IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 29, 2018, accepted November 26, 2018, date of publication December 4, 2018,
date of current version December 27, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2884793*

# Towards Robust Human-Robot Collaborative Manufacturing: Multimodal Fusion

**HONGYI LIU[1], TONGTONG FANG[2], TIANYU ZHOU[2], AND LIHUI WANG[1]**

[1]Department of Production Engineering, KTH Royal Institute of Technology, SE-10044 Stockholm, Sweden
[2]Department of Software and Computer Systems, KTH Royal Institute of Technology, SE-10044 Stockholm, Sweden

Corresponding author: Lihui Wang (lihuiw@kth.se)

**ABSTRACT** Intuitive and robust multimodal robot control is the key toward human–robot collaboration (HRC) for manufacturing systems. Multimodal robot control methods were introduced in previous studies. The methods allow human operators to control robot intuitively without programming brand-specific code. However, most of the multimodal robot control methods are unreliable because the feature representations are not shared across multiple modalities. To target this problem, a deep learning-based multimodal fusion architecture is proposed in this paper for robust multimodal HRC manufacturing systems. The proposed architecture consists of three modalities: speech command, hand motion, and body motion. Three unimodal models are first trained to extract features, which are further fused for representation sharing. Experiments show that the proposed multimodal fusion model outperforms the three unimodal models. This paper indicates a great potential to apply the proposed multimodal fusion architecture to robust HRC manufacturing systems.

**INDEX TERMS** Deep learning, human-robot collaboration, multimodal fusion, intelligent manufacturing systems.

## I. INTRODUCTION

Recently, human-robot collaboration (HRC) has emerged as a research spotlight in several manufacturing industries [1]. Compared with traditional manufacturing systems where human operators and robots are strictly separated due to safety reasons, HRC manufacturing systems allow human operators and robots to work together in a shared environment. By utilising the advantages from both human operators and robots, HRC manufacturing systems empower human operators to actively assign repetitive and dangerous tasks to robots [2]–[4], while human operators can focus on more interesting and challenging tasks. One of the most significant differences between traditional manufacturing systems and HRC manufacturing systems is the control interface of the industrial robot [4]. Since the strict separation between human operators and industrial robots is required, traditional industrial robot controllers adopt model-based control methods [5] and brand-specific robot control code. However, HRC manufacturing systems are normally deployed in dynamic environments where human operators and industrial robots are coexisted. Control commands therefore need to be assigned to industrial robots actively. Consequently, in HRC manufacturing systems, the human operators should be able to access and control the robots intuitively and effortlessly without writing and debugging brand-specific code [6].

Multimodal HRC was proposed for intuitive robot control [2], [4], [7]. It allows human operators to control the robots with human-friendly methods such as hand and body motion [1], speech command [2], haptic touch [8], etc. However, due to the unstable sensors and limited recognition accuracy, most of the multimodal HRC systems are intuitive but unreliable. Thus, current multimodal HRC systems still cannot be directly applied in manufacturing environments.

A simple way to achieve robust multimodal HRC is to add more modalities and utilise the representations from different modalities. Instead of hard-coded rules or simple logic gates between different modalities [2], [4], [6], more advanced information fusion method should be available to fully share and utilise the embedded hidden data patterns from multimodalities.

The recent advancements of deep learning have had a profound impact on industry and society [9], [10]. Deep learning algorithms not only outperform human experts in recognition and strategy-related tasks [9], [11], but also demonstrate the potential in multimodal fusion [12]. Deep learning and multimodal fusion provide outstanding flexibility and capability in capturing hidden patterns from high dimensional multimodal data, which can be potentially utilised to analyze the multimodalities to make comprehensive decisions for robust multimodal HRC.

As shown in Figure 1, in this paper, the authors propose a multimodal fusion architecture for intuitive and robust HRC systems with microphone-based speech command recognition, Leap Motion-based hand motion recognition [13], and camera-based body motion recognition. To design the proposed architecture, the authors review related literature in multimodal HRC and multimodal deep learning. Different deep neural networks are adopted to process the input sequential data from each unimodality. The trained three unimodal models are further fused by multimodal fusion architecture to facilitate robust decision making. The authors also compare and visualize the results of multimodal fusion. Discussions and future directions are given before concluding the paper.

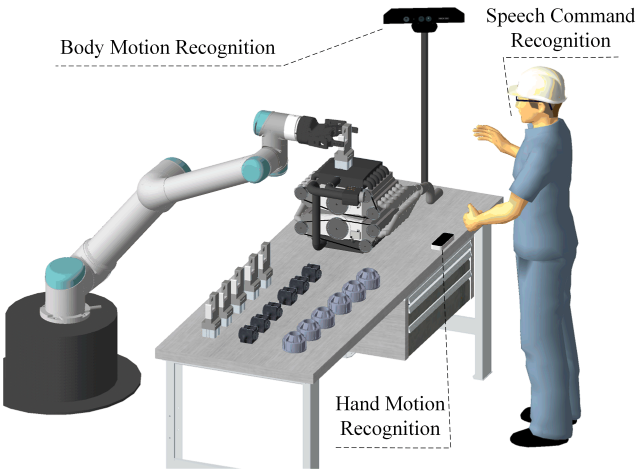


**FIGURE 1. Multimodal human-robot collaboration enabled by speech command recognition, hand motion recognition, and body motion recognition.**

## II. RELATED WORK

### A. MULTIMODAL HUMAN-ROBOT COLLABORATION

For intuitive robot control, multimodal HRC has been researched extensively [7], [14], [15]. As one of the most effective tools for human-human communication, speech commands have been applied to robot control by many researchers [4], [7], [14], [16]. Without the noisy environment, the speech command recognition can reach relatively high accuracy [4], [14]. In the manufacturing context, the noisy environment can be a challenging factor [2], [4]. Several researchers also explored the possibility to adopt body motion recognition for robot control [1], [4], [14]. In body motion recognition, both 2-D camera and 3-D camera can be applied [1]. Recent advancements of human tracking with 2-D camera improved the reliability of body motion tracking [17], which can be adopted to further increase the body motion recognition accuracy. Another popular modality for multimodal HRC is hand motion [1], [7], [14], [15]. Many different sensors can be utilised for hand motion recognition, such as: 2-D camera [7], 3-D camera [4], [15], electromyographic (EMG) band [1], and gyroscope glove [1]. With the development of sensor technologies, the applicability of hand motion recognition could be further extended. In some literature, gaze is adopted as a modality for multimodal HRC robot control [14], [18]. Defined as the direction of eyes pointing in space, gaze can potentially provide timely contextual information for HRC manufacturing applications. With new sensor technologies, some literature suggested that facial expressions and emotional information can also be used as a modality in multimodal HRC [14]. In short, the emerging sensor technologies and new modalities provide new possibilities for HRC manufacturing applications.

