Active Self-Paced Learning for Cost-Effective and Progressive Face Identification

Liang Lin [ID], Keze Wang, Deyu Meng [ID], Wangmeng Zuo, and Lei Zhang [ID]

Abstract—This paper aims to develop a novel cost-effective framework for face identification, which progressively maintains a batch of classifiers with the increasing face images of different individuals. By naturally combining two recently rising techniques: active learning (AL) and self-paced learning (SPL), our framework is capable of automatically annotating new instances and incorporating them into training under weak expert recertification. We first initialize the classifier using a few annotated samples for each individual, and extract image features using the convolutional neural nets. Then, a number of candidates are selected from the unannotated samples for classifier updating, in which we apply the current classifiers ranking the samples by the prediction confidence. In particular, our approach utilizes the high-confidence and low-confidence samples in the self-paced and the active user-query way, respectively. The neural nets are later fine-tuned based on the updated classifiers. Such heuristic implementation is formulated as solving a concise active SPL optimization problem, which also advances the SPL development by supplementing a rational dynamic curriculum constraint. The new model finely accords with the "instructor-student-collaborative" learning mode in human education. The advantages of this proposed framework are two-folds: i) The required number of annotated samples is significantly decreased while the comparable performance is guaranteed. A dramatic reduction of user effort is also achieved over other state-of-the-art active learning techniques. ii) The mixture of SPL and AL effectively improves not only the classifier accuracy compared to existing AL/SPL methods but also the robustness against noisy data. We evaluate our framework on two challenging datasets, which include hundreds of persons under diverse conditions, and demonstrate very promising results. Please find the code of this project at: http://hcp.sysu.edu.cn/projects/aspl/

Index Terms—Cost-effective model, active learning, self-paced learning, incremental processing, face identification

✦

1 INTRODUCTION

WITH the growth of mobile phones, cameras and social networks, a large amount of photographs is rapidly created, especially those containing person faces. To interact with these photos, there have been increasing demands of developing intelligent systems (e.g., content-based personal photo search and sharing from either his/her mobile albums or social network) with face recognition techniques [1], [2], [3]. Thanks to several recently proposed pose/expression normalization and alignment-free approaches [4], [5], [6], identifying face in the wild has achieved remarkable progress. As for the commercial product, the website "Face.com" once provided an API (application interface) to automatically detect and recognize faces in photos. The main problem in such scenarios is to identify individuals from images under a relatively unconstrained environment. Traditional methods usually handle this problem by supervised learning [7], while it is typically expensive and time-consuming to prepare a good set of labeled samples. Since only a few data are labeled, Semi-supervised learning [8] may be a good candidate to solve this problem. But it has been pointed out by [9]: Due to large amounts of noisy samples and outliers, directly using the unlabeled data may significantly reduce learning performance.

This paper targets on the challenge of incrementally learning a batch of face recognizers with the increasing face images of different individuals.[1] Here we assume that the person faces can be basically detected and localized by existing face detectors. However, to build such a system is quite challenging in the following aspects.

- Person faces have large appearance variations (see examples in Fig. 1a) caused by diverse views and expressions as well as facial accessories (e.g., glasses and hats) and aging. The different lighting condition is also required to be considered in practice.
- It is possible that only a few labeled samples are accessible at first, and the changes of personal faces are rather unpredictable over time, especially under the current scenarios that there are large amount of images swarmed into Internet every day.

- L. Lin and K. Wang are with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, China and the Engineering Research Center for Advanced Computing Engineering Software of Ministry of Education, China. E-mail: linliang@ieee.org, kezewang@gmail.com.
- D. Meng is with the School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xian Shi, Shaanxi Sheng 710048, P.R. China. E-mail: dymeng@mail.xjtu.edu.cn.
- W. Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, P.R. China. E-mail: cswmzuo@gmail.com.
- L. Zhang is with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong. E-mail: cslzhang@comp.polyu.edu.hk.

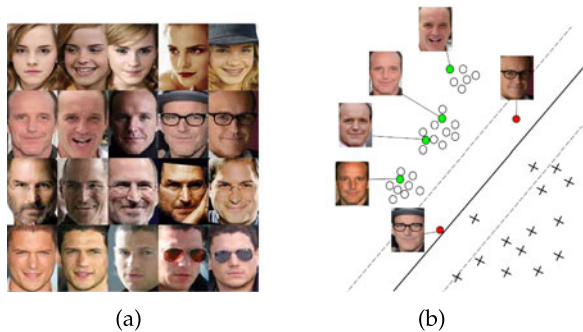1. http://hcp.sysu.edu.cn/projects/aspl/

Fig. 1. Illustration of high- and low-confidence samples in the feature space. (a) Shows a few face instances of different individuals, and these instances have large appearance variations. (b) Illustrates how the samples distribute in the feature space, where samples of high classification confidence distribute compactly to form several clusters and low confidence samples are scattered and close to the classifier decision boundary.

- Even though a few user interventions (e.g., labeling new samples) could be allowed, the user effort is desired to be kept minimizing over time.

