IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 20, 2021, accepted June 28, 2021, date of publication July 13, 2021, date of current version July 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095934

Hierarchical Human Action Recognition to Measure the Performance of Manual Labor

JEFFERSON HERNANDEZ[1], GABRIELA VALAREZO[2], RICHARD COBOS[1], JOO WANG KIM[3,1], RICARDO PALACIOS[1], AND ANDRES G. ABAD[1]

[1]Industrial Artificial Intelligence (INARI) Research Lab, Escuela Superior Politécnica del Litora (ESPOL), Guayaquil 090902, Ecuador
[2]Facultad de Ingenieria en Mecanica y Ciencias de la Produccion (FIMCP), Escuela Superior Politecnica del Litoral (ESPOL), Guayaquil 090902, Ecuador
[3]School of Civil, Environmental and Land Management Engineering, Politecnico di Milano, 20133 Milano, Italy

Corresponding author: Andres G. Abad (agabad@espol.edu.ec)

This work was supported in part by the Tiendas Industriales Asociadas Sociedad Anonima (TIA S.A.).

ABSTRACT Measuring manual-labor performance has been a key element of work scheduling and resource management in many industries. It is performed using a standard data system called Time and Motion Study (TMS). Many industries still rely on direct human effort to execute the TMS methodology which can be time-consuming, error-prone, and expensive. In this paper, we introduce an automatic replacement of the TMS technique that works at two levels of abstraction: primitive and activity actions. We leverage on recent advancements in deep learning methods and employ an encoder-decoder based classifier to recognize primitives and a continuous-time hidden Markov model to recognize activities. We show that our system yields results competitive with those obtained with several common human action recognition models. We also show how our proposed system can help operational decisions by computing productivity indicators such as worker availability, worker performance, and overall labor effectiveness.

INDEX TERMS Time and motion study, deep learning, action recognition, manual labor, performance, human effort.

I. INTRODUCTION

Measuring the performance of manual labor has been a key element of work scheduling and resource management in many industries, with particular relevance in areas such as manufacturing, construction, and logistics, where human labor can account for up to 50% of the total project cost [14]. Furthermore, it has been proposed that the biggest inefficiency losses in the workplace are due to human-effort waste [34]. Continuously monitoring the workforce to precisely quantify and benchmark labor productivity allows us to take corrective actions to mitigate these losses.

Measuring the performance of manual tasks can improve their efficiency and effectiveness by characterizing and simplifying their design; assessing and measuring productivity; and assisting in ergonomic evaluations and in calculations of the distribution of manual tasks [25].

Attempts to carry out this measurement usually rely on defining a hierarchy of worker tasks, abstracting them into levels, which allows a more organized observation. One of such abstractions consists of hierarchically differentiating between *primitives* and *activities* [26]: the former defined as the movement of a specific body part, and the latter being the combination of successive primitives. A worker (e.g., during a time and motion study [11], [24]) or a group of workers (e.g., during work sampling [10]) is selected to collect data, with post processing often being required. Variations of this methodology have been applied in various fields such as manufacturing [1], [6] and health [10], [17].

Many industries still rely on human effort to measure manual-labor performance by direct and detailed observation using a stopwatch and determining the times and motions used for specific tasks. Clearly, this method can be time-consuming, error-prone, and overall expensive.

This *human-in-the-loop* approach presents three problems: (P1) it is not effective for recording activities of great complexity as those commonly seen in industry [19]; (P2) it is frequently difficult for analysts to distinguish between primitives and activities [25]; and (P3) it is a costly effort prohibiting continuous measurement.

The associate editor coordinating the review of this manuscript and approving it for publication was Constantinos Marios Angelopoulos.

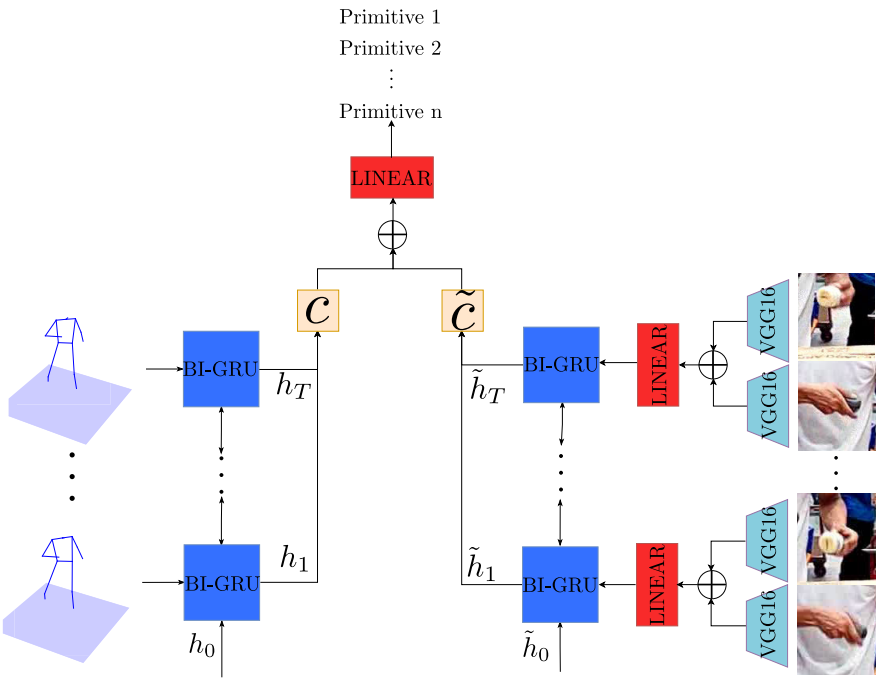


FIGURE 1. Proposed worker primitives recognition system consisting of two RNNs: one composed by a Bi-GRU network that processes skeletal data; and the other by a VGG16 network that extracts features around the left and right hands, which are concatenated and passed through a linear layer before being processed by a Bi-GRU network. The vectors c and $\tilde{c}$ extracted from both RNNs are concatenated and fed to a linear classifier.

A. LITERATURE REVIEW

In order to tackle the aforementioned problems, research has aimed at replacing human observers with automatic procedures providing a continuous stream of information (attenuating P1), a finer understanding of worker activities (solving P2), and reduced costs and increased speeds (attenuating P3). These works have primarily focused on the use of automatic sensor- and vision-based solutions [18], [20], [35] and since each primitive and activity produces characteristic patterns, machine-learning algorithms are used to learn and identify worker tasks.