To recognise the sequential data, various machine learning models can be applied. Hidden Markov model (HMM) is suitable for modelling sequences while maintaining the spatiotemporal characteristics within the sequences. Filler-based HMM [19]–[21] was the earliest algorithm aiming at speech command recognition, i.e., Keyword Spotting (KWS). HMM was also regarded as a promising approach for hand motion recognition [22], [23]. Large margin-based classification algorithms such as Support Vector Machine (SVM) was utilised to maximise the detection rate of keywords [24]. Furthermore, SVM can be adopted for problems such as hand motion recognition [25], facial expression recognition [25], and body motion recognition [1], [26]. Other commonly used algorithms for motion recognition tasks include Ensemble Method [27] and Dynamic Time Warping [28], [29].

As discussed in Section I, deep learning has become one of the most effective machine learning models. Given large amount of data, deep learning can provide human-level performance on various tasks [10]. Convolutional Neural Network (CNN) is a popular deep learning model that can be adopted for sequential data recognition tasks such as speech command recognition [4], [30]–[32]. Fernández *et al.* [33] applied Recurrent Neural Network (RNN) as a discriminative deep learning model for KWS tasks. Results turned out that RNN outperforms traditional HMM approach by a large margin. RNN has been used in other sequence recognition tasks such as speech command recognition [4], [34] and motion recognition [4], [35], [36]. As a special type of RNN, Long short-term memory (LSTM) is especially suitable for long sequential data recognition tasks [37]–[39]. Compare with basic RNN, LSTM selectively remember or forget information in training, which is beneficial for learning the long-term dependencies. Meanwhile, deep learning boosted the advancement of transfer learning [40]–[42]. With the possibility to reuse feature extraction capabilities of other powerful networks, the model training time and difficulty of the image or video-related tasks can be greatly reduced [40], [43]. The above mentioned scientific investigations support the argument that the deep learning models outperform the traditional machine learning models.

### B. MULTIMODAL DEEP LEARNING

As pointed out in Sections I and II-A, deep learning has demonstrated the potential in feature representations for multimodal fusion. The multimodal data sources consist of a wide range of modality combinations such as text and images [44], RGB-D images and RGB images [45], [46], audio and video [12], [47], audio and text [48], [49], audio and images [50]. Ngiam *et al.* [12] and Dorfer *et al.* [50] pointed out that multimodal fusion models generally outperform unimodal models.

Further analysis shows that multimodal fusion has two different paradigms. With Restricted Boltzmann Machines (RBMs) or deep autoencoders, models in the first paradigm can learn a joint feature representation from a multimodal dataset [12], [44]. In contrast, models in the second paradigm concatenate the feature representation of each modality [45], [46], [48]–[50]. As pointed out by Lenz *et al.* [51], the first paradigm has better performance in cases where the modalities have significant differences, e.g., text and images, whereas the second paradigm tends to perform better when the modalities are similar, e.g., RGB-D images and RGB images. In the first paradigm, as suggested by Oramas *et al.* [52], the model training process could be slowed down by complex modalities such as audio and video. Moreover, models in the first paradigm may put greater emphasis on one particular modality so other modalities are underutilised. The second paradigm provides a parameter-efficient solution, as the size of the weight matrices only doubles. Therefore, learning feature representations separately from unimodalities and fusing them afterwards ensure a cost-effective solution with full use of inputs.

After the above comprehensive analysis, in this work, the authors adopt the second paradigm. Firstly, the authors train three independent unimodality models. Secondly, the feature representations from speech command, hand motion, and body motion are concatenated into a single feature representation. Further details are presented in Sections III and IV.

## III. METHODOLOGY

This section presents the problem definition and the explanation of the adopted solution. As reviewed in Section II, the multimodal HRC recognition problem can be treated as a typical machine learning problem. Therefore, the authors formulate the problem to facilitate a machine learning solution. The formulated problem is further solved with three separated unimodal models. Lastly, the multimodal fusion is presented.

### A. PROBLEM STATEMENT

In this section, the multimodal HRC problem is formulated as a multiclass classification problem. $\mathcal{D}$ represents the collected data from sensors. The data consists of $m$ samples drawn from an unknown distribution of the feature space and class label space $\mathcal{X} \times \mathcal{C}$, denoted by $\mathcal{D} = \{(\mathbf{X_i}, c_i)\}_{i=1}^{m}$. $\mathbf{X}$ is the feature vector where $\mathbf{X_i} = \{x_1, x_2, \ldots, x_m\}$ and $c$ is a finite set of categorical labels with $k$ categories. $\mathbf{X_i}$ is assigned to a certain categorical label $c_i$ according to an underlying unknown function $f : \mathbf{X_i} \rightarrow c_i$.

This paper aims to find a hypothesis $g$, a classifier, from the hypothesis space $\mathcal{H}$ that best approximates the true function $f$. Since $f$ is unknown, the true error in approximation is unavailable. However, the classification error $\mathcal{J}$ can be measured empirically by running classifier $g$ over the data samples $\mathcal{D}$. Thus, the optimisation problem can be solved by minimising the empirical error,

$$\min_{g \in \mathcal{H}} \mathcal{J}_{\mathcal{D}}(g) \tag{1}$$

where

$$\mathcal{J}_{\mathcal{D}}(g) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{g(\mathbf{X_i}) \neq c_i\} \tag{2}$$

where $\mathbb{I}$ is a characteristic function.

### B. UNIMODALITY REPRESENTATION

After the problem statement, the authors select three deep learning models to recognise three modalities collected from sensors. As introduced in Section II, there are different deep learning solutions available for all the three modalities. The authors select deep learning models according to their performance on the unimodal dataset. The three selected deep learning models are CNN for speech command recognition, LSTM for hand motion recognition and transfer learning-enabled body motion recognition.

#### 1) CNN FOR SPEECH COMMAND RECOGNITION

As introduced in Section II-A, CNN is a neural network consisting of several convolutional and pooling layers (max-pooling). CNN outperforms traditional rule-based feature extraction approaches in terms of robustness and efficiency. It has been successfully applied to a wide range of applications including speech command recognition [53]–[55]. As suggested by Sainath and Parada [54], the spectrum of audio input has strong correlations in both time and frequency axis. Capturing local correlations with CNN through weight sharing has been shown to be favourable in many applications [55], whereas models such as SVM and Multilayer Perceptron (MLP) ignore the temporal dependency of the speech signal. Therefore, in this study, the authors adopt CNN for speech recognition.

