Learning Representative Deep Features for Image Set Analysis

Zifeng Wu, Yongzhen Huang, *Member, IEEE*, and Liang Wang, *Senior Member, IEEE*

***Abstract*—This paper proposes to learn features from sets of labeled raw images. With this method, the problem of over-fitting can be effectively suppressed, so that deep CNNs can be trained from scratch with a small number of training data, i.e., 420 labeled albums with about 30 000 photos. This method can effectively deal with sets of images, no matter if the sets bear temporal structures. A typical approach to sequential image analysis usually leverages motions between adjacent frames, while the proposed method focuses on capturing the co-occurrences and frequencies of features. Nevertheless, our method outperforms previous best performers in terms of album classification, and achieves comparable or even better performances in terms of gait based human identification. These results demonstrate its effectiveness and good adaptivity to different kinds of set data.**

***Index Terms*—Album classification, deep learning, gait recognition, image set.**

I. Introduction

In THE context of computer vision, besides isolated still images, various kinds of image sets compose the rest body of raw data. These include videos [1]–[3][1] and collections of images [4], [5]. They can have temporal structures, e.g., gait sequences [6] and albums [4], or not, e.g., faces in videos [7], photos of objects from multiple viewpoints [5]. Compared to a single image, a set of them can convey much richer and sometimes higher level semantic information. For example, a video can better express the shape and appearance of an object than a still image, since it can record that object in all possible viewpoints [5]. And a collection of photos can tell the whole story of some event, i.e., how it happened, developed and concluded [4].

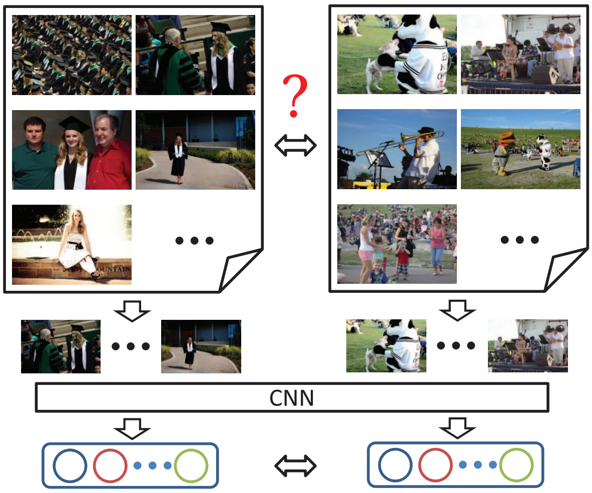


Fig. 1. In order to analyze sets of images, we learn deep features which are directly comparable and can fit into any off-the-shelf classifiers. The photos in this figure are taken from the personal event classification (PEC) dataset [4].

To deal with general image set data, some approaches resort to subspace learning to obtain set-level features [8], [9], others adopt metric learning to evaluate the similarity between pairs of image sets [10], and some approaches learn structured models to explore hierarchical concepts [11]. To deal with temporal sequential data, many approaches resort to sophisticated models such as hidden Markov models (HMMs) [12] and conditional random fields (CRFs) [13]. With these graph-based models, they can model the temporal structure and the transitions between states in a sequence of images. The limitation is that most of the above mentioned methods rely on handcrafted low-level features, which might not be optimal for a specific task.

Recently, the convolutional neural network (CNN) [14] has been widely accepted as the most powerful tool for feature learning from still images [15]. The target of this paper is to learn deep features from labeled sets of images. In this way, we can replace the handcrafted features with these on-line learned features, which can hopefully boost the performances of approaches to image set analysis, just as they did for those approaches to image analysis [15]. On the other hand, by turning image sets into features with a fixed dimension, we can analysis sets of images more conveniently, for example as illustrated in Fig. 1, evaluating the similarity between two albums directly with the Euclidean distance or classify them with off-the-shelf tools such as support vector machines. In the course to that goal, we will be faced with at least two problems. Take the personal event classification (PEC) dataset [4] for example. First, the number of contained photos can vary

Manuscript received March 15, 2015; revised July 02, 2015; accepted August 24, 2015. Date of publication September 10, 2015; date of current version October 20, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316300, by the National Natural Science Foundation of China under Grant 61135002 and Grant 61420106015, by the CCF-Tencent Open Fund, and by the 360 OpenLab Program. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Guo-Jun Qi.

Z. Wu is with the Australian Centre for Visual Technologies, University of Adelaide, Adelaide, SA 5005, Australia (e-mail: zifeng.wu@adelaide.edu.au).

Y. Huang and L. Wang are with the National Laboratory of Pattern Recognition, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yzhuang@nlpr.ia.ac.cn; wangliang@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2477681

[1] “The first international workshop on action recognition with a large number of classes,” [Online]. Available: http://crcv.ucf.edu/ICCV13-Action-Workshop

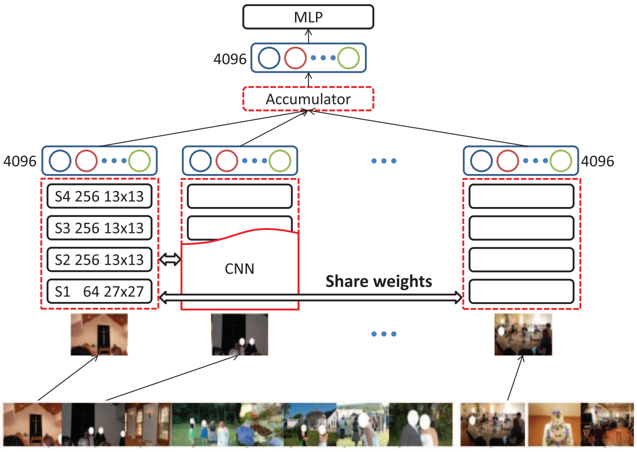


Fig. 2. Illustration of our approach to feature learning from image sets. The example is actually a temporal sequence of photos about a wedding event taken from the personal event classification (PEC) dataset [4]. See the text for details. Better viewed in color.

across different albums, i.e., from about a dozen to more than a hundred. Second, the number of training data is limited, i.e., 420 albums composed of about 30,000 photos. To this end, we in this paper propose to train deep CNNs from randomly sampled subsets of images. In spite of the small number of training data, the method can still effectively train very deep models from scratch.

