Learning Adaptive Hidden Layers for Mobile Gesture Recognition

Ting-Kuei Hu, Yen-Yu Lin, Pi-Cheng Hsiu
Research Center for Information Technology Innovation, Academia Sinica, Taiwan
Email: {tkhu, yylin, pchsiu}@citi.sinica.edu.tw

Abstract

This paper addresses two obstacles hindering advances in accurate gesture recognition on mobile devices. First, gesture recognition performance is highly dependant on feature selection, but optimal features typically vary from gesture to gesture. Second, diverse user behaviors and mobile environments result in extremely large intra-class variations. We tackle these issues by introducing a new network layer, called an *adaptive hidden layer* (AHL), to generalize a hidden layer in deep neural networks and dynamically generate an activation map conditioned on the input. To this end, an AHL is composed of *multiple neuron groups* and *an extra selector*. The former compiles multi-modal features captured by mobile sensors, while the latter adaptively picks a plausible group for each input sample. The AHL is end-to-end trainable and can generalize an arbitrary subset of hidden layers. Through a series of AHLs, the great expressive power from *exponentially* many forward paths allows us to choose proper multi-modal features in a sample-specific fashion and resolve the problems caused by the unfavorable variations in mobile gesture recognition. The proposed approach is evaluated on a benchmark for gesture recognition and a newly collected dataset. Superior performance demonstrates its effectiveness.

Introduction

Gesture recognition aims to identify categories of handed gestures, which are sequences of distinct hand shapes. It provides an intuitive and convenient means for human-machine interaction in various mobile applications. Conventional vision-based approaches have delivered promising results in controlled settings. However, in the context of mobile and wearable devices, accurate gesture recognition becomes quite challenging due to frequently large and unfavorable variations caused by diverse user behaviors and ubiquitous environments.

Such challenges could potentially be resolved through using extra information from sensing modalities (Jong et al. 2016), such as IR proximity sensors (Butler, Izadi, and Hodges 2008), magnetic field sensors (Hwang, Ahn, and Wohn 2013), and accelerometers (Zaen et al. 2014) on mobile devices. Approaches based on multi-modal learning, e.g. (Ngiam et al. 2011), can enhance recognition accuracy by leveraging complementary multi-modal information. However, two major issues hinder the development of these approaches for mobile gesture recognition. First, these approaches seek an immutable combination of multi-modal information, but the optimal modality typically varies from gesture to gesture, as an example shown in Figure 1. Second, mobile contexts suffer from large modality-specific environment variations. For example, visual features from cameras are sensitive to lighting conditions or viewing angles. By contrast, motion features from accelerometers are easily affected by the plane on which the gesture is performed.

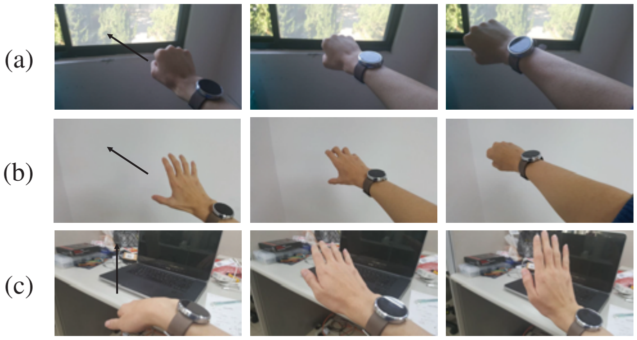


Figure 1: Gestures of categories (a) *fist*, (b) *fetch*, and (c) *palm-up*. Visual modalities are effective to separate gestures of category *fist* from the rest owing to their distinctive hand shapes. In contrast, motion-based modalities are discriminative for gestures of category *palm-up* due to the consistent moving trajectories. Meanwhile, gestures of class *fetch* are better predicted by using both types of modalities jointly.

In this work, we propose an approach based on *deep neural networks* (DNNs), which are characterized by the effectiveness in joint feature extraction and nonlinear classifier learning. Specifically, we introduce a new network layer, called an *adaptive hidden layer* (AHL), which generalizes a hidden layer in DNNs and can dynamically generate an appropriate activation map for a given input. Unlike a conventional hidden layer equipped with a single group of neurons, an AHL is composed of *multiple neuron groups* and *an extra selector*. The neuron groups in AHL extract diverse features and are optimized so that each training data can be well processed by at least one group, while the selector that implements softmax normalization can dynamically pick a plausible group for each input sample.

The proposed AHL is a differential module which makes the overall network end-to-end trainable. It is general in the sense that it can generalize a subset or all of the hidden layers in an arbitrary DNNs framework. Each AHL maintains multiple neuron groups and picks a group for an input. The number of parameters increases linearly in the resultant network. However, for an input passing through a series of AHLs, the number of possible *paths* grows *exponentially*. This great flexibility allows us to extract numerous combinations of intra-modal and inter-modal features and address large variations caused by the intrinsic diversity of user behaviors and the external variety of environments for mobile gesture recognition.

The proposed AHL is evaluated on one benchmark dataset for gesture recognition, the *ChaLearn LAP large-scale isolated gesture dataset* (IsoGD) (Wan et al. 2016), and a new dataset collected by us. Both datasets contain multi-modal gestures for recognition. The primary goal of evaluation on IsoGD is to compare the proposed approach with the existing ones in the same experimental settings, such as the adopted features and data splits. The experimental results show that using AHLs results in the state-of-the-art performance. We collected a new multi-modal gesture dataset, another contribution of this work, where the gesture samples were recorded with two different modalities and collected with different types of variations such as the lighting condition and the gesture performing plane. We aim at analyzing how AHL leverages multi-modal features to improve recognition on this dataset. The proposed AHL is applied to a DNNs-based approach (Ngiam et al. 2011) for multimodal learning. It turns out that AHL can make the most of the complementary multi-modal information in an input-specific fashion, leading to a significant boost in mobile gesture recognition performance.

