Pattern Recognition 81 (2018) 23–35

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning content and style: Joint action recognition and person identification from human skeletons

Hongsong Wang, Liang Wang*

Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 100190, P.R. China

ARTICLE INFO

Article history:
Received 15 May 2017
Revised 20 March 2018
Accepted 27 March 2018
Available online 27 March 2018

Keywords:
Content and style
Action recognition
Person identification from motions
Skeleton transformation
Multi-task RNN

ABSTRACT

Humans are able to simultaneously identify a person and recognize his or her action based on biological motions. Previous work usually treats action recognition and person identification from motions as two separate tasks with different objectives. In this paper, we present an end-to-end framework to perform these two tasks together. Inspired by the recent success of deep recurrent neural networks (RNN) for skeleton based action recognition, we propose a new pipeline to recognize both actions and persons from skeletons extracted by RGBD sensors. The structure includes two subnets and is end-to-end trainable. The former is skeleton transformation, which accommodates viewpoint changes and noise. The latter is multi-task RNN for joint learning and various architectures are explored including a novel architecture that learns the joint probability between the two output variables. Experiments on 3D action recognition benchmark datasets demonstrate the benefits of multi-task learning and our method dramatically outperforms the existing state-of-the-art in action recognition.

1. Introduction

Human visual system can quickly and efficiently detect another living being performing some actions in a visual scene and recognize many aspects of biological, psychological, and social significance [1]. Biological motion contains information about actions as well as the identity of persons. The motion patterns are decomposed into *content* and *style* [2,3]. The *content* represents the temporal dynamics of body poses and the *style* indicates the personalized style of actions which can be used for person identification. What our visual system seems to solve so effortlessly is still an unsolved problem in computer vision.

Learning *content* and *style* corresponds to two important tasks for vision based human motion understanding, i.e., action recognition and person identification from biological motion. Due to different goals, the existing methods treat them as two separate or even mutually exclusive tasks. Action recognition is concerned with what is the performed action, regardless of human subjects. The difference that different persons do the same action in various ways is the inter-class difference that has to be reduced. While person identification from biological motion addresses the question of who is the person performing the action. It aims to seek distinguishable variations between the same actions performed by different persons, allowing for an arbitrary type of actions.

Most previous approaches recognize human actions from videos. Johansson's experiments [4] show that a large set of actions can be recognized from motions of the main joints of skeletons, which have inspired most of the literature about human body pose estimation and action recognition. Recently, skeleton based action recognition gains more popularity due to the advent of cost-effective depth sensors (e.g., Microsoft Kinect) and fast and accurate skeleton estimation algorithms from a single depth image [5]. These depth sensors support real-time non-invasive pose estimation. Currently, the Kinect v2 can physically sense depth and estimate reliable skeletons by 8 m. The area of human pose estimation in videos is also developing fast, and there are several popular benchmarks and effective methods. Compared with the video data, skeletons are more succinct and explicitly depict the dynamics of actions.

In this paper, we aim to simultaneously recognize both *content* and *style* from movements of human. We opt to consider RGBD data and learn representations from human skeletons. A novel and unified framework is proposed to conduct action recognition and person identification from human skeletons. The proposed method inherits the merits of deep recurrent neural networks (RNN) for skeleton based action recognition [6–8]. Fig. 1 shows an architecture of our method. It first learns representations from the raw

* Corresponding author.
E-mail addresses: hongsong.wang@nlpr.ia.ac.cn (H. Wang), wangliang@nlpr.ia.ac.cn (L. Wang).

https://doi.org/10.1016/j.patcog.2018.03.030
0031-3203/

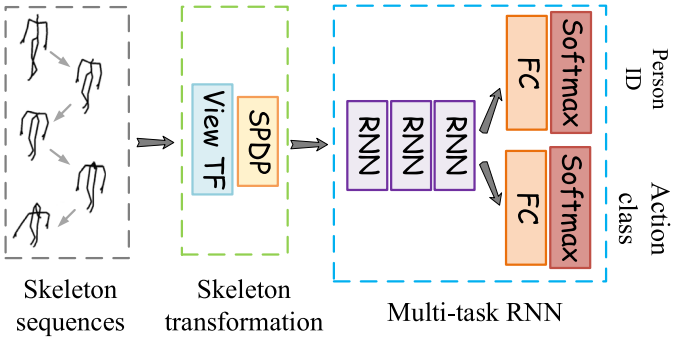


Fig. 1. Pipeline of our approach for joint learning *content* and *style*. Here, View TF, SPDP, FC denote the viewpoint transformation, spatial dropout, and fully-connected layer, respectively. The whole architecture is end-to-end trainable and could simultaneously predict action class and person ID from raw skeletons.

skeletons, and then performs the two tasks together using the shared representations based on multi-task learning.

The proposed pipeline consists of two components: skeleton transformation for robust representation and multi-task RNN for joint learning. The former aims to address the problem of viewpoint changes and noise by the proposed viewpoint transformation layer and spatial dropout layer, respectively. The latter extends the generic RNN in a multi-task learning manner, which comprises of shared layers and task-specific layers. The shared RNN layers learn the commonalities across tasks and the task-specific RNN layers model the differences for the corresponding task. To investigate the ability of different shared and task-specific representations, we enumerate seven architectures with different amounts of sharing layers. We also examine the two special architectures. One is equivalent to two separate networks and has no shared parameter. The other is a novel architecture with no task-specific parameter, and learns the joint probability between the two output variables. We apply our model to skeleton based action recognition with cross-view evaluation to compare with the existing approaches.

In summary, the main contributions of this paper are listed as follows:

- To the best of our knowledge, we are the first to pair action recognition and person identification inspired by the fact that our visual system can simultaneously recognize *content* and *style* from biological motions.
- We propose a new end-to-end trainable pipeline, which consists of skeleton transformation and multi-task RNN.
- We propose multi-task RNN with different amounts of sharing layers as well as a novel architecture that learns the joint probability between the two output variables.
- We obtain state-of-the-art results in skeleton based action recognition. Experiments show that for these two tasks, learning one task would benefit from learning another task.
- For person identification, we achieve a accuracy of 65.2% from novel viewpoints within 40 categories solely based on skeletons.

2. Related work

Learning *content* and *style* from skeletons is related to a range of topics, e.g., skeleton based action recognition and multi-task learning. Here we briefly review representative work on those topics.

2.1. Skeleton based action recognition

Skeleton based action recognition becomes popular due to the advances of pose estimation. There are many recent accurate pose estimation and pose tracking algorithms. For example, a deep structure is proposed to represent the human body in a coarse-to-fine procedure [9]. A max-margin Markov based model is presented to track human pose [10] and a united graphical model is developed to integrate the problems of pose estimation and visual tracking [11].

Traditional approaches can be divided into two categories: joint based approaches [12,13] and body part based approaches [14,15]. Readers are referred to these survey papers [16,17]. Joint based approaches consider the human skeletons simply as a set of points. These approaches use various features (e.g., joint positions [18], joint orientations [19], pairwise relative joint positions [20,21]) to represent the motion of either individual joints or combinations of joints. While body part based approaches regard the human skeleton as a connected set of rigid segments. They directly model the temporal evolution of individual body parts [22] or connected pairs of body parts [23,24].

There is a growing trend of using recurrent neural networks (RNN) for skeleton based action recognition. An end-to-end hierarchical RNN architecture is the first attempt towards this task [6,25]. Afterwards, a fully-connected deep long short term memory (LSTM) network with regularization terms to learn co-occurrence features of joints [7] is proposed, and a part-aware extension of LSTM is presented to make use of the physical structure of the human body [8]. In addition, Veeriah et al. [26] propose a differential gating scheme for LSTM to emphasize the salient motions between successive frames. A spatio-temporal long short term memory (STLSTM) network is proposed to model the contextual dependencies of joints both in the temporal and spatial domain [27]. Recently, Song et al. [28] design a spatial and temporal attention based RNN structure to learn discriminative spatial and temporal features. Wang and Wang [29] present a two-stream RNN architecture to leverage both temporal dynamics and spatial configurations. These approaches merely predict the actions and we exploit action recognition jointly learned with person identification.

2.2. Person identification from motions

Person identification from motion is a widely studied topic in the computer vision community. Many work focus on motion of a particular type, i.e., locomotions. A typical example is gait recognition [30], which aims to discriminate individuals by the way they walk and is particularly suitable for long-distance human identification. Methods of gait recognition can be roughly divided into two categories, model based methods [31,32] and appearance based methods [33–35]. Model based methods model the underlying structure of human body to measure physical gait parameters such as trajectories, limb lengths, and angular speeds. In contrast, appearance based methods analyze gait sequences and extract gait representations directly from videos.