CNN composes the basic and kernel parts of all the champion algorithms in the ImageNet large scale visual recognition challenge (ILSVRC)[2] since 2012, when Krizhevsky *et al.* won the first place with the AlexNet [15]. And last year, the extremely deep GoogLeNet [16] with 22 trainable layers has reduced the top-5 classification error rate down to 6.67% on this challenging dataset with one thousand categories. Besides the image classification tasks, features learned by CNNs can also boost the performances of various methods for other image-based tasks, e.g., CNN features of regions for object detection [17] and hierarchical features for scene labeling [18]. However, in the context of image set data analysis, the development of CNN-based feature learning seems less satisfactory. Some methods just tune the features learned from still images with single frames,[1] which can not make full use of the characteristics of set data. One recent CNN-based feature learning method proposed by Simonyan and Zisserman [19] let one CNN learn from raw frames, and another one learn from multiple optical flow maps between frames. Score-level fusion was applied so that the two can be trained separately. As a result, their method combined the appearances and motions in videos. In this sense, our method is very different from theirs, which ignores the motion information in image sets. By equally treating randomly picked images from each set, our method focuses on the frequencies and co-occurrences of discriminative features.

The first contribution of this paper is the proposed network architecture as illustrated in Fig. 2, within which the features of randomly picked images are accumulated (by summation) to compute set-level representations. The network can be trained as a whole, which is different from those approaches applying afterwards score-level fusion , [19]. Besides, as the second contribution, we verify our method and its variations on two kinds of image set data, i.e., the personal event classification (PEC) dataset [4] and the CASIA-B gait database [6]. Particularly, on the PEC dataset, our method outperforms previous methods with a large margin.

The next section will uncover more details of the proposed method. After that, we will apply it to album classification in Section III and cross-view gait based human identification in Section IV, before we conclude this paper in Section V.

II. Method

A. Overview

The main idea of our method is illustrated in Fig. 1. Given image sets with category labels, the whole network, including the convolutional neural network (CNN) for feature extraction and the multi-layer perceptron (MLP) for set classification, can be trained in an end-to-end manner with the back-propagation algorithm. In each round of training or testing:

- randomly pick a given number T of images from a set;
- feed each of them respectively into the same CNN which extracts 4,096-dimensional frame-level features;
- third, accumulate these features to obtain a global representation for the set;
- and finally, feed the representation into the MLP for classification.

At this point, consider T to be a number around a dozen, e.g., eight. Also note that the extracted features can be of any tractable dimension. We just occasionally use this setting for both of the two tasks in this paper.

By averaging multiple randomly-picked image-level features within the network, our method is supposed to be advantageous in at least three aspects.

- Uses more reliable labels compared with the approaches considering images separately. Take the PEC dataset [4] for example, in an album, there might be a number of photos which almost have nothing to do with the album's category label. Fitting models for these photos has no merits and results in over-fitting. However, T photos in an album, in most cases, can compose a large part of the whole story of a personal event.
- Works well for relatively small training dataset. In the PEC dataset, there are about 75 photos in each album on average. Either using each photo separately or using each album as a whole leads to a small set of training data. However, if we pick T photos from each album, there will be a huge number of possible combinations of subsets. It can help us with combating over-fitting, which is vital for training deep CNNs.
- Ensures the validity of subsequent sum operations during training. This is convenience since we can thus sum any number of frame-level features to obtain the representation of an image set, and directly use the MLP spontaneously learned with the CNN to predict the category of that set. In approaches considering images separately, image-level features are also summed for each set to obtain a global

[2] "ImageNet large scale visual recognition challenge 2012 (ILSVRC 2012)," [Online]. Available: http://www.image-net.org/challenges/LSVRC/2012/

representation [20]. However, the features are not inherently learned for such purpose. Accordingly, they can rely on subsequently trained shallow classifiers, usually SVMs, to select discriminative features from the global representations.

B. Network Details

The whole network is composed of three parts, i.e., the feature extractor, the accumulator and the classifier, as illustrated in Fig. 1. We will introduce them respectively below.

The feature extractor, i.e., the CNN in Fig. 1, is shared among different images. Usually, it is composed of several stages. The key component of these stages could be any one of the three, i.e., a fully-connected layer, a locally-connected layer or a convolution layer. Usually, there should be a following non-linear activation function. Besides, optional components can include a spatial pooling layer and a normalization layer.

A convolution layer can be derived from a fully-connected layer via two steps of simplifications in order to reduce the number of parameters. In a fully-connected layer, as the name tells, its nodes are fully connected to those of its previous layer. The first simplification is to impose the spatial locality so that the nodes are only connected to local regions of the previous layer. This kind of layers are also known as the locally-connected layers. The next simplification is to share the weights across all spatial locations. Recently, a typical CNN usually is composed of several convolution and fully-connected layers, which works soundly in most cases. However, when a layer is supposed to be spatially sensible and its input layer has a large number of nodes, it had better be locally-connected for the sake of less parameters. The classic non-linear activation functions include the hyperbolic tangent function $\tanh(x)$ and the logistic function $f(x) = 1/(1 + e^{-x})$. However, Krizhevsky *et al.* [15] pointed out that networks with the rectified linear unit (ReLU) [21] $f(x) = max(0, x)$ can be trained several times faster due to ReLU's non-saturating characteristic. ReLU might not be the optimal choice for performance, but it is favorable for the sake of efficiency. The spatial pooling amounts to down-sampling by preserving only one activity for each local region of a feature map. The preserved value can either be the maximum or the average activity within that region. Empirical results show that max pooling performs better in most cases.

