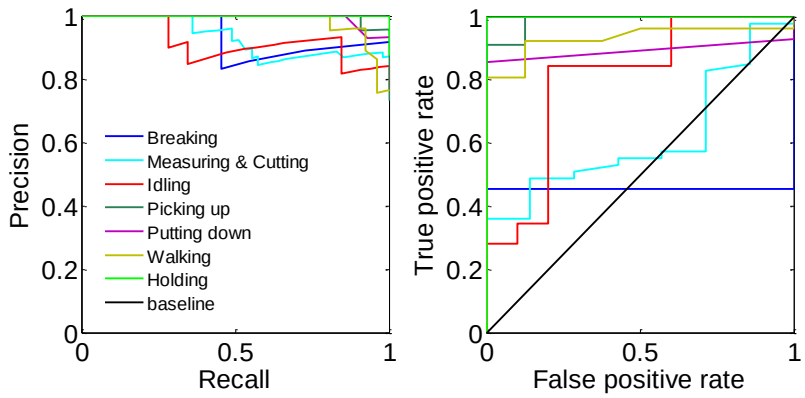


Figure 16: Precision-Recall curve and ROC on atomic activity classifiers.

As demonstrated in Figure 16, there are three categories (Breaking, Idling, and Measuring & Cutting) which tend to show a low precision when the threshold has been set high to maintain a low rate of FP. This type of behavior is not favorable as we are willing to have a high precision even when the rate of recall is low. Moreover, the graphs' slopes in ROC curve are extremely gradual in Breaking and Measuring & Cutting activity categories which describes a bias behavior toward threshold variance; i.e., the predicted scores are distributed densely in two separate regions and there is a considerable gap in between. Nevertheless, Figure 7 shows a distinctly lower score density for Breaking and Measuring & Cutting categories compared to other classes. In the next section, one dimensional KDEs of each category will be displayed and discussed further.

5.2. Long Sequence Action Recognition Results and Discussions

The performance of our long sequence activity recognition algorithm is mainly dependent on three components: KDE, windows length resolution- uniformly distributed time duration steps within $[\mu_k - 2\sigma_k, \mu_k + 2\sigma_k]$-, and HMM. As the first step, estimated probability densities executed by atomic activity recognition algorithm will be displayed using KDE function for all the categories. Afterwards, we present HMM performance with respect to variable windows length

resolutions for bag-of-poses histogram. And finally, HMM performance will be discussed by means of confusion matrix and sequential predicted actions versus ground truth. The atomic activity classifier infers each activity by generating bag-of-poses histogram based on various lengths in the range of 4σ and uses the maximum obtained scores to estimate probability density by KDE for each individual class. Figure 17 illustrates the estimated probability density of each category based on maximum scores obtained by atomic activity classifier. Expectedly, all distributions have a higher density in the range of $S \geq 1$ however, for some categories such as: Measuring & Cutting, Breaking, and Idling local maximums could be observed at a relatively lower score in addition to the global maxima in $S \geq 1$ range. This addresses the precision-recall and ROC curves behavior for these categories as discussed in section 5.1. The red lines in each graph represent classification score of 1 and are used as a datum to interpret the quality of score distribution.