The growing popularity of Kinect has led to the recent work of person identification from depth and skeletons. For example, Barbosa et al. [36] exploit several soft-biometrics features extracted from range data and find that height and torso/legs ratio are the most informative cues. Munsell et al. [37] use a two-step approach that first classifies the locomotion type then applies a locomotion-specific identity classifier to identify the individual. Wu and Konrad [38] use a dynamic time warping framework based on skeletons for both user identification and user authentication. The follow-up work investigates the potential performance and robustness gains in user authentication using multiple Kinects [39]. A generative model for person identification based on motion patterns of skeletons from an arbitrary predefined set of action types is presented [40]. Recently, Haque et al. [41] present an attention based model that reasons on human body shape and motion dynamics to identify individuals given only depth images. Pala et al. [42] investi-

gate anthropometric measures in unconstrained settings and show the improvement of re-identification performance of the widely used clothing appearance cue. There approaches usually identify the persons from the sequence of a given action (e.g., walking), while we handle the input sequence with an unknown type of action and investigate person identification jointly learned with action recognition.

2.3. Analyzing actions and persons

An action is something which is done by a person. Actions and persons are intertwined with each other. Unveiling the relationships between action recognition and person identification is a meaningful research topic and there exists several previous work. For example, Kobayashi and Otsu [43] present a general method which could be applied to action recognition and multiple-person identification appeared in a motion image sequence. As an action often involves many people active in the scene, Ramanathan et al. [44] introduce an attention based model to identify the key person responsible for the action without being explicitly trained with such annotations. Xie et al. [45] propose a novel method to mine representative actions of each person as complementary features for person identification from video data. Wang and Wang [46] introduce the new problem of cross-agent recognition which could recognize the actions of a particular person while the training data comes from another different person. Although the relationships between actions and persons are considered, these approaches focus on either action recognition or person identification. In contrast, we simultaneously address the two tasks by using a single network.

The work similar to ours is [47], which proposes an unified framework for person identification and activity recognition based on graph signal processing by using the same set of features. However, the train/test splits of the two tasks in [47] are different, and we are the first to pair activity recognition and person identification by multi-task learning.

2.4. Multi-task learning

Multi-task learning [48], i.e., multiple learning tasks are solved at the same time to exploit commonalities and differences across tasks, has a rich history in machine learning. It has widespread applications in computer vision [49–52], natural language processing [53,54], genomics [49], etc. Given such a broad scope, we only focus on multi-task learning in the context of deep neural networks used in computer vision.

Multi-task learning based on deep neural networks has displayed remarkable success in computer vision. For example, the well-known Fast R-CNN [55] for object detection uses a multi-task loss jointly train for classification and bounding-box regression. For multi-view face detection, Zhang and Zhang [56] present a multi-task deep CNN to jointly learn face detection together with facial pose estimation and facial landmark localization. For facial landmark detection, Zhang et al. [57] achieve robust results by joint learning with correlated tasks, such as appearance attribute, expression, demographic, and head pose. For attribute prediction, Abdulnabi and Wang [58] learn binary semantic attributes through a multi-task CNN. For action recognition, a deep CNN is investigated for the tasks of pose estimation and action classification in unconstrained images [59]. There are also some RNN based approaches. For instance, a multi-task RNN is used to refine coarse predictions through multiple steps for immediacy prediction [60]. Different from the previous approaches, we investigate different structures of multi-task RNN based on the input sequence.

3. Preliminary

Different from feedforward neural networks that map from one input vector/matrix to one output vector/matrix, recurrent neural networks (RNN) have an internal state to exhibit dynamic temporal behavior. They can process arbitrary sequences and map an input sequence to another output sequence. The hidden state representation h_t at each time step t of a simple and popular RNN model on account of the input x_t at the current step and the state representation h_{t-1} of the previous step:

$$\begin{aligned} h_t &= \sigma(W_x x_t + W_h h_{t-1} + b_h) \\ y_t &= \sigma(W_o h_t + b_o) \end{aligned} \tag{1}$$

where W_x, W_h, W_o, b_h, b_o are parameters and σ is a nonlinear activation function.

The standard RNN cannot store information for long periods of time due to the vanishing and exploding gradient problem. To address this problem, long short-term memory (LSTM) [61] is proposed by using additional gates to determine when the input is significant enough to remember, when it should continue to remember or forget the value, and when it should output the value. The structure of an LSTM unit is shown in Fig. 2(a). The hidden state representation h_t of a LSTM unit is updated as:

$$\begin{aligned} i_t &= \sigma(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f) \\ c_t &= f_t c_{t-1} + i_t \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \\ h_t &= o_t \tanh(c_t) \end{aligned} \tag{2}$$

where x_t denotes the input at each time step t, and i, f, o correspond to the input gate, forget gate and output gate, respectively. All the matrices W are the connection weights and all the variables b are biases.

For many sequence recognition tasks, both past context and future context are useful. Take action recognition as an example, the anticipatory action of present depends not only on the past but also on the expectations of the future. Bidirectional Recurrent Neural Networks (BRNN) [62] elegantly combines both forward and backward dependencies by using two separate recurrent hidden layers to present the input sequence. An example of unfolded BRNN is shown in Fig. 2(b). Every time step in the output sequence provides complete historical and future context.

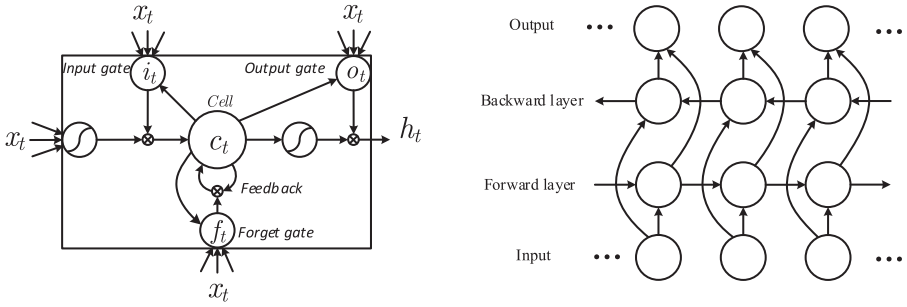
RNN architectures are naturally suitable for the sequence classification, e.g., skeleton based action classification, where an input sequence is assigned with a single class label. Deep RNN can be constructed by stacking multiple recurrent hidden layers on top of each other. The deep stacked RNN model can deal with multiple time scales in the input sequence. The formulation of stacked RNN is:

$$h_t^{(l)} = f_h^{(l)}(h_t^{(l-1)}, h_{t-1}^{(l)}) \tag{3}$$

where $h_t^{(l)}$ is the hidden state of the lth level at time step t, and $f_h^{(l)}$ nonlinear function of the RNN unit. When $l = 1$, the state is computed using x_t instead of $h_t^{(l-1)}$.

4. Joint learning content and style

For joint learning *content* and *style* from sequences of human skeletons, the learning system observes two supervised learning tasks, i.e., action recognition and person identification. The goal is to simultaneously address both tasks by sharing information between them. The pipeline of learning *content* and *style* is shown in Fig. 1. The structure consists of two components: skeleton transformation for robust representation and multi-task RNN for joint learning.



(a) LSTM (b) BRNN

Fig. 2. (a) A LSTM block with input, output, and forget gates [61]. (b) An unfolded bidirectional network [62]. The solid line denotes the weighted connection between units and the weights are reused at every time step. The outputs of the forward and backward layers are concatenated to present the output sequence.

4.1. Skeleton transformation

In practical scenarios, the capturing viewpoints of cameras differ among different sequences and the estimated skeletons are noisy. Accordingly, the skeleton transformation step addresses the problems of viewpoint changes and noise. It comprises a viewpoint transformation layer and a spatial dropout layer. After transformation, the sequence has two major changes. First, the skeleton sequence observed from a horizontal or vertical angle is transformed to a standard skeleton observed from the frontal view. Second, the coordinates of joints of a particular time step are randomly dropped with a certain rate. The details are described below.

Skeletons may be observed from arbitrary camera viewpoints in a realistic scenario. To reach a view-invariant representation, a viewpoint transformation layer is constructed to transform the coordinates of input skeletons. Let $X = \{x_1, x_2, \ldots, x_n\}$ denote a sequence of skeletons, for time step t, $x_t = [p_1, p_2, \ldots, p_m]$, where $p_i = [x_i, y_i, z_i]$ is the coordinates of the joint i. Given three angles α, β, γ about the x, y, z axis in the 3D coordinate system, for each joint and each time step, the new coordinates can be obtained by $\widetilde{p} = p \cdot R$, where $R = R_z(\gamma)R_y(\beta)R_x(\alpha)$. Details about the formulation of the three basic rotation matrices in terms of angles are presented in [25,29,63]. This operation simulates the viewpoint changes of the camera and improves the robustness of our model.