Given the above stated settings, the l-th convolution stage's activities $\boldsymbol{H}_l$ can be concisely formulated as

$$\boldsymbol{H}_l = \mathrm{pool}(\max(0, \boldsymbol{W}_l \otimes \boldsymbol{H}_{l-1} + \boldsymbol{b}_l)) \tag{1}$$

wherein $\otimes$ denotes the convolution operation, $\boldsymbol{H}_{l-1}$ is the input given by the previous layer (so $\boldsymbol{H}_0$ is the original input data), $\boldsymbol{W}_l$ contains a number of filters, and $\boldsymbol{b}_l$ contains a number of biases shared across different spatial locations. To compute each feature map in $\boldsymbol{H}_l$, accordingly, there will be one filter in $\boldsymbol{W}_l$ and one entry in $\boldsymbol{b}_l$ (the bias). Note that the ReLU is also included in (1), as well as the spatial max pooling.

Considering that ReLU never saturates in $[0, +\infty)$, it is safe to feed data into networks with no local contrast normalization, as long as there are some examples producing positive activities [15]. However, Krizhevsky *et al.* [15] also reported that their proposed cross-map local response normalization can aid generalization. It implements a form of lateral inhibition, introducing competition among the big activities on adjacent feature maps.

For an activity a_i at certain spatial location on the i-th feature map, the cross-map normalized activity b_i can be computed as [15]

$$b_i = a_i / \left(\gamma + \alpha \sum_{j \in \mathrm{nb}(k,i)} (a_j) \right)^{\beta} \tag{2}$$

wherein α, β, γ and k are all configurable parameters. Once a network gets initialized, its feature maps will be arranged in certain order. Let there be N feature maps, and the k neighbors of the ith feature map $\mathrm{nb}(k, i)$ will be $\{j | j = \max(0, i - n/2), \cdots \min(N - 1, i + k/2)\}$. Notably, only the activities at the same spatial location participate in this kind of normalization.

There are millions of parameters in a deep CNN. Usually, for a specific task, the given data can not afford to train the model due to over-fitting. Besides increasing the number of training data or applying data augmentation, Krizhevsky *et al.* [15] reported that the dropout technique [22] is often helpful for combating over-fitting. It amounts to dropping nodes with a rate of 50% during training. Dropped nodes will neither contribute to the forward nor the backward propagation. Accordingly, the activities should be multiplied by 0.5 during testing. Dropout has been explained as an efficient way for combining multiple networks [22], which reduces co-adaptations of nodes and forces networks to learn more robust features. In this paper, we do not apply dropout in the feature extractor. Instead, we apply it to the accumulated set-level representations, as illustrated in Fig. 3 and 5.

The feature extractor in our method are composed of the above mentioned components. For example in Fig. 2, there are four convolution stages and one fully-connected layer. It extracts 4,096-dimensional image-level features from images, which are then fed into the accumulator.

An accumulator adds up but not concatenates the activations. This is the reason why it has a 4,096-dimensional output. Suppose that the number of images picked out from each image set is T. The accumulator can be concisely formulated as

$$\boldsymbol{r} = \boldsymbol{H}^1 \oplus \boldsymbol{H}^2 \oplus \cdots \oplus \boldsymbol{H}^T \tag{3}$$

wherein $\oplus$ denotes the element-wise sum operation, $\boldsymbol{H}^t$ is the output of the last layer in the feature extractor for the t-th image picked from a set, and $\boldsymbol{r}$ is the global representation of that set.

Together with the CNN for feature extraction, we train a multi-layer perceptron (MLP) for classification with the Logistic regression loss. During test, we sample M groups of images from each set, feed them into the network to compute M scores, and finally predict the label with the average score. Or in formulate, we find

$$\operatorname*{arg\,max}_{i=1,\cdots,N_c} \sum_{m=1}^{M} s_i^m \tag{4}$$

wherein N_c is the number of possible categories, and s_i^m is the score for the ith category and the mth sampled group of images.

C. Implementation Details

To combat over-fitting, one can apply data augmentation to the training data, i.e., transforming the original examples in various ways, e.g., rescaling, rotating, flipping and cropping. For the sake of efficiency, we only apply the last two kinds of transforms in this paper. During training, we first randomly pick out a number of images from each set, then we take out a random crop from each of the images, and third, we also flip the crop with a probability of 50%. During testing, we always take the central crop and never flip it. For either case, we subtract the mean of all the images in the training set from each of the sampled crops.

Following the suggestion by Krizhevsky *et al.* [15], we set the four configurable parameters in (2) as $\alpha = 10^{-4}$, $\beta = 0.75$, $\gamma = 2$ and $k = 5$. We train the networks using back-propagation with the logistic regression loss, and update the weights with a mini-batch size of 128. We initialize the weights of each layer using a Gaussian distribution with a mean of zero and a standard deviation of 0.01, and the biases of nodes in all layers with the constant zero (if not specified). We start with a learning rate of 0.01, and reduce it to 0.001 when the accuracy on the evaluation set stops improving. For all layers, the momentums for weights and biases are 0.9, and the weight decay is 0.0005.

We following a simple strategy for boot-strapping in this paper based on two considerations. On one hand, the random sampling strategy results in a large number of possible combinations of images for each set. It is sometimes intractable to cover all of them. On the other hand, although the number of combinations is very large, many of them share notable overlaps. A deep network can quickly fit a great part of the combinations, but at the same time leaves a number of hard samples. It will become more and more hard for the networks to be fed with these samples, since they only compose a small proportion of all the combinations. To this end, we keep the wrongly classified training samples in a pool and sort them according to their losses in descending order. Within each mini-batch, we pick at most half of the samples from the pool. This strategy is supposed to speed up the training process.

III. Album Classification

Album classification amounts to predicting the category of a group of photos. Or in formulation, suppose there is an album $\boldsymbol{x}$ and a set of possible labels $\{c_i | i = 1, \cdots, N_c\}$, the category of the album $y(\boldsymbol{x})$ is to be predicted. To train models in a supervised manner, a set of labeled albums are given as $\{(\boldsymbol{x}_i, y_i) | i = 1, \cdots, N\}$.

