A Semi-Supervised Human Action Learning

Mohsen Tavana¹, Mohammad Mohammadi², Hamid Parvin³

1) Department of Computer Engineering, Mamasani Branch, Islamic Azad University, Mamasani, Iran

2) Young Researchers and Elite Club, Mamasani Branch, Islamic Azad University, Mamasani, Iran
tavana@mamasaniiau.ac.ir; mhp63@iauyasooj.ac.ir; parvin@iust.ac.ir

Received: 2015/10/13; Accepted: 2016/01/02

Abstract

Exploiting multimodal information like acceleration and heart rate is a promising method to achieve human action recognition. A semi-supervised action recognition approach AUCC (Action Understanding with Combinational Classifier) using the diversity of base classifiers to create a high-quality ensemble for multimodal human action recognition is proposed in this paper. Furthermore, both labeled and unlabeled data are applied to obtain the diversity measure from multimodal human action recognition. Any classifiers can be applied by AUCC as its base classifier to create the human action recognition model, and the diversity of classifier ensemble is embedded in the error function of the model. The model’s error is decayed and back-propagated to the basic classifiers through each iteration. The basic classifiers’ weights are acquired during creation of the ensemble to guarantee the appropriate total accuracy of the model. Considerable experiments have been done during creation of the ensemble. Extensive experiments show the effectiveness of the offered method and suggest its superiority of exploiting multimodal signals.

Keywords: Action Learning, Ensemble Learning, Machine Learning.

1. Introduction

One of the most considerable and valuable issues of ever-present computing is human action recognition, especially for user-centric cell phone applications including gaming, healthcare, and rehabilitation. Researchers have been attempting to recognize human action only from accelerometer data [1, 2, 3] or other certain environment data like electrocardiogram [4]. However, those approaches hardly attain acceptable performance in real-world applications because of the constraint of single circumstance information.

In general, we can divide offered attempts at using multimodal information for human action recognition into two groups: attribute based ensemble [5, 6, 7, 8] and classifier based ensemble [9, 10]. The first group has mostly focused on appending multimodal handcrafted attributes into a lengthy attribute vector, and then used a single basic classifier to achieve human action recognition. However, some of modality data have not been examined on attribute construction for human action recognition. Thus, the researchers have to either design new attributes [7] or employ related measures from other research fields [11]. Additional physiological or environmental cues for a pattern may be brought by multimodal attributes, but the attribute compatibility issue is left to
be solved [9]. This makes it difficult for a single classifier to differentiate the patterns across modalities.

The latter group combines the base classifiers constructed on different modalities separately as an alternative solution. Classifier based ensemble aims to enhance the generalization ability and accuracy of human action recognition. On the other hand, the base classifiers are usually over-fitted to generate correlated errors [12]. Moreover, most researches in classifier based ensemble still have to exploit hand-crafted attributes for the base classifiers. Diversity learning of classifier ensemble has shown its strong ability to ensure that all the base classifiers craft uncorrelated errors [13, 14]. Both diversity and accuracy have been considered in the error functions of most of works [15, 16, 17] to extract the diversity of classifier ensemble.

In this paper, AUCC (Action Understanding with Combinational Classifier), a neural network based semi-supervised action recognition approach is presented which exploits the diversity of classifier ensemble. The diversity measure, embedded in the error function of the model, can be obtained from both labeled and unlabeled data. The model’s error is decomposed and back-propagated to base classifiers in each iteration. Consequently, all the networks would be able to be trained simultaneously and interactively on the training dataset. The trained NNs can automatically find appropriate attribute representation. To further enhance the total recognition performance, multimodal networks are fused at classifier level with different weights. The major contributions of this work are summarized as follows:

1) Presenting a semi-supervised multimodal human action recognition approach which exploits the diversity of classifier ensemble based on neural network. So, the model benefits can be gained from both the most advances in neural networks and learning to diversity techniques in this method.

2) Utilizing stacking framework to realize classifier level fusion, which can ensure the total classification performance by assigning different weights to different base classifiers.

3) Appraising this approach on two benchmarked datasets and perform extensive comparison with other methods. The experimental results suggest that AUCC is able to yield a competitive human action recognition performance, and has its superiority on exploiting multimodal signals.

The rest of the paper is organized as follows. Section II describes the existing approaches related to multimodal human action recognition and learning to diversity. The proposed method is explained in section III. Section IV presents experimental results. Finally, Section V is the conclusion.