In CNN speech command recognition, the speech command dataset $X^S$ is transformed into 2-dimensional spectrograms by fast Fourier transform (FFT) [53] before being processed by CNN. The convolution operation is formulated as

$$y^{(j)} = ReLU\Big(\sum_{i} a^{(ij)} \cdot x^{(i)} + b^{(j)}\Big) \tag{3}$$

where $x^{(i)}$ and $y^{(j)}$ denote the $i$-th input map and the $j$-th feature map respectively. $x^{(i)}$ is a local region where weights are shared among each convolution neuron $a^{(ij)}$. $a^{(ij)}$ denotes the convolution neuron between the $i$-th input map and the $j$-th feature map. $b^{(j)}$ denotes the bias of convolution neuron $a^{(ij)}$. *ReLU* ($y = \max(0, x)$) activation function was proved to be better than *sigmoid* activation [56]. Max-pooling outputs the maximum value of each of the local neighbour (e.g., a 2 × 2 pixel grid of an image). Max-pooling makes each feature map invariant to local translations in the input map, which also proved to be useful in CNN [57], [58].

In model training, the authors adopt categorical cross-entropy as the cost function $\mathcal{J}$, defined as:

$$\mathcal{J} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\log\left(p_{ij}\right) \tag{4}$$

where $y_{ij}$ is the binary indicator of whether the observation $X_i^S$ is of class $c_i$, $p_{ij}$ is the predicted probability of whether observation $X_i^S$ corresponds to class $c_i$, $n$ is the number of training samples, $k$ is the number of categorical labels.

### 2) LSTM FOR HAND MOTION RECOGNITION

In hand motion recognition, the dataset consists of hand motion time series with strong sequential dependencies within each hand motion. The details of the data format are explained in Section IV-A2.

LSTM is an RNN that controls how information flows across the internal states through multiplicative gate units in the neural networks [37]. Due to its advantage of learning long-term dependencies, LSTM is currently the desired model for several sequence processing tasks such as machine translation [59] and text generation [60]. At time $t$, a typical LSTM cell $c_t$ contains three gate units: input gate $i_t$, forget gate $f_t$ and output gate $o_t$, connected with recurrent and feed-forward links. The final state $h_t$ is dominated by the cell output gate $o_t$. The passed by information is selectively accumulated in the cell, which enables the possibility to remember and refer to previous information. Our implemented LSTM, closely follows [61], is showed below:

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}\circ c_{t-1} + b_i\right) \tag{5}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}\circ c_{t-1} + b_f\right) \tag{6}$$

$$c_t = f_t\cdot c_{t-1} + i_t\cdot\tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{7}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}\circ c_t + b_o\right) \tag{8}$$

$$h_t = o_t\circ\tanh\left(c_t\right) \tag{9}$$

where $\circ$ is Hadamard product, $\sigma$ is the sigmoid function, $W_{ij}$ is the $i$ to $j$ weight matrix.

By stacking LSTM layers on top of one another, the output of the previous LSTM layer can be the input of the next LSTM layer. The stacked LSTM model learn the temporal features at a higher level [62]. Due to such advantage, the authors adopt the stacked LSTMs approach.

### 3) TRANSFER LEARNING-ENABLED BODY MOTION RECOGNITION

The body motion dataset is available in the format of video clips. To recognise body motion from the video clips, there are many different sequence recognition models available. As mentioned in Section II-A, recently, transfer learning has emerged as an effective approach for image recognition. By reusing the knowledge from other networks, the training time of a new model can be reduced. Since a video clip can also be sampled as image sequences, it is possible to utilise transfer learning to recognise body motion from video clips.

Transfer learning can be defined by a domain $\mathcal{I}$, a task $\mathcal{T}$, a learning source $S$, and a target source $T$ [41], [42]. The domain $\mathcal{I}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(\mathbf{X})$. Given a domain $\mathcal{I} = \{\mathcal{X}, P(\mathbf{X})\}$, the task $\mathcal{T} = \{\mathcal{C}, f(\cdot)\}$ is represented by a label space $\mathcal{C}$ and a predictive function $f(\cdot)$, which can be learned from the training dataset $\{(\mathbf{X_i}, c_i)\}_{i=1}^{n}$, where $n$ is the number of training samples. The learning of the target predictive function $f_T(\cdot)$ in the target domain $\mathcal{I}_T$ can be improved by utilising the knowledge learned from the source domain $\mathcal{I}_S$ and the source learning task $\mathcal{T}_S$.

In the specific case of body motion recognition, a function $f_S(\cdot)$ is trained from the source domain $\mathcal{I}_S$ and the source task $\mathcal{T}_S$ with a large amount of labelled images. Since the source domain $\mathcal{I}_S$ is in image format and the target domain $\mathcal{I}_T$ is in video format, the data of the target domain $\mathcal{I}_T$ can be sampled as sequences of images where the pre-trained network $f_S(\cdot)$ can act as a generic feature extractor to transfer the knowledge represented by parameters learned from the source domain $\mathcal{I}_S$ [40]. After the transfer learning-enabled feature extraction, the representation of training dataset can be denoted as $r^{B_1}\left(X_i^B;\theta^{B_1}\right)$ where $\theta^{B_1}$ are the parameters transfered from $f_S(\cdot)$ and $X_i^B$ denotes the training samples from body motion dataset. To fine-tune the target function $f_T(\cdot)$, a fully connected layer is added. The network is further trained by minimising the cross-entropy loss, before being fed into a softmax layer for nomalisation [63]:

$$\min_{W^f,\theta^{B_1}}\sum_{i=1}^{n}\mathcal{L}\left(\mathit{softmax}\left(W^f r^{B_1}\left(X_i^B;\theta^{B_1}\right)\right), c_i\right) \tag{10}$$

where $W^f$ are the weights of the softmax layer, $c_i$ are the categorical labels from the training dataset.