To deal with photo collections, many approaches focused on exploiting the collection structure that is often found in personal and professional photo archives. To name a few, Cao *et al.* [23] reduced the complexity of propagating labels between images organized within collections, observing that image in the same collection tend to depict similar scenes. They [24] further extended this idea into a hierarchical model which split a photo collection into a sub-sequence of *events*, composed of images depicting similar scenes. Mattivi *et al.* [25] proposed to aggregate the SVM scores of each photo in a collection, and use that score for classification into eight social classes. As for videos, Izadinia and Shah [26] exploited known sub-events as an intermediate representation of collections for event classification. They discarded time information in favor of co-occurrence of sub-events. Instead, Tang *et al.* [12] treated sub-events as unobserved latent variables. They associated these sub-events with explicit durations. Transitions from one sub-event to another can only occur when the previous one has expired, which requires sub-events and their boundaries to be fully observed. Due to the sparsely sampled photos in albums, these boundaries are often missing in the context of album classification. As a result, Bossard *et al.* [4] had to adapt that model. Inspired by discretely observed Markov jump processes, they proposed a Markov model where transition probabilities are functions of the temporal gap between images as if it were measure by a stopwatch.

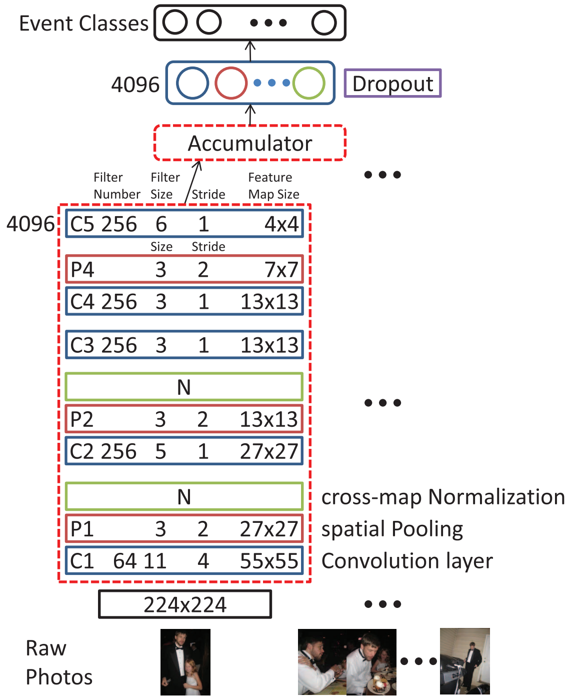


Fig. 3. Details of the network for feature learning in terms of album classification.

A. Learning Features From Albums

A typical method for album classification often resorts to sophisticated models such as HMMs to exploit the transmissions between underlying states [4]. However, since the method proposed in this paper focuses on feature learning, we can directly train networks for album classification. The details of the used network is illustrated in Fig. 3. During training, we randomly pick out a number of photos from each album, and feed them into the network as a single sample. We randomly take out crops in size 224 × 224 from each of the photos. During testing, we always take the central crop. We anticipate the network to learn complex object and scene patterns. To this end, the network is derived from the very deep AlexNet [15]. We remove the first

TABLE I
IMPACT OF T ON CLASSIFICATION ACCURACY (%) EVALUATED WITH OUR VALIDATION SET OF THE PEC DATASET [4]

	M=1	M=2	M=4	M=8	M=16	M=32	M=64	M=128	M=256
T=1	35.6±5.6	44.2±5.3	50.4±4.6	54.6±4.7	55.8±7.0	58.8±4.2	60.2±4.3	59.9±6.7	53.6±3.3
T=2	44.9±4.0	51.7±5.0	53.9±6.6	56.0±8.9	58.7±3.5	61.4±3.5	57.7±4.0	57.1±5.1	–
T=4	52.1±4.2	60.7±4.6	59.4±3.5	66.3±4.9	65.4±4.2	63.5±5.4	64.5±5.4	–	–
T=8	59.4±5.5	64.4±4.6	64.6±5.5	64.5±5.6	68.8±5.0	67.6±3.6	–	–	–
T=16	65.6±6.4	65.5±6.2	66.5±5.8	**70.1**±2.7	67.7±4.7	–	–	–	–

TABLE II
COMPARISON OF OUR METHOD WITH PREVIOUS ONES ON THE TEST SET OF THE PEC DATASET BY CLASS-WISE AND AVERAGE ACCURACIES (%). $M = N_{\text{im}}$ MEANS APPLYING SCORE-LEVEL FUSION ON IMAGES IN EACH SET. A-NET MEANS THAT THE NETWORK IS INITIALIZED WITH PRETRAINED ALEXNET [15]

Method	Birthday	Children's birthday	Christmas	Concert	Cruise	Easter	Exhibition	Mean (%)	Recall@2 (%)
No motion [5]	0	30	50	100	80	50	70	51.43	70.00
With motion [5]	10	30	70	100	50	50	70	55.71	72.86
T=1, M=N_{im}	0	30	100	100	80	10	40	56.43	77.14
T=1, M=64	4.0±4.9	28.0±6.0	95.0±5.0	98.0±4.0	67.0±7.8	16.0±4.9	36.0±4.9	55.71±1.53	74.71±1.97
T=16, M=8	0.0±0.0	56.0±6.6	77.0±6.4	89.0±3.0	68.0±7.5	27.0±9.0	79.0±3.0	**63.36**±1.50	77.07±2.01
T=1, M=64, A-net	6.0±6.6	49.0±3.0	96.0±4.9	99.0±3.0	75.0±8.1	44.0±4.9	69.0±10.4	71.71±1.29	85.36±0.86
T=16, M=8, A-net	12.0±7.5	57.0±4.6	89.0±3.0	100.0±0.0	82.0±4.0	44.0±4.9	75.0±5.0	**73.43**±1.27	**90.43**±1.25