Vision-based approaches have gained interest over wearable sensor-based ones, since they provide data with a less intrusive collection methodology [12], [31], [39]. These methods have been predominantly based on the use of depth sensors (e.g., Kinect) to build a 3D skeleton model [8], [15]. However, the use of such devices admits only minimal disturbances, which can result in high implementation costs and noisy measurements. The aforementioned methods have only been shown to be successful in controlled-laboratory environments and, therefore, not suited for tackling the intricacies of industrial workplaces.

The methods described in [40] and [38] are the most similar to our own. In [40] an automatic system is proposed to monitor construction activities: they performed action recognition using dense trajectories and achieved an accuracy of 59% but their method is computationally intensive and may not scale well to real-life scenarios. In [38] an automatic system is proposed to monitor piece assembling activities: they performed action recognition using a hierarchical-clustering based convolutional neural network (CNN) model and achieved an accuracy of 56%; however, the dataset used to evaluate their method was fairly limited, containing only hands, and the method required extensive preprocessing, limiting its real-life use. Existing models in the literature do not incorporate the context of the tools the worker is using, which can help to discriminate actions with high inter-class similarity. We believe that, to the best of our knowledge, our work is the first one to measure manual-labor performance using two levels of abstractions, thus increasing accuracy and scalability and facilitating the calculation of performance metrics.

B. RESEARCH CONTRIBUTION

In this paper, we propose an automatic vision-based approach for worker-task recognition *in situ*, at two levels of abstraction: primitives and activities. Our recognition system (Section II) integrates a 2D human-pose estimation neural network with a 3D human-pose inference module, to extract skeletal data from RGB videos of workers performing various tasks. Then, it uses this pose data to extract features around the workers hands by using a CNN. Finally, it combines the 3D pose and extracted features into a sequence that is fed to a Gated Recurrent Unit (GRU) to classify the primitives (see Figure 1). These primitives are passed to a continuous-time hidden Markov model (CT-HMM) to classify the activities.

We show that our proposed system is competitive with commonly used models for human action recognition; and that it can be used to measure the performance of manual labor (Section III), becoming a suitable replacement for the human-in-the-loop methodology.

The contribution of this paper is threefold. (1) We present a hierarchical human recognition methodology that works at two levels of abstraction (primitives and activities). (2) We integrate context from objects, differentiating our methodology from other approaches that have primarily focused on skeletal or sensor data. (3) We show how our proposed system can be used to make operations-engineering decisions by estimating *standard times* and calculating performance indicators such as: *worker availability*, *worker performance*, and *overall labor effectiveness*.

II. METHODOLOGY

In this section, we introduce our methodology, in which primitives are segmented and learned by an RNN classifier, while activities are learned by a CT-HMM classifier modeling the relationship among primitives.

There are two general assumptions for our methodology: (1) only one person must be performing an action at a time, and (2) actions must be recorded in the same vicinity. For primitive recognition, the assumption is that all actions performed by workers can be represented with our primitives taxonomy (described in Section III). For activity recognition, the assumption is that the distribution of times and sequences for primitives is time invariant.

A. PRIMITIVES RECOGNITION

Primitives recognition is comprised of three steps: 3D human pose estimation, object identification for context feature extraction, and primitives classification.

1) 3D HUMAN POSE ESTIMATION

Human pose consists of 2D or 3D coordinates of human joints or keypoints—such as elbows, wrists, and shoulders—resulting in a compact and lightweight representation of the human body.

We estimate 2D pose data for every video frame using the OpenPose framework [4]: a bottom-up approach which first detects key body parts and then uses Part Affinity Fields (PAF) to associate the detected body parts. We infer 3D poses from 2D poses to obtain a richer representation, using a residual feedforward network, based on the work in [23]. To maintain the aspect ratio of the poses, the 2D inputs are normalized by subtracting the mean of frame at time t and dividing by the maximum standard deviation $\sigma^{(t)} = \max(\sigma_x^{(t)}, \sigma_y^{(t)})$ before entering the network.

2) OBJECT IDENTIFICATION FOR CONTEXT FEATURE EXTRACTION

Skeleton-based recognition usually struggles to classify actions related to the use of objects. To overcome this issue, we use the pose to locate the worker hands and create a box of 100×100 pixels around them from which we extract features that represent the objects the worker is holding. We use the VGG-16 architecture [32], a commonly-used feature extractor composed of a stack of convolutional filters with small receptive fields of steadily increasing depth. Figure 2 shows the t-SNE [22] embeddings for the features extracted around the left and the right hands. This graph shows that the features of primitives in which the worker is holding a product or a tool are clustered together. For visualization purposes, we removed from the graph those features representing primitives in which nothing is being held, since they appeared dispersed.

3) PRIMITIVES CLASSIFICATION

In order to classify primitives, we use an RNN classifier based on the *encoder–decoder* framework [7], [33]. For the encoder we use two RNNs: one for the human pose data and the other for the extracted hand features. For succinctness, we only describe the procedure for one encoder and use the sequence element x_t for interchangeably representing the hand features or the human pose data at time t. The encoder reads the input sequence $\mathbf{x} = (x_1, \cdots, x_T)$ and transforms it into a vector c, such that

$$\begin{aligned} h_t &= f(x_t, h_{t-1}), \text{ and} \\ c &= q(\{h_1, \cdots, h_T\}), \end{aligned} \tag{1}$$

where $h_t \in \mathbb{R}^n$ is the hidden state at time t, c is a vector generated from the sequence of hidden states, and f and q are nonlinear functions [2]. In this work, we use the GRU architecture as f and the self-attention mechanism devised in [37] as q, unlike other works that use $q(\{h_1, \cdots, h_T\}) = h_T$ [33], which forces the network to compress all information about the sequence into a single vector. In order to reduce compression of information in the vector c, we use a self-attention mechanism that works by calculating energy vector α_i as

$$\alpha_i = \text{SOFTMAX}\left(\frac{h_T^{\mathsf{T}} W h_i}{n}\right), \tag{2}$$

where matrix W is a square matrix whose dimensions correspond to the feature size of h_i and the energy α_i reflects the importance of the annotation h_i with respect to the final hidden state of the sequence used for classification. To obtain the vector c, we perform the following linear combination $c = \sum_i \alpha_i \tilde{W} h_i$. Intuitively, this process implements an attention mechanism in the decoder, that allows it to decide to which parts of the sequence it pays attention to. The matrices W and $\tilde{W}$ give extra capacity to the model, relieving the encoder from having to compress all information of the sequence into a fixed-length vector.

The decoder is trained to predict the class y_t given the vector c, which is modeled as $p(y_t|c) = g(c)$; here g is a nonlinear, potentially multi-layered, function that outputs the conditional distribution of y_t.