## C. MULTIMODAL FUSION

As introduced in Section II-B, multimodal models can potentially generate more knowledge on a classification problem and outperform unimodal models. To achieve a better classification result, the authors decide to fuse the above trained unimodal models. An example of multimodal fusion is illustrated in Figure 2. The speech command recognition model, hand motion recognition model, and body motion recognition model without the last layer can be represented as $r^S\left(X_i^S;\theta^S\right)$, $r^H\left(X_i^H;\theta^H\right)$, and $r^{B_2}\left(X_i^B;\theta^{B_2}\right)$, where $X_i^S$, $X_i^H$, and $X_i^B$ are the training samples from speech command recognition dataset, hand motion dataset, and body motion dataset. $\theta^S$, $\theta^H$, and $\theta^{B_2}$ are the network parameters from speech command recognition model, hand motion recognition model, and body motion recognition model.

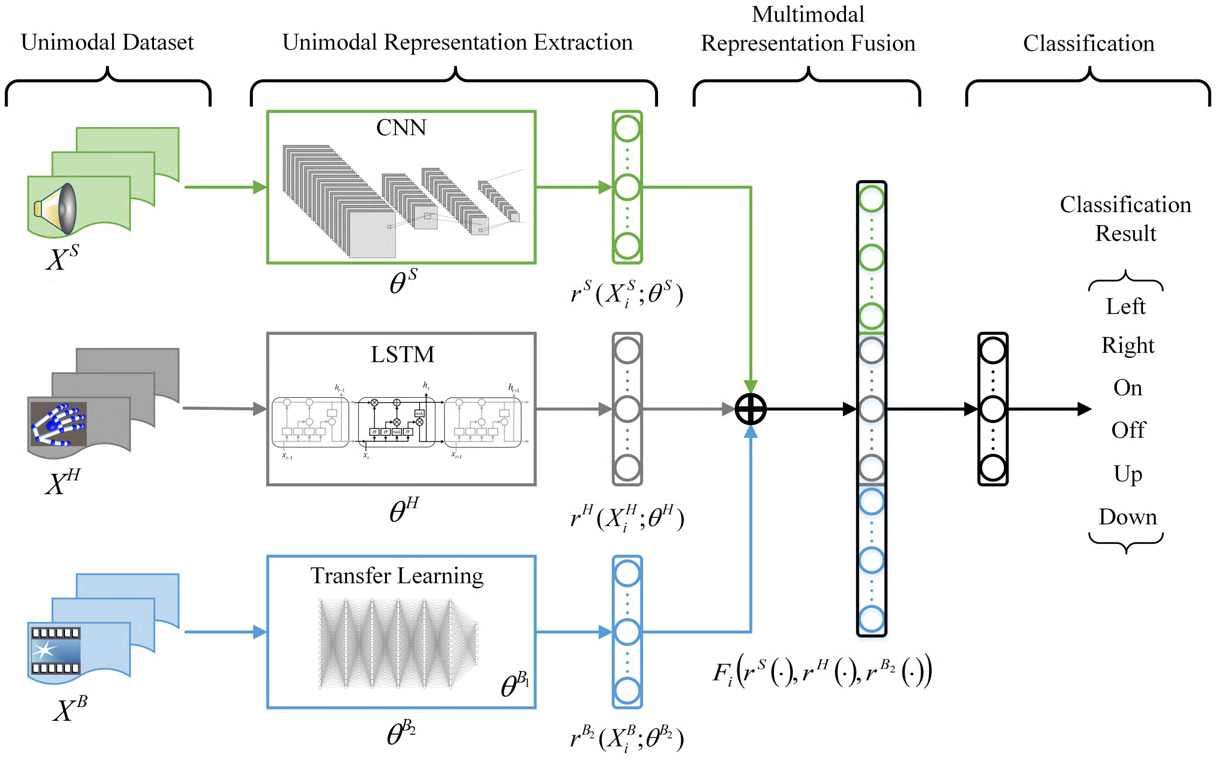


**FIGURE 2.** Illustration of the proposed multimodal fusion architecture.

The three trained models are fused by a concatenate function $F$:

$$\mathcal{G} = F\left(r^S\left(X_i^S;\theta^S\right), r^H\left(X_i^H;\theta^H\right), r^{B_2}\left(X_i^B;\theta^{B_2}\right)\right) \quad (11)$$

where $\mathcal{G}$ is the representation after the concatenation. The fused model can be further trained by minimising a loss function defined by cross-entropy. Finally, the optimised network is connected to a softmax function to normalise the output result. The training process of the fused model is represented by:

$$\min_{W^F,\theta^S,\theta^H,\theta^{B_2}} \sum_{i=1}^{n} \mathcal{L}\left(softmax\left(W^F\mathcal{G}\right), c_i\right) \quad (12)$$

where $W^F$ denotes the weights of the softmax layer after multimodal fusion.

## IV. EXPERIMENT

In this section, the authors first describe the dataset from each modality and the necessary data cleaning procedures. The authors then illustrate the multimodal dataset preparation and multimodal model training process. Furthermore, the training results of unimodal models are compared with their baseline models. At last, the authors present the result of multimodal fusion and the comparison of which with other unimodal models.

### A. CUSTOMISED MULTIMODAL DATASET AND PREPROCESSING

#### 1) SPEECH COMMAND RECOGNITION

The Speech Command Dataset [64] is chosen as the dataset for the training of speech command recognition. The dataset, created by TensorFlow and AIY teams at Google, consists of over 65000 one-second audio recordings of 30 short words. A subset of 6 categories of speech commands is selected and labelled specifically under the context of multimodal HRC, including *left, right, on, off, up,* and *down*. The total number of the selected audio recordings is 14178.

Each audio recording is essentially a 1-D vector of strength signals. To facilitate CNN model training, each audio recording is preprocessed and transformed into a 2-D matrix that can be treated as a single-channel image, as shown in Figure 3. Each speech command is segmented into 0.02s pieces with an overlapping of 0.01s. Each segment is then turned into a 1-D mel-frequency cepstral coefficient vector (MFCC). Stated differently, the MFCC vectors are generated along the time-axis of the audio input as an additional dimension.