2. Related Work

The majority of early researches seek to recognize human action, being as a key field in ubiquitous computing, only from accelerometer data [1, 2, 3, 18, 19, 20, 21] or other data such as electrocardiogram [4]. Bao et al. [1] pioneered in measuring the performance of human action recognition system from acceleration sensors. Krishnan et al. [18] applied two accelerometers to recognize five activities. Likewise, [3, 21] also gave their human action recognition method based on accelerometers. CenceMe application [19] offered an action recognition engine with Nokia N95 phone. Kwapisz et al. [2] presented an exhaustive discussion on action recognition using phone-based accelerometers. Guo et al. [20] used coordination mapping and structural motifs to
perform a device orientation and placement invariant human action recognition based on smartphone. The main flaw embedded into those approaches is that acceleration or physiological data only is hardly to reach reliable performance in real-world applications. Hence, the way of exploiting multimodal information for action recognition becomes a hot research topic.

A. Multimodal human action recognition

Available efforts to utilize multimodal information for action recognition can be generally divided into two categories, i.e., attribute based ensemble [5, 6, 7, 8, 11, 22] and classifier based ensemble [9, 10].

Feature based ensemble is a straightforward consideration that concatenates multimodal information in attribute level. Kunze et al. [5] realized that combining the data of gyroscopes and accelerometers can effectively remove the negative impacts of sensor variability. Subramanya et al. [11] exploited data from accelerometer, microphone, phototransistor, temperature, and barometric pressure to recognize six activities. Combining accelerometers with physiological sensors [7, 8] or location sensors [22] is also investigated to improve recognition accuracy. Lara et al. [6] presented Centinela a system that fused acceleration data with physiological data to achieve the recognition of five ambulation activities. Centinela also conventionally combines different modalities in attribute level by handcrafting transient attributes from physiological data.

On the other hand, some modality data have not been investigated on attribute construction for human action recognition. Thus, the researchers have to either design new attributes [7] or employ related measures from other research fields [11]. In addition, most of these researchers have applied a single classifier to attain action recognition, and usually a single classifier cannot perform well with attributes from different sensors.

Classifier based ensemble aims to enhance the generalization ability and accuracy of action recognition by combining the base classifiers constructed on different modalities separately. Li et al. [9] combined the classification scores provided by support vector machines (SVM) and Gaussian mixture models (GMM), and applied them on acceleration and electrocardiogram data. Guo et al. [10] employed Heart Rate Variability (HRV) attributes extracted from physiological signals in action recognition and present a modified classifier level combination method based on the stacking framework. Due to uniform expressions (i.e., predictive scores) outputted from different classifiers, fusing in classifier level can address the attribute compatibility issues arising from different time shifts, window length configurations, and sampling frequencies. More importantly, classifier level fusion can adjust the classifier ensemble weights according to their classification confidence (accuracy) [10].

Unfortunately, most researches in classifier based ensemble still have to extract handcrafted attributes for the base classifiers. In addition, most classifier based ensemble methods in human action recognition focus their efforts on minimizing the classification loss. However, only optimizing accuracy would inevitably lead the over-fitting of the base classifiers, i.e., generating correlated errors.

B. Learn to Diversity of Ensemble

Diversity learning of classifier ensemble has shown its powerful ability to ensure that all the individual classifiers craft uncorrelated errors. There exists two ways to exploit diversity of classifier ensemble: only optimizing diversity measure [13] and trading off
between diversity and accuracy measures [15, 16, 17]. Since no research evidence has theoretically shown the clear correlation between diversity and accuracy in ensembles, most work in literature tend to make the tradeoff between diversity and accuracy measures, instead of only optimizing diversity measure. For instance, Yin et al. [15] attempted to utilize diversity for classifier selection and combination heuristically and iteratively. Yu et al. [23, 24] proposed the diversity regularized machine, which efficiently generates an ensemble of assorted SVMs. These work has effectively justified the usefulness of diversity learning to improve the performance of classifier ensemble.

In order to exploit further diversity measure, some other research is very insightful. Chen et al. [17] took diversity as regularized item to improve neural network ensembles. Unlabeled data were used in [12, 16] to enhance the diversity of classifier ensemble.

Inspired by the literatures, we exploit the diversity of classifier ensemble based on neural network in action recognition. AUCC firstly constructs a semi-supervised neural network ensemble, which would be able to not only automatically discover appropriate attribute representation but also utilize unlabeled data to generate diverse base classifiers. Based on the generated base classifiers, classifier level fusion is then employed to improve the generalization ability and accuracy of action recognition.