Method	Graduation	Halloween	Hiking	Road trip	Saint Patrick's day	Skiing	Wedding	F1 score (%)	
No motion [5]	40	0	90	20	60	80	50	50.63	
With motion [5]	40	30	80	40	30	100	80	56.16	
T=1, M=N_{im}	30	20	0	100	100	100	80	50.01	
T=1, M=64	32.0±4.0	30.0±7.7	1.0±3.0	97.0±6.4	98.0±4.0	100.0±0.0	78.0±6.0	50.58±1.76	
T=16, M=8	60.0±0.0	41.0±8.3	27.0±4.6	98.0±4.0	100.0±0.0	88.0±4.0	77.0±4.6	**60.79**±1.57	
T=1, M=64, A-net	76.0±8.0	74.0±12.8	27.0±10.0	100.0±0.0	96.0±4.9	100.0±0.0	93.0±7.8	68.93±0.98	
T=16, M=8, A-net	69.0±8.3	82.0±4.0	52.0±7.5	91.0±5.4	98.0±4.0	100.0±0.0	77.0±4.6	**71.56**±1.54	

two fully-connected layers from the network in order to alleviate the problem of over-fitting.

B. Experiments

Album classification amounts to assigning category labels to collections of photos. The most recent dataset to this end is the personal event classification (PEC) dataset. There are 809 albums in total, composed of more that 61,000 images, belonging to 14 social events such as boat cruise, graduation, wedding and birthday.

We in this paper follow the standard protocol proposed by the authors of this dataset [4]. Specifically, we will keep their specified 140 albums (ten per class) for test, randomly pick out 84 albums (six per class) for validation and train networks with the rest albums. And finally, class-wise and average classification accuracies are reported. Since we can sample different subsets of photos from an album, the results can vary across different runs. For this reason, we test the trained models for ten times, and report the means and standard deviations.

1) Impact of T : We first verify the impact of the hyper parameter T on performance, i.e., the number of accumulated images in each sample. The results on our validation set are listed in Table I, wherein M is the number of views during testing. Namely, we randomly sample M group of photos and average their scores to predict the album's label. For the cases with a small T, the problem of over-fitting hinders networks from learning discriminative features. Although the performance can be improved by increasing M, but it will saturate at a unsatisfactory level soon. According to the results, the optimal settings are $T = 16$ and $M = 8$. We follow this setting in the rest experiments of this paper.

Note that our method will approximately degrade into a trivial approach to image set classification. It amounts to feeding a network with single images and the labels of their sets during training and applying score-level fusion during testing. The minor different part is that M should be the very number of images for each set. Results obtained with this strategy on the test set are given in Table II. Although there is an improvement smaller than 1%, i.e., from 55.71% to 56.43, the accuracy is still worse than our best case (63.36%). This result shows the importance of feeding a network with sets of images.

2) Comparison With Previous Methods: The comparison of our method with those in the literature is presented in Table II. On average, our method (63.36 ± 1.50%) outperforms the best previous performer [4] (55.71%) by more than 7%. Besides, the recall@2 rate of our method (77.07 ± 2.01%) is higher than the

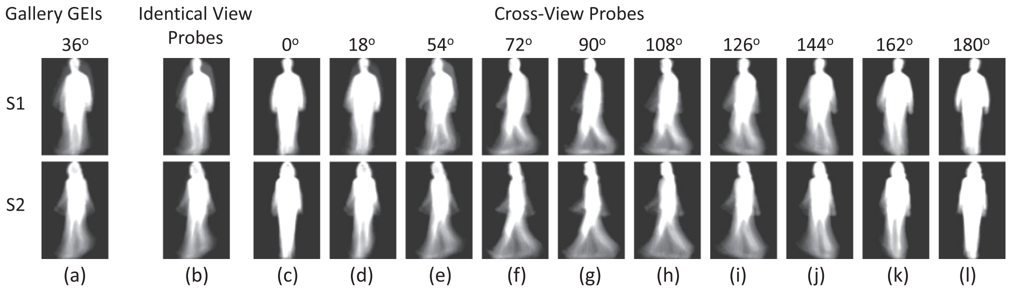


Fig. 4. Example GEIs of two subjects (S1-S2) in the CASIA-B gait dataset [6]. Column A: GEIs in the gallery, with view angle 36°. Column B: probes with the same view angle. Columns C–L: probes with view angle variations. Gait recognition amounts to identifying the most similar gallery GEI for each probe.

previous best result (72.86%) by more than 4%, and the F1 score (57.68%) is also better than 56.16. Note that our method does not make use of any motion information at all. Nevertheless, the obvious improvement in performance shows the effectiveness of our learned features.

To initialize the network with the pre-trained 1000-way AlexNet [15] can further improve the performance, as shown in Table II. In this case, the training of a network can start from a much better initial state, which can somehow suppress over-fitting. As a result, the improvement gained by feeding a network with sets of images becomes smaller.

IV. Gait Recognition

Gait recognition amounts to predicting the identity of a probe sample, given a gallery which is composed of gait samples registered with identities in advance. Or in formulation, suppose there is one probe sample $\boldsymbol{x}$ and N_{g} samples in the gallery $\{(\boldsymbol{x}_i, y_i = i) | i = 1, \cdots, N_{\mathrm{g}}\}$, where y_i denotes the identity of sample $\boldsymbol{x}_i$. Given the above data, the identity of probe $y(\boldsymbol{x})$ is to be predicted.