#### 2) HAND MOTION RECOGNITION

The hand motion dataset includes 1183 sequences of hand motions with six different categorical labels. The dataset is HRC customised and collected by Leap Motion Controller [13]. The Leap Motion Controller captures the direction and orientation of key hand joints and bones in a

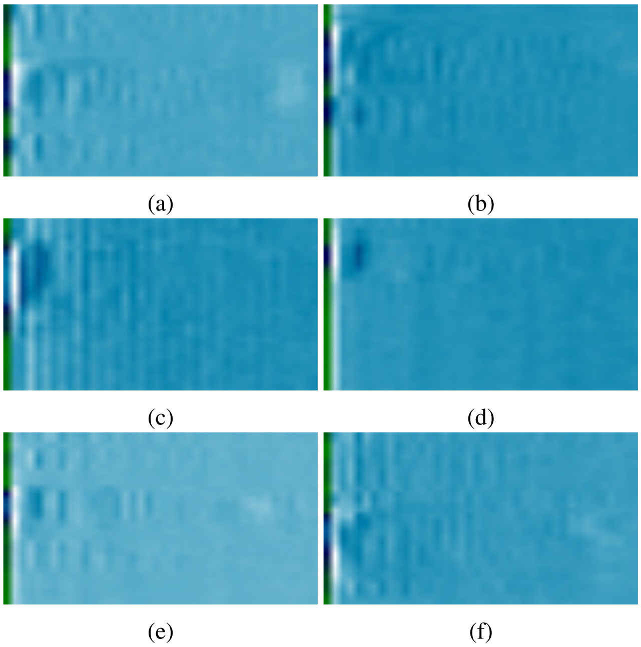


**FIGURE 3.** Visualized MFCC representations of the six selected speech commands. (a) Left. (b) Right. (c) On. (d) Off. (e) Up. (f) Down.

frequency of 100Hz. The tracked hand bones are metacarpal, proximal, intermediate, distal and the tips of five fingers. With the bones and joints captured, robust hand skeleton models can be built during the hand motion. The parameters from the skeletons are regarded as hand motion features. In each time stamp of the hand motion, 64 hand motion features are captured. To facilitate multimodal fusion, the six hand motions are defined the same as that in speech command recognition dataset and body motion recognition dataset. An example of collected hand motions is showed in Figure 4.

### 3) BODY MOTION RECOGNITION

The body motion recognition dataset consists of 1379 HRC customised body motion video clips with six different categorical labels. The defined six motions are exactly the same as that in hand motion recognition dataset, shown in Figure 4. For the purpose of transfer learning, the collected video clips are further sampled into sequences of images. The transfer learning process extracts features from the image sequences by Inception-v3 [43] pre-trained model. The processed image sequences are stored and prepared as the input data for training multimodal fusion model. As introduced in Sections II-A and III-B3, the transfer learning approach to reusing a pre-trained model as the feature extractor will greatly reduce the model training time.

## B. MULTIMODAL FUSION TRAINING

As mentioned above, the experiment contains three modalities: speech command, hand motion and body motion. To train a fused model, data from the same label within the three unimodal datasets are sampled randomly without semantic change. Therefore, the potential size of the training set could be extremely large. As previously mentioned, three unimodal models are trained to extract the representations from the second-last layer of each unimodal model. Thereafter, the representations of the three modalities are concatenated and fed into another MLP, resulting in a network architecture shown in Figure 2.

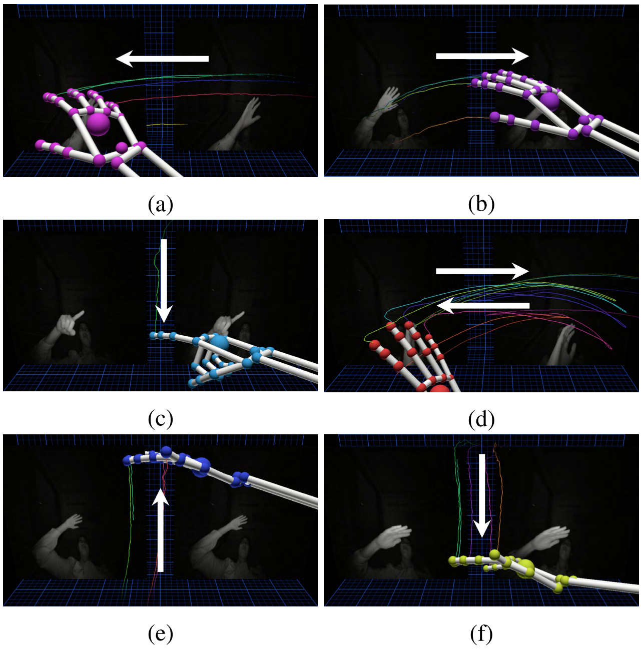


**FIGURE 4.** Visualisations of the six selected hand motion commands. The hand skeletons are representations of captured hands in motion, where the colourful joints are the tracked hand joints and the white arrows point out the directions of hand motion. (a) Left. (b) Right. (c) On. (d) Off. (e) Up. (f) Down.

The multimodal fusion architecture can be trained in two different manners: the weights of the network is marked either "trainable" (i.e., without weight lock) or "non-trainable" (i.e., with weight lock). Trainable weights in the multimodal fusion model are updated during the training of the final multimodal MLP classifier. In the case of non-trainable weights, the weights in unimodal models are fixed and only the weights in the multimodal MLP classifier can get updated during the training of the multimodal fusion model.

## C. RESULTS OF UNIMODAL MODELS

In this section, the authors present the results of unimodal models in comparison with their baseline models. The baseline model for speech command, hand motion and body motion recognition is SVM, Random Forest and Random Forest respectively, whereas the adopted approach is CNN, LSTM and MLP via transfer learning. As shown in Table 1,

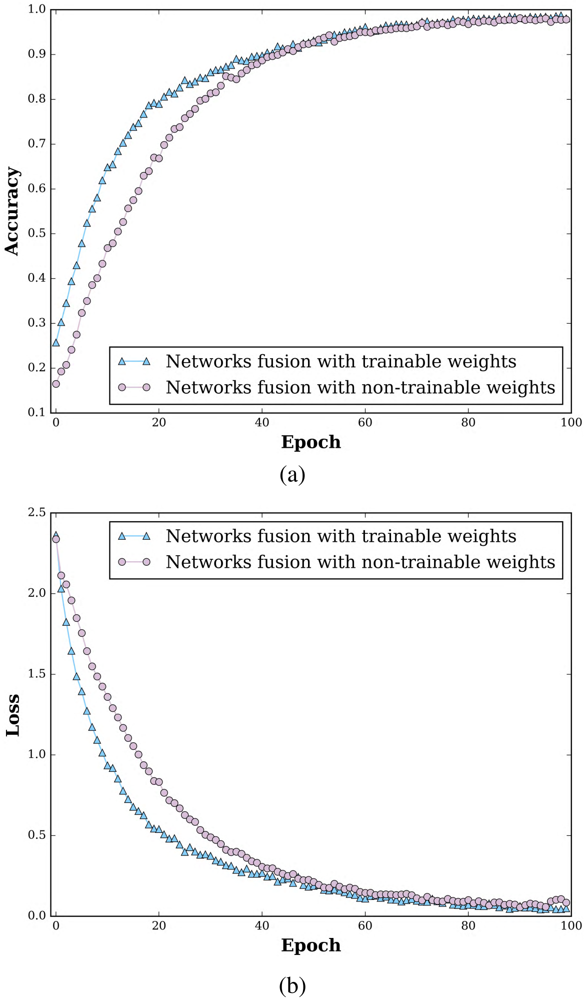
**TABLE 1.** Test accuracy of each unimodality with baseline models and adopted models.