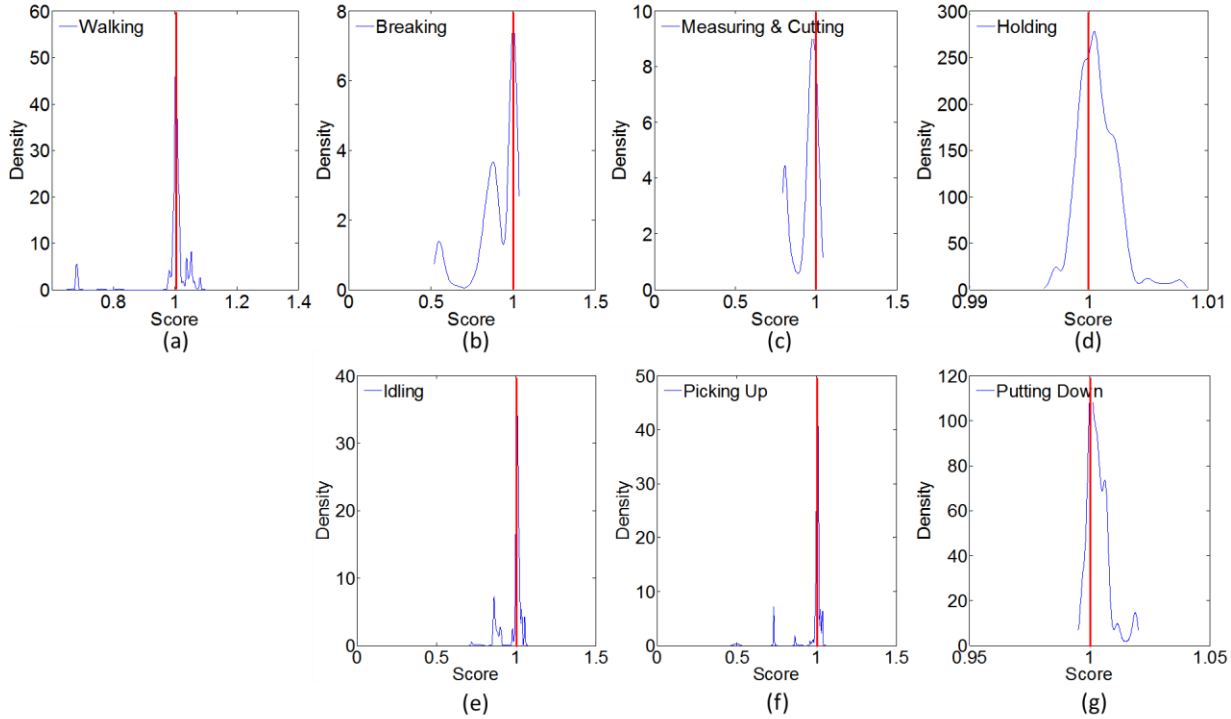


Figure 17: Kernel Density Function classification scores for all atomic activity categories.

The next parameter is the resolution of windows that create the bag-of-poses histogram in order to find the best duration representing each action through along sequence videos. We have chosen a resolution range of $[\frac{4\sigma}{7}, \frac{4\sigma}{9}, \ldots, \frac{4\sigma}{19}]$ to evaluate our HMM performance based on a 5-fold cross validation. Figure 18 depicts the correspondence of HMM accuracy and windows resolution. An overall increasing pattern in HMM accuracy by increasing windows resolution could be observed. However, this graph is not monotonically increasing and at some points the accuracy drops opposed the expectation. Since the differences are not high, this behavior could be potentially due to the cross validation approach where all training and testing examples are chosen randomly. Based on Figure 18, we have selected a $\frac{4\sigma}{17}$ resolution for the rest of our experiments.

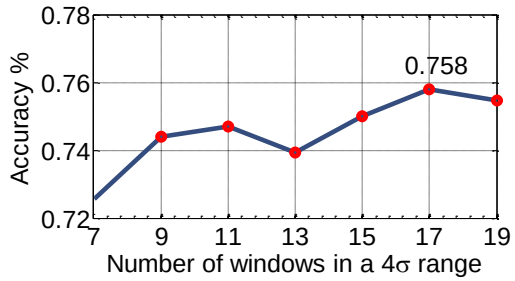


Figure 18. HMM performance with respect to various windows resolutions.

After all parameters set to their optimum values, we evaluate the entire algorithm by analyzing long sequences of images and comparing the results to the ground-truthed data. Hence, a 5-fold cross validation on a dataset of 11 long sequence videos is performed and an average accuracy of 76% with the maximum accuracy of 91% are achieved. Figure 19 demonstrates the result of one of the folds for long sequence activity recognition.