Many of the widely-used gait recognition datasets provide gait energy images (GEI) [27], which are the average silhouettes along the temporal dimension. For example, some GEIs of two subjects in the CASIA-B gait dataset [6] are shown in Fig. 4. There are eleven viewpoint considered in this dataset, i.e., from 0° to 180° with a step of 18°. In the easiest case, probe GEIs and those in the gallery are in an identical viewpoint. Computing the similarities based on the Euclidean distance achieved pretty good results [6]. This paper considers the cross-view cases, which are much harder to deal with [6]. In these cases, probe GEIs and those in the gallery are in different viewpoints. There are many alternatives for GEIs, e.g., chrono-gait images [28] and gait flow images [29]. However, a recent empirical study by Iwama *et al.* [30] shows that GEI, despite of its simplicity, is the most stable and effective kind of features for gait recognition on their proposed dataset with 4,007 subjects.

Cross-view gait recognition methods can be roughly divided into three categories. The first category is based on 3D model of human body [31]–[33], while the second category is based on handcrafted view-invariant features. The methods, most related to this paper, belong to the third category, which amounts to learning the projections across different viewpoints. These methods rely on the training data to cover the views which appear in the gallery and probe samples. With learned mapping matrices, gait features in different views can be projected into certain common subspace for better matching. Compared with the first two categories of cross-view gait recognition methods, the third category can be applied for scenarios with no explicit action by subjects, and can also be directly applied to views which are significantly different from the side view, e.g., frontal or back view.

To name a few methods in the third category, Makihara *et al.* [34] proposed an SVD-based view transformation model (VTM) to project gait features from one view into another. Kusakunniran *et al.* [35] used truncated SVD to avoid the oversizing and over-fitting problem of VTMs. After pointing out the limitations of SVD-based VTMs, they reformulated the VTM reconstruction problem as a support vector regression (SVR) problem [36]. They selected local regions of interests based on local motion relationships, instead of global features [34], [35], to build VTMs through support vector regression. They further improved the performance by introducing sparsity to the regression [37]. Instead of projecting gait features into one common space, Bashir *et al.* [38] used canonical correlation analysis (CCA) to project each pair of gait features into two subspaces with maximal correlation. Kusakunniran *et al.* [39] claimed that there may exist some weakly-correlated or non-correlated information in global gait features across views and carried out motion co-clustering to partition the global gait features into multiple groups of segments. They applied CCA on these segments, instead of using the global gait features as Bashir *et al.* did in [38]. Most of the above mentioned methods trained multiple mapping matrices, one for each pair of viewpoints. Recently, Hu *et al.* [40] proposed to apply a unitary linear projection, named as view-invariant discriminative projection (ViDP). The unitary nature of ViDP enabled cross-view gait recognition to be conducted without knowing the query gait views. On the other hand, Hu [41] designed a kind of gait feature named as enhanced Gabor gait (EGG), which encodes both statistical and structural characteristics with a non-linear mapping. The regularized local tensor discriminant analysis (RLTDA) was applied for dimensionality reduction. RLTDA was supposed to be able to capture the nonlinear manifolds which are robust against view variations, but it is sensitive to initialization. For that reason, a number of RLTDA learners were accordingly fused for obtaining better performance.

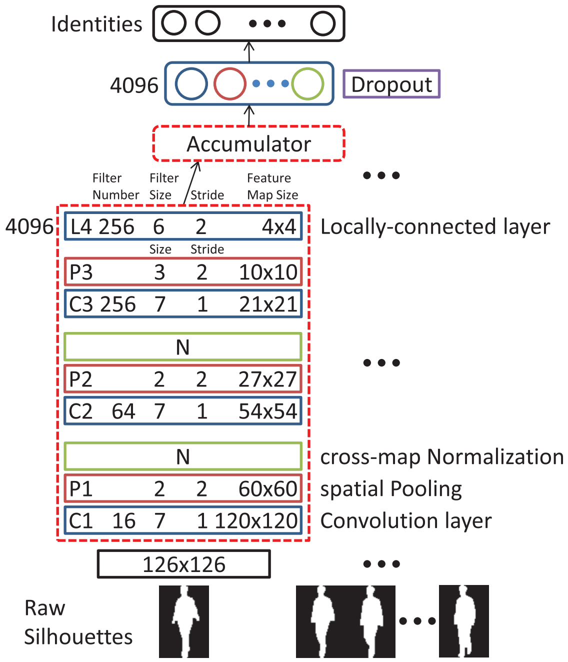


Fig. 5. Details of the network for feature learning in terms of identity classification.

A. *Learning Features From Silhouettes*

Being different from the above mentioned methods, our method in this paper uses features learned from raw silhouettes instead of GEIs. These features are trained in terms of person classification on a training dataset. In the next step, we directly use the Euclidean distance to measure the similarity between each pair of gait features, which is necessary for the subsequent identification step. This measure is not supposed to be optimal. Nevertheless, we here focus on feature learning, and leave more stronger methods as our future work. For example, we can train similarity models with pairs of gait features, labeled as *identical* when they are from the same person or *different* when from two persons. Similar methods have already been presented in the literature [42], [43].

The used network is illustrated in Fig. 5 in detail. During training, we randomly pick out a number of raw silhouettes from each gait sequence, and feed them into the network as a single sample. The silhouettes are cropped and rescaled into size 128 × 88 in advance. We pad them with zeros into size 128 × 128, and randomly take out crops in size 126 × 126 from them. During testing, we always take the central crop. Considering that the local patterns in these binary silhouettes are simpler than those in colored images, we use smaller number of filters in the first two convolution stages. Considering that the silhouettes are roughly aligned in advance, we use a locally-connected layer (L4) to compute the frame-level features. In this way, we can make use of the global structure of gait silhouettes. Apparently, it is reasonable to compute features for heads and legs in different manners. Considering that comparison of local details is vital for gait matching, we make smaller the size of receptive field, i.e., 94 pixels for the nodes on the L4 layer. For comparison, the one on the C5 layer in Fig. 3 is 224.