Related Work

Two components, *feature extraction* and *classifier learning*, are crucial to the establishment of gesture and action recognition systems. In vision-based approaches, spatio-temporal features, e.g., (Feng, Xu, and Shapiro 2012; Song et al. 2014; Taylor et al. 2014; Tang et al. 2015), are widely adopted to characterize hand shapes and motion trajectories. Methods based on graphical models such as (Lee and Kim 1999; Chen, Fu, and Huang 2003; Lin et al. 2017) encode temporal dynamics and facilitate recognition. Non-vision-based approaches to gesture recognition, e.g. (Gupta et al. 2012; Pu et al. 2013), often rely on features generated via domain knowledge. The adopted features in the aforementioned approaches serve as the inputs to learn the classifiers for gesture recognition. However, these features are handcrafted and may provide suboptimal performance.

Deep learning offers the superior capability of extracting useful features from data and has been successfully applied to broad applications, e.g., activity recognition (Simonyan and Zisserman 2014; Hammerla, Halloran, and Plötz 2016), speech recognition (Hinton et al. 2012), saliency detection (Hsu, Lin, and Chuang 2017), object recognition (Shih et al. 2017), and ubiquitous computing applications (Kim, On, and Zhang 2016; Song, Kanasugi, and Shibasaki 2016), demonstrating high adaptability to various data statistics. Features in deep learning are derived to best optimize the objective of the coupled classifier, thus leading to superior performance.

Research efforts, e.g., (Ngiam et al. 2011; Srivastava and Salakhutdinov 2012; Sohn, Shang, and Lee 2014), have been made to generalize *deep autoencoders* (Hinton and Salakhutdinov 2006) and *deep Boltzman machines* (Salakhutdinov and Hinton 2009) to multi-modal learning. Empirical studies have shown that the mid-level features generated on modality-specific layers typically have lower within-modality correlation than the raw features, facilitating cross-modal fusion. Thus, these approaches first build the modality-specific layers, and then fuse the mid-level features across modalities on higher layers.

This aforementioned strategy is also commonly used for multi-modal gesture recognition. Li *et al*. (Li et al. 2016) proposed a *C3D-based model* (Tran et al. 2015) to train two seperate sub-networks from RGB and depth modalities respectively, and used the concatenated mid-level features as the input to the SVM classifier for final prediction. Zhu *et al*. (Zhu et al. 2017) presented a gesture recognition approach combining the C3D-based model and convolutional LSTM, and fused different modalities with equal weights. Neverova *et al*. (Neverova et al. 2016) proposed the *ModDrop* model by introducing a random dropping technique on different channels to learn cross-modal correlations. However, most methods of this class use fixed weights for multi-modal feature fusion, and neglect the fact that optimal features for recognition differ from gesture to gesture. To address this issue, the proposed AHL learns multiple neuron groups for diverse feature fusions, and employs a selector to adaptively select a plausible group condition input.

Recent studies have addressed the issue of large data variations by predicting the output conditioned on the input. For example, the tree-structured *convolutional neural networks* (CNNs) was presented in (Li et al. 2015) for facial trait recognition, where one of the learned sub-networks at a tree node is dynamically selected for each local facial patch. In (Xiong et al. 2015), the tree-structured CNNs were applied to face recognition, where the sub-networks and the split functions at the tree nodes can be jointly optimized. The drawback of the tree-structured CNNs models is that the number of sub-networks increases exponentially with the tree level, making training a deep architecture infeasible. Our approach provides a feasible means of leveraging the expressive power enabled by the exponentially many paths like the tree-structured models because its parameters grow linearly, with a leading coefficient of around 2 in our case.

The Proposed Approach

We introduce our approach *adaptive hidden layer* (AHL) in this section. Given a target DNNs architecture, our goal is to generalize a subset or all of the hidden layers to AHLs so that the adaptation power of AHLs can be leveraged to better handle the large variations in mobile gesture recognition. In

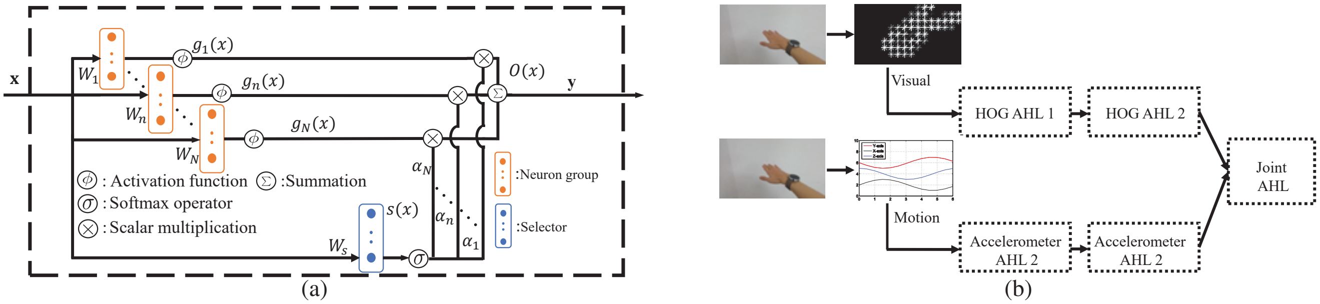


Figure 2: (a) AHL. (b) The network architecture generalized by AHLs for gesture recognition on the dataset we collected.

the following, the forward pass and the backward propagation of an AHL are firstly described. Then, we show that an AHL can serve as the building block for network construction. Finally, some implementation details are specified.

Forward Pass

Consider a hidden layer of M neurons. It maps an input $\mathbf{x} \in \mathbb{R}^D$ to the output $\mathbf{y} \in \mathbb{R}^M$, and can be formulated by

$$\mathbf{y} = \phi(W * \mathbf{x} + \mathbf{b}) \in \mathbb{R}^M, \tag{1}$$

where $W \in \mathbb{R}^{M \times D}$ is the learnable weight matrix, $\mathbf{b} \in \mathbb{R}^M$ is the bias, and $*$ is the matrix multiplication operator. The activation function ϕ is the rectified linear unit in this work.