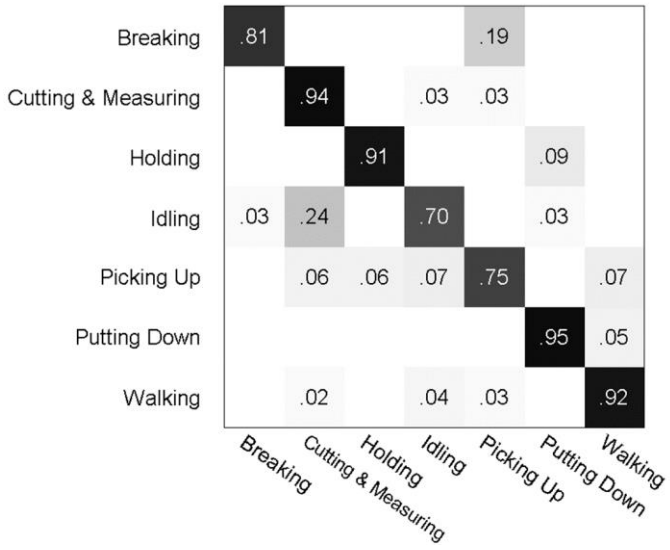


Figure 19. Confusion matrix illustrating long sequence recognition performance with 750 cluster centers, histogram intersection kernel, and $\frac{4\sigma}{17}$ windows resolution.

Furthermore, we are interested in investigating the type of transitions or action categories in which our algorithm seems not to perform well; thus, visualizing predictions versus manually labeled ground truths is a comprehensive method to better interpret the limitations and weakness of our approach. Figure 20 represents sequence predictions versus ground truths for two depth sequences with various durations and sequence of actions.

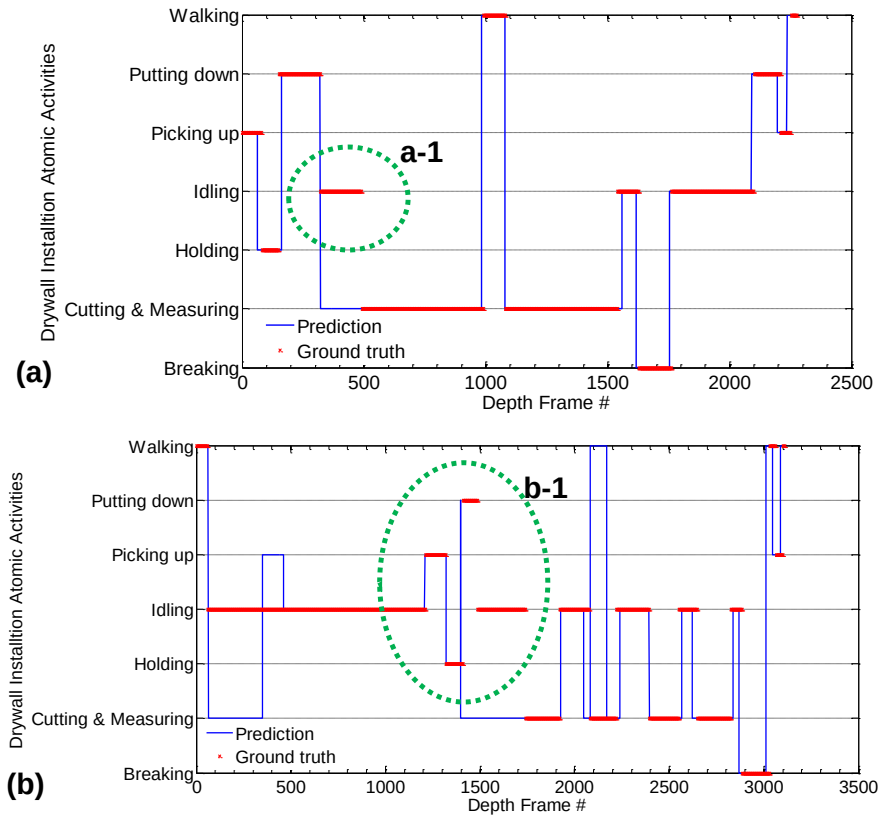


Figure 20. Long sequence actin recognition prediction (blue lines) versus ground truth (red lines).

Based on Figures 21(a) and (b), it could be understood that our algorithm is prone to erroneous in Idling category prediction as shown in circled regions (21(a)-1 and 21(b)-2). One of the most important contributing factors is an excessively high intra-class duration variability in Idling category. Since construction operations are highly interdependent, idle times are not easily predictable and numerous factors which might delay the construction operations are involved such as: waiting for material delivery, waiting on other crews to finish a predecessor activity, or even preparatory works such as measurements which could make other crew members idle. Moreover, in a previous discussion on atomic action recognition we concluded that due to a high intra-class pose variability of idle actions, we are having a relatively low accuracy even in atomic action recognition which intensifies the deviation in long sequence action recognition (Figure 19).