3. Methodology

Maximizing the generalization performance and minimizing the classification loss of base classifiers are two goals which multimodal human action recognition naturally bears as an ensemble learning process. To accomplish the goals, AUCC has two distinct process stages according to Fig. 1. Firstly, based on neural network ensemble, both the diversity and accuracy measures are employed to train a good ensemble (with accurate and diverse base classifiers). The diversity of base classifiers is able to enhance their generalization performance by minimizing uncorrelated errors. Secondly, the base classifiers are incorporated by classifier level fusion to ensure the overall classification performance.

![Fig. 1. The framework of AUCC.](image-url)
A. Problem Definition

Let $\mathcal{X}$ and $\gamma$ denote the space of inputs and the set of class labels, respectively. $y = \{y_i | 1 \leq i \leq c\}$, where $c$ is the number of class labels. Given training data set $\mathcal{S} = \mathcal{L} \cup \mathcal{U}$, where $\mathcal{L} = \{(x_i, y_i) | 1 \leq i \leq L\}$ contains $L$ labeled training examples and $\mathcal{U} = \{x_i | L+1 \leq i \leq L+U\}$ contains $U$ unlabeled training examples, $\mathcal{X}_f$ and $\gamma_f$. In addition, we use $\mathcal{D} = \{x_i | 1 \leq i \leq L\}$ to denote the unlabeled dataset derived from $\mathcal{L}$.

Suppose that the classifier ensemble is composed of $m$ base classifiers $f = \{f_k | 1 \leq k \leq m\}$, corresponding to different modal data. AUCC maximizes the diversity and accuracy measures of the classifiers on the labeled data $\mathcal{L}$, as well as the diversity of the classifiers on the unlabeled data $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$. The loss function of AUCC is as follows:

$$E(f, \mathcal{L}, \mathcal{D}) = E_y(f, \mathcal{L}) - \lambda E_d(f, \mathcal{D})$$  \hspace{1cm} (1)

Where, the first term $E_y(f, \mathcal{L})$ corresponds to the classification loss for label prediction (e.g., logistic); the second term $E_d(f, \mathcal{D})$ corresponds to the diversity loss of $f$ on a specified data set $\mathcal{D}$ (e.g., $\mathcal{D} = \mathcal{U}$); the non-negative parameter $\lambda$ controls the trade-off between the two terms. Similar loss functions with the combination of a classification loss and a penalty term (i.e., diversity) have been well investigated in ensemble learning [12, 25]. Furthermore, $e(f_k, \mathcal{L})$ denotes the empirical loss of the $k$th base classifiers $f_k$ on the labeled dataset $\mathcal{L}$. Then,

$$E_y(f, \mathcal{L}) = \frac{1}{m} \sum_{k=1}^{m} e_k(f_k, \mathcal{L})$$ \hspace{1cm} (2)

Many diversity measures are extensively acknowledged in classifier ensemble, e.g., Disagreement [26], Q-Statistics, Double Fault [27], Kappa [28], and Prediction Confidence [16]. However, until now no diversity measure has research evidences theoretically showing its correlation with the ensemble accuracy except negative correlation learning (NCL) [17]. In this paper, $E_d(f, \mathcal{D})$ is calculated by NCL as follows:

$$E_d(f, \mathcal{D}) = \frac{1}{m} \sum_{k=1}^{m} e_d(f_k, \mathcal{D})$$ \hspace{1cm} (3)

Where,

$$e_d(f_k, \mathcal{D}) = \frac{1}{2|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left( f_k(x) - \bar{f} \right)^2$$ \hspace{1cm} (4)

And $\bar{f} = \frac{1}{m} \sum_{k=1}^{m} f_k$ is the combination formulation of $\bar{f}$, which consists of accurate and diverse base classifiers, by minimizing the loss function in Equation (1) as follow:
B. Semi-supervised Neural Network Ensemble

Existing efforts to utilize multimodal information for action recognition have to encounter the attribute extraction step. However, some modality data have not been thoroughly investigated on attribute construction for human action recognition. Hence, the researchers have to either design new attributes [7] or employ related measures from other research fields [11]. To address the problem, AUCC uses NN to implement the base classifiers. It has been extensively demonstrated that the most advances in neural networks (i.e., deep neural networks) can automatically discover appropriate attribute representation.