Built on a hidden layer, an AHL illustrated in Figure 2(a) is composed of N neuron groups and one selector. The neuron groups are designed to account for the large variations among training data by compiling different features, i.e.,

$$g_n(\mathbf{x}) = \phi(W_n * \mathbf{x} + \mathbf{b}_n) \in \mathbb{R}^M, \text{ for } n = 1, 2, ..., N, \tag{2}$$

where $W_n \in \mathbb{R}^{M \times D}$ and $\mathbf{b}_n \in \mathbb{R}^M$ are the weight matrix and the bias of the nth neuron group g_n. When jointly learning the neuron groups and the selector, the learnable parameters $\{W_n, \mathbf{b}_n\}$ of each group g_n will be optimized for data falling into this group.

The selector adaptively selects an appropriate neuron group for input $\mathbf{x}$. It takes the input features, i.e. $\mathbf{x}$, as the input, and generates a vector $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_N]^\top \in \mathbb{R}^N$, where α_n can be interpreted as the degree of fitness of applying neuron group g_n to input $\mathbf{x}$. The selector, which implements softmax normalization as an approximation of selection, is defined by

$$\boldsymbol{\alpha} = \sigma(W_s * \mathbf{x} + \mathbf{b}_s) \in \mathbb{R}^N, \tag{3}$$

where $W_s \in \mathbb{R}^{N \times D}$ and $\mathbf{b}_s \in \mathbb{R}^N$ are the parameters of the selector. Softmax normalization operator σ is used, therefore $\alpha_n \geq 0$ and $\sum_{n=1}^N \alpha_n = 1$. It is worth mentioning that few neuron groups suffice for a deep network architecture, since stacking AHLs has made the number of the forward paths grow exponentially. The performance gain by setting N larger 2 is minor in the empirical test. Thus, we set N to 2 in this work. The number of the learnable parameters becomes roughly doubled, because the number of the parameters in the selector is far less than that in a neuron group and is neglectable.

By jointly considering the N neuron groups and the selector, the output of $\mathbf{x}$ in the AHL is a weighted combination of the group-specific responses followed by computing the activation function, namely

$$\mathbf{y} = \sum_{n=1}^N \alpha_n g_n(\mathbf{x}) \in \mathbb{R}^M. \tag{4}$$

Backward Propagation

We show that an AHL consisting of multiple neuron groups and an extra selector can be learned by stochastic gradient descent. To this end, the gradient of the objective function ℓ with respect to the parameters, i.e., $(\frac{\partial \ell(\mathbf{x})}{\partial W_n}, \frac{\partial \ell(\mathbf{x})}{\partial \mathbf{b}_n})$ for the nth neuron group and $(\frac{\partial \ell(\mathbf{x})}{\partial W_s}, \frac{\partial \ell(\mathbf{x})}{\partial \mathbf{b}_s})$ for the selector, is required. The objective ℓ for network learning is set to maximize the multinomial logistic regression in this work.

The gradient for the nth neuron group is firstly considered. We have $\frac{\partial \ell(\mathbf{x})}{\partial W_n} = \frac{\partial \ell(\mathbf{x})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial o(\mathbf{x})} \frac{\partial o(\mathbf{x})}{\partial W_n}$ by using the Chain rule, where $o(\mathbf{x}) = \sum_{i=1}^N \alpha_i g_i(\mathbf{x})$. Terms $\frac{\partial \ell(\mathbf{x})}{\partial \mathbf{y}}$ and $\frac{\partial \mathbf{y}}{\partial o(\mathbf{x})}$ can be derived by referring to the literature (Rumelhart, Hinton, and Williams 1988) and Eq. (4), so we detail the derivation of $\frac{\partial o(\mathbf{x})}{\partial W_n}$ as follows:

$$\frac{\partial o(\mathbf{x})}{\partial W_n} = \frac{\partial \sum_{i=1}^N \alpha_i g_i(\mathbf{x})}{\partial W_n} \tag{5}$$

$$= \alpha_n \frac{\partial g_n(\mathbf{x})}{\partial W_n} \tag{6}$$

The derivation from Eq. (5) to Eq. (6) is based on the product rule. The term $\frac{\partial g_n(\mathbf{x})}{\partial W_n}$ in Eq. (6) can be computed according to the definition of $g_n(\mathbf{x})$ in Eq. (2).

The process of derivation of $\frac{\partial \ell(\mathbf{x})}{\partial W_n}$ is completed. The derivation of $\frac{\partial \ell(\mathbf{x})}{\partial \mathbf{b}_n}$ can be obtained similarly. It follows that the gradient of the nth neuron group $\{\frac{\partial \ell(\mathbf{x})}{\partial W_n}, \frac{\partial \ell(\mathbf{x})}{\partial \mathbf{b}_n}\}$ is available. We repeat this procedure for each neuron group, and update all the N neuron groups accordingly.

The gradient for updating the selector is also derived by using the Chain rule, namely $\frac{\partial \ell(\mathbf{x})}{\partial W_s} = \frac{\partial \ell(\mathbf{x})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial o(\mathbf{x})} \frac{\partial o(\mathbf{x})}{\partial W_s}$. The part $\frac{\partial \ell(\mathbf{x})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial o(\mathbf{x})}$ has been obtained when the gradient for the neuron groups was computed. With auxiliary variable $s(\mathbf{x}) = W_s * \mathbf{x} + \mathbf{b}_s$, we compute $\frac{\partial o(\mathbf{x})}{\partial W_s}$ by

$$\frac{\partial o(\mathbf{x})}{\partial W_s} = \frac{\partial \sum_{i=1}^{N} \alpha_i g_i(\mathbf{x})}{\partial W_s} = \sum_{i=1}^{N} \frac{\partial \alpha_i}{\partial W_s} g_i(\mathbf{x}) \quad (7)$$

$$= \sum_{i=1}^{N} \frac{\partial \alpha_i}{\partial s(\mathbf{x})} \frac{\partial s(\mathbf{x})}{\partial W_s} g_i(\mathbf{x}) \quad (8)$$

$$= \sum_{i=1}^{N} \frac{\partial \alpha_i}{\partial s(\mathbf{x})} \frac{\partial (W_s * \mathbf{x} + \mathbf{b}_s)}{\partial W_s} g_i(\mathbf{x}), \quad (9)$$

where term $\frac{\partial \alpha_i}{\partial s(\mathbf{x})}$ in Eq. (9) is attainable by referring to Eq. (3) while the rest can be computed straightforward. The gradient of the bias $\frac{\partial \ell(\mathbf{x})}{\partial \mathbf{b}_s}$ can be similarly derived. After we get the gradient $(\frac{\partial \ell(\mathbf{x})}{\partial W_s}, \frac{\partial \ell(\mathbf{x})}{\partial \mathbf{b}_s})$, the selector is updated with it.