CHAPTER 6

CONCLUSION

This paper presents a novel method for activity analysis of construction workers using inexpensive RGB-D sensors, an approach which not only is invariant to changes in the field of view, view angle, and anthropomorphic differences, but also, is not affecting worker privacy due to the nature of only capturing depth images. Our experiments with an average accuracy of 76% and a maximum accuracy of 91% hold promise in an applicable solution to automated activity analysis for interior construction operations. Furthermore, followed by the huge leaps in imaging technologies, state-of-the-art depth cameras with Time of Flight (TOF) systems are holding promise to a great future in activity recognition not only for indoor activities, but also for outdoors.

To further improve our system, we would focus on implementing TOF-based depth sensors with higher resolutions and higher depth range (3.5 – 7 meters). These sensors (now at the same price as Kinect) holds promise to increase the accuracy of initial stages such as skeleton detection and tracking to a great extent. Moreover, we would like to fuse the RGB and depth images in the frames which the existing algorithms struggle in skeleton detection and calibration, and by identifying the objects which the worker are interacting with, create a more accurate segmentation algorithm which provides a fully robust skeleton extraction. As another step to enhance our algorithm, we would like to create a more informative model by learning various workers' interactions with existing construction entities such as: tools, materials, equipment, location, etc.

Thus, we will focus on finding a comprehensive method to combine the workers' activity information with tools and equipment details. There has not been any previous research on tool detection in construction area and this challenging research problem could either be addressed

through 2D or 3D-based detection algorithms (based on either RGB images or depth images). One of the most important benefits of modeling these interactions is improving the activity recognitions where there are similarities among activity poses. However, there are numerous challenges associated with tool detection process which need to be developed as the next step. The first and the most important challenge is to solve the occlusion problem for tool detection; due to the nature of construction operations and the low chance that worker's tools are visible to the camera's or sensor's view, a complementary approach might be required to detect the workers' tools and recognize their identities. Other probable challenges might be visual clutter and photometric visual variability which requires more digging and exploration. Finally, we will propose a comprehensive dynamic network model which blends our pose information with obtained tools and equipment details to learn and infer short and long sequence construction activities with a higher performance.

REFERENCES

[1] P Goodrum, C Haas. Partial Factor Productivity and Equipment Technology Change at Activity Level in U.S. Construction Industry, J.Constr.Eng.Manage. 128 (2002) 463-472.

[2] MC Gouett, CT Haas, PM Goodrum, CH Caldas. Activity Analysis for Direct-Work Rate Improvement in Construction, J.Constr.Eng.Manage. 137 (2011) 1117-1124.

[3] H Nasir, CT Haas, DA Young, SN Razavi, C Caldas, P Goodrum. An implementation model for automated construction materials tracking and locating, Canadian Journal of Civil Engineering. 37 (2010) 588-599.

[4] Construction Industry Institute (CII). Guide to activity analysis, Construction Industry Institute. (2010) 1-76.

[5] M Shahtaheri. Setting Target Rates for Construction Activity Analysis Categories, (2012).1-156

[6] H Nasir, H Ahmed, C Haas, PM Goodrum. An analysis of construction productivity differences between Canada and the United States, Constr.Manage.Econ. (2013) 1-13.

[7] T Cheng, J Teizer, GC Migliaccio, UC 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.

[8] V Escorcia, MA Dávila, M Golparvar-Fard, JC Niebles, Automated Vision-based Recognition of Construction Worker Actions for Building Interior Construction Operations Using RGBD Cameras, Proc. Construction Research Congress. (2012) 879-888.

[9] MC Gouett, Activity Analysis for Continuous Productivity Improvement in Construction, (2010). 1-159

[10] NIST, Criteria for Performance Excellence, National Institute of Science and Technology (2001-2012). 1-88.

[11] T Cheng, M Venugopal, J Teizer, P Vela. Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments, Autom.Constr. 20 (2011) 1173-1184.

[12] J Teizer, D Lao, M Sofer, Rapid automated monitoring of construction site activities using ultra-wideband, Proc. Automation and Robotics in Construction. 24 (2007) 23-28.