Skeletons collected by sensors like Kinect may not always be reliable due to noise and occlusion. To handle the influence of noise and occlusion, HBRNN [6] uses the Svaitzky–Golay filter in the temporal domain to smooth the skeletons and Spatio-Temporal LSTM [27] adds a new gate to LSTM unit to analyze the reliability of the input. Here, we adopt an alternative approach based on spatial dropout [64]. The standard dropout [65] independently sets the neurons of network activations with probability p_{drop} during training. For testing, all activations are used and $1 - p_{drop}$ is multiplied to account for the increase in expected bias. Given a matrix of the sequence representation of size $n \times d$, where n is the length of the sequence and d is the feature dimension at each time step, the spatial dropout layer performs only n dropout trials and extends the dropout value across the feature dimension.

4.2. Multi-task RNN

Multi-task learning methods use a shared representation to train different tasks in parallel. For deep neural networks, architectures based on multi-task learning share the hidden layers at the bottom of the networks. Although the RNN structures are widely used, the multi-task RNN structures have been rarely explored before. Motivated by multi-task learning [48], we investigate different multi-task RNN architectures.

The resulting networks consist of shared layers and task-specific layers. The shared layers are directly connected to the common input and the task-specific layers are attached to the last shared layer. For each task, the basic architecture has three RNN layers and one fully-connected layer with softmax activation. This allows us to build seven end-to-end trainable architectures with multi-task learning following a similar line of reasoning, as shown in Fig. 3. Details of structures are described below.

Late Split. The *Late Split* shares all the layers across different tasks except the fully-connected layer with softmax activation. It possesses the same features learned from the deep neural networks before classification. This architecture is succinct and widely used for multi-task learning.

Middle Split. Compared with the *Late Split*, the *Middle Split* makes the last shared RNN layer independent for each task. The shared layers are two stacked layers of RNN and the task-specific layers consist of one RNN along with the classification layer. The independent RNN layer models the temporal evolution of features for the corresponding task.

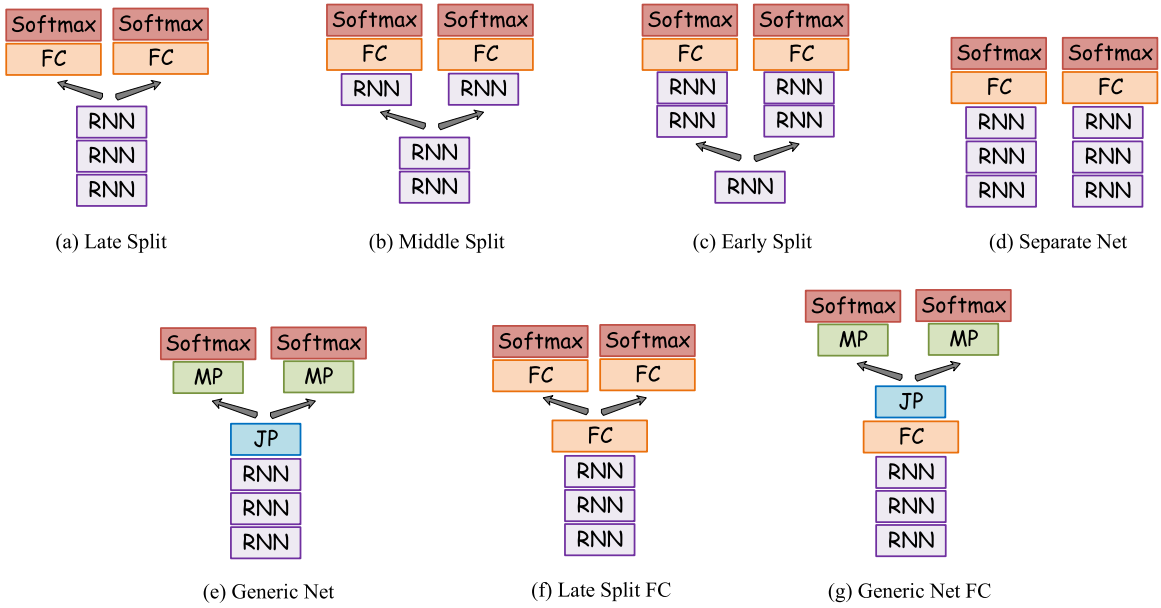
Early Split. The *Early Split* incorporates two RNN layers for each task and one shared RNN layer. It assumes the two tasks are less related and uses stacked RNN to adapt the features from the shared RNN. Compared with the above two structures, this architecture has more parameters thereby making it flexible to fit each task.

Separate Net. One extreme case is splitting at the lowest layer, which is equivalent to two separate networks altogether and is called *Separate Net*. This architecture has no shared parameters. The assumption behind the structure is that the two tasks are totally unrelated and no features are shared for any layer. This architecture is regarded as a baseline method without multi-task learning.

Generic Net. Another extreme architecture does not contain parameters specifically for a particular task. Although the above four architectures allow for a varying amount of shared layers for multi-task learning, it also poses a question that at which layer of the network should one split, especially for deep neural networks with a large number of layers. Rather than enumerating all the possibilities, we propose a *Generic Net* model for multi-task learning with no task-specific parameters. We begin by introducing two new layers: joint probability (JP) layer and marginal probability (MP) layer.

Multi-task learning needs to predict multiple output variables from a single input. In the case, we aim to predict the classes of both actions and persons. Given the learned representation h of the input sequence by the top layer of stacked RNN, the JP layer is designed to compute the joint probability between the two output variables:

$$p(i, j|h; \theta) = \phi(Wh + b) \tag{4}$$



(a) Late Split (b) Middle Split (c) Early Split (d) Separate Net

(e) Generic Net (f) Late Split FC (g) Generic Net FC

Fig. 3. An overview of seven multi-task RNN architectures for modeling sequences. The recurrent layer, fully-connected layer and softmax activation are denoted by RNN, FC, Softmax, respectively. To illustrate the proposed architecture, we introduce two new layers: JP and MP, which denote the joint probability layer and the marginal probability layer, respectively.

where i, j are are indexes of categories for actions and persons, respectively, $\phi(\cdot)$ is the activation function and $\theta \equiv \{W, b\}$ are the parameters. The JP layer models the correlation between categories of actions and persons. It is similar to the fully-connected layer and only has two parameters, i.e., the weights W and the biases b.

The MP layer calculates the probability for each variable by summing the joint probability over other variables. In the case of joint action recognition and person identification, we have two MP operations. We place a softmax function on top of the MP layer to normalize the real values in the range (0, 1) that add up to 1. The formulations of two MP layer with softmax activation are:

$$\begin{cases} p(i) = softmax(\sum_j p(i,j)) \\ q(j) = softmax(\sum_i p(i,j)) \end{cases} \tag{5}$$

where $p(i)$, $q(j)$ denote the predicted probabilities of classes of actions and persons, respectively.

Note that both the MP layer and softmax activation have no parameters. Although this model is presented with two predicted discrete output variables, it can be easily generalized to multiple variables of corresponding tasks.

Late Split FC. The dimensionality of features plays an important role for accurate classification. A low dimensionality may not be adequately enough to represent the features of interest. To better adapt to this problem, the *Late Split FC* uses another fully-connected layer to transform the output features of the RNN layers for the *Late Split*.

Generic Net FC. Similar to the *Late Split FC*, the *Generic Net FC* also adopts a fully-connected layer to transform the features learned from stacked RNN layers based on the structure of *Generic Net*.

4.3. Training methods

Let the predicted probability distribution over the class labels of actions and persons be p and q, respectively. There are N samples in the training set, and $(x_n, y_n^{(1)}, y_n^{(2)})$ is the training sample indexed by $n \in \{1, 2, \ldots, N\}$, where $y_n^{(1)}, y_n^{(2)}$ are ground truth labels of actions and persons, respectively. Considering the cross-entropy loss between the predicted class probabilities and the true probabilities, the two objective functions to be minimized are:

$$\begin{cases} L^{(1)} = -\frac{1}{N}\sum_n \log(p(y_n^{(1)})) + \mu \left\| \theta^{(1)} \right\| \\ L^{(2)} = -\frac{1}{N}\sum_n \log(q(y_n^{(2)})) + \mu \left\| \theta^{(2)} \right\| \end{cases} \tag{6}$$

where $\|\theta^{(1)}\|$, $\|\theta^{(2)}\|$ are regularization terms over the parameters of the last fully-connected layer before softmax function, and μ is a hyperparameter.