| Model/Accuracy | Speech command | Hand motion | Body motion |
|---|---|---|---|
| Baseline | 71.00% | 94.98% | 93.06% |
| Adopted model | 93.83% | 98.24% | 95.95% |

the performance of adopted approaches is better than the baseline models in all unimodalities. In hand motion recognition, the predicted accuracy on the test dataset is improved from 94.98% to 98.24%, and in body motion recognition, the accuracy is improved from 93.06% to 95.95%. Particularly for speech command recognition, the adopted approach significantly boosts the accuracy from 71.00% to 93.83%. Possible reasons for the results are further discussed in Section V.

### D. RESULT OF MULTIMODAL MODEL

Figure 5 presents the training accuracy and loss for the fused model with trainable and non-trainable weights. As can be seen from Figure 5a and Figure 5b, the accuracy and loss for fused model with trainable weights improves faster than the model with non-trainable weights. After around 40 epochs, both networks converge to the same level, around 98.07% in training accuracy and 0.0496 in loss. For the test dataset, the accuracy is 99.58%. To illustrate the performance on test dataset with different categorical labels, the authors plot the confusion matrix in Figure 6, where the differences between the true labels and the predicted labels derived from the fused model are shown. As shown in Figure 6, the fused model achieves an accuracy of 100% in all labels except that in the label *right* where 4% of the true label *right* are predicted as *left*.



(a)

(b)

**FIGURE 5. The performance of multimodal fusion training including the comparison of training progress with trainable and non-trainable weights. The training accuracy converge to 98.07%. The test accuracy is 99.58%. (a) Training accuracy for multimodal fusion. (b) Training loss for multimodal fusion.**

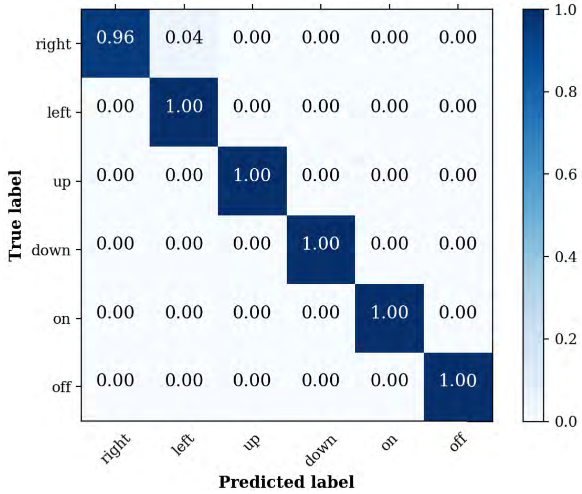
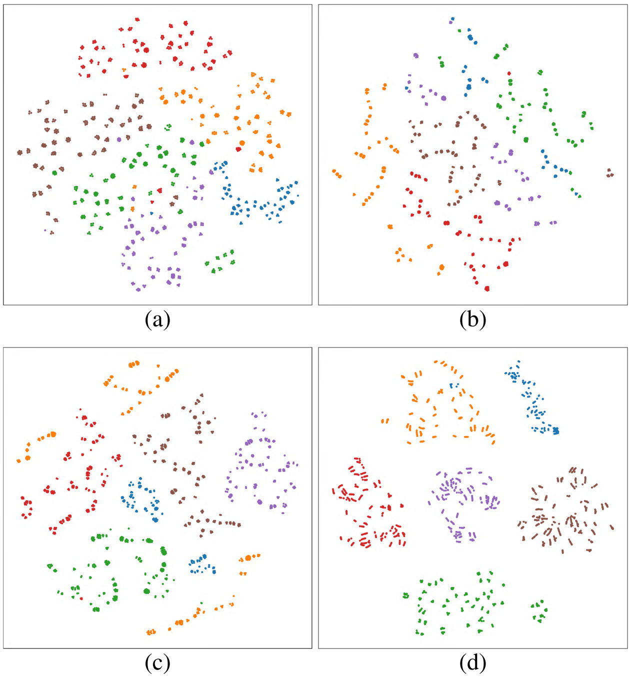


**FIGURE 6. Confusion matrix of the fused model.**

The authors further visualize the hidden representations with the test dataset using t-SNE [65] for both unimodal models and the fused model in Figure 7. The hidden representations refer to the hidden distribution of the test dataset when the test dataset is applied to the trained model for prediction



(a)

(b)

(c)

(d)

**FIGURE 7. t-SNE visualisations of the hidden representations of the test dataset, where the six different colours represent predicted different labels. (a) Speech command model. (b) Hand motion model. (c) Body motion model. (d) Fused model.**

and evaluation. Each plotted test dataset corresponds to its trained model. For instance, the speech command test dataset is used on the trained speech recognition CNN to extract the hidden representations, whereas the multimodal representations are obtained from the fused model where the multimodal test dataset is applied. In Figure 7, each point denotes a data point from its corresponding test dataset and the six colours represent six categorical labels. Clearly, compared with the other hidden representations, the multimodal fusion representations are better separated. This reflects the fact that the multimodal fusion model outperforms all unimodal models in the experiment.

## V. DISCUSSIONS

Table 1 shows that the test accuracy of the adopted CNN, LSTM and MLP via transfer learning model surpass the SVM and Random Forest baseline models. The above-mentioned improvements confirm the literature review in Section II: the deep learning models outperform traditional machine learning models. One of the interesting observations is that the accuracy of speech command recognition significantly improves from 71.00% to 93.83%, whereas the accuracy of the other two modalities only improves by less than 5%. Part of the reason for the accuracy improvement differences can be the different data sample size. The speech recognition dataset involves 10 times more data samples than the other two modalities. Deep learning models capture some of the hidden patterns that cannot be recognised by traditional machine learning models in a large dataset. In the other two data modalities, since the datasets are not so large, most of the patterns can already be recognised by traditional machine learning models. Therefore, the improvement is not so obvious. However, as shown in Figure 7, the data inputs are not well classified in the unimodal models. The achieved accuracy still cannot guarantee a robust HRC system.