We use the bi-directional variant of RNNs (Bi-RNN) [30] because we want vector h_t to summarize the preceding and

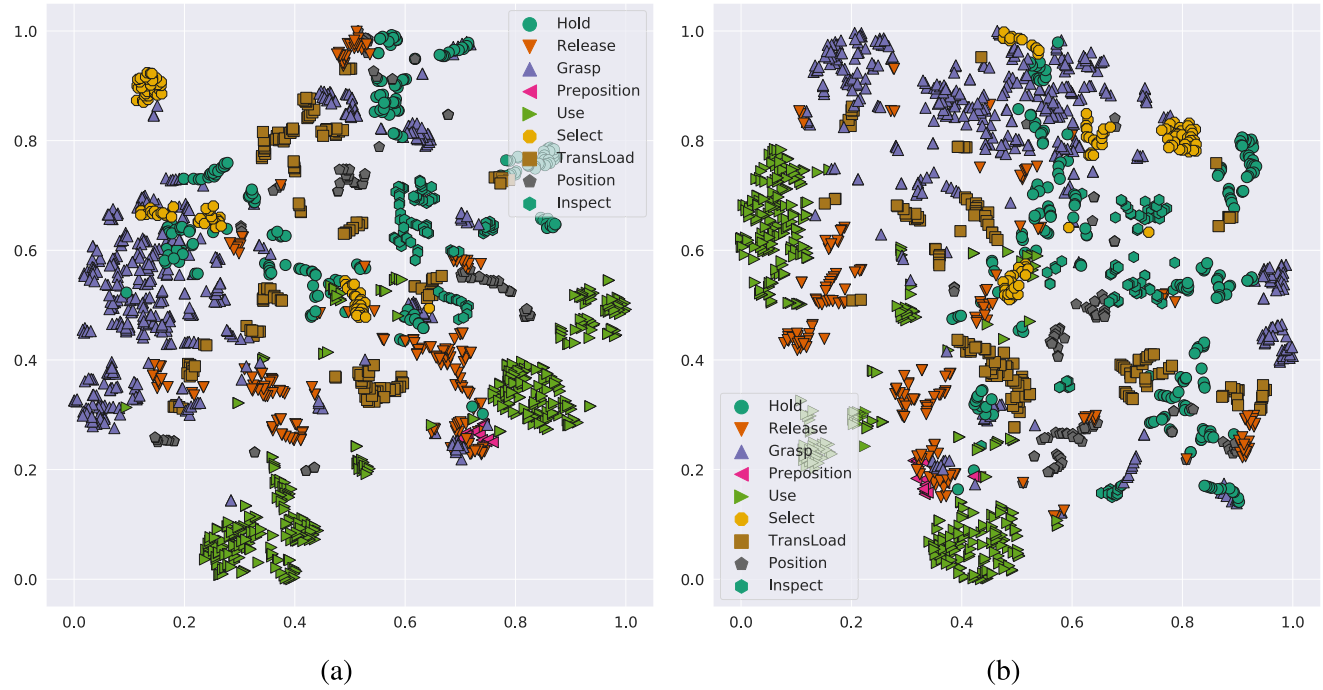


FIGURE 2. (a) t-SNE embedding of VGG16 features extracted around the left hand. (b) Idem for the right hand. (Pre-print note: Best viewed in color.)

the following elements. A Bi-RNN consists of forward and backward RNNs: the former reads the input sequence in ascending order (from x_1 to x_T) and outputs forward states $(\overrightarrow{h_1}, \cdots, \overrightarrow{h_T})$; while the latter reads the input in descending order (from x_T to x_1) and outputs backward states $(\overleftarrow{h_1}, \cdots, \overleftarrow{h_T})$. The final state of the Bi-RNN is obtained concatenating the forward and backward states as $h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]$. Figure 1 depicts the proposed methodology.

B. ACTIVITY RECOGNITION

While primitives are being extracted from video data, the system concurrently learns hierarchical-structure information and the way activities are composed by the sequencing of these primitives.

1) MODEL DESCRIPTION

We model the higher-level structure of activities by using a CT-HMM, i.e. a HMM in which the transitions between hidden states as well as the observations can occur at arbitrary times (see Figure 3).

In our model observable data o depends on the hidden state s via an observation model $p(o|s)$, where the observations $O = \{o_{t_0}, \cdots, o_{t_V}\}$ are obtained at irregularly sampled continuous points in time $\{t_0, \cdots, t_V\}$; and multiple transitions between hidden states can occur before an observation is obtained. A CT-HMM is defined by the set $\lambda = \{\mathbf{b}, \boldsymbol{\pi}, \mathbf{Q}\}$; where $\mathbf{b}$ is the observation model $p(o|s)$; $\boldsymbol{\pi}$ is the initial hidden state distribution; and $\mathbf{Q}$ is a state transition rate matrix whose elements q_{ij} describe the rate at which the process transitions

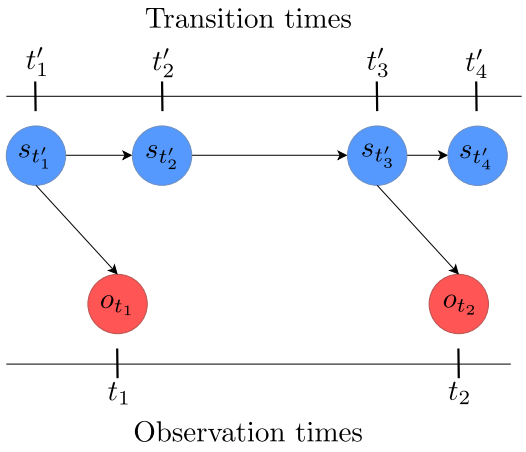


FIGURE 3. In a continuous time hidden Markov model both the hidden states and the transition times are unobserved. Furthermore, multiple hidden states transitions can occur before an observation is obtained.

from state i to j for $i, j \in \{1, 2, \cdots, S\}$ and $i \neq j$, and whose diagonal elements are set such that $q_i = \sum_{i \neq j} q_{ij}$ and $q_{ii} = -q_i$.

The transient behavior of the hidden states is as follows: the CT-HMM stays in the hidden state $s_{t_k} = i$ for time $\Delta_k = t_{k+1} - t_k$ exponentially distributed with rate parameter $1/q_i$. Upon transitioning, the hidden state shifts to state $s_{t_{k+1}} = j$ with probability q_{ij}/q_i. The transitions over hidden states factor into two distributions: an exponential distribution for the inter-occurrence time when the next transition occurs, and a multinomial distribution to determine where the state

transitions. The distribution $P(t)$, over hidden states at some future time t, can be computed as $\exp(\mathbf{Q}t)$, which already takes into account all possible intermediate state transitions between unobserved i and j.