We use the combination of above two losses to simultaneously optimize multiple objectives:

$$L = \lambda L^{(1)} + (1-\lambda) L^{(2)} \tag{7}$$

where λ is the weight coefficient, $0 \le \lambda \le 1$. The whole network is trained based on stochastic gradient descent by the combined objective function.

5. Experiments and analysis

In this section, we first describe the datasets and the implementation details including the experimental setup. Then, we compare the results of different structures and analyze the distinctions between different actions. Our results of action recognition are also compared with the previous state-of-the-art results. Finally, we analyze different training methods and evaluate the parameters to draw further insights of the proposed model.

5.1. Datasets

NTU RGB+D. The NTU RGB+D dataset [8] is currently the largest depth-based action recognition dataset. It is captured by Kincet v2 in various background conditions with 3D coordinates of 25 joints. The dataset contains more than 56 thousand sequences and 4 million frames. There are 60 different action classes including daily, mutual, and health-related actions. The actions are performed by 40 different human subjects, whose age range is from 10 to 35. We follow the cross-view evaluations protocol as in [8].

Northwestern-UCLA Multiview. The Northwestern-UCLA Multiview Action3D dataset [66] contains RGB, depth and skeletons cap-

tured simultaneously by three Kinect cameras. It includes 10 action categories: *pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry*. Each action is performed by 10 different human subjects. We use the same experimental setting [66] that uses samples from 2 cameras as the training data, and use samples from 1 camera as the testing data.

UWA3D Multiview II. The UWA3D Multiview Activity II dataset [67] is collected in the lab using Kinect. It consists of 30 human activities performed by 10 different human subjects. Each subject performed the same actions four times from four different views: front view, left and right side views, and top view. This dataset is very challenging due to heavy occlusion and the point that the start and end positions of human body for the same actions are different. We follow the cross-validation [67] that use the samples from two views as the training data and the samples from the two remaining views as the testing data.

5.2. Implementation details

For joint learning *content* and *style*, two requirements should be satisfied. First, labels of both actions and persons are required for multi-task learning. Second, there exist no unseen novel class for both actions and persons on the test set.

However, most action recognition datasets with videos from the web only include annotation of human actions. Lots of RGBD datasets with the launch of Microsoft Kinect provide IDs of human subjects, which can be regarded as labels of persons. Numerous experimental evaluations of these datasets adopt the cross-subject protocol, i.e., subjects of the dataset are split into training and test groups and the human subjects between the training set and test set are different. Fortunately, there are some evaluations of the RGBD datasets that satisfy both requirements of joint learning *content* and *style*, e.g., cross-view evaluation, which aims to recognize actions from views that are unseen in the training data. To establish the effectiveness of our model, we exploit multi-task learning using the same splitting protocols and compare our results against the existing results reported on these datasets.

We consider a sequence of 3D skeletons as the input. For preprocessing, we normalize skeletons by subtracting the central joint, which is the average of 3D coordinates of the hip center, hip left and hip right. To allow for batch learning, we convert sequences to a fixed length T by sampling and zero padding. Here, T should be larger than the length of most sequences to reduce loss of information caused by sampling. We set $T = 100$ for the NTU RGB+D dataset and $T = 80$ for the other two datasets.

We adopt bidirectional LSTM unit for all recurrent layers due to its excellent performance. The number of neurons of the above datasets are 512, 512, 256, respectively. For *Late Split FC, Generic Net FC*, the number of neurons of the fully-connected layer is twice that of the corresponding RNN layer. The networks are trained using stochastic gradient descent. The initial learning rate is set to 0.02. We decrease the learning rate by 30% after every 40 epochs. For the weight parameters, we take the following default values, $\lambda = 0.5, \mu = 0.0001$, and the sensitivity is evaluated in Section 5.7. Our implementation is based on Theano and the NVIDIA TITAN X GPU is used.

5.3. Experimental results

The experimental results of the proposed methods on the three datasets are shown in Table 1. It should be noted that *Separate Net* which uses separate networks for action recognition and person identification serves as the baseline method without multi-task learning. The details of the seven proposed structures are described in Section 4.2.

In terms of the accuracy of action recognition on the three datasets, the methods based on multi-task learning (*Late Split, Middle Split, Early Split*) outperform *Separate Net* by more than 1.4%, 2.0% and 1.8%, respectively. Additionally, *Middle Split* yields the best performance on two datasets. For example, on the NTU RGB+D dataset and the UWA3D Multiview II dataset, *Middle Split* outperforms the other three methods by more than 2.0%, 1.7%, respectively. For the Northwestern-UCLA dataset, the result of *Middle Split* is slightly inferior to that of *Early Split*. Note that the only differences between the four methods are the different levels of shared layers. The results demonstrate that the proposed multi-task learning methods are effective for skeleton based action recognition. Due to the excellent performances, we recommend the structure of *Middle Split* which uses two common RNN layers to learn the shared representation and one independent RNN layer to adjust the learned representation for each task.

We also observe that *Generic Net* consistently performs better than *Separate Net* on the three datasets. *Generic Net*, in which all parameters are shared between the two tasks, is the opposite extreme of *Separate Net*. The result is consistent with our discussion that action recognition benefits from joint learning *content* and *style*. Comparing *Late Split FC* to *Late Split*, we find that increasing the feature dimension by placing another fully-connected layer on top of the RNN layers does not necessarily improve the performance. Similar results can be obtained by comparing *Generic Net FC* to *Generic Net*. We conclude that the learned features of RNN layers can be directly used for action classification.

For the accuracy of person identification, the results of *Late Split, Middle Split, Early Split* are much higher than those of *Separate Net*. For example, on the NTU RGB+D dataset, *Middle Split* outperforms *Separate Net* by 15.7%. Other multi-task RNN structures also beat the baseline *Separate Net* by considerable margins. We also find that *Late Split FC* obtains the best results on two dataset, i.e., the NTU RGB+D dataset and the UWA3D Multiview II dataset. We argue that additional fully-connected layer to transform the features learned from RNN layer might be useful for person identification. It should be noted that by leveraging the benefits of multi-task RNN, we achieve a person identification accuracy of 65.2% within 40 categories solely based on skeletons.

5.4. Analysis of different actions

In this section, we aim to analyze the accuracies of both action recognition and person identification for different actions. Here we show the results of *Middle Split* and similar results can be achieved for other architectures.

Fig. 4 illustrates the recognition accuracies of different actions of joint learning *content* and *style* on the NTU RGB+D dataset. For some actions, e.g., *use a fan feeling warm, staggering, salute, taking a selfie, hand waving, cheer up, throw, brushing hair*, both accuracies of action recognition and person identification are very high. These actions involve certain movements of hands which might be discriminative for recognition of *content* and *style*. For some other actions, e.g., *walking apart from each other, hugging other person*, the accuracies of action recognition are very high but the accuracies of person identification are very low. These actions are interactive activities between two persons and the movements of the individual person are relatively small and rigid. There are also some actions, e.g., *shake head, writing, reading*, both accuracies of action recognition and person identification are relatively low. These actions only allow a slight degree of movement thus conveying limited information for recognition.

The recognition accuracies of different actions on the Northwestern-UCLA dataset is shown in Fig. 5. We can see that for *drop trash, pick up with two hands*, both accuracies of action recognition and person identification are 100%. And for the

Table 1
Accuracies of action recognition and person identification of different architectures on the three datasets with cross-view evaluation.