One of the topics worth discussing is the different approaches for multimodal fusion. For instance, the authors explore multimodal fusion in different network layers. The fused models' performance is extremely different in terms of accuracy, and the reason for the differences is unclear. However, the authors found out that with the current dataset, the fusion at the second-last layer still generates a parameter-efficient and well-performed model. As can be seen clearly in Figure 6, the test accuracy is further improved with the fused model. By comparing the hidden representations of test dataset of Figure 7d with Figure 7a, Figure 7b, and Figure 7c, the test set is best classified by the fused model shown in Figure 7d. With further observation, the 4% wrongly classified *right* data points can actually be found at the top middle part of Figure 7d, where several blue data points (in *right* label) are surrounded by a cluster of yellow data points (in *left* label). The reason for the accuracy improvement can be attributed to the richer knowledge representations learned by the multimodal representations fusion process. Although each unimodality cannot provide enough knowledge for a decision with high accuracy, multimodality collectively provide enough knowledge to make a much better decision with higher accuracy. In multimodal HRC, the result of this paper suggests that the sensor instability can be offset by adding more sensors and data modalities. Eventually, a robust HRC system can be achieved with several sensors. With further real-time HRC system implementation, the fused model can potentially provide better accuracy and be used in a robust multimodal HRC system.

The authors also explore the fusion process with trainable and non-trainable weights. As illustrated in Figure 5, during the multimodal fusion, the accuracy improves faster with trainable weights where the parameters of the unimodal models change together with the fusion. Eventually, the two approaches converge to the same accuracy level. The reason is that the search space of the classification problem increases, as the weights can change. Since the two approaches eventually find similar solutions, it is natural that the two approaches converge to the same accuracy level.

The proposed model fusion approach can potentially be beneficial for future multimodal HRC manufacturing system applied in industry. As discussed in Section I, one of the biggest differences between industrial machinery and consumer electronics is the reliability or robustness. For manufacturing systems, safety and reliability share the highest priority. In future HRC manufacturing systems, intuitive multimodal robot control is required. The multimodal fusion approach proposed in this paper can serve as a robust multimodality HRC system beyond intuitive requirement. The human operator can control the robot not only intuitively with human-friendly modalities but also robustly. The robot control command will be active only when all three modalities are in place and the fusion model generates a right output. In the case of missing modalities, the proposed approach will not output any result. In order to realise the above mentioned robust HRC manufacturing system, a real-time HRC system and sensors need to be developed.

## VI. CONCLUSIONS AND FUTURE WORKS

In this paper, the authors propose a multimodal fusion architecture for robust multimodal HRC manufacturing systems. To prepare the model fusion, three unimodal models are firstly trained to extract the representations from three modalities, i.e. speech command, hand motion and body motion. In particular, three selected unimodal models are trained to fit three modalities dataset: a CNN model for speech command dataset, an LSTM model for hand motion dataset and an MLP via transfer learning model for body motion dataset. After the unimodality model training, the three models are further concatenated and fused. Experiments demonstrate the accuracy of the fused model in comparison with the unimodal models. The discussions indicate a great potential to apply the proposed approach in future multimodal HRC manufacturing systems.

As future works, possible directions can be: (i) to implement the proposed architecture in a real-time scheme; (ii) to further investigate theoretical aspects and the model architecture to explain the neural networks; and (iii) to encompass different modalities for HRC robot control.

## ACKNOWLEDGMENT

The authors express great appreciation to Prof. Magnus Boman for his valuable and constructive suggestions during the planning and development of this work.

## REFERENCES

[1] H. Liu and L. Wang, "Gesture recognition for human-robot collaboration: A review," *Int. J. Ind. Ergonom.*, vol. 68, pp. 355–367, Nov. 2017.

[2] C. Kardos, Z. Kemény, A. Kovács, B. E. Pataki, and J. Váncza, "Context-dependent multimodal communication in human-robot collaboration," *Procedia CIRP*, vol. 72, pp. 15–20, May 2018.

[3] H. Liu and L. Wang, "Human motion prediction for human-robot collaboration," *J. Manuf. Syst.*, vol. 44, pp. 287–294, Jul. 2017.

[4] H. Liu, T. Fang, T. Zhou, Y. Wang, and L. Wang, "Deep learning-based multimodal control interface for human-robot collaboration," *Procedia CIRP*, vol. 72, pp. 3–8, May 2018.

[5] T. Brogårdh, "Present and future robot control development—An industrial perspective," *Annu. Rev. Control*, vol. 31, no. 1, pp. 69–79, 2007.

[6] W. Ji, Y. Wang, H. Liu, and L. Wang, "Interface architecture design for minimum programming in human-robot collaboration," *Procedia CIRP*, vol. 72, pp. 129–134, May 2018.

[7] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-robot interface," *IEEE Intell. Syst.*, vol. 16, no. 1, pp. 16–21, Jan. 2001.

[8] B. Yao, Z. Zhou, L. Wang, W. Xu, Q. Liu, and A. Liu, "Sensorless and adaptive admittance control of industrial robot in physical human-robot interaction," *Robot. Comput.-Integr. Manuf.*, vol. 51, pp. 158–168, Jun. 2018.

[9] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[11] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[13] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[14] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 116–134, Oct./Nov. 2007.

[15] A. Cherubini, R. Passama, A. Meline, A. Crosnier, and P. Fraisse, "Multimodal control for human-robot cooperation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2013, pp. 2202–2207.

[16] T. Belpaeme *et al.*, "Multimodal child-robot interaction: Building social bonds," *J. Hum.-Robot Interact.*, vol. 1, no. 2, pp. 33–53, 2013.

[17] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. (2016). "Realtime multi-person 2D pose estimation using part affinity fields." [Online]. Available: https://arxiv.org/abs/1611.08050

[18] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 5048–5054.

[19] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1990, pp. 129–132.

[20] M.-C. Silaghi and H. Bourlard, "Iterative posterior-based keyword spotting without filler models," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop*, 1999, pp. 213–216.

[21] M.-C. Silaghi, "Spotting subsequences matching an HMM using the average observation probability criteria with application to keyword spotting," in *Proc. AAAI*, 2005, pp. 1118–1123.

[22] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A hidden Markov model-based continuous gesture recognition system for hand motion trajectory," in *Proc. 19th Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2008, pp. 1–4.

[23] C. Keskin, A. Erkan, and L. Akarun, "Real time hand tracking and 3D gesture recognition for interactive interfaces using HMM," in *Proc. ICANN/ICONIPP*, 2003, pp. 26–29.