TABLE III
COMPARISON OF OOR METHOD WITH PREVIOUS ONES ON CASIA-B BY AVERAGE ACCURACIES (%)

Gallery NM #1-4	0°-180°			36°-144°		
Probe NM #5-6	54°	90°	126°	54°	90°	126°
CCA [40]	–	–	–	66	66	67
ViDP [42]	64.2	60.4	65.0	87.0	87.7	89.3
Ours	**77.7**	59.9	**75.0**	**89.2**	81.5	**90.0**

B. *Experiments*

We verify our method with the CASIA-B gait dataset [6]. There are 124 subjects in total, and 110 sequences per subject. Specifically, there are eleven views $(0^\circ, 18^\circ, \cdots, 180^\circ)$ and ten sequences per subject for each view. Among the ten, six are taken under normal walking conditions (NM). Four of the six are in the gallery (NM #1-4) and the rest two are kept as probes (NM #5-6). Another two are taken when the subjects are in their coats (CL), kept as probes (CL #1-2), and the remaining two are taken with bags (BG), also kept as probes (BG #1-2). Example GEIs extracted from this dataset can be found in Fig. 4. Cross-view gait recognition on this dataset is challenging, especially when the cross-view angle is larger than 36° [6], [39].

We apply the non-overlapping test strategy. Specifically, we train networks with all the gait sequences of the first 74 identities, and keep those of the remaining 50 identities for test. The results on the normal walking subset are reported. Namely, we evaluate our method with NM #5-6 as probes and with NM #1-4 in the gallery. Unlike many previous works, here the gait sequences are not split according to gait cycles. Instead, they are treated as a single image set. So, in total there are about 5k (74 × 11 × 6) image sets for training. Also note that the task is not multi-view but cross-view gait recognition. We do have access to all view angles during training the networks. However, in each test, there will only be the GEIs in one view angle involved in the gallery. We have to iterate the possible probe and gallery view angles so as to cover all the cross-view combinations (identical view cases are excluded). The reported results are the average recognition rates obtained by fixing the probe view angle and varying the gallery view angle.

1) Results and Comparison With Previous Methods: The comparison of our method with those in the literature is presented in Table III. The two methods are listed here because they are the most recent and best performers, and the results were obtained with the same division of training and testing data as ours. Our method performs better than previous methods in four out of the six reported cases. Also note that, our method directly uses the Euclidean distance of learned gait features to measure similarities, while ViDP [40] involves discriminative training of similarity models. Nevertheless, our overall performance is better than theirs.

Someone might expect that the 90° view, with the richest gait information, should be easier than other view angles such as 36°. One of the causes of this result is the cross-view setting in our experiment. To explain that, recall the GEIs given in Fig. 4. Besides the 0° and 180° views, the profile view (90°) is the most different one from the rest views. The number of samples fed into our networks are roughly the same for different view

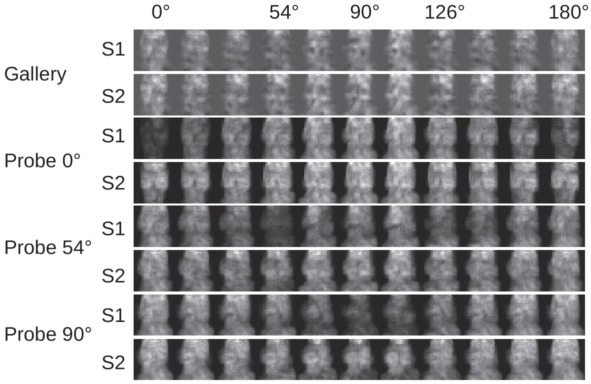


Fig. 6. Qualitative results obtained on CASIA-B. See the text for details.

angles. As a result, the trained features will lean on the oblique views, e.g., 36°, 54°, 126° and 144°.

The average activities on Layer C3 is shown in Fig. 6. Two subjects, i.e., S1 and S2, are involved. The first two rows show gait features in the gallery set corresponding to different view angles respectively. The rest rows show the element-wise absolute differences between probe and gallery gait features. For example, the third and forth row corresponding to an S1's probe in view angle 0°. The third row shows the differences between this probe and the eleven gallery gait features in the first row. Similarly, the forth row shows the differences between it and the features in the second row, and so on for the next four rows. Apparently, it is easier to identify gait features in adjacent (0° and 18°) and partly symmetric (0° and 180°, see Fig. 4) views. Note that this qualitative result also support our previous claimed reason for the relative low performance when the probe view angle is 90°. Unlike the other views, there is no symmetric view for 90°. An equal sampling rate will let our network lean on the other views.

V. Conclusion

In this paper, we have proposed a feature learning method for image sets, which can effectively handle less large data by suppressing over-fitting. It focuses on the co-occurrences and frequencies of discriminative features, and ignores motions between adjacent frames. Nevertheless, the proposed method has achieved convincing performances on two kinds of temporal sequential data, i.e., albums and gait videos. Especially, in the album classification task, it has outperformed previous methods with a significant margin.

The learned features are supposed to perform well in those tasks when the motion information is not that vital, e.g., face recognition in videos. Besides, even in the motion-related scenarios, they can also hopefully enhance classic models, due to their richer representations to appearances.

References

[1] G. Jun Qi *et al.*, "Correlative multi-label video annotation," in *Proc. ACM Multimedia*, 2007, pp. 17–26.

[2] G. Jun Qi, X. Sheng Hua, and H. Jiang Zhang, "Learning semantic distance from community-tagged media collection," in *Proc. ACM Multimedia*, 2009, pp. 243–252.

[3] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human action classes from videos in the wild," Center for Res. in Comput. Vis., Univ. of Central Florida, Orlando, FL, USA, Tech. Rep. CRCV-TR-12-01, 2012.

[4] L. Bossard, M. Guillaumin, and L. van Gool, "Event recognition in photo collections with a stopwatch hmm," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1193–1200.

[5] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2003, vol. 2, pp. II-409–II-415.

[6] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. IEEE Int. Conf. Pattern Recog.*, Aug. 2006, pp. 441–444.

[7] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2003, vol. 1, pp. I-313–I-320.

[8] G. Jun Qi, C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, May 2012.