We summarize the gradient derivation by showing the dependence relationships among variables in Figure 2(a), including input $\mathbf{x}$, learnable parameters $\{W_n\}$ and W_s, auxiliary variables $s(\mathbf{x})$ and $o(\mathbf{x})$, and output $\mathbf{y}$. Since the gradient of all parameters in the AHL is available for an input $\mathbf{x}$, the network with the integration of the AHLs remains end-to-end trainable and can be efficiently optimized by stochastic gradient descent.

Discussion. In forward pass, the selector dynamically selects a suitable neuron group for each training data. In backward propagation, all neuron groups and the selector are optimized according to a joint objective. While each of the neuron groups is optimized for the training data assigned to it, the selector is derived to re-assign the training data to the updated neuron groups. In this way, each sample would be better taken into account by a specific neuron group.

An issue of learning the mixture of neuron groups is regarding the balancing of the clustered data sizes. We introduce an entropy-based function, called *selection balancing regularizer* (SBR), to regularize the training of the selector. SBR penalizes the cases where the selector assigns most data to a small subset of neuron groups. Consider each training sample $\mathbf{x}$ in a batch. Its distribution $\boldsymbol{\alpha} = [\alpha_n | 1 \leq n \leq N]$ computed by the selector via Eq. (3) indicates the neuron group to which it is assigned. Let $p(n)$ denote the probability that a data sample is assigned to the nth neuron group, for $n \in \{1, 2, ..., N\}$. $p(n)$ is estimated by summing α_n of every sample in the batch, and is normalized so that $\{p(n) | 1 \leq n \leq N\}$ is a probability distribution. Then, the SBR is defined by

$$-\beta \sum_{n=1}^{N} (p_n + \epsilon) log(p_n + \epsilon), \quad (10)$$

where ϵ is a small positive constant used for robust entropy computation. The SBR in Eq. (10) is added to the loss function for learning the selector. Parameter β controls the strength of this term. In this work, we set β to 1 in all the experiments. Because the SBR is differentiable with respect to all learnable parameters, the whole network is still end-to-end trainable and can be optimized by stochastic gradient descent.

Generalization

Recent studies of deep learning showed that the earlier hidden layers in deep models tend to extract low-level features while the later layers tend to detect high-level concepts (Zhou et al. 2015). The proposed AHL picks an appropriate neuron group conditioned on the input. For further generalization, the AHL can act as the building block to generalize multiple hidden layers of a DNNs model so that both low-level and high-level features are adaptively selected to enhance recognition.

In this work, we improve mobile gesture recognition with multi-modal signals captured by mobile sensors. Hence, the proposed AHLs are used in multi-modal deep learning network architectures, and adaptively fuse the multi-modal signals. For evaluation on the IsoGD dataset, we consider Zhu *et al.*'s network (Zhu et al. 2017), which is the current state-of-the-art method for this dataset. An AHL is added to the last layer of each modality-specific sub-network, and one AHL is further constructed on the top of the modality-specific sub-networks to capture the cross-modal information. For evaluation on the dataset we collected, AHLs are applied to the model in (Ngiam et al. 2011). Figure 2(b) shows the resultant network architecture, where two modality-specific sub-networks are firstly constructed and one joint sub-network is then built on the top of them.

Both architectures form directed acyclic graphs, so the back-propagation algorithm is applicable for efficient gradient computation. With AHLs, not only intra-modal but also cross-modal adaption is carried out. As shown in the experiments, the adaptive flexibility for sample-specific feature selection makes the most of complementary multi-modal features, and greatly facilitates mobile gesture recognition.

Implementation Details

This work is implemented based on the *Theano* library, and is evaluated on two different datasets, i.e., the IsoGD dataset and the dataset we collected. On the IsoGD dataset, we follow the setting of the applied network (Zhu et al. 2017). The number of neurons in each group is 512 for the AHL applied to the last layer of each modality-specific sub-network, and it is 249 for AHL on the top of modality-specific sub-networks. The batch size and weight decay are set to 13 and 0.00004, respectively. The base learning rate is initialized to 0.1 and dropped by a factor of 2 every 5 epochs. At most 100 epochs are conducted.

On the dataset we collected, the two modality-specific sub-networks only connect to their respective signal modalities. One is the visual signal extracted by the *histogram of oriented gradient* (HOG), and the other is the motion signal recorded by 3-axis accelerometers. The modality-specific mid-level features are concatenated and act as the input to the joint sub-network. Table 1 summarizes the sizes of the five AHLs in the three sub-networks shown in Figure 2(b). In optimization, the batch size and the weight decay are set to 128 and 0.0001 respectively. The initial learning rate is

layer	# of neuron groups	# of neurons in each group
HOG layer 1	2	64
HOG layer 2	2	128
accelerometer layer 1	2	64
accelerometer layer 2	2	128
joint layer	2	256

Table 1: The size of each adaptive hidden layer (AHL).

set to 0.1 and decreased by a factor of 2 every 10 epochs. The learning procedure stops at the 100th epoch.

For network initialization, the parameters of the neuron groups are initialized by using the method in (Ngiam et al. 2011), while the parameters of the selector are initialized by using the method in (Glorot and Bengio 2010). In addition, we use k-means clustering to partition the training data at the first epoch for learning the neuron groups.