The sufficient statistics of the model are the cumulative amount of time τ_i the model is in hidden state i and the number of times n_{ij} the model transitions from hidden state i to j. Given a current estimate of the model parameters $\lambda = \{\mathbf{b}, \boldsymbol{\pi}, \mathbf{Q}\}$ and observations O, we can write the log-likelihood for the model as [21], [27]:

$$\begin{aligned}\log\mathcal{L}(\lambda; O) = &\sum_{i=1}^{S}\sum_{j=1, i\neq j}^{S}\Big\{\log(q_{ij})\mathbb{E}_{s_t}\left[n_{ij}|O, \mathbf{Q}\right] \\ &- q_i\mathbb{E}_{s_t}\left[\tau_i|O, \mathbf{Q}\right]\Big\} + \sum_{i=1}^{S}\mathbb{E}_{s_t}\left[\mathbb{1}_{\{s_{t_0}=i\}}\right]\log\pi_i \\ &+ \sum_{v=0}^{V}\sum_{i=1}^{S}\mathbb{E}_{s_t}\left[\mathbb{1}_{\{s_{t_v}=i\}}\right]\log p(o_{t_v}|\mathbf{b}), \end{aligned} \tag{3}$$

where $\mathbb{1}_{\{A\}}$ is the indicator function for event A and expectations are computed using Monte Carlo methods. The log-likelihood in Eq. 3 is composed of four terms: a multinomial distribution to determine the state transitions n_{ij}; an exponential distribution for the inter-occurrence time τ_i, an initial state distribution π, and an observation model $\log p(o_{t_v}|\mathbf{b})$. The first two terms are unobserved, since a realization of the CT-HMM is observed only at discrete and irregular times—which are distinct from the inter-occurrence times—and since multiple hidden states transitions can occur before an observation is obtained. The last two terms can be estimated using the discrete time HMM formulation (for a detailed treatment the reader is referred to [3]). In the next two sub-sections, we focus on the estimation from data of the first two terms n_{ij} and τ_i, as well as on the transition matrix $\mathbf{Q}$.

2) MODEL TRAINING

We describe the procedure to train the CT-HMM following the formulation in [21], [27], which finds maximum-likelihood estimates for the parameters using the expectation-maximization (EM) algorithm.

M-step: similar to the formulation of the discrete time HMM (the reader is referred to [3]) except for element

$$q_{ij} = \frac{\mathbb{E}_{s_t}\left[n_{ij}|O, \mathbf{Q}\right]}{\mathbb{E}_{s_t}\left[\tau_i|O, \mathbf{Q}\right]}. \tag{4}$$

E-step: consists in estimating expected values for the sufficient statistics and the indicator function, which we calculate using a continuous variant of the Baum-Welch algorithm [21]. Let $\alpha_{t_v}(i)$ and $\beta_{t_v}(i)$ be the forward and backward likelihood vectors for observed data, respectively; defined as follows:

$$\begin{aligned}\alpha_{t_0}(i) &= \pi_i b_i(o_{t_0}), \\ \alpha_{t_{v+1}}(i) &= \left[\sum_{j}^{S}\alpha_{t_v}(j)e^{\mathbf{Q}\Delta_k}\right] b_i(o_{t_{v+1}}),\end{aligned}$$

$$\begin{aligned}\beta_{t_V}(i) &= 1, \text{ and} \\ \beta_{t_v}(i) &= \sum_{j}^{S} e^{\mathbf{Q}\Delta_k} b_j(o_{t_v})\beta_{t_{v+1}}(j). \end{aligned} \tag{5}$$

By using the forward and backward likelihood vectors, the expected value for the indicator function becomes

$$\mathbb{E}_{s_t}\left[\mathbb{1}_{\{s_{t_v}=i\}}\right] = \frac{\alpha_{t_v}(i)\beta_{t_v}(i)}{\sum_{i=1}^{S}\alpha_{t_v}(i)\beta_{t_v}(i)}. \tag{6}$$

The expected value for the total amount of time the model remains in a hidden state can be calculated by dividing the time-frame in $V-1$ intervals $[t_v, t_{v+1})$, and computing the expected time spent per interval

$$\begin{aligned}&\mathbb{E}^{v}_{s_t}\left[\tau_i|O, \mathbf{Q}\right] \\ &\quad= \frac{1}{\alpha_{t_v}(i)\beta_{t_v}(i)}\int_{t_v}^{t_{v+1}}\alpha_{t_v}(i)e^{\mathbf{Q}(t-t_v)}e^{\mathbf{Q}(t_{v+1}-t)}\beta_{t_v}(i)dt, \end{aligned} \tag{7}$$

and adding those expected values together. Moreover, a formula similar to Eq. 7 but scaled by q_{ij} can be found for the expected number of times the model transitions from hidden state i to j. Classification of an activity is done using an ensemble-like methodology by training a CT-HMM for each category and computing the optimal weights for the ensemble using a held-out dataset.

III. CASE STUDY: ORDER PREPARATION IN A DISTRIBUTION CENTER

Order preparation in a distribution center is the operational process associated with packing and shipping orders; it can be done in a variety of ways [29]. For our purposes, we chose a system consisting of aisles of racks with tote containers in which workers place items into, following instructions from light indicators. Order preparation commonly involves whole-body motions, such as repetitive bending and material or tool handling. We recorded RGB videos data of these complex actions and used them to test our proposed methodology. This research considers various primitives—predefined in [11] and shown in Figure 4—and activities—*Label*, *Scan*, *Search*, *Putting*, *Confirm*.

Our case study considered six workers with different proficiency levels. Each subject was recorded for 40-60 min with a relative view-point variation of 60°, 120°, 240°, and 300°; these views were selected to minimize hand occlusion. RGB videos were recorded at a resolution of 1280 × 720 pixels at 30 frames per second; this high frame-rate was selected to minimize hand occlusion due to dense sampling. When occlusion did occur the sample was simply discarded.