[9] S. Chen, C. Sanderson, M. Harandi, and B. Lovell, "Improved image set classification via joint sparse approximated nearest subspaces," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 452–459.

[10] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 329–336.

[11] S. Chang, G. Jun Qi, J. Tang, Q. Tian, Y. Rui, and T. S. Huang, "Multimedia LEGO: Learning structured model by probabilistic logic ontology tree," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2013, pp. 979–984.

[12] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1250–1257.

[13] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell, "Hidden-state conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1853, Oct. 2007.

[14] Y. LeCun, B. Boser, J. Denker, and D. Henderson, "Handwritten digit recognition with a back-propagation network," in *Proc. NIPS*, 1990, pp. 396–404.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1106–1114.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Robinovich, "Going deeper with convolutions," *CoRR*, 2014 [Online]. Available: http://arxiv.org/abs/1409.4842

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013 [Online]. Available: http://arxiv.org/abs/1311.2524

[18] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *CoRR*, 2014 [Online]. Available: http://arxiv.org/abs/1406.2199

[20] Y. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," in *Proc. ECCV'14 Int. Workshop Competition Action Recog. Large Number Classes*, 2014 [Online]. Available: http://crcv.ucf.edu/THUMOS14/

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[22] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012 [Online]. Available: http://arxiv.org/abs/1207.0580

[23] L. Cao, J. Luo, and T. Huang, "Annotating photo collections by label propagation according to multiple similarity cues," in *Proc. ACM Multimedia*, 2008, pp. 121–130.

[24] L. Cao, J. Luo, H. Kautz, and T. Huang, "Image annotation within the context of personal photo collections using hierarchical event and scene models," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 208–219, Feb. 2009.

[25] R. Mattivi, J. Uijlings, F. de Natale, and N. Sebe, "Exploitation of time constraints for (sub-)event recognition," in *Proc. ACM J-MRE*, 2011, pp. 7–12.

[26] H. Izadinia and M. Shah, "Recognizing complex events using large margin joint low-level event model," in *Proc. ECCV*, 2012, pp. 430–444.

[27] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[28] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *Proc. ECCV*, 2010, pp. 257–270.

[29] T. Lam, K. Cheung, and J. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recog.*, vol. 44, no. 4, pp. 973–987, Apr. 2010.

[30] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database: Comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.

[31] G. Zhao, G. Liu, H. Li, and M. Pietikainen, "3D gait recognition using multiple cameras," in *Proc. Int. Conf. Automat. Face Gesture Recog.*, Apr. 2006, pp. 529–534.

[32] R. Bodor, A. Drenner, D. Fehr, O. Masoud, and N. Papanikolopoulos, "View-independent human motion classification using image-based reconstruction," *Image Vision Comput.*, vol. 27, no. 8, pp. 1194–1206, Jul. 2009.

[33] G. Ariyanto and M. Nixon, "Model-based 3D gait biometrics," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–7.

[34] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. ECCV*, 2006, pp. 151–163.

[35] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *Proc. Workshop Tracking Humans Evaluation Motion Image Sequences*, 2009, pp. 1058–1064.

[36] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 974–981.

[37] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circits Syst. Video Technol.*, vol. 22, no. 6, pp. 966–980, Jun. 2012.

[38] K. Bashir, T. Xiang, and S. Gong, "Cross-view gait recognition using correlation strength," in *Proc. BMVC*, 2010, pp. 1–11.

[39] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 696–709, Feb. 2014.

[40] M. Hu, Y. Wang, Z. Zhang, J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2034–2045, Dec. 2013.

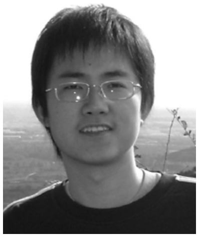
[41] H. Hu, "Enhanced Gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1274–1286, Jul. 2013.

[42] G. Jun Qi, X. Sheng Hua, Y. Rui, J. Tang, and H. Jiang Zhang, "Two-dimesional active learning for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[43] G. Jun Qi, X. Sheng Hua, Y. Rui, J. Tang, and H. Jiang Zhang, "Two-dimensional multi-label active learning with an efficient online adaptation model for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1880–1897, Oct. 2009.

Zifeng Wu received the B.Sc. and M.Sc. degrees in mechanical engineering and automation from Beihang University, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015.

He is currently a Researcher with the Australian Centre for Visual Technologies, University of Adelaide, Adelaide, SA, Australia. His research interests include computer vision and deep learning.



Yongzhen Huang (S'08–M'10) received the B.E. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2006, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2011.

In July 2011, he joined the National Laboratory of Pattern Recognition, CASIA, where he is currently an Associate Professor. He has authored or coauthored more than 30 papers in the areas of computer vision and pattern recognition at international journals, such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, and conferences, such as the Conference on Computer Vision and Pattern Recognition, the Conference on Neural Information Processing Systems, the International Conference on Image Processing, and the International Conference on Pattern Recognition. His current research interests include pattern recognition, computer vision, machine learning, and biologically inspired vision computing.

Liang Wang (M'09–SM'09) received the B. Eng. and M. Eng. degrees from Anhui University, Hefei, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2004.

From 2004 to 2010, he was a Research Assistant with Imperial College London, London, U.K., and Monash University, Melbourne, Australia; a Research Fellow with the University of Melbourne, Melbourne, Australia; and a Lecturer with the University of Bath, Bath, U.K. He is currently a full Professor of the Hundred Talents Program with the National Lab of Pattern Recognition, CASIA. He has authored or coauthored papers that appeared in highly-ranked international journals, such as the IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Image Processing, and leading international conferences such as the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, and the International Conference on Data Mining. His major research interests include machine learning, pattern recognition, and computer vision.

Prof. Wang is a Fellow of the IAPR. He is an Associate Editor of the IEEE Transactions on Cybernetics, the *International Journal of Image and Graphics*, *Signal Processing*, *Neurocomputing*, and the *International Journal of Cognitive Biometrics*.