Experimental Results

The performance of the proposed adaptive hidden layer (AHL) is evaluated in this section. We first describe the datasets used for performance evaluation, which are IsoGD dataset and the gesture dataset that we collected on mobile devices. Then the quantitative results are reported and analyzed with the comparison of our approach to some baselines and other competing approaches on both datasets. Finally, the effect of stacking multiple AHLs and the analysis of how the AHLs behavior with different degrees of data variations are explored on the dataset we collected.

Dataset Description

IsoGD Dataset It is a large-scale collection of 47,933 RGB-D gesture videos of 249 classes performed by 21 individuals. Each video represents a gesture sample and contains two types of modalities, i.e., the RGB data and depth data. This dataset is divided into the training subset, validation subset, and testing subset respectively. Due to the lack of the data labels in the testing subset, our approach and the competing approaches are trained on the training subset and evaluated on the validation subset.

Our Gesture Dataset It contains gestures of 14 common categories as shown in Fgiure 3. Because the public datasets for gesture recognition were collected in controlled environments or lacked multi-modal information captured by mobile sensors, we collected a new multi-modal gesture dataset. Seven participants, six right-handers and one left-hander, were recruited for dataset construction. The participants wore smart watches and performed the gestures in front of mobile devices. These gestures were recorded with both the videos captured by the cameras of the smart phones and the 3-axis acceleration (ACCE) captured by the accelerometers of the smart watches.

Each participant performed each class of the gestures three times. In addition to the variations yielded by users (the seven participants), we simulated the real-world situations by introducing four types of environmental variations in dataset collection. Two types of variations related to the visual modality are the lighting condition and the background. The other two types of variations related to the motion modality are the angle of the performing plane and the degree of wear tightness. Each type of variations has two different states, such as clean background vs. cluttered background. Thus, there are totally 16 environmental settings when the four types of variations are jointly involved. It follows that the dataset contains 4,704 (= 7 participants ×14 classes ×16 settings ×3 times) gestures. We randomly split the gestures of a class into two equal-size groups, one for training and one for testing, 10 times. The average performance is measured with the 10 splits. The large variations make this dataset quite challenging. Nevertheless, a problem like this can serve as a good testbed to evaluate the proposed approach for adaptive multi-modal learning.

approach	accuracy
C3D (RGB only)	37.30%
C3D+ConvLSTM (RGB only)	43.88%
DEEP C3D+ConvLSTM (RGB only)	43.56%
ours (RGB only)	**44.88%**
C3D (Depth only)	40.50%
C3D+ConvLSTM (Depth only)	44.66%
DEEP C3D+ConvLSTM (Depth only)	47.99%
ours (Depth only)	**48.96%**
C3D (RGB+Depth)	49.20%
C3D+ConvLSTM (RGB+Depth)	51.02%
DEEP C3D+ConvLSTM (RGB+Depth)	51.40%
ours (RGB+Depth)	**54.14%**

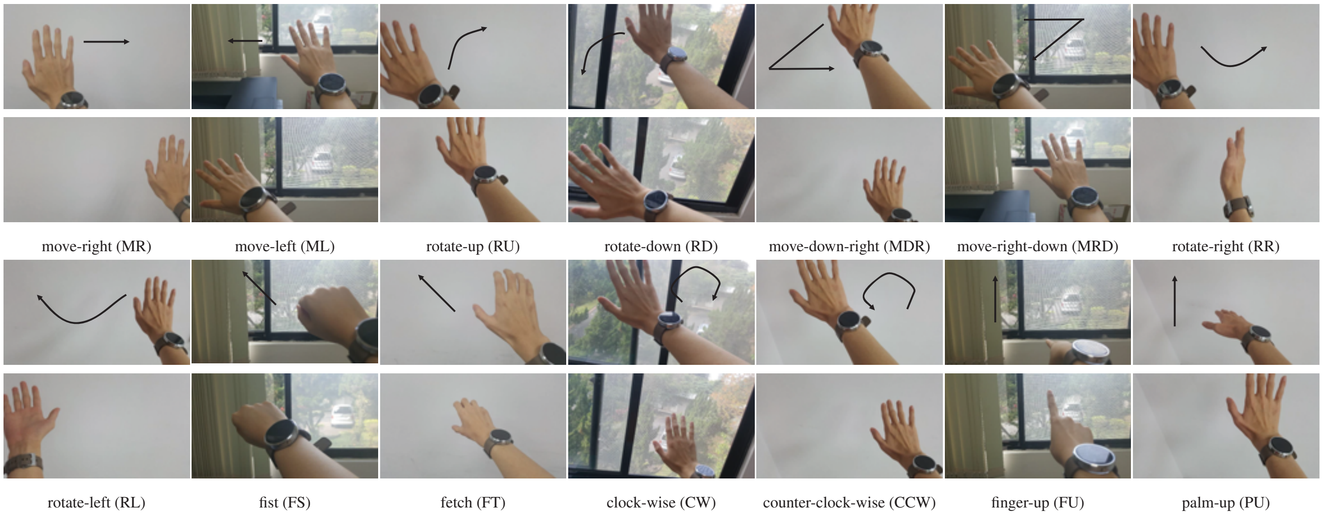
Table 2: Recognition rates on the IsoGD dataset.

For compact representations of raw data, we extracted the features from both the visual and motion modalities of the gestures. For the visual modality, we uniformly sampled five frames from each video, and computed the HOG features of the five frames. The numbers of cells and orientations and the block size are set to 15, 4, and 4 respectively. The HOG features of the five frames are concatenated, yielding a 1,200-dimensional feature vector. For the motion modality, we kept the raw signal of the five sampled frames, and computed the mean and the standard deviation for each temporal window of size 3 along each axis. After further including the magnitudes and the absolute differences, the resultant feature vector is of dimension 53.