Method	NTU RGB+D dataset		Northwestern-UCLA		UWA3D Multiview II	
	Action	Person	Action	Person	Action	Person
Late Split	80.6	59.0	86.6	38.4	48.3	23.7
Middle Split	**82.6**	63.3	87.3	40.6	**50.9**	23.2
Early Split	80.4	57.3	**88.1**	41.9	49.3	21.6
Separate Net	79.0	47.6	84.6	36.9	46.5	20.1
Generic Net	80.1	56.4	88.3	**43.4**	48.7	23.0
Late Split FC	80.8	**65.2**	87.9	41.0	47.7	**23.7**
Generic Net FC	80.4	58.2	87.2	41.7	47.4	23.1

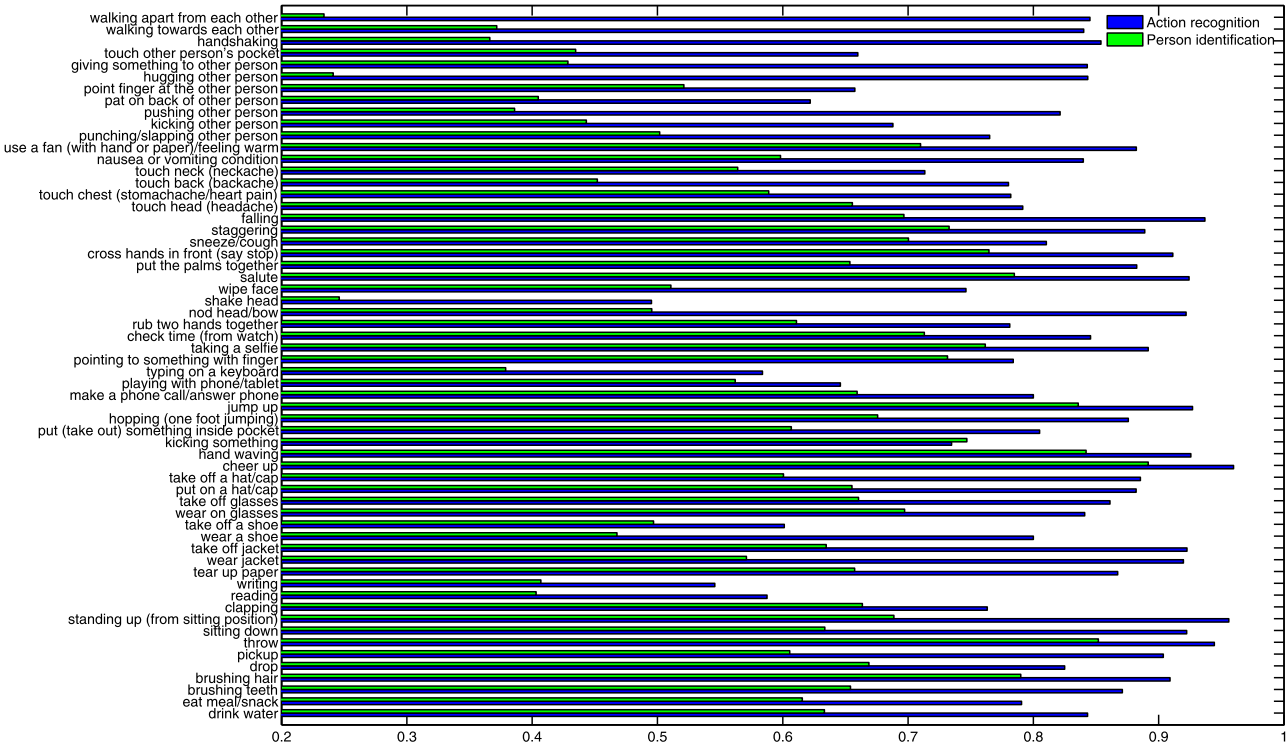


Fig. 4. Recognition accuracies of different actions on the NTU RGB+D dataset.

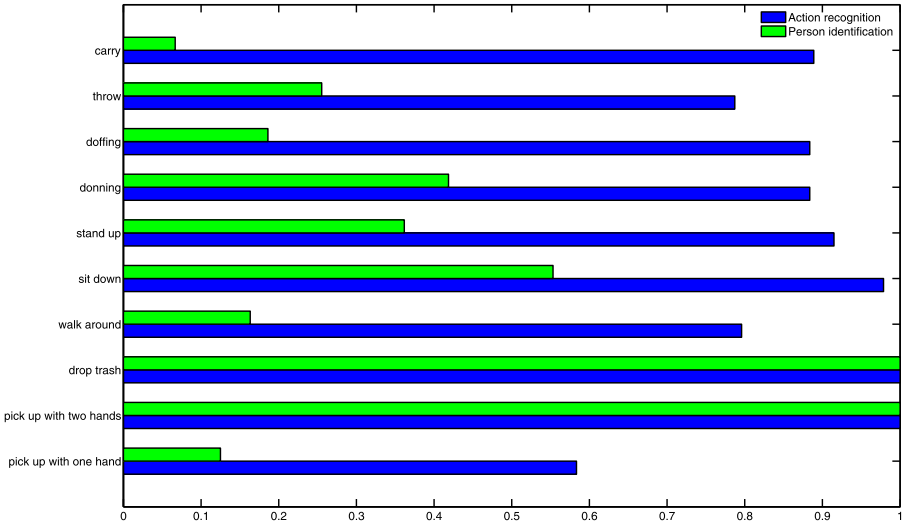


Fig. 5. Recognition accuracies of different actions on the Northwestern-UCLA dataset.

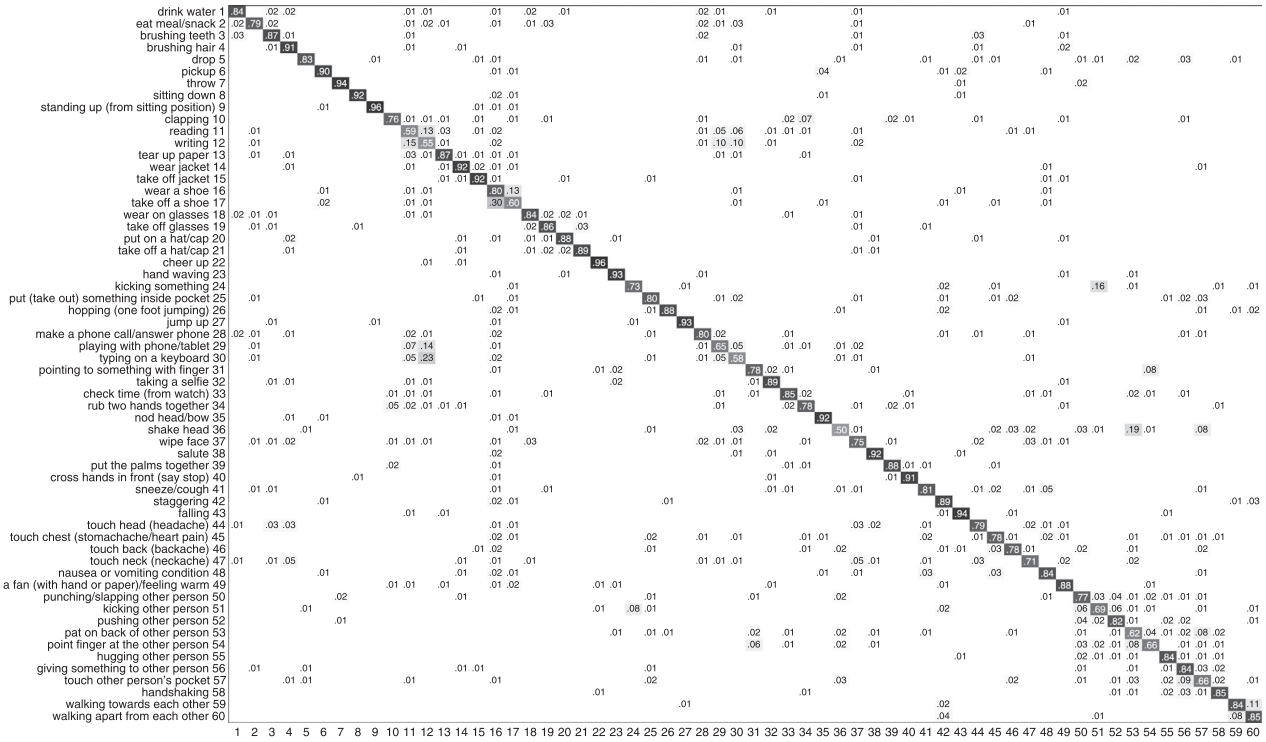


Fig. 6. The confusion matrix of action recognition on the NTU RGB+D dataset.

actions like *carry, doffing*, the accuracies of action recognition are nearly 90%, but the accuracies of person identification are below 20%. The results are consistent with the discussion above that some actions might not suitable for person identification as the involving movements are small and rigid.