The collected data was manually labeled selecting from thirteen primitives (see Figure 4) and five activities. Actions irrelevant to the activities were considered as delays and labeled into one category. These actions include taking irregular rest time, walking to grab a box, and disposing of boxes. Other actions not taken into account included communicating with coworkers and using the restroom. The final dataset was partitioned into training and validation sets, discarding

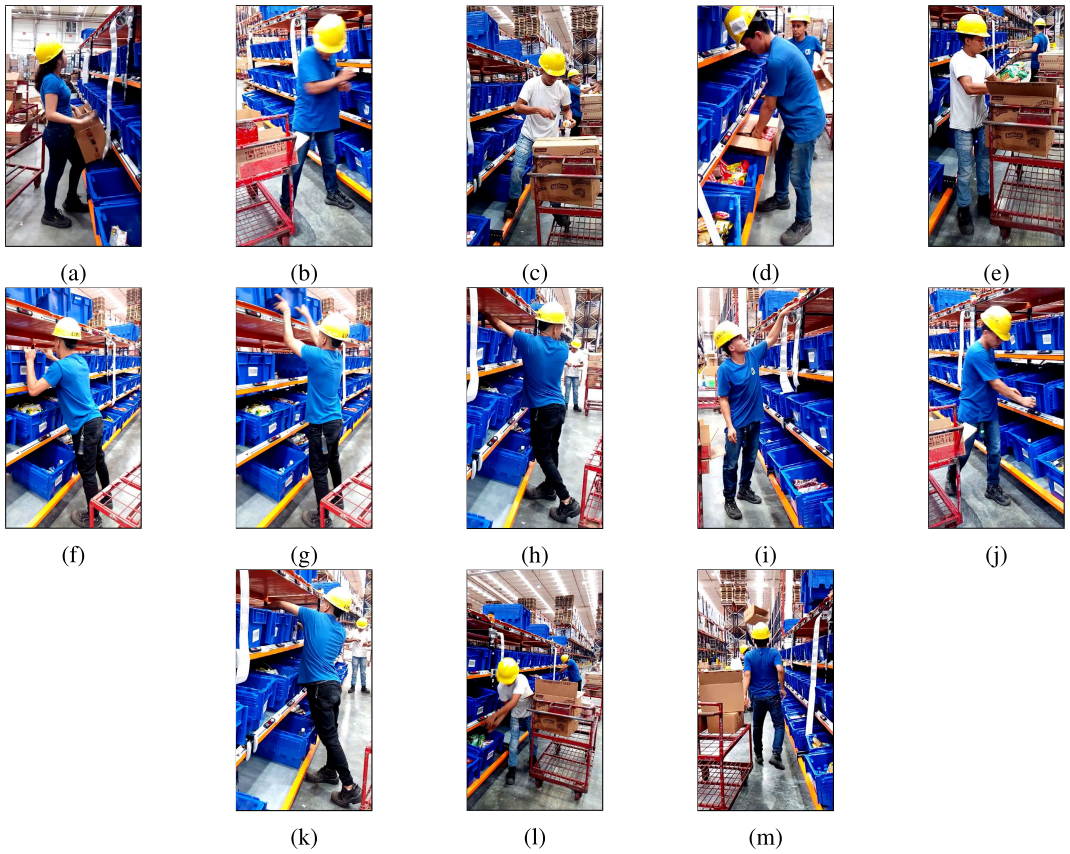


FIGURE 4. Samples from the selected primitives for the sorting activities in the distribution center. The primitives are: (a) search, (b) find, (c) select, (d) grasp, (e) hold, (f) transport loaded, (g) transport empty, (h) position, (i) use, (j) inspect, (k) preposition, (l) release load, (m) and delays.

10 min between sets to ensure data independence. Data was partitioned again using a moving-window approach with a size of 1 s. This process produced 5, 000 samples which we partitioned into training and validation splits of sizes 80% and 20%, respectively. All experiments were run on an Intel i9 3.3 GHz processor with 20 cores, 48 GB of RAM memory, and an Nvidia GTX 1080 Ti with 12 GB of vRAM using the PyTorch deep-learning framework [28].

TABLE 1. Primitives recognition results from different methods, using the cross-subject and cross-person evaluation criteria on our dataset.

	Cross-View		Cross-Person	
Model name	**Top-1 Acc**	**Top-3 Acc**	**Top-1 Acc**	**Top-3 Acc**
R(2+1)D [36]	68.50	76.00	**37.61**	**60.68**
I3D [5]	67.50	78.50	35.04	49.57
ResneXt3D [16]	67.50	76.50	34.19	55.56
SlowFast [9]	66.50	77.00	35.04	49.57
Ours	**78.36**	**81.70**	35.70	49.60

A. BENCHMARK EVALUATIONS

We assessed the performance of our system, on both primitives and activities recognition, by comparing it against several models: R(2+1)D [36], I3D [5], ResneXt3D [16] and SlowFast [9]. To compare models we devised two protocols for primitives and activities classification.

For primitives classification, the protocols were: cross-view—designed to test if the models can generalize to view variation—corresponding to angles 60°, 120° for training and 240°, 300° for testing; and cross-subject—designed to test if the models can generalize to subject variation—corresponding to workers 0, 1, 2 and 3 for training, and 4 and 5 for testing.

For activities classification, the protocol was to compare the classical end-to-end classification framework with the CT-HMM model framework. In the end-to-end framework, we used a variable frameskip to reduce the duration of activities to 64 frames and no information of the primitives was provided. In the CT-HMM model framework, we maintained the duration of each video and used the logits produced by the models in the cross-view primitives classification protocol as input to the model.

All models were trained for 100 epochs, with learning rate of 10^{-3}, reducing it by a factor of 10 in the epochs 60 and 80, as a learning rate schedule. We used standard data augmentation techniques for human action recognition in videos.

Table 1 shows that our system, which combines the pose data and the RGB data around the subject hand, has