Performance Comparison

Our approach is compared to those of the following two categories for performance analysis:

Single-modal learning: Approaches of this categories recognize gestures by using features generated from a single modality. `C3D+ConvLSTM` (Zhu et al. 2017) is the state-of-the-art method for the IsoGD dataset. It combines 3D-CNNs and convolutional LSTM to train two modality-specific sub-networks. We applied the AHL to the last layer of each of its sub-networks. To make a fair comparison in terms of the number of learnable parameters, we also established `DEEP C3D+ConvLSTM` which is the same as `C3D+ConvLSTM`, except that the modality-specific sub-network is learned with one additional conventional hidden



move-right (MR) move-left (ML) rotate-up (RU) rotate-down (RD) move-down-right (MDR) move-right-down (MRD) rotate-right (RR)

rotate-left (RL) fist (FS) fetch (FT) clock-wise (CW) counter-clock-wise (CCW) finger-up (FU) palm-up (PU)

Figure 3: The gesture dataset we collected. This figure shows one example from each of the fourteen gesture categories, including the first frame, the motion direction, and the category and abbreviation.

layer. The number of neurons of this additional hidden layer is twice that of each group of the added AHL in our proposed network. Hence, the learnable parameters of DEEP C3D+ConvLSTM and our proposed network are nearly the same. For our collected dataset, we learned the *deep autoencoder* (DAE) (Vincent et al. 2010) with three hidden layers and used the algorithm of stochastic gradient descent to fine-tune the network. The resulting approaches are denoted by DAE+HOG and DAE+ACCE respectively, when the visual and motion modalities are considered individually. These single-modal approaches act as the competing baselines to evaluate the importance of using multi-modal features.

Multi-modal learning: Approaches of this category recognize gestures by using features of multiple modalities. On IsoGD, two approaches C3D+ConvLSTM (RGB+Depth) and DEEP C3D+ConvLSTM (RGB+Depth) are used. Both approaches accomplish multi-modal learning by late fusion. That is, the final prediction is obtained by averaging the predictions of the modality-specific subnetworks. On the dataset that we collected, the baseline, DAE+conc. feat., is constructed by using DAE with the concatenated visual and motion features as input. Ngiam et al. (Ngiam et al. 2011) proposed a DAE-based architecture for multi-modal learning, multi-modal DAE, where modality-specific sub-networks are firstly built, and a joint sub-network is then established on the top of the modality-specific sub-networks. Approach multi-modal DAE is a representative and powerful framework for multi-modal learning. Although recent advances in multi-modal DAE such as (Srivastava and Salakhutdinov 2012; Sohn, Shang, and Lee 2014) have been made, they focused on text features in form of sparse or binary vectors. Thus, these approaches are not applicable to our cases. As mentioned previously, the proposed AHL generalizes each hidden layer of multi-modal DAE. Thus, the number of learnable parameters becomes double. For fair comparison, a variant of multi-modal DAE is yielded by doubling the neuron number of each layer.

Table 2 reports the recognition rates of our approach and the competing methods on the IsoGD dataset. Although the methods, C3D+ConvLSTM (RGB+Depth) and DEEP C3D+ConvLSTM (RGB+Depth), can effectively improve the accuracy by fusing the evidence computed with individual modalities, it still leaves space for improvement since equal-weight fusion is not adequate to account for data-specific, intra- and inter-modal variations. The proposed approach carries out adaptive feature selection for each input. It turns out that our approach not only improves single-modal learning accuracy, i.e., 44.88% for RGB and 48.96% for Depth, but also achieves the superior recognition rate of 54.14% for RGB+Depth to the state-of-the method C3D+ConvLSTM (RGB+Depth).

Table 3 reports the recognition rates of our approach and the competing methods on the collected gesture dataset. Methods DAE+HOG and DAE+ACCE give the recognition rates of 81.52% and 76.24%, respectively. The results indicate that none of the visual features and the motion features dominates the other. The baseline DAE+conc. feat. fuses the multi-modal information by feature concatenation. It results in a higher accuracy of 82.34%. The approach multi-modal DAE achieves much better accuracy of 86.48%. It implements modality-specific subnetworks to avoid the problems caused by the divergence among modalities. Meanwhile, it leverages both the intramodality and inter-modality correlations captured by different sub-networks to enhance the performance. The numbers of neurons in multi-modal DAE are tuned and set to those reported in Table 1. Hence, its double-sized variant does not increase the accuracy. It indicates that simply adding more learnable parameters does not help.

The proposed approach enables adaptive multi-modal learning by creating exponentially many forward paths for feature extraction. It achieves the recognition rate of 90.57%, and gives a significant performance gain of 4.09%

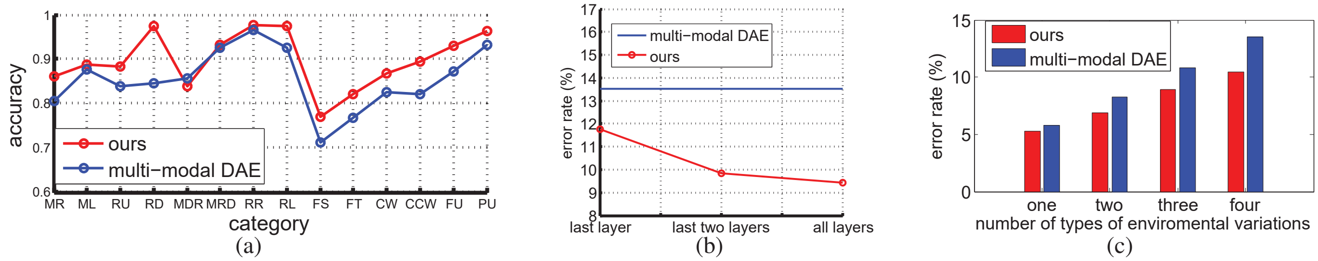


Figure 4: (a) The category-wise comparison between our approach and the competing method `multi-modal DAE`. (b) The error rates of our approach with different numbers of AHLs. (c) The performances of our approach and the competing method `multi-modal DAE` on data with different degrees of environmental variations.

approach	accuracy
DAE + HOG	81.52%
DAE + ACCE	76.24%
DAE + conc. feat.	82.34%
multi-modal DAE	86.48%
double-sized multi-modal DAE	86.02%
ours	**90.57%**

Table 3: Recognition rates on the dataset we collected.