Conventional incremental face recognition methods such as incremental subspace approaches [10], [11] often fail on complex and large-scale environments. Their performances could be dropped drastically when the initial training set of face images is either insufficient or inappropriate. In addition, most of existing incremental approaches suffer from noisy samples or outliers in the model updating. In this work, we propose a novel active self-paced learning framework (ASPL) to handle the above difficulties, which absorbs powers of two recently rising techniques: active learning (AL) [12], [13] and self-paced learning (SPL) [14], [15], [16]. In particular, our framework tends to conduct a "Cost-less-Earn-more" working manner: as much as possible pursuing a high performance while reducing costs.

The basic approach of the AL methods is to progressively select and annotate most informative unlabeled samples to boost the model, in which user interaction is allowed. The sample selection criteria is the key in AL, and it is typically defined according to the classification uncertainty of samples. Specifically, the samples of low classification confidence, together with other informative criteria like diversity, are generally treated as good candidates for model retraining. On the other hand, SPL is a recently proposed learning regime to mimic the learning process of humans/animals that gradually incorporates easy to more complex samples into training [17], [18], where an easy sample is actual the one of high classification confidence by the currently trained model. Interestingly, the two categories of learning methods select samples with the opposite criteria. This finding inspires us to investigate the connection between the two learning regimes and the possibility of making them complementary to each other. Moreover, as pointed out in [3], [19], learning based features are considered to be able to exploit information with better discriminative ability for face recognition, compared to the handcrafted features. We thus utilize the deep convolutional neural network (CNN) [20], [21] for feature extraction instead of using handcraft image features. In sum, we aim at designing a cost-effective and progressive learning framework, which is capable of automatically annotating

new instances and incorporating them into training under weak expert recertification. In the following, we discuss the advantage of our ASPL framework in two aspects: "Cost-less" and "Earn-more".

(I) *Cost less*: Our framework is capable of building effective classifiers with less labeled training instances and less user efforts, compared with other state-of-the-art algorithms. This property is achieved by combining the active learning and self-paced learning in the incremental learning process. In certain feature space of model training as Fig. 1b illustrates, samples of low classification confidence are scattered and close to the classifier decision boundary while high confidence samples distribute compactly in the intra-class regions. Our approach takes both categories of samples into consideration for classifier updating. The benefit of this strategy includes: i) High-confidence samples can be automatically labeled and consistently added into model training throughout the learning process in a self-paced fashion, particularly when the classifier becomes more and more reliable at later learning iterations. This significantly reduce the burden of user annotations and make the method scalable in large-scale scenarios. ii) The low-confidence samples are selected by allowing active user annotations, making our approach more efficiently pick up informative samples, more adapt to practical variations and converge faster, especially in the early learning stage of training.

(II) *Earn more*: The mixture of self-paced learning and active learning effectively improves not only the classifier accuracy but also the classifier robustness against noisy samples. From the perspective of AL, extra high-confidence samples are automatically incorporated into the retraining without cost of human labor in each iteration, and faster convergence can be thus gained. These introduced high-confidence samples also contribute to suppress noisy samples in learning, due to their compactness and consistency in the feature space. From the SPL perspective, allowing active user intervention generates the reliable and diverse samples that can avoid the learning been misled by outliers. In addition, utilizing the CNN facilitates to pursue a higher classification performance by learning the convolutional filters instead of hand-craft feature engineering.

In brief, our ASPL framework includes two main phases. At the initial stage, we first learn a general face representation using an architecture of convolutional neural nets, and train a batch of classifiers with a very small set of annotated samples of different individuals. In the iteration learning stage, we rank the unlabeled samples according to how they relate to the current classifiers, and retrain the classifiers by selecting and annotating samples in either active user-query or self-paced manners. We can also make the CNN fine-tuned based on the updated classifiers.

The key point in designing such an effective interactive learning system is to make an efficient labor division between computers and human participants, i.e., we should possibly feed computable and faithful tasks into computers, and to possibly arrange labor-saving and intelligent tasks to humans [22]. The proposed ASPL framework provides a rational realization to this task by

automatically distinguishing high-confidence samples, which can be easily and faithfully recognized by computers in a self-paced way, and low-confidence ones, which can be discovered by requesting user annotation.

The main *contributions* of this work are several folds. i) To the best of our knowledge, our work is the first one to make a face recognition framework capable of automatically annotating high-confidence samples and involve them into training without need of extra human labor in a purely self-paced manner under weak recertification of active learning. Especially in that along the learning process, we can achieve more and more pseudo-labeled samples to facilitate learning totally for free. Our framework is thus suitable in practical large-scale scenarios. The proposed framework can be easily extended to other similar visual recognition tasks. ii) We provide a concise optimization problem and theoretically interpret that the proposed ASPL is an rational implementation for solving this problem. iii) This work also advances the SPL development, by setting a dynamic curriculum variation. The new SPL setting better complies with the "instructor-student-collaborative" learning mode in human education than previous models. iv) Extensive experiments on challenging CACD and CASIA-WebFace datasets show that our approach is capable of achieving competitive or even better performance under only small fraction of sample annotations than that under overall labeled data. A dramatic reduction ($>$ 30 percent) of user interaction is achieved over other state-of-the-art active learning methods.