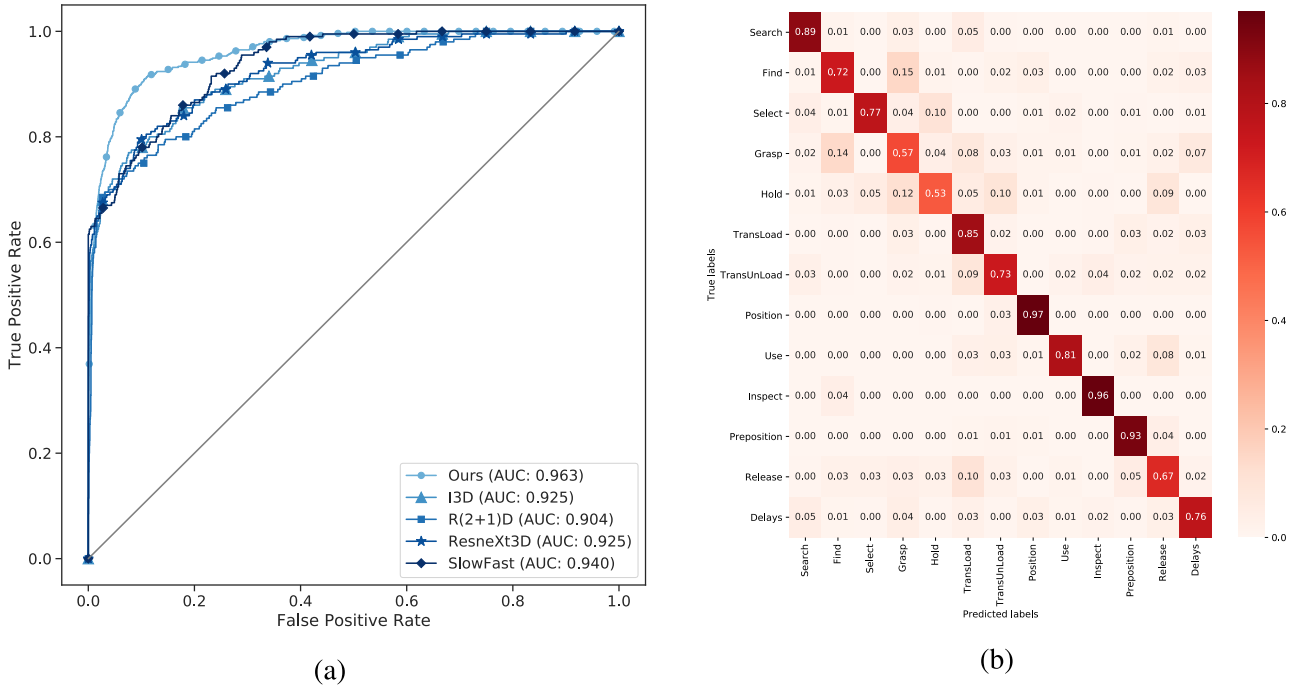


FIGURE 5. (a) ROC curves for primitives recognition of all models following the cross-view protocol. (b) Confusion matrix of primitives recognition of our model.

TABLE 2. Activity recognition results from different methods, using the end-to-end and CT-HMM evaluation criteria on our dataset.

	End-to-End		CT-HMM	
Model name	**Top-1 Acc**	**Top-3 Acc**	**Top-1 Acc**	**Top-3 Acc**
R(2+1)D [36]	79.62	**97.55**	90.00	98.75
I3D [5]	69.84	95.92	83.75	98.75
ResneXt3D [16]	78.26	95.38	**92.50**	98.75
SlowFast [9]	74.73	95.11	88.75	97.50
Ours	**80.46**	96.09	**92.50**	**100**

the highest accuracy among the compared models in the cross-view experiments; while R(2+1)D performed better than all other models in the cross-person experiments. Even though our model does not achieve the highest test accuracy in the cross-person primitives classification protocol, the results show that it is also highly competitive, among the compared models, in this task. The ROC curves for the cross-view task are shown in Figure 5a with our model achieving the highest AUC (0.963).

Table 2 shows that all models improve their performance when the activities recognition is managed by the CT-HMM, instead of the classical end-to-end approach. One of the potential reasons for these results is the fact that activities have a higher duration and the end-to-end approach is limited to a relatively low number of frames. Also, the CT-HMM is more suitable to small datasets, since models with large capacity tend to overfit. The use of the CT-HMM model increases the activities classification accuracy by 17% in all models. This improvement on performance is also seen in Figure 6a, with our model attaining the highest AUC in both end-to-end and CT-HMM methodologies, 0.953 and 0.989, respectively.

Our system achieved an average accuracy of 78.36% for the task of primitives recognition. The primitives categories with the highest accuracy were *Position*, *Inspect* and *Preposition*, while the categories with the lowest accuracy were *Grasp*, *Hold* and *Release* (see Figure 5b). Likewise, our proposed system achieved an average accuracy of 92.50% for the task of activities recognition. The activities categories with the highest accuracy were *Label* and *Putting*, while those with the lowest accuracy were *Scan*, *Search* and *Confirm* (see Figure 6b).

B. LABOR PRODUCTIVITY METRICS

We used our proposed system for *standard times* calculation (see Table 3) following the procedure described in [11]. This procedure considers Personal, Fatigue, and Delay (PFD) allowances. We have considered a base allowance of 5% for the five activities, which represents the personal needs allowance. For the *Search* activity, we consider an extra allowance of 5% since it requires a high level of attention. Finally, for the *Putting* activity we consider an extra allowance of 12% since it generates mental stress and requires a high level of attention. These allowance values are an industry standard and the reader is refereed to [11] for more information.

We used our proposed system to calculate labor productivity metrics. We considered worker availability, performance, and overall labor effectiveness (OLE). *Worker availability*

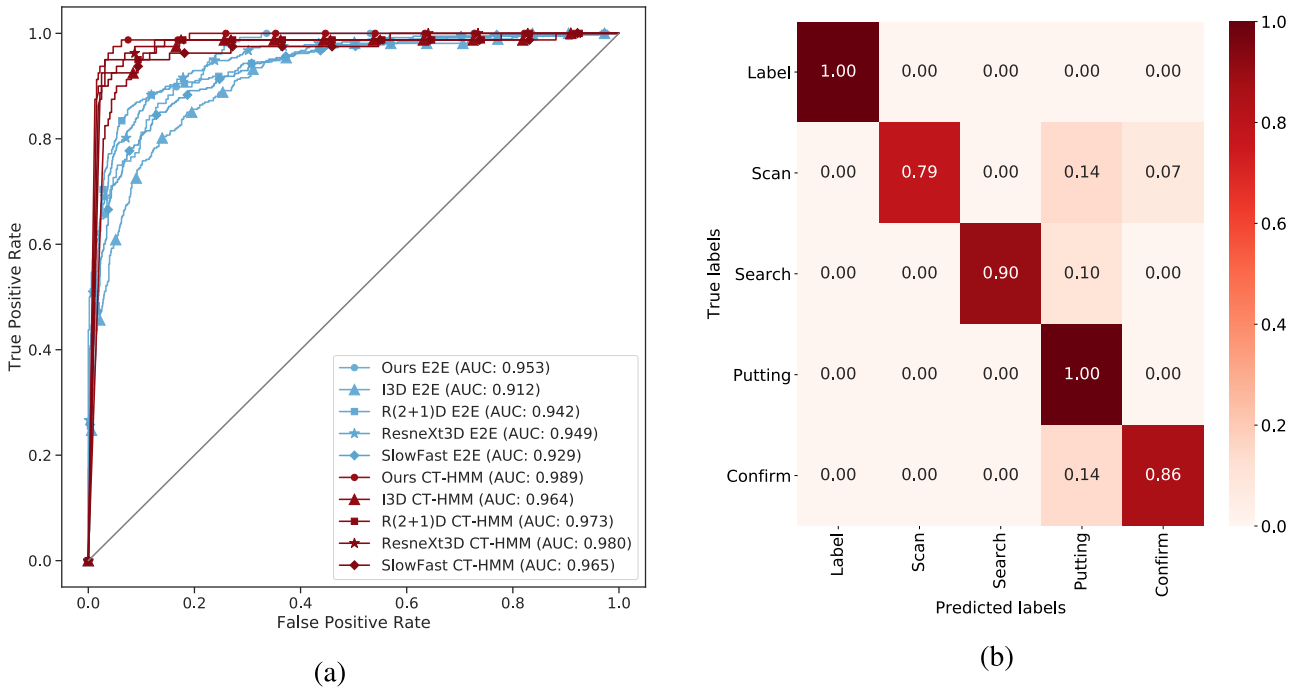


FIGURE 6. (a) ROC curve for activities recognition of all models. (b) Confusion matrix of activities recognition of our model.