$(= 90.57\% - 86.48\%)$ over the best competing approach `multi-modal DAE`. Note that the performance gain does not come from the increase of the network size in AHL, since our approach is also superior to the double-sized variant of `multi-modal DAE`, where the number of the learnable parameters is similar to that in our approach. The quantitative results confirm that our approach can make the most of the complementary multi-modal information, and remarkably outperform all competing methods.

To gain insight into the average accuracy, we report the category-wise recognition rates of our approach and `multi-modal DAE` in Figure 4(a). It is interesting to observe that the category-wise performance gains of our approach are quite diverse, namely from 12.94% in category *rotate-down* (RD) to -1.95% in category *move-down-right* (MDR). In general, both the visual and motion features are crucial for most of the 14 categories. Thus, `multi-modal DAE` learns a fixed set of multi-modal features to characterize most data. The two categories where our approach gives the most gains are RD and FT. The discriminative features in the two categories are quite different to those in the other categories. While the motion features are discriminative for category RD due to its distinctive moving pattern, the visual features are effective for category FT owing to the consistent hand shape. Our approach carries out adaptive multi-modal learning, and selects proper features for each gesture. Thus, it is considerably and consistently superior to `multi-modal DAE` in most categories.

The training and testing costs of AHL are about N times higher than those of a hidden layer, where N is the number of neuron groups. Nevertheless, stacking multiple AHLs results in exponentially many forward paths. Thus, on IsoGD, we stack just three AHLs with $N = 2$ on Zhu *et al.*'s network (Zhu et al. 2017) with only 10.1% and 15.8% extra computational costs in training and testing, respectively.

Network Generalization

Passing a data sample through a series of AHLs in a network enables the adaptive selection of both low-level and high-level features. We analyze the effect of stacking multiple AHLs to generalize the given network shown in Figure 2(b), which consists of two parallel modality-specific sub-networks. Therefore, we consider it a three-layer network. The proposed AHLs are applied to its last k hidden layers. By varying the value of k from 1 to 3, the error rates of the resultant networks are reported in Figure 4(b). For the ease of comparison, Figure 4(b) also plots the error rate of `multi-modal DAE`, which can be considered the special case when the value of k is set to 0.

Some observations can be found in the figure. First, when the joint layer (the last layer) of the network is replaced by an AHL, the error rate is greatly decreased from 13.52% to 11.75%. It points out that adaptive multi-modal feature selection is crucial for mobile gesture recognition. Second, when the modality-specific layers are also replaced by AHLs, the error rate is further reduced from 11.75% to 9.43%. It demonstrates that adaptive feature selection within individual modalities is helpful in dealing with the large data variations. Third, the monotonically decreasing error rates confirm that the features extracted by different AHLs are complementary. Therefore, stacking multiple AHLs allows dynamically selecting both intra-modal and inter-modal features, and results in an extra performance boost.

Effect of Data Variations

We address the issue of large data variations by adaptive multi-modal learning. Therefore, it is worthy to analyze how the proposed approach behaviors on data with different degrees of variations. Recall that the collected gesture dataset involves four types of environmental variations, including the lighting condition, the background, the gesture performing plane and the degree of wear tightness. Each variation type has two different states. A larger number of variation types leads to a higher degree of data variations.

We adjust the number of variation types to control the degrees of data variations. Suppose the number is set to

datsetset	approach	accurancy
IsoGD	C3D+ConvLSTM	51.02%
	Ours (without SBRs)	51.87%
	Ours (with SBRs)	**54.14%**
Our collected dataset	multi-modal DAE	86.48%
	Ours (without SBRs)	90.06%
	Ours (with SBRs)	**90.57%**

Table 4: The effects of using the proposed AHL and SBRs on the IsoGD dataset and our collected gesture dataset.

m. Namely, m out of the four variation types are considered while each of the rest is fixed to one of its two states. There are totally $C_m^4 \times 2^{4-m}$ such configurations, each of which corresponds to one particular data subset. We evaluate the average performance of our approach and `multi-modal DAE` on these configurations. By varying the value of m from 1 to 4, the error rates of the two approaches are shown in Figure 4(c). It is clear that the more the types of environmental variations, the larger the error rates of both approaches, but the performance gain of our approach over `multi-modal DAE` becomes more evident. The results show that adaptive multi-modal learning enabled by our approach is crucial in mobile applications where large variations are frequently present.

Ablation Study

To evaluate the effects of using the proposed AHL and the SBRs, we conduct the ablation study via removing SBRs and replacing AHLs with conventional hidden layers on the IsoGD dataset and our collected gesture dataset. Note that both the training and testing settings of each ablation experiment are kept exactly the same for fair comparison. We report the results when multi-modal data are used, as illustrated in Table 4.

Our observations are given as follows: 1) Apparent performance gains can be obtained with the aid of SBRs on both the IsoGD dataset and our collected dataset. 2) The effect of adding SBRs is dataset-dependent. While the performance of our approach on the IsoGD dataset drops remarkably when SBRs are removed, we can still achieve a significant performance gain by merely using AHL on our collected dataset. Our explanation is that the IsoGD datset has only visual modalities recorded by cameras and depth sensors. Thus, data of different visual modalities are sensitive to the same factors, and the selector tends to assign most data to a small subset of neuron groups due to random initialization. By adding SBRs, we penalize the unbalanced data distribution in terms of α, which helps to regularize the training process.

The gesture dataset that we collected and the source codes of the proposed AHL will be available at http://cvlab.citi.sinica.edu.tw/ProjectWeb/AHL/.