[24] S. Tabibian, A. Akbari, and B. Nasersharif, "An evolutionary based discriminative system for keyword spotting," in *Proc. Int. Symp. Artif. Intell. Signal Process. (AISP)*, Jun. 2011, pp. 83–88.

[25] K.-P. Feng and F. Yuan, "Static hand gesture recognition based on HOG characters and support vector machines," in *Proc. 2nd Int. Symp. Instrum. Meas., Sensor Netw. Autom. (IMSNA)*, Dec. 2013, pp. 936–938.

[26] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," in *Proc. 9th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, May/Jun. 2012, pp. 28–32.

[27] A. S. Micilotta, E.-J. Ong, and R. Bowden, "Detection and tracking of humans by probabilistic body part assembly," in *Proc. BMVC*, no. 1, 2005, pp. 429–438.

[28] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping," in *Proc. VISAPP*, vol. 1, 2013, pp. 620–625.

[29] T. Arici, S. Celebi, A. S. Aydin, and T. T. Temiz, "Robust gesture recognition using feature pre-processing and weighted dynamic time warping," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 3045–3062, 2014.

[30] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4277–4280.

[31] D. Palaz and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proc. INTERSPEECH*, no. EPFL-CONF-210029, 2015.

[32] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Proc. INTERSPEECH*, 2013, pp. 1173–1175.

[33] S. Fernández, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *Proc. Int. Conf. Artif. Neural Netw.* Springer, 2007, pp. 220–229.

[34] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[35] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. Int. Workshop Hum. Behav. Understand.* Springer, 2011, pp. 29–39.

[36] J. Nagi *et al.*, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Nov. 2011, pp. 342–347.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 194–197.

[39] W. Zhu *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI*, vol. 2, no. 5, 2016, pp. 3697–3703.

[40] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[41] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[42] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1855–1862.

[43] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

[44] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[45] R. Socher, B. Huval, B. Bhat, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 656–664.

[46] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for RGB-D based object recognition," in *Experimental Robotics*. Springer, 2013, pp. 387–402.

[47] A. Schindler and A. Rauber, "An audio-visual approach to music genre classification through affective color features," in *Proc. Eur. Conf. Inf. Retr.* Springer, 2015, pp. 61–67.

[48] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Proc. 7th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2008, pp. 688–693.

[49] R. Neumayer and A. Rauber, "Integration of text and audio features for genre classification in music information retrieval," in *Proc. Eur. Conf. Inf. Retr.* Springer, 2007, pp. 724–727.

[50] M. Dorfer, A. Arzt, and G. Widmer. (2016). "Towards score following in sheet music images." [Online]. Available: https://arxiv.org/abs/1612.05050

[51] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, nos. 4–5, pp. 705–724, 2015.

[52] S. Oramas, O. Nieto, F. Barbieri, and X. Serra. (2017). "Multi-label music genre classification from audio, text, and images using deep features." [Online]. Available: https://arxiv.org/abs/1707.04916

[53] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 559–563.

[54] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1478–1482.

[55] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun./Jul. 2004, pp. 97–104.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[57] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2559–2566. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2010.5539963

[58] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2146–2153.

[59] 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.

[60] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.

[61] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[62] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, *Long Short Term Memory Networks for Anomaly Detection in Time Series*. Leuven, Belgium: Presses Universitaires de Louvain, 2015, p. 89.

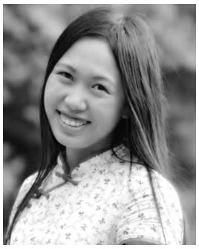
[63] P. Wang, H. Liu, L. Wang, and R. X. Gao, "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration," *CIRP Ann.*, vol. 67, no. 1, pp. 17–20, 2018.

[64] P. Warden. (2018). "Speech commands: A dataset for limited-vocabulary speech recognition." [Online]. Available: https://arxiv.org/abs/1804.03209

[65] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

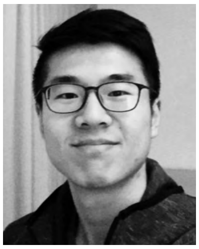
**HONGYI LIU** received the B.S. degree from Southeast University, Nanjing, China, in 2012, and the M.S. degree from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2015, where he is currently pursuing the Ph.D. degree. In 2013, he was a Research Assistant with ETH Zurich, Zurich, Switzerland. From 2013 to 2014, he was a Data Scientist with ABB Robotics, Västerås, Sweden.

His research interests are focused on human–robot collaboration, deep learning, human-motion perception, and data analytic in manufacturing. During his Ph.D. study, he has been actively involved in several EU and Sweden funded projects, such as SYMBIO-TIC and COROMA. He has authored 10 scientific publications.



**TONGTONG FANG** received the B.S. degree in statistics from Southwest University, Chongqing, China, in 2016. She is currently pursuing the double M.S. degree in data science with the KTH Royal Institute of Technology and Nice Sophia Antipolis University. She was exchanged to study at Paul Sabatier University, Toulouse, France, from 2014 to 2015.

In 2015, she was a Research Intern with Paul Sabatier University, independently responsible for measuring determinants in households' livelihood strategy by statistical modeling. Since 2017, she has been involved with the research project of integrating up-to-date deep learning algorithms in human–robot collaboration. In the near future, she will be a Ph.D. candidate motivated to solve fundamental problems in machine learning, particularly in variational inference, few-shot learning, meta-learning, and so on.



**TIANYU ZHOU** received the B.S. degree in software engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016, and the M.S. degree in data science from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2018. He is currently a Data Scientist with Scania CV AB, Stockholm, Sweden.

From 2017 to 2018, he was a Research Assistant with the Department of Production Engineering, KTH Royal Institute of Technology. His research interests include human–robot collaboration, deep learning, and machine learning.

**LIHUI WANG** is currently a Chair Professor with the KTH Royal Institute of Technology, Stockholm, Sweden. He has published eight books and authored in excess of 450 scientific publications. He is actively engaged in various professional activities. His research interests are focused on cyber-physical systems, cloud manufacturing, predictive maintenance, real-time monitoring and control, human–robot collaboration, and sustainable manufacturing systems. He is a fellow of CIRP, SME, and ASME, a registered Professional Engineer in Canada, and the Board Director of the North American Manufacturing Research Institution of SME. He is also the Editor-in-Chief of the *International Journal of Manufacturing Research* and the *Robotics and Computer-Integrated Manufacturing*, the Editor of the *Journal of Intelligent Manufacturing*, and the Associate Editor of the *Journal of Manufacturing Systems* and the *International Journal of Production Research*.