TABLE 3. Standard times calculation.

Activity	Normal Times (s)	% PFD Allowance	Standard Time (s)
Label	2.45	1.05	2.57
Scan	1.34	1.05	1.41
Search	1.18	1.10	1.30
Putting	3.47	1.17	4.06
Confirm	1.17	1.05	1.23

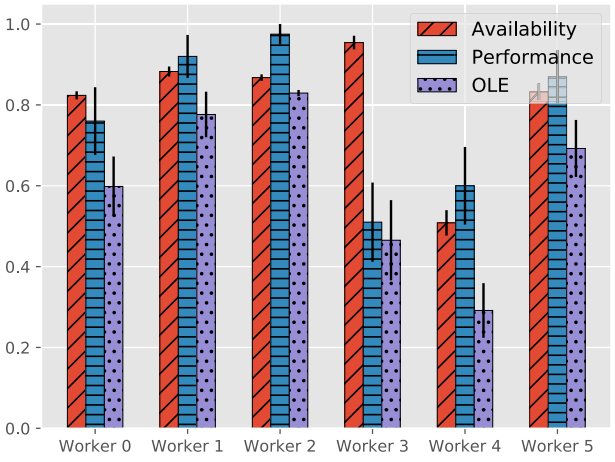


FIGURE 7. Worker availability, performance, and overall labor effectiveness measured by our proposed system.

measures the percentage of time spent doing productive (value-adding) activities. Likewise, *Performance* compares execution times of workers to standard times, and *OLE* is calculated as the product of these two metrics and the quality of work obtained from historical data. These metrics allow managers to make operational decisions, providing needed information to analyze their combined effect over the activities [13]. For instance, they can help locate areas in which an optimal work schedule can be critical to productivity. Figure 7 details the obtained metrics for each worker.

The calculated metrics can help to analyze productivity, detect opportunities to improve proficiency at the individual level, and identify corrective actions such that all operations become up to standards.

IV. CONCLUSION AND FUTURE WORK

In this paper, we present an automatic replacement for the human-in-the-loop methodology for measuring the performance of manual labor using skeletal data and features extracted around the hands. We focused on task recognition at two levels of abstraction–primitives and activities–using an encoder-decoder based classifier and a continuous time hidden Markov (CT-HMM) classifier, respectively. As a case study, we collected RGB video data from order preparation tasks in a distribution center (DC) and labeled it selecting from 13 predefined primitives and 5 activities. This dataset had several challenging characteristics such as: view angle and illumination changes, and interrupted work-flow. The achieved accuracy was 78.36% for primitives recognition and 92.50% for activities recognition.

The main contribution of this paper is the hierarchical human action recognition methodology, working at two levels

of abstraction. The experimental results suggest that leveraging upon the relation between primitives and activities significantly improves the accuracy of the system; other methods that employ end-to-end frameworks are less suitable for the task, especially with limited data and training time.

These results demonstrate that video data along with deep-learning techniques can take advantage of characteristic patterns according to the type of primitives and activities performed; and show that our methodology is capable of reaching human-level proficiency in measuring the performance of manual labor. Since we have used standard and predefined primitives, the proposed approach has potential to be adapted to various industrial settings.

Current limitations of our system include: it requires very little intra-class variability, meaning primitives must be performed in a very similar way; it identifies objects using pre-trained networks, which could confuse the system when similar-looking objects are used; and it needs dense annotations at the primitives level.

Directions for future work include: incorporating context using an object recognition neural network trained on objects commonly found in DCs; testing the proposed methodology in other areas; using context free grammars instead of the CT-HMM, which may allow to represent higher levels of abstractions; and performing unsupervised learning of the primitives and activities taxonomy.

ACKNOWLEDGMENT

The authors would like to thank Tiendas Industriales Asociadas Sociedad Anonima (TIA S.A.), a leading grocery retailer in Ecuador, for providing funding for this research effort and for granting access to their Distribution Center, particularly, to their put-to-light system, providing a rich environment for developing our Case Study.

REFERENCES

[1] K. S. Al-Saleh, "Productivity improvement of a motor vehicle inspection station using motion and time study techniques," *J. King Saud Univ., Eng. Sci.*, vol. 23, no. 1, pp. 33–41, 2011.
[2] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Jan. 2015, pp. 1–15.
[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.
[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7291–7299.
[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.
[6] K. Chatzis, "Searching for standards: French engineers and time and motion studies of industrial operations in the 1950s," *Hist. Technol., Int. J.*, vol. 15, no. 3, pp. 233–261, 1999.
[7] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.
[8] D. I. Van Blommestein, A. F. Van Der Merwe, S. Matope, and A. D. Swart, "Automation of work studies: An evaluation of methods for a computer based system," in *Proc. CIE*, Jul. 2012, pp. 212–226.
[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6202–6211.
[10] S. A. Finkler, J. R. Knickman, G. Hendrickson, M. Lipkin, Jr., and W. G. Thompson, "A comparison of work-sampling and time-and-motion techniques for studies in health services research," *Health Services Res.*, vol. 28, no. 5, pp. 577–597, 1993.
[11] A. Freivalds, *Niebel's Methods, Standards, and Work Design*, vol. 700. Boston, MA, USA: McGraw-Hill, 2009.
[12] Z. Gao, H. Z. Xuan, H. Zhang, H. Wan, and K.-K. R. Choo, "Adaptive fusion and category-level dictionary learning model for multiview human action recognition," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9280–9293, Dec. 2019.
[13] G. Gordon, *Lean Labor: A Survival Guide for Companies Facing Global Competition*. Lowell, MA, USA: Kronos, 2011.
[14] M. C. Gouett, C. T. Haas, P. M. Goodrum, and C. H. Caldas, "Activity analysis for direct-work rate improvement in construction," *J. Construct. Eng. Manage.*, vol. 137, no. 12, pp. 1117–1124, Dec. 2011.
[15] S. Han, M. Achar, S. Lee, and F. Peña-Mora, "Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring," *Vis. Eng.*, vol. 1, no. 1, Dec. 2013.
[16] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.
[17] G. C. Harewood, K. Chrysostomou, N. Himy, and W. L. Leong, "A 'time-and-motion' study of endoscopic practice: Strategies to enhance efficiency," *Gastrointestinal Endoscopy*, vol. 68, no. 6, pp. 1043–1050, 2008.
[18] L. Joshua and K. Varghese, "Accelerometer-based activity recognition in construction," *J. Comput. Civil Eng.*, vol. 25, no. 5, pp. 370–379, Sep. 2011.
[19] G. Kanawaty, *Introduction to Work Study*. Geneva, Switzerland: International Labour Organization, 1992.
[20] A. Ladjailia, I. Bouchrika, H. F. Merouani, N. Harrati, and Z. Mahfouf, "Human activity recognition via optical flow: Decomposing activities into basic actions," *Neural Comput. Appl.*, vol. 32, no. 21, pp. 16387–16400, Nov. 2020.
[21] Y.-Y. Liu, S. Li, F. Li, L. Song, and J. M. Rehg, "Efficient learning of continuous-time hidden Markov models for disease progression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3600–3608.
[22] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[23] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2640–2649.
[24] H. B. Maynard, G. J. Stegemerten, and J. L. Schwab, *Methods-Time Measurement*, vol. 292. New York, NY, USA: McGraw-Hill, 1948.
[25] F. Meyers and F. Stewart, *Motion and Time Study for Lean Manufacturing*. Hoboken, NJ, USA: Prentice-Hall, 2002.
[26] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, 2006.
[27] U. Nodelman, C. R. Shelton, and D. Koller, "Expectation maximization and complex duration distributions for continuous time Bayesian networks," in *Proc. 21st Conf. Uncertainty Artif. Intell. (UAI)*, 2005, pp. 421–430.
[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS-W*, 2017, pp. 1–4.
[29] G. Richards, *Warehouse Management: A Complete Guide to Improving Efficiency and Minimizing Costs in the Modern Warehouse*. London, U.K.: Kogan Page, 2017.
[30] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[31] J. Seo, S. Han, S. Lee, and H. Kim, "Computer vision techniques for construction safety and health monitoring," *Adv. Eng. Inform.*, vol. 29, no. 2, pp. 239–251, 2015.
[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–14.
[33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[34] F. W. Taylor, *The Principles of Scientific Management*, vol. 202. New York, NY, USA, 1911.