Conclusions

We have presented a new network layer, called the adaptive hidden layer (AHL), which is composed of multiple neuron groups and an extra selector. The former compiles different features to better characterize data of high variations, while the latter dynamically picks a plausible group for each input. We leverage the proposed AHL to enable adaptive multi-modal learning in DNNs, and illustrate it on mobile gesture recognition. The experimental results on two datasets show that by stacking multiple AHLs, our approach significantly outperforms all competing methods, and reaches the state-of-the-art performance. In the future, we will apply AHLs to generalize other network layers, such as the convolutional layers of CNNs for visual data processing and the internal layers of RNNs or LSTM for structured data prediction.

Acknowledgement. This work was supported by Ministry of Science and Technology under grants MOST 104-2628-E-001-003-MY3, MOST 105-2221-E-001-030-MY2, and MOST 105-2218-E-001-006.

References

Butler, A.; Izadi, S.; and Hodges, S. 2008. Sidesight: Multi-”touch” interaction around small devices. In *UIST*, 201–204.

Chen, F.-S.; Fu, C.-M.; and Huang, C.-L. 2003. Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Video Computing* 745–758.

Feng, A. W.; Xu, Y.; and Shapiro, A. 2012. An Example-Based Motion Synthesis Technique for Locomotion and Object Manipulation. In *SIGGRAPH*, 95–102.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 249–256.

Gupta, S.; Morris, D.; Patel, S.; and Tan, D. 2012. Soundwave: Using the doppler effect to sense gestures. In *CHI*, 1911–1914.

Hammerla, N. Y.; Halloran, S.; and Plötz, T. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *IJCAI*, 1533–1540.

Hinton, G., and Salakhutdinov, R. 2006. Reducing the dimensionality of data with neural networks. *Science* 504 – 507.

Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.-R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; and Kingsbury, B. 2012. Deep neural networks for acoustic modeling in speech recognition. *SPM* 82–97.

Hsu, K.-J.; Lin, Y.-Y.; and Chuang, Y.-Y. 2017. Weakly supervised saliency detection with a category-driven map generator. In *BMVC*.

Hwang, S.; Ahn, M.; and Wohn, K.-Y. 2013. Maggetz: Customizable passive tangible controllers on and around conventional mobile devices. In *UIST*, 411–416.

Jong, Y.-W.; Hsiu, P.-C.; Cheng, S.-W.; and Kuo, T.-W. 2016. A semantics-aware design for mounting remote sensors on mobile systems. In *DAC*.

Kim, E.; On, K.; and Zhang, B. 2016. Deepschema: Automatic schema acquisition from wearable sensor data in restaurant situations. In *IJCAI*, 834–840.

Lee, H.-K., and Kim, J. H. 1999. An HMM-based threshold model approach for gesture recognition. *TPAMI* 961–973.

Li, S.; Xing, J.; Niu, Z.; Shan, S.; and Yan, S. 2015. Shape driven kernel adaptation in convolutional neural network for robust facial trait recognition. In *CVPR*, 220–230.

Li, Y.; Miao, Q.; Tian, K.; Fan, Y.; Xu, X.; Li, R.; and Song, J. 2016. Large-scale gesture recognition with a fusion of RGB-D data based on the C3D model. In *ICPR*, 25–30.

Lin, S.-Y.; Lin, Y.-Y.; Chen, C.-S.; and Hung, Y.-P. 2017. Recognizing human actions with outlier frames by observation filtering and completion. *ACM TOMM*.

Neverova, N.; Wolf, C.; Taylor, G.; and Nebout, F. 2016. Moddrop: Adaptive multi-modal gesture recognition. *TPAMI* 1692–1706.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*, 689–696.

Pu, Q.; Gupta, S.; Gollakota, S.; and Patel, S. 2013. Whole-home gesture recognition using wireless signals. In *MobiCom*, 27–38.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1988. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*.

Salakhutdinov, R., and Hinton, G. 2009. Deep Boltzmann machines. In *AISTATS*, 448–455.

Shih, Y.-F.; Yeh, Y.-M.; Lin, Y.-Y.; Weng, M.-F.; Lu, Y.-C.; and Chuang, Y.-Y. 2017. Deep co-occurrence feature learning for visual object recognition. In *CVPR*.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 568–576.

Sohn, K.; Shang, W.; and Lee, H. 2014. Improved multimodal deep learning with variation of information. In *NIPS*, 2141–2149.

Song, J.; Sörös, G.; Pece, F.; Fanello, S. R.; Izadi, S.; Keskin, C.; and Hilliges, O. 2014. In-air gestures around unmodified mobile devices. In *UIST*, 319–329.

Song, X.; Kanasugi, H.; and Shibasaki, R. 2016. Deeptransport: Prediction and simulation of human mobility transportation mode at a citywide level. In *IJCAI*, 2618–2624.

Srivastava, N., and Salakhutdinov, R. R. 2012. Multimodal learning with deep boltzmann machines. In *NIPS*, 2222–2230.

Tang, N. C.; Lin, Y.-Y.; Hua, J.-H.; Wei, S.-E.; Weng, M.-F.; and Liao, H.-Y. M. 2015. Robust action recognition via borrowing information across video modalities. *IEEE TIP*.

Taylor, S.; Keskin, C.; Hilliges, O.; Izadi, S.; and Helmes, J. 2014. Typehoverswipe in 96 bytes: A motion sensing mechanical keyboard. In *CHI*, 1695–1704.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 4489–4497.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR* 3371–3408.

Wan, J.; Li, S. Z.; Zhao, Y.; Zhou, S.; Guyon, I.; and Escalera, S. 2016. Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In *CVPRW*, 761–769.

Xiong, C.; Zhao, X.; Tang, D.; Jayashree, K.; Yan, S.; and Kim, T.-K. 2015. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, 3667–3675.

Zaen, J. V.; Hausmann, J.; Salvi, K.; and Deriaz, M. 2014. Gesture recognition for interest detection in mobile applications. In *ICMCS*, 345–350.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2015. Object detectors emerge in deep scene CNNs. In *ICLR*.

Zhu, G.; Zhang, L.; Shen, P.; and Song, J. 2017. Multimodal gesture recognition using 3-d convolution and convolutional lstm. *IEEE Access* 4517–4524.