Next we analyze the confusion matrix of action recognition. For person identification, as it makes no sense to analyze different accuracies of different person subjects, we do not give the corresponding confusion matrix. The confusion matrix of action recognition on the NTU RGB+D dataset is shown in Fig. 6. We observe that there are six pairs of actions with high misclassified rate, i.e., *(reading, writing), (palying with phone/tablet, writing), (typing on a keyboard, writing), (take off a shoe, wear a shoe), (kicking something, kicking other person), (shake head, pat on back of other person)*. For example, samples of *take off a shoe* are misclassified as *wear a shoe* with a rate of 30%. Generally, *take off a shoe* is the opposite action of *wear a shoe*. A sequence of *take off a shoe* can be considered as *wear a shoe* if we watch the sequence backwards. As the bidirectional LSTM which models both forward and backward dependencies is adopted as the recurrent unit, it might be difficult for our model to distinguish between samples that belong to an action and its opposite action. Additionally, samples from actions like *reading, writing* are much similar and hard to discriminate solely based on skeletons. It should be noted that the confusion matrix is not symmetric. For example, samples of *shake head* are misclassified as *pat on back of other person* with a high rate of 19%, but samples of *pat on back of other person* are only misclassified as *shake head* with a low rate of 2%.

The confusion matrix of action recognition on the Northwestern-UCLA dataset is shown in Fig. 7. We find that there are five pairs of actions with the misclassified rate more than 5%, i.e., *(pick up with one hand, pick up with two hands), (throw, walk around), (pick up with one hand, throw), (throw, carry)*. One interesting observation is that samples of *pick up with one hand* can be easily misclassified as *pick up with two hands*, but no samples of *pick up with two hands* is misclassified as *pick up with one hand*. Some actions are harder to recognize than some others and the confusion matrix of action recognition is not symmetric.

Table 2
Comparison of the proposed approach with the previous methods for action recognition on the NTU RGB+D dataset.

Method	Accuracy
Lie group [68]	52.8
Skeletal quads [69]	41.4
FTP dynamic [70]	65.2
HBRNN [6]	64.0
Part-aware LSTM [8]	70.3
Trust gate ST-LSTM [27]	77.7
Two-stream RNN [29]	79.5
Multi-task RNN	**82.6**

5.5. Comparison with the state-of-the-art

Previous efforts have been made on robust representation of action for performance improvement, and the person who performs the action is often neglected in action recognition research. To demonstrate the benefits of joint learning for action recognition, we compare our results with the reported state-of-the-art results on three public benchmark datasets. Here, we choose the architecture of *Middle Split* due to the excellent performances. To make fair comparison with the other state-of-the-art methods, we do not choose the best architecture for different datasets as the ground truth of testing data is not always available.

Table 2 shows the results on the NTU RGB+D dataset. Our method significantly outperforms three methods based on handcrafted features, i.e., 3D skeletons representation in a Lie group [68], Fisher vector encoding of skeletal quads [69] and FTP dynamic [70]. Comparing with other RNN based approaches (e.g., [6,8,27,29]), our result is also considerably higher. Finally, our

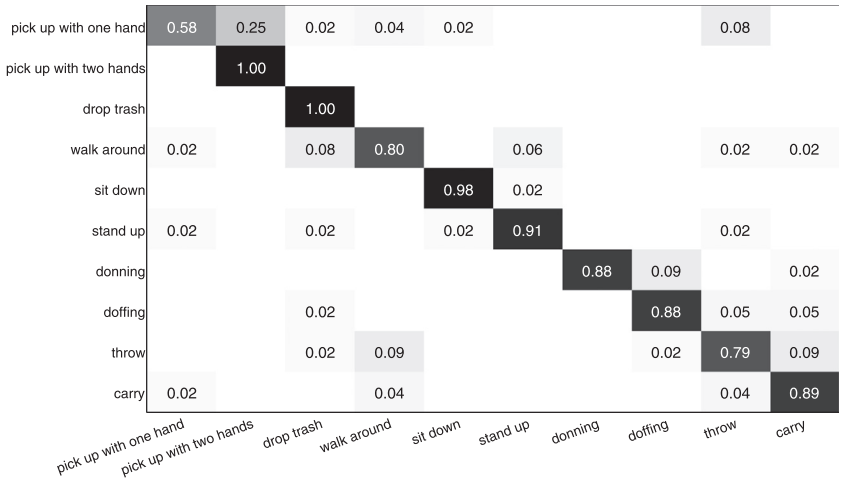


Fig. 7. The confusion matrix of action recognition on the Northwestern-UCLA dataset.

Table 3
Comparison of the proposed approach with the previous methods based on skeletons for action recognition on the Northwestern-UCLA Multiview Action3D dataset.

Method	Accuracy
Hankelet [71]	54.2
HOJ3D [19]	54.5
Actionlet [72]	69.9
MSTAOG [66]	73.3
Lie group [68]	74.2
HBRNN [25]	80.5
Two-stream RNN [29]	85.5
Multi-task RNN	**87.3**

method beats the newest spatio-temporal LSTM with trust gates [27] and two-stream RNN [29] by 4.9% and 3.1%, respectively.

Table 3 summarizes the results on the Northwestern-UCLA Multiview Action3D dataset. Our performance is much higher than the results based on handcrafted features as well as those of RNN based methods. For example, our method outperforms HBRNN [25] and two-stream RNN [29] by 6.8% and 1.8%, respectively.

The results on the UWA3D Multiview Activity II dataset are summarized in Table 4. Here our methods are only compared with the approaches solely based on skeletons. Nearly for all the five splits, our results are significantly higher than the previous reported results. For the mean accuracy, our method is 7.5% higher than the method [68].

For person identification, most of the approaches build on appearance based features and techniques based on skeletons from RGBD data are seldom provided. In addition, person identification experiments are often performed based on a particular action (e.g., walking). Our experiments of joint action recognition and person identification require an input sequence with an unknown action type in unconstrained environments. Thus, there are no public datasets suitable for our task with reported results of person identification. To further demonstrate the effectiveness of our approach, we implement some state-of-the-art methods based on skeleton based features and compare the accuracies of person identification with ours. The results on the challenging NTU RGB+D dataset are provided in Table 5. Similar to the discussion of action recognition, we choose *Middle Split* as the architecture of the multi-task RNN. We first compare our method with methods based on handcrafted descriptors, i.e., the exhaustive combination of distances among joints, distances between the floor plane and all the possible joints [36], and a set of anthropometric measures extracted from poses [42]. We observe that our performance is significantly higher, which shows the superiority of deep learning methods over the methods based on handcrafted features. Our result is also much higher than when compared with multi-layer LSTM network. The results show the benefits of our approach of joint learning content and style for person identification based on skeletons.

5.6. Comparison of training methods

In Section 4.3, we just use a linear combination of the losses of *content* and *style*. As multi-task learning methods are popular for deep neural networks (e.g., [53]), we provide an alternative method by looping between the two tasks. The training procedure are as follows:

1. Select the next task.
2. Select a random training sample.
3. Update the networks for the corresponding objective function by stochastic gradient descent with respect to this sample.
4. Go to 1.

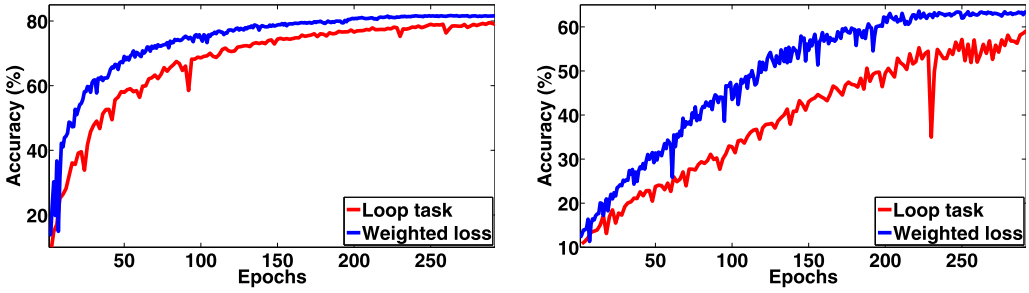
The results on the NTU RGB+D dataset using *Late Split* are shown in Fig. 8. Here, *Weighted loss, Loop task* denote the training method in Section 4.3 and the above alternative approach, respectively. We can observe that, for both tasks, the performances of *Weighted loss* are better than those of *Loop task*. In addition, the results *Weighted loss* exhibits the faster convergence and smaller fluctuation. For example, after 50 epochs, for both action recognition and person identification *Weighted loss* outperforms *Loop task* by 11.1% and 7.6%, respectively. This can be interpreted that for *Loop task*, different losses during training iterations introduce a high level of random fluctuations. The results indicate that it is unnecessary to design training strategies to loop between tasks for joint learning *content* and *style* when the number of objective functions are small.

5.7. Evaluation of parameters

The multi-task RNN has two parameters, i.e., the weight coefficient of the loss of actions and the regularization hyperparameter, denoted by λ and μ, respectively. Here we evaluate the impact of parameters on the performance, and provide the results on the NTU RGB+D dataset using *Late Split*. It should be noted that similar results are observed for other datasets.