![Fig. 2. The NN architecture of AUCC.](image)

A multilayer network with M layers of hidden units is employed in Fig. 2, for each modality of neural network ensemble. Thus, the base classifier $f_k^m (1 \leq k \leq m)$ is modeled as:

$$f_k^m (x) = S \left( w_k^M h_k^M (x) + b_k^M \right)$$  \hspace{1cm} (6)

Where, $w_k^M$ and $b_k^M$ are the weights and bias value for the output (Mth) layer, respectively. $S$ is a non-linear squashing function, e.g., sigmoid $(t) = 1 / (1 + e^{-t})$ and tanh$(\cdot)$. Typically the j-th layer $(1 \leq j \leq M)$ can be defined as:

$$h_k^j (x) = S \left( w_k^{j-1} h_k^{j-1} (x) + b_k^{j-1} \right), j > 1$$  \hspace{1cm} (7)

Where,
Here, a standard fully connected multilayer network as an example is taken. Other advance NN designs combined with prior knowledge about a particular problem (i.e., convolutional networks [29, 30]) can also be employed in AUCC.

Correspondingly, the first term $E_y (f, \mathcal{L})$ in Equation (1) is set to be the mean square error function $MSE(f_k, \mathcal{L})$ on the labeled dataset, which is commonly used to measure the empirical loss of NN. By decomposing the loss function $E(f, \mathcal{L}, \mathcal{D})$ into each neural network, the loss function of $f_k$ is as follows:

$$e_d (\mathcal{L}, \mathcal{D}) = MSE(f_k, \mathcal{L}) - \lambda e_d (f_k, \mathcal{D})$$

Where,

$$MSE(f_k, \mathcal{L}) = \frac{1}{2} \sum_{i=1}^{L} (f_k(x_i) - y_i)^2$$

Algorithm 1 Semi-supervised NN learning process of AUCC

**Input:** the training data set $TS = \mathcal{L} \cup \mathcal{U}$, the trade-off parameter $\lambda > 0$, the number of iterations $T > 0$.

**Output:** the set of modality-specific NNs: $MNN$

1. initialize $MNN$ as given or default network structures
2. $\mathcal{L} \leftarrow$ remove the labels of $\mathcal{L}$
3. repeat
4. Each $f_k$ in $MNN$ executes **Forward Propagation**:
5. use (6), (7) and (8).
6. calculate $\bar{t} = \frac{1}{m} \sum_{k=1}^{m} f_k$
7. Each $f_k$ in $MNN$ executes **Backward Propagation**:
8. If $\mathcal{D} = \mathcal{L}$
9. $\nabla w_k \leftarrow$ use (11), (12) and (13)
10. Else if $\mathcal{D} = \mathcal{U}$
11. $\nabla w \leftarrow$ use (11), (13) where $k$
12. until converge or reach max iterations $T$.
13. **return** $MNN$
Given the parameters of the neural networks, a forward propagation is first performed for one or a mini-batch of multimodal example(s), then the errors propagate backwards from the output nodes to the input nodes regarding the network’s modifiable weights, and finally the weights are updated by a gradient descent step [30]. Taking derivatives of \( e_k (\mathcal{L}, \mathcal{D}) \) with respect to \( w_k \), the gradient is as follows:

\[
\nabla w_k = \frac{\partial e_k (\mathcal{L}, \mathcal{D})}{\partial w_k} = \frac{\partial \text{MSE} (f_k, L)}{\partial w_k} - \lambda \frac{\partial e_d (f_k, \mathcal{D})}{\partial w_k}
\]

(11)

Where

\[
\frac{\partial \text{MSE} (f_k, L)}{\partial w_k} = \sum_{i=1}^{L} (f_k (x_i) - y_i)^2 \frac{\partial f_k (x_i)}{\partial w_k}
\]

(12)

Where

\[
\frac{\partial e_d (f_k, \mathcal{D})}{\partial w_k} = \left(1 - \frac{1}{M} \right) \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} (f_k (x) - \hat{f} (x))^2 \frac{\partial f_k (x)}{\partial w_k}
\]

(13)

Note that we can deduce \( \frac{\partial f_k (x)}{\partial w_k} \) via classical back-propagation algorithm depending on the structure of the modality-specific NN. Therefore, we can evaluate the required derivatives \( \nabla w_k \), and apply the gradient descent step. In addition, Equation (13) is defined on unlabeled dataset, e.g., \( \mathcal{D} = \mathcal{L} \cup \mathcal{U} \). Therefore, the diversity loss \( E_d (f, \mathcal{D}) \) would be obtained from unlabeled data. The semi-supervised NN learning process of AUCC is described in Algorithm 1. After this process, AUCC not only can learn an optimal combination of multimodal representation, but also more importantly the ensemble error (related to diversity and accuracy) is back-propagated to effectively and adaptively train the modality-specific neural networks.