[13] A Giretti, A Carbonari, B Naticchia, M DeGrassi. Design and first development of an automated real-time safety management system for construction sites, Journal of Civil Engineering and Management. 15 (2009) 325-336.

[14] A Costin, N Pradhananga, J Teizer. Leveraging passive RFID technology for construction resource field mobility and status monitoring in a high-rise renovation project, Autom.Constr. 24 (2012) 1-15.

[15] N Pradhananga, J Teizer. Automatic spatio-temporal analysis of construction site equipment operations using GPS data, Autom.Constr. 29 (2013) 107-122.

[16] J Hildreth, M Vorster, J Martinez. Reduction of short-interval GPS data for construction operations analysis, J.Constr.Eng.Manage. 131 (2005) 920-927.

[17] A Peddi, L Huan, Y Bai, S Kim, Development of human pose analyzing algorithms for the determination of construction productivity in real-time, Construction Research Congress, ASCE (2009) 11-20.

[18] J Teizer, PA Vela. Personnel tracking on construction sites using video cameras, Advanced Engineering Informatics. 23 (2009) 452-462.

[19] ER Azar, B McCabe. Vision-based recognition of dirt loading cycles in construction sites, Construction Research Congress (2012) 1042-1051.

[20] L Joshua, K Varghese. Accelerometer-based activity recognition in construction, J.Comput.Civ.Eng. 25 (2010) 370-379.

[21] L Joshua, K Varghese, Classification of bricklaying activities in work sampling categories using accelerometers, Construction Research Congress (2012) 919-928.

[22] L Joshua, K Varghese, Construction activity classification using accelerometers, Construction Research Congress (2010) 61-70.

[23] L Joshua, K Varghese. Selection of Accelerometer Location on Bricklayers Using Decision Trees, Computer-Aided Civil and Infrastructure Engineering. (2013).

[24] J Dai, PM Goodrum, WF Maloney. Construction craft workers' perceptions of the factors affecting their productivity, J.Constr.Eng.Manage. 135 (2009) 217-226.

[25] PM Goodrum, CT Haas, C Caldas, D Zhai, J Yeiser, D Homm. Model to predict the impact of a technology on construction productivity, J.Constr.Eng.Manage. 137 (2010) 678-688.

[26] PM Goodrum, D Zhai, MF Yasin. Relationship between changes in material technology and construction productivity, J.Constr.Eng.Manage. 135 (2009) 278-287.

[27] TR Taylor, M Brockman, D Zhai, PM Goodrum, R Sturgill, Accuracy analysis of selected tools for estimating contract time on highway construction projects, Construction Research Congress (2012) 217-225.

[28] M Memarzadeh, A Heydarian, M Golparvar-Fard, J Niebles, Real-time and automated recognition and 2D tracking of Construction workers and equipment from Site video streams, Computing in Civil Engineering (2012) 429-436.

[29] J Yang, O Arif, PA Vela, J Teizer, Z Shi. Tracking multiple workers on construction sites using video cameras, Advanced Engineering Informatics. 24 (2010) 428-434.

[30] S Chi, CH Caldas. Automated object identification using optical video cameras on construction sites, Computer-Aided Civil and Infrastructure Engineering. 26 (2011) 368-380.

[31] M Park, I Brilakis. Construction worker detection in video frames for initializing vision trackers, Autom.Constr. 28 (2012) 15-25.

[32] J Yang, P Vela, J Teizer, Z Shi. Vision-Based Tower Crane Tracking for Understanding Construction Activity, J.Comput.Civ.Eng. (2012).

[33] M Park, C Koch, I Brilakis. Three-dimensional tracking of construction resources using an on-site camera system, J.Comput.Civ.Eng. 26 (2011) 541-549.

[34] ER Azar, S Dickinson, B McCabe. Server-customer interaction tracker; a computer vision-based system to estimate dirt loading cycles, J.Constr.Eng.Manage. 139(7) (2012) 785-794

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

[36] J Gong, CH Caldas. Computer vision-based video interpretation model for automated productivity analysis of construction operations, J.Comput.Civ.Eng. 24 (2009) 252-263.

[37] JY Kim, CH Caldas. Vision-based action recognition in the internal construction site using interactions between worker actions and construction objects, Automation and Robotics in Construction 30 (2013) 661-668.