[35] J. Teizer, "Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites," *Adv. Eng. Informat.*, vol. 29, no. 2, pp. 225–238, Apr. 2015.
[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[38] Z. Wang, R. Qin, J. Yan, and C. Guo, "Vision sensor based action recognition for improving efficiency and quality under the environment of industry 4.0," *Procedia CIRP*, vol. 80, pp. 711–716, 2019.
[39] J. Yang, M.-W. Park, P. A. Vela, and M. Golparvar-Fard, "Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future," *Adv. Eng. Informat.*, vol. 29, no. 2, pp. 211–224, Apr. 2015.
[40] J. Yang, Z. Shi, and Z. Wu, "Vision-based action recognition of construction workers using dense trajectories," *Adv. Eng. Informat.*, vol. 30, no. 3, pp. 327–336, Aug. 2016.

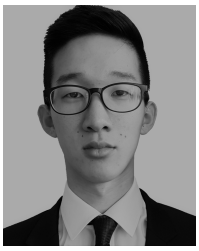
JEFFERSON HERNANDEZ was born in Guayaquil, Ecuador, in 1995. He received the degree in industrial engineering from the Escuela Superior Politecnica del Litoral, in 2019.

He is currently a Researcher with the Industrial Artificial Intelligence Lab (INARI), ESPOL. His research interests include artificial intelligence, deep learning, reinforcement learning, and applications of artificial intelligence to industrial settings.

GABRIELA VALAREZO was born in Guayaquil, Ecuador, in 1993. She received the degree in industrial engineering from the Escuela Superior Politecnica del Litoral, in 2019. From 2017 to 2019, she did multiple internships, one of which was in the multinational company Schlumberger, as a Field Engineer Trainee.

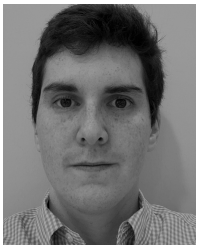
RICHARD COBOS received the B.Sc. degree in chemical engineering from the Escuela Superior Politecnica del Litoral (ESPOL), Guayaquil, Ecuador, in 2018, where he is currently pursuing the M.Eng. degree in industrial automation and control.

In 2018, he was a Research Assistant with the ESPOL Physics Department. He is currently a Researcher with the Industrial Artificial Intelligence Lab (INARI), ESPOL. His research interests include artificial intelligence and process control and modeling of chemical plants. His awards and honors include the Best Chemical Engineering Graduate for his Diploma at ESPOL, the Medal for Best Natural and Mathematical Sciences Graduate at ESPOL, the Philanthropic Society Medal, Guayaquil, and the Medal for Best Chemical Engineering Student at ESPOL.



JOO WANG KIM was born in Seoul, Republic of Korea, in 1995. He received the degree (Hons.) in civil engineering from the Escuela Superior Politecnica del Litoral, in 2019. He is currently pursuing the master's degree in civil engineering with the Politecnico di Milano.

From 2017 to 2019, he did multiple internships, two of which were in the multinational companies Schlumberger and Veolia, as a Field Engineer Trainee and a Thesis Intern, respectively. Since 2019, he has been working as a Research Assistant at INARI Lab, where he did multiple projects in computer vision, mainly in the pose-based action recognition field.



RICARDO PALACIOS was born in Guayaquil, Ecuador, in 1994. He received the B.S. degree in mechanical engineering from the University of Miami, Florida, in 2016, and the M.S. degree in mechanical engineering from Columbia University, in 2019.

In 2015, he was a Research Assistant with the Center for Advanced Multiscale Studies, University of Miami. From 2013 to 2016, he was a member of Engineers Without Borders. From 2018 to 2019, he was a Research Assistant with the Musculoskeletal Biomechanics Laboratory, Columbia. He is in the process of obtaining a patent developed while at the University of Miami. His research interest includes structure analysis of hollow composites. His research interests include the mechanical analysis of cartilage in cows and its degradation through time.

ANDRES G. ABAD received the B.Sc. degree in statistics from ESPOL, in 2005, and the M.Sc. and Ph.D. degrees in industrial and operations engineering from the University of Michigan, Ann Arbor, in 2008 and 2010, respectively.

He is currently an Associate Professor with the Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador. He is the Director of industrial artificial intelligence with the INARI Research Lab, ESPOL, focused on exploring the use of deep learning to solve real industrial problems. In recent years, he served as the Technical Advisor for the Ecuadorian Minister of Industry and a Data Scientist Consultant for the Import Retail Firm, Ecuador. His research and professional interests include the development of industrial solutions based on deep learning, machine learning, data science, and mathematical optimization.