C. Multiple Modalities Combination

Accuracy and diversity measures are employed to train a good ensemble in the first process stage of AUCC. Note that the ensemble is measured by summing up the outputs of all the base classifiers (line 6 in Algorithm 1), which is used for traditional classifier ensemble, where all classifiers are built on exactly the same modality dataset. On the other hand, in AUCC, multimodal networks are built on modality-specific datasets respectively. Simply summing up the outputs of all the base classifiers would degrade the final classification performance. Therefore, the second process stage of AUCC aims to adjust the classifier ensemble weights according to the classification confidence (accuracy) of the base classifiers. For instance, acceleration-NN is usually effective to discriminate “walking” and “running”, and therefore it is going to have larger ensemble weight than other modality-specific NNs (e.g., electrocardiogram-NN).
Classifier level fusion is employed to ensure the overall classification performance. Considering the compatibility issues arising from different time shifts, window length configurations, and sampling frequencies, fusion multiple modalities in attribute level would weaken the utility of other modality attributes. Fusing in classifier level, the problem can be addressed because the outputs from different classifiers have uniform expressions, i.e., predictive scores. More importantly, classifier level fusion can adjust the classifier ensemble weights according to their classification confidence (accuracy). We utilize a-stack, an adaptive stacking framework [10], to combine the predictions of the base classifier \( f = \{ f_k \mid 1 \leq k \leq m \} \) obtained from the AUCC first stage.

Given training dataset \( TS = \mathcal{L} \), and the base classifiers \( f = \{ f_k \mid 1 \leq k \leq m \} \) with \( f_k : \chi \rightarrow \gamma \), the output label is \( y \leftarrow \arg \max \max \arg \max f_{k,1}(x), f_{k,1}(x) \) is the predictive score returned by the classifier \( f_k \) when the input \( x \) is labeled with \( y \). And the score \( f_{IL}(x) \) is usually the conditional probability \( p_j(y \mid x) \) in statistical machine learning methods. As shown in Fig. 3, A-stack tries to find a meta-classifier \( M : S \rightarrow \gamma \), where \( S = \{ f_1(x), f_2(x), ... , f_{IL}(x), ... , f_{n1}(x), ... , f_{nL}(x) \} \), \( f_{ij}(x) \) is the predictive scores returned by the \( k \)th base classifier for the class label \( j \) for \( 1 \leq k \leq m \) and \( 1 \leq j \leq L \). To combine multimodal information properly, the base classifiers are constructed on their data separately and then fuse their scores in a meta level classifier.

4. Experiment

In this section, the experiments are presented to evaluate the proposed method. Firstly, the experimental setup and the utilized datasets are described. Secondly, the impact of parameters like the diversity control parameter \( \lambda \), on the classification performance is studied. Thirdly, our method is compared with several state-of-the-art action recognition methods. Fourthly, experiments are further conducted to show whether unlabeled data benefits the performance of action recognition. Finally, we study the effects of other popular diversity measures such as disagreement and double-fault measures, on the performance of AUCC.

Two benchmarked real-world datasets on action recognition are applied in our experiments. To the best of our knowledge, they are the latest available multimodal datasets with complete annotation process. Both of which contain multiple modality
data (e.g., physiological and inertial data). The statistics of the two datasets are summarized in Table 1.

1) PAMAP2 Dataset [31] contains data of 18 different human activities (e.g., walking, cycling, and playing soccer, etc.), performed by 9 subjects wearing 3 inertial measurement units (IMU) and a heart rate monitor. Most of the activities were performed over approximately 3 minutes. Over 10 hours of data were collected altogether, from which nearly 8 hours were labeled as one of the 18 activities. The three IMUs are attached over the chest, dominant wrist, and dominant ankle of subjects, respectively. Each IMU contains two 3-axis accelerometers, a 3-axis gyroscope, and a 3-axis magnetic sensor, all sampled at 100Hz. The heart rate chest strap records heart rate values with approximately 9Hz.

Table 1. The statistics of the utilized datasets

<table>
<thead>
<tr>
<th></th>
<th>PAMAP2</th>
<th>MHEALTH</th>
</tr>
</thead>
<tbody>
<tr>
<td># of action</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td># of subjects</td>
<td>1 Female/8 Males</td>
<td>10 volunteers</td>
</tr>
<tr>
<td>Sensors</td>
<td>3 IMUs and a heart rate monitor</td>
<td>3 IMUs and a 2-lead ECG monitor</td>
</tr>
<tr>
<td># of modalities</td>
<td>4(acceleration, angular velocity, magnetic field, and heart rate)</td>
<td>4(acceleration, angular velocity, magnetic field, and ECG)</td>
</tr>
<tr>
<td>Labeled/total duration</td>
<td>27248.27/38504.91 seconds</td>
<td></td>
</tr>
</tbody>
</table>