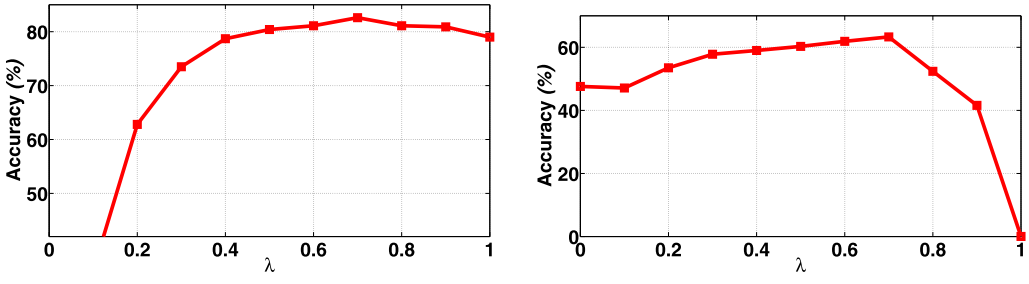
Table 4
Comparison of the proposed approach with the previous methods based on skeletons for action recognition on the UWA3D Multiview Activity II dataset. Each time two views are used for training and the remain two views are individually used for testing.

Training views	V_1 and V_2		V_1 and V_3		V_1 and V_4		V_2 and V_3		V_2 and V_4		V_3 and V_4		Mean
Test view	V_3	V_4	V_2	V_4	V_2	V_3	V_1	V_4	V_1	V_3	V_1	V_2	
HOJ3D [19]	15.3	28.2	17.3	27.0	14.6	13.4	15.0	12.9	22.1	13.5	20.3	12.7	17.7
Actionlet [72]	45.0	40.4	35.1	36.9	34.7	36.0	49.5	29.3	57.1	35.4	49.0	29.3	39.8
Lie group [68]	49.4	42.8	34.6	39.7	38.1	44.8	53.3	33.5	53.6	41.2	**56.7**	32.6	43.4
Multi-task RNN	**50.2**	**50.9**	**51.5**	**52.7**	**50.0**	**46.8**	**56.5**	**47.2**	**55.8**	**48.7**	56.5	**44.4**	**50.9**



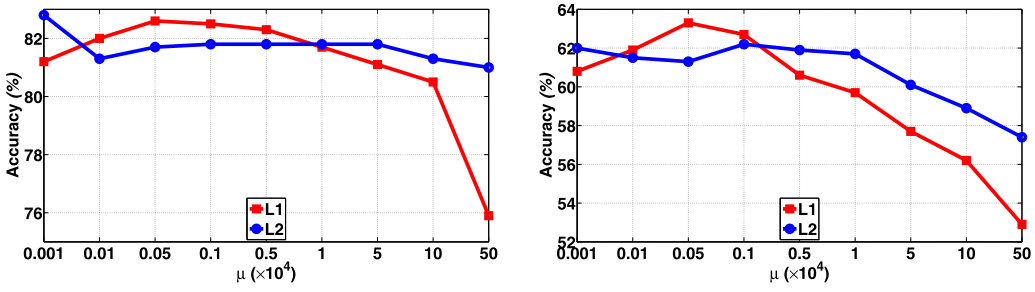
(a) Action (b) Person

Fig. 8. Comparison of training methods for joint learning *content* and *style*. Here, *Weighted loss* is the method which uses a weighted combination of the two losses, and *Loop task* is the alternative method by looping between the tasks.



(a) Action (b) Person

Fig. 9. Sensitivity of the weight coefficient of the loss of action recognition.



(a) Action (b) Person

Fig. 10. Sensitivity of the regularization hyperparameter. We use both L1 regularization and L2 regularization. Here, the scale of the horizontal axis is changed to suit the needs.

Fig. 9 shows the accuracy of action recognition and person identification w.r.t. the parameter λ. We observe that both tasks obtain the best results when $\lambda = 0.7$. The accuracy of action increases markedly with the increase of λ when $\lambda < 0.4$, and maintains a high accuracy when $\lambda > 0.4$. For person identification, the accuracy increases steadily when $\lambda < 0.7$, but starts to decrease sharply when $\lambda > 0.7$. The two extreme points ($\lambda = 1, \lambda = 0$) represent the two separate networks for action and person in the architecture of *Separate Net*, respectively. The accuracy of persons increases as λ increases from 0 to 0.7 owing to the benefits of multi-task learning, and decreases as λ increases from 0.7 to 1.0 because of the shrinking weight coefficient for the loss of person identification. Similar explanation can be achieved by analyzing the accuracy of action recognition. We can find that a good choice of λ for both tasks is $\lambda \in [0.3, 0.7]$.

For the regularization hyperparameter, we choose $\mu \in \{1 \times 10^{-7}, 1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 0.001, 0.005\}$ and plot the results in Fig. 10. For L1 regularization, the best value is $\mu = 5 \times 10^{-6}$ where both accuracies achieve the best performance. The accuracy of action recognition is very promising when $1 \times 10^{-7} <$

Table 5
Comparison of the proposed approach with the state-of-the-art methods for person identification on the NTU RGB+D dataset.

Method	Accuracy
Skeleton features [36]	21.0
Anthropometric measures [42]	28.6
1 layer LSTM	36.5
2 layer LSTM	42.6
3 layer LSTM	43.5
Multi-task RNN	**63.3**

$\mu < 0.001$, but drops sharply when $\mu > 0.001$. While for person identification, the accuracy begins to decrease steadily with the increase of μ when $\mu > 1 \times 10^{-4}$. The results reveal that a smaller μ is more preferred for L1 regularization. For L2 regularization, the accuracy of both tasks maintains a high performance for a much wider range, but the peak values are lower than those of L1 regularization. We conclude that for L1 regularization, the accuracy is not sensitive to μ in a long range (e.g., $\mu < 0.001$), and it is even less sensitive for L2 regularization.

6. Conclusion and future work

In this paper, we present an end-to-end RNN architecture based on multi-task learning to simultaneously conduct action recognition and person identification. The structure consists of two components: skeleton transformation and multi-task RNN. For skeleton transformation, viewpoint transformation and spatial dropout are utilized to learn robust representation. For multi-task RNN, different architectures with different amounts of sharing layers are investigated. We apply the proposed model to skeleton based action recognition with cross-view evaluation and achieve state-of-the-art performances on three benchmark datasets. The experiments demonstrate that for both tasks of person identification and action recognition, learning one task would benefit from learning another task. One potential drawback of the proposed multi-task RNN is that it is impractical to enumerate all possible architectures for each set of tasks to find the best architecture, especially for very deep neural networks. In the future, we will build a model to automatically learn an optimal combination of shared and task-specific representations for multi-task learning.

Acknowledgment

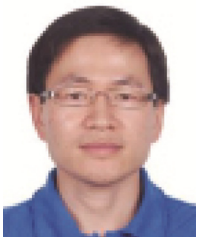
This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004 and 61420106015), Capital Science and Technology Leading Talent Training Project (Z181100006318030) and Beijing Natural Science Foundation (4162058).