[38] HS Koppula, R Gupta, A Saxena. Learning human activities and object affordances from RGB-D videos, arXiv preprint arXiv:1210.1207. (2012).

[39] J Sung, C Ponce, B Selman, A Saxena, Unstructured human activity detection from rgbd images, IEEE transaction on Robatics and Automation (2012) 842-849.

[40] G Ballin, M Munaro, E Menegatti, Human Action Recognition from RGB-D Frames Based on Real-Time 3D Optical Flow Estimation, Biologically Inspired Cognitive Architectures 2012, Springer, 2013, pp. 65-74.

[41] Y Zhao, Z Liu, L Yang, H Cheng, Combing rgb and depth map features for human activity recognition, IEEE in Signal & Information Processing. (2012) 1-4.

[42] J Han, L Shao, D Xu, J Shotton. Enhanced Computer Vision with Microsoft Kinect Sensor: A Review, IEEE Transactions on Cybernetics. (2013) 1-17.

[43] S Han, S Lee, F Peña-Mora. Vision-Based Detection of Unsafe Actions of a Construction Worker: A Case Study of Ladder Climbing, J.Comput.Civ.Eng. 27 (2012) 635–644.

[44] SJ Ray, J Teizer. Real-time construction worker posture analysis for ergonomics training, Advanced Engineering Informatics. 26 (2012) 439-455.

[45] S Han, M Achar, S Lee, F Peña-Mora. Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring, Visualization in Engineering. 1 (2013) 1-13.

[46] B Yao, L Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, (2010) 17-24.

[47] HS Koppula, A Saxena, Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation, International Conference on Machine Learning (2013).

[48] PrimeSense, Nite middleware. http://www.primesense.com/solutions/nite-middleware/. 2013.

[49] H Lou. Implementing the Viterbi algorithm, Signal Processing Magazine, IEEE. 12 (1995) 42-52.

[50] M Brand, N Oliver, A Pentland, Coupled hidden Markov models for complex action recognition, Proc. IEEE Computer Society Conference (1997) 994-999.

[51] M Brand, V Kettnaker. Discovery and segmentation of activities in video, Pattern Analysis and Machine Intelligence, IEEE Transactions on. 22 (2000) 844-851.

[52] N Ikizler, D Forsyth, Searching video for complex activities with finite state models, IEEE on Computer Vision and Pattern Recognition (2007) 1-8.

[53] N İkizler, DA Forsyth. Searching for complex human activities with no visual examples, International Journal of Computer Vision. 80 (2008) 337-357.

[54] D Ramanan, D Forsyth, A Zisserman, Tracking people and recognizing their activities, IEEE Computer Society Conference 2 (2005) 1194 vol. 2.

[55] N Oliver, A Garg, E Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels, Comput.Vision Image Understanding. 96 (2004) 163-180.

[56] T Mori, Y Segawa, M Shimosaka, T Sato, Hierarchical recognition of daily human actions based on continuous hidden markov models, Proc. IEEE Automatic Face and Gesture Recognition. (2004) 779-784.

[57] O Cappé, E Moulines, T Ryden. Inference in hidden Markov models , Springer. (2005).

[58] M Piccardi, O Perez, Hidden Markov Models with Kernel Density Estimation of Emission Probabilities and their Use in Activity Recognition, Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. (2007) 1-8.

[59] GD Forney Jr. The viterbi algorithm, Proc IEEE. 61 (1973) 268-278.

[60] S Russell, Artificial intelligence: A modern approach, 2/E, Pearson Education India 2003.

[61] LR Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition, Proc IEEE. 77 (1989) 257-286.

[62] C Chang, C Lin. LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology. 2 (2011) 27:1-27:27.

[63] A Barla, F Odone, A Verri, Histogram intersection kernel for image classification, Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on. 3 (2003) III-513-16 vol.2.

[64] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." In Proceedings of the 23rd international conference on Machine learning, pp. 233-240. ACM, 2006.

[65] MATLAB 2013a and Mixture of Gaussians and Hidden Markov Model Matlab toolbox (version 0.9), The MathWorks, Inc., Natick, Massachusetts, United States.

[66] A Ihler, M Mandel. Kernel Density Estimation Toolbox for MATLAB (R13), (2003).