A. Experimental Setting and Datasets

2) MHEALTH [32] dataset comprises inertial and physiological sensor recordings for 10 subjects while performing 10 human activities. Sensors placed on the subject’s chest, right wrist, and left ankle are used to measure the motion experienced by diverse body parts, i.e., acceleration, angular velocity, and magnetic field orientation. The sensor positioned on the chest also provides 2-lead ECG measurements, which can be potentially used for basic heart monitoring, checking for various arrhythmias, or looking at the effects of exercise on the ECG. All sensing modalities are recorded at a sampling rate of 50Hz. The activities were collected in an out-of-lab environment with no constraints on the execution way of these activities, with the exception that the subjects should try their best when executing an action.

We employ the datasets to perform background action recognition task [31], which categorizes the activities into 6 classes, including lying, sitting/standing, walking, running, cycling, and other (the remaining activities except the above six classes). The idea behind the definition of this task is that users always perform meaningful activities, and ignoring these other activities would limit the applicability of action recognition methods. More importantly, the complexity of the classification problem would
significantly increase in this way [31]. Thus, the task is employed to evaluate our method.

A sliding window strategy with fixed overlap is utilized to extract subsequences from the original data. Based on the experience gained in [10, 31, 32], the window size and sliding step are set as 5.12s and 1s, respectively, which are fixed in the following experiments. Note that the raw subsequences extracted by sliding window are used as the input of AUCC. Logistic regression is employed as the meta level classifier of AUCC.

Four commonly used performance measures are used for evaluating classification algorithm: precision, recall, f-measure, and accuracy. Since the f-measure combines precision and recall in a single value, we only take f-measure and accuracy as the performance measures. All results are evaluated by Leave-One-Subject-Out (LOSO) validation, which is regarded as the standard evaluation strategy of action recognition methods. LOSO validation can ensure subject independent in the evaluation process.

B. Experiment 1: the impact of parameters

We evaluate the impact of the maximum gradient descent steps $T$ on the performance of AUCC. The setting containing all modalities in MHEALTH (i.e., “Acc+Gyro+Magn+ECG”) is used and the trade-off parameter $\lambda$ is set to 0. Fig. 4 gives the training error of AUCC varying along with the epoch of the training stage. The training error is gradually stable when the epoch is large enough (epoch>50). This usually indicates that the parameters of the network have been close to converge. Thus, we take $T = 50$ in our following experiments.

To investigate the impact of the trade-off parameter $\lambda$ on the performance of AUCC, we increase the value of $\lambda$ from 0 to 2 in this experiment, with a fixed step of 0.1. The setting containing all modalities in MHEALTH (i.e., “Acc+Gyro+Magn+ECG”) is selected to demonstrate the impact of parameter $\lambda$. The data of three subjects is used for the LOSO validation. The classification results (i.e., accuracy and f-measure) on the dataset are depicted in Fig. 5. It can be observed that in every dataset, there is a value of $\lambda$ under which the best classification performance can be achieved.
The parameter $\lambda$ is used to control the diversity of the classifier ensemble. It can be found that when $\lambda$ increases, there is an increasing phase for the classification performance and then a decreasing phase is followed. This might happen when $\lambda$ is very small, so all base classifiers tend to generate correlated errors. In an extreme case, ensembling all “same” base classifiers would not lead to performance improvement. On the other hand, when $\lambda$ is very large, the diversity measure is dominant over the accuracy measure. In an extreme case, the final decision is “confused” when all base classifiers vote different opinions. Fortunately, there is a preferable balance point, i.e., $\lambda = 0.7$.

C. Experiment 2: comparison with other methods

To evaluate the effectiveness and show its competitive performance of the proposed approach, AUCC is compared with the other two state-of-the-art action recognition approaches (i.e., FE [6] and CE [10]). These comparative methods are elaborately chosen for fair comparisons. The comparison between classical attribute based ensemble method (i.e., FE) and AUCC aims to test the ability to learn useful attribute representations. While, the comparison between classifier level ensemble method (i.e., CE) and AUCC intends to show the ability of diversity learning to ensure that all the base classifiers craft uncorrelated errors.