References

[1] N.F. Troje, Decomposing biological motion: a framework for analysis and synthesis of human gait patterns, J. Vis. 2 (5) (2002) 2.
[2] J.B. Tenenbaum, W.T. Freeman, Separating style and content with bilinear models, Neural Comput. 12 (6) (2000) 1247–1283.
[3] C.-S. Lee, A. Elgammal, Gait style and gait content: bilinear models for gait recognition using gait re-sampling, in: Proceedings of the 2004 IEEE Conference on Automatic Face and Gesture Recognition, 2004, pp. 147–152.
[4] G. Johansson, Visual perception of biological motion and a model for its analysis, Percept. Psychophys. 14 (2) (1973) 201–211.
[5] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (1) (2013) 116–124.
[6] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the 2015 IEEE Conference on Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.
[7] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, et al., Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, in: Proceedings of the 2015 AAAI, 2, 2016, p. 8.
[8] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.
[9] L. Zhao, X. Gao, D. Tao, X. Li, A deep structure for human pose estimation, Signal Process. 108 (2015) 36–45.
[10] L. Zhao, X.E.A. Gao, Tracking human pose using max-margin Markov models, IEEE Trans. Image Process. 24 (12) (2015) 5274–5287.
[11] L. Zhao, X. Gao, D. Tao, X. Li, Learning a tracking and estimation integrated graphical model for human pose tracking, IEEE Trans. Neural Netw. Learn. Syst. 26 (12) (2015) 3176–3186.
[12] R. Qiao, L. Liu, C. Shen, A. van den Hengel, Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition, Pattern Recognit. 66 (2017) 202–212.
[13] F. Patrona, A. Chatzitofis, D. Zarpalas, P. Daras, Motion analysis: action detection, recognition and evaluation based on motion capture data, Pattern Recognit. 76 (2018) 612–622.
[14] Y. Guo, Y. Li, Z. Shao, DSRF: a flexible trajectory descriptor for articulated human action recognition, Pattern Recognit. 76 (2018) 137–148.
[15] W. Ding, K.E.A. Liu, Tensor-based linear dynamical systems for action recognition from 3D skeletons, Pattern Recognit. 77 (2018) 75–86.
[16] L.L. Presti, M. La Cascia, 3D skeleton-based human action classification: a survey, Pattern Recognit. 53 (2016) 130–147.
[17] J. Zhang, W. Li, P.O. Ogunbona, P. Wang, C. Tang, RGB-D-based action recognition datasets: a survey, Pattern Recognit. 60 (2016) 86–105.
[18] M.E. Hussein, M. Torki, M.A. Gowayyed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in: Proceedings of the 2013 International Joint Conference on Artificial Intelligence, 2013.
[19] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27.
[20] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1290–1297.
[21] X. Yang, Y.L. Tian, Eigenjoints-based action recognition using Naive-Bayes-nearest-neighbor, in: Proceedings of the 2012 IEEE Conference on Computer vision and Pattern Recognition Workshops, 2012, pp. 14–19.
[22] Y. Yacoob, M.J. Black, Parameterized modeling and recognition of activities, in: Proceedings of the 1998 IEEE International Conference on Computer Vision, 1998, pp. 120–127.
[23] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition, J. Vis. Commun. Image Represent. 25 (1) (2014) 24–38.
[24] E. Ohn-Bar, M. Trivedi, Joint angles similarities and HOG2 for action recognition, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 465–470.
[25] Y. Du, Y. Fu, L. Wang, Representation learning of temporal dynamics for skeleton-based action recognition, IEEE Trans. Image Process. 25 (7) (2016) 3010–3022.
[26] V. Veeriah, N. Zhuang, G.-J. Qi, Differential recurrent neural networks for action recognition, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015, pp. 4041–4049.
[27] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal LSTM with trust gates for 3D human action recognition, in: Proceedings of the 2016 European Conference on Computer Vision, Springer, 2016, pp. 816–833.
[28] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Proceedings of the 2017 AAAI, 1, 2017, p. 7.
[29] H. Wang, L. Wang, Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3633–3642.
[30] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1505–1518.
[31] A.F. Bobick, A.Y. Johnson, Gait recognition using static, activity-specific parameters, in: Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition, 1, 2001, pp. 423–430.
[32] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic body biometrics for gait recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (2) (2004) 149–158.
[33] Z. Liu, S. Sarkar, Improved gait recognition by gait dynamics normalization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (6) (2006) 863–876.
[34] J. Man, B. Bhanu, Individual recognition using gait energy image, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 316–322.
[35] M. Deng, C. Wang, Q. Chen, Human gait recognition based on deterministic learning through multiple views fusion, Pattern Recognit. Lett. 78 (2016) 56–63.
[36] I.B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, V. Murino, Re-identification with RGB-D sensors, in: Proceedings of the 2012 European Conference on Computer Vision, Springer, 2012, pp. 433–442.

[37] B. Munsell, A. Temlyakov, et al., Person identification using full-body motion and anthropometric biometrics from kinect videos, in: Proceedings of the 2012 European Conference on Computer Vision Workshops, Springer, 2012, pp. 91–100.
[38] J. Wu, J.E.A. Konrad, Dynamic time warping for gesture-based user identification and authentication with kinect, in: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 2371–2375.
[39] J. Wu, J. Konrad, P. Ishwar, The value of multiple viewpoints in gesture-based user authentication, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 90–97.
[40] I. Kviatkovsky, I. Shimshoni, E. Rivlin, Person identification from action styles, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 84–92.
[41] A. Haque, A. Alahi, F.F Li, Recurrent attention models for depth-based person identification, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1229–1238.
[42] F. Pala, R. Satta, G. Fumera, F. Roli, Multimodal person reidentification using RGB-D cameras, IEEE Trans. Neural Netw. Learn. Syst. 26 (4) (2016) 788–799.
[43] T. Kobayashi, N. Otsu, Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation, in: Proceedings of the 2004 IEEE International Conference on Pattern Recognition, 4, 2004, pp. 741–744.
[44] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, F.F Li, Detecting events and key actors in multi-person videos, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3043–3053.
[45] W. Xie, H. Yao, X. Sun, S. Zhao, T. Han, C. Pang, Mining representative actions for actor identification, in: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 1253–1257.
[46] H. Wang, L. Wang, Cross-agent action recognition, IEEE Trans. Circuits Syst. Video Technol. (2018), doi:10.1109/TCSVT.2017.2746092. in press.
[47] T. Batabyal, A.E.A. Vaccari, UGRASP: a unified framework for activity recognition and person identification using graph signal processing, in: Proceedings of the 2015 IEEE International Conference on Image Processing, 2015, pp. 3270–3274.
[48] R. Caruana, Multitask Learning, in: Learning to Learn, Springer, 1998, pp. 95–133.
[49] G. Obozinski, B. Taskar, M.I. Jordan, Joint covariate selection and joint subspace selection for multiple classification problems, Stat. Comput. 20 (2) (2010) 231–252.
[50] A. Quattoni, M. Collins, T. Darrell, Transfer learning for image classification with sparse prototype representations, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[51] X. Zhang, M.H. Mahoor, Task-dependent multi-task multiple kernel learning for facial action unit detection, Pattern Recognit. 51 (2016) 187–196.
[52] Y. Zheng, J.E.A. Fan, Hierarchical learning of multi-task sparse metrics for large-scale image classification, Pattern Recognit. 67 (2017) 97–109.
[53] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of the 2008 International Conference on Machine Learning, ACM, 2008, pp. 160–167.
[54] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (Aug) (2011) 2493–2537.
[55] R. Girshick, Fast R-CNN, in: Proceedings of the 2015 IEEE Conference on International Conference on Computer Vision, 2015, pp. 1440–1448.
[56] C. Zhang, Z. Zhang, Improving multiview face detection with multi-task deep convolutional neural networks, in: Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1036–1041.
[57] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial landmark detection by deep multi–task learning, in: Proceedings of the 2014 European Conference on Computer Vision, Springer, 2014, pp. 94–108.
[58] A.H. Abdulnabi, G.E.A. Wang, Multi-task CNN model for attribute prediction, IEEE Trans. Multimed. 17 (11) (2015) 1949–1959.
[59] G. Gkioxari, B. Hariharan, R. Girshick, J. Malik, R-CNNs for pose estimation and action detection, CoRR (2014). abs/406.5212
[60] X. Chu, W. Ouyang, et al., Multi-task recurrent neural network for immediacy prediction, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015, pp. 3352–3360.
[61] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[62] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[63] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognit. 68 (2017) 346–362.
[64] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 648–656.
[65] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, CoRR (2012). abs/1207.0580.
[66] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656.
[67] H. Rahmani, A. Mahmood, D. Huynh, A. Mian, Histogram of oriented principal components for cross-view action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (12) (2016) 2430–2443.
[68] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a Lie group, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 588–595.
[69] G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: human action recognition using joint quadruples, in: Proceedings of the 2014 IEEE International Conference on Pattern Recognition, 2014, pp. 4513–4518.
[70] J.-F. Hu, W.-S. Zheng, J. Lai, J. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5344–5352.
[71] B. Li, O.I. Camps, M. Sznaier, Cross-view activity recognition using Hankelets, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1362–1369.
[72] J. Wang, Z. Liu, Y. Wu, Learning actionlet ensemble for 3D human action recognition, in: Human Action Recognition with Depth Cameras, Springer, 2014, pp. 11–40.

Hongsong Wang received the B.S. degree in automation from Huazhong University of Science and Technology 2013. He is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include action recognition, video classification, and deep learning.



Liang Wang (SM'09) received both the B.S. and M.S. degrees from Anhui University in 1997 and 2000 respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he has been working as a Research Assistant at Imperial College London, United Kingdom and Monash University, Australia, a Research Fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full Professor of Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition and computer vision. He has widely published at highly-ranked international journals such as IEEE TPAMI and IEEE TIP, and leading international conferences such as CVPR, ICCV and ICDM. He is an associate editor of IEEE Transactions on SMC-B. He is currently an IAPR Fellow and Senior Member of IEEE.