The rest of the paper is organized as follows. Section 2 presents a brief review of related work. Section 3 overview the pipeline of our framework, followed by a discussion of model formulation and optimization in Section 4. The experimental results, comparisons and component analysis are presented in Section 5. Section 6 concludes the paper.

2 RELATED WORK

In this section, we first present a review for the incremental face recognition, and then briefly introduce related developments on active learning and self-paced learning.

Incremental Face Recognition. There are two categories of methods addressing the problem of identifying faces with incremental data, namely incremental subspace and incremental classifier methods. The first category mainly includes the incremental versions of traditional subspace learning approaches such as principal component analysis (PCA) [23] and linear discriminant analysis (LDA) [11]. These approaches map facial features into a subspace, and keep the eigen representations (i.e., eigen-faces) updated by incrementally incorporating new samples. And face recognition is commonly accomplished by the nearest neighbor-based feature matching, which is computational expensive when a large number of samples are accumulated over time. On the other hand, the incremental classifier methods target on updating the prediction boundary with the learned model parameters and new samples. Exemplars include the incremental support vector machines (ISVM) [24] and the online sequential forward neural network [25]. In addition, several attempts have been made to absorb advantages from both of the two categories of methods. For example, Ozawa et al., [26] proposed to integrate the

Incremental PCA with the resource allocation network in an iterative way. Although these mentioned approaches make remarkable progresses, they suffer from low accuracy compared with those of batch-based state-of-the-art face recognizers, and none of these approaches have been successfully validated on large-scale datasets (e.g., more than 500 individuals). And these approaches are basically studied in the context of fully supervised learning, i.e., both initial and incremental data are required to be labeled.

Active Learning. This branch of works mainly focus on actively selecting and annotating the most informative unlabeled samples, in order to avoid unnecessary and redundant annotation. The key part of active learning is thus the selection strategy, i.e., which samples should be presented to the user for annotation. One of the most common strategies is the certainty-based selection [27], [28], in which the certainties are measured according to the predictions on new unlabeled samples obtained from the initial classifiers. For example, Lewis et al., [27] proposed to take the most uncertain instance as the one that has the largest entropy on the conditional distribution over its predicted labels. Several SVM-based methods [28] determine the uncertain samples as they are relatively close to the decision boundary. The sample certainty was also measured by applying a committee of classifiers in [29]. These certainty-based approaches usually ignore the large set of unlabeled instances, and are thus sensitive to outliers. A number of later methods present the information density measure by exploiting the information of unlabeled data when selecting samples. For example, the informative samples are sequentially selected to minimize the generalization error of the trained classifier on the unlabeled data, based on a statistical approach [30] or prior information [31]. In [32], [33], instances are taken to maximize the increase of mutual information between the candidate instances and the remaining ones based on Gaussian Process models. The diversity of the selected instance over the unlabeled data has been also taken into consideration [34]. Recently, Elhamifar et al., [12] presented a general framework via convex programming, which considered both the uncertainty and diversity measure for sample selection. However, these mentioned active learning approaches usually emphasize those low-confidence samples (e.g., uncertain or diverse samples) while ignoring the other majority of high-confidence samples. To enhance the discriminative capability, wang et al. [8] proposed a unified semi-supervised learning framework, which incorporates the high confidence coding vectors of unlabeled data into training under the proposed effective iterative algorithm, and demonstrate its effectiveness in dictionary-based classification. Our work inspires by this work, and also employs the high-confidence samples to improve both accuracy and robustness of classifiers.

Self-Paced Learning. Inspired by the cognitive principle of humans/animals, Bengio et al. [17] initialized the concept of curriculum learning (CL), in which a model is learned by gradually including samples into training from easy to complex. To make it more implementable, Kumar et al. [18] substantially prompted this learning philosophy by formulating the CL principle as a concise optimization model named self-paced learning. The SPL model includes a weighted loss term on all samples and a general SPL regularizer imposed on sample weights. By sequentially optimizing

the model with gradually increasing pace parameter on the SPL regularizer, more samples can be automatically discovered in a pure self-paced way. Jiang et al. [14], [16], [35] provided more comprehensive understanding for the learning insight underlying SPL/CL, and formulated the learning model as a general optimization problem as:

$$\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \sum_{i=1}^{n} v_i L(\mathbf{w}; \mathbf{x}_i, y_i) + f(\mathbf{v}; \lambda) \qquad (1)$$
$$\text{s.t. } \mathbf{v} \in \Psi,$$

where $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ corresponds to the training dataset, $L(\mathbf{w}; \mathbf{x}_i, y_i)$ denotes the loss function which calculates the cost between the objective label y_i and the estimated one, $\mathbf{w}$ represents the model parameter inside the decision function, $\mathbf{v} = [v_1, v_2, \ldots, v_n]^T$ denote the weight variables reflecting the samples' importance. λ is a parameter for controlling the learning pace, which is also referred as "pace age".

In the model, $f(\mathbf{v}; \lambda)$ corresponds to a self-paced regularizer. Jiang et al. abstracted three necessary conditions it should be satisfy [14], [16]: (1) $f(v; \lambda)$