In the comparative studies, only labeled samples are utilized, i.e., $\mathcal{T}_s = \mathcal{L}, \mathcal{U} = \emptyset$. Total six kinds of combination settings are presented for each dataset (Acc, Gyro, Magn, HR/ECG denote the four kinds of modalities as shown in Table 1). For example, “Acc+Magn” means both acceleration and magnetic field are utilized to recognize activities. For AUCC, the maximum number of gradient descent steps and the trade-off parameter $\lambda$ are set to 50 and 0.7, respectively. The employed attributes of FE include time-domain attributes, e.g., mean, variance, standard deviation, median, maximum, minimum, root mean square, correlation between two axes, zero crossing rate, skewness, and kurtosis, as well as frequency-domain attributes, e.g., entropy and spectral entropy. For the other compared methods, default parameters suggested in literatures are adopted. For fair comparison, logistic regression is employed as the classifier of all compared methods.
The classification performances of these methods on two datasets with six kinds of combination settings are shown in Table 2. On each setting of these methods, the mean as well as the standard deviation of accuracy and F-measure are recorded. Furthermore, to statistically measure the significance of performance difference, pairwise t-tests at 95% significance level are conducted between the methods. From Table 2, following tendencies could be discerned for all settings:

Table 2. The classification performance results (MEAN±STD.).

<table>
<thead>
<tr>
<th>Modalities in datasets</th>
<th>AUCC</th>
<th>Acc</th>
<th>ACC</th>
<th>FE</th>
<th>ACC</th>
<th>CE</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAMAP2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc</td>
<td>0.687±0.0</td>
<td>0.625±0.1</td>
<td>0.676±0.18</td>
<td>0.640±0.19</td>
<td>0.693±0.10</td>
<td>0.655±0.11</td>
<td></td>
</tr>
<tr>
<td>Acc+HR</td>
<td>0.718±0.1</td>
<td>0.688±0.1</td>
<td>0.705±0.10</td>
<td>0.674±0.10</td>
<td>0.689±0.20</td>
<td>0.685±0.20</td>
<td></td>
</tr>
<tr>
<td>Acc+Magn</td>
<td>0.767±0.0</td>
<td>0.730±0.1</td>
<td>0.744±0.11</td>
<td>0.711±0.12</td>
<td>0.725±0.11</td>
<td>0.708±0.12</td>
<td></td>
</tr>
<tr>
<td>Acc+Gyro</td>
<td>0.787±0.0</td>
<td>0.728±0.0</td>
<td>0.710±0.09</td>
<td>0.680±0.10</td>
<td>0.697±0.11</td>
<td>0.669±0.11</td>
<td></td>
</tr>
<tr>
<td>Acc+Gyro+Magn+HR</td>
<td>0.836±0.0</td>
<td>0.801±0.0</td>
<td>0.736±0.11</td>
<td>0.705±0.11</td>
<td>0.758±0.02</td>
<td>0.745±0.12</td>
<td></td>
</tr>
<tr>
<td>MHEALTH:</td>
<td>0.844±0.0</td>
<td>0.810±0.0</td>
<td>0.740±0.04</td>
<td>0.708±0.12</td>
<td>0.760±0.11</td>
<td>0.746±0.11</td>
<td></td>
</tr>
<tr>
<td>Acc</td>
<td>0.825±0.0</td>
<td>0.843±0.0</td>
<td>0.835±0.05</td>
<td>0.821±0.08</td>
<td>0.811±0.06</td>
<td>0.846±0.05</td>
<td></td>
</tr>
<tr>
<td>Acc+Magn</td>
<td>0.844±0.0</td>
<td>0.857±0.0</td>
<td>0.838±0.05</td>
<td>0.819±0.07</td>
<td>0.846±0.08</td>
<td>0.837±0.05</td>
<td></td>
</tr>
<tr>
<td>Acc+Gyro</td>
<td>0.894±0.0</td>
<td>0.900±0.0</td>
<td>0.873±0.04</td>
<td>0.870±0.07</td>
<td>0.876±0.06</td>
<td>0.863±0.04</td>
<td></td>
</tr>
<tr>
<td>Acc+Gyro+Magn+ECG</td>
<td>0.861±0.0</td>
<td>0.887±0.0</td>
<td>0.857±0.05</td>
<td>0.861±0.05</td>
<td>0.878±0.05</td>
<td>0.885±0.06</td>
<td></td>
</tr>
<tr>
<td>win/tie/loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/</td>
<td>/</td>
<td>10/1/1</td>
<td>83/1</td>
<td>74/1</td>
<td>83/1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Six kinds of combination settings are presented for each dataset (Acc, Gyro, Magn, HR/ECG denote the four kinds of modalities as shown in Table 1); for example, “Acc+Magn” means both acceleration and magnetic field are utilized to recognize activities. *b*/# indicates whether AUCC is statistically superior/inferior to the compared method (pairwise t-test at 95% significance level).

1) First, the classification performance of AUCC is statistically superior to that of FE in all multimodal settings. It suggests that AUCC has the ability to automatically learn discriminative attribute representations for human action recognition.
2) Second, the classification performance of CE is higher than that of FE in most cases but still statistically inferior to that of AUCC. This result justifies the conclusion
of [10] that classifier level fusion is able to ensure the overall classification performance by avoiding attribute compatibility issues. However, CE usually over-fits to generate correlated errors. On the contrary, AUCC can ensure that all the individual classifiers craft uncorrelated errors by exploiting the diversity of classifier ensemble.

3) Third, the superiority of AUCC over the compared methods (i.e., FE and CE) increases along with the number of modalities (from one to four kinds of modalities). This indicates that, as the number of modalities increases, both the attribute compatibility issues (in FE) and the classifier correlated errors (in CE) deteriorate gradually. And AUCC could effectively utilize the diversity of classifier ensemble to benefit action recognition from the multimodal data.

D. Experiment 3: the benefit from unlabeled data

To find out the benefits obtained from the diversity of classifier ensemble in multimodal action recognition, especially from unlabeled data, LOSO validation is adapted by splitting out partial labeled training data as unlabeled training data \( U \), thus \( TS = L \cup U \). The unlabeled training data \( U \) aims to augment the diversity among base classifiers in AUCC. The maximum number of gradient descent steps and the trade-off parameter \( \lambda \) are set to 50 and 0.7, respectively. Fig. 6 reports the average performance improvement (i.e., accuracy and f-measure) obtained from unlabeled data under various unlabeled data sizes, i.e., \( D = U, \eta = |U|/(|L| + |U|) \). The range of \( \eta \) is \( \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\} \).

![Fig. 6. The benefit obtained from unlabeled data under various unlabeled data sizes.](image)

As illustrated in Fig. 6, following points could be discerned from the results:

1) The performance increases gracefully along with the increase of unlabeled data sizes (\( \eta \) range from 0.1 to 0.3). This suggests that the unlabeled data do help to improve the performance in some way.

2) However, the performance begins to degrade when \( \eta \) is greater than 0.4. This indicates that the diversity measure can only ensure the overall performance in some way, which justifies the necessary to optimize diversity and accuracy measures in the training process simultaneously.

E. Experiment 4: other diversity measures
3) We adapt three typical diversity measures (i.e., Disagreement, Double Fault, and Prediction Confidence [16]) on our method to evaluate the usefulness of different diversity measures. The benefit obtained from the three measures is compared with NCL, which is used in this paper. The maximum number of gradient descent steps and the trade-off parameter $\lambda$ are set to 50 and 0.7, respectively. Fig. 6 reports the average accuracy improvement (by subtracting the accuracy under $\eta=0$, $\lambda=0$) obtained from different diversity measures under unlabeled data $\eta = |d| / (|c| + |d|) = 0.3$.

![Fig. 6. The average performance improvement obtained from different diversity measures.](image)

Fig. 6. The average performance improvement obtained from different diversity measures.

As shown in Fig. 7, all of those diversity measures help improve the classification performance of multimodal action recognition. However, NCL tends to achieve the best result, followed by Prediction Confidence. This might be because most existing diversity measures (e.g., Disagreement and Double Fault) are calculated based on the binary (correct/incorrect) outputs of the base classifiers. Differently, both NCL and Prediction Confidence are able to utilize the prediction difference calculated based on the concrete output (score). However, none of those diversity measures has research evidences theoretically showing its correlation to the ensemble accuracy except NCL (the theoretical analysis is in [17]).

5. Conclusion and future work

Utilizing multimodal information is a promising issue in human action recognition. Existing works attempt to combine multiple modalities by attribute or classifier ensemble. However, the attribute construction and classifier combination strategy are still intractable. In this paper, we propose AUCC, a novel action recognition approach exploiting the diversity of classifier ensemble based on neural network to address this problem. Our approach exploits the diversity of the base classifiers to construct a good ensemble for multimodal human action recognition by leaning to diversity techniques. Firstly, the diversity of classifier ensemble is embedded in the error function of neural network (NN) and back-propagated to enhance the diversity and accuracy of the learned multimodal representation. Next, the base classifiers are incorporated by classifier level
fusion to adjust the classifier ensemble weights. Comprehensive experiments indicate that AUCC is able to yield a competitive classification performance in human action recognition task, and has its superiority on exploit the benefit of multimodal signals.

Then, we will extend our work in the following directions: Firstly, utilize unsupervised techniques like stacked convolutional auto-encoders to further improve the generalization and recognition performance of our model; secondly, investigate the effects of multiple diversity measures on ensemble (combining multiple diversity measures) on the performance of action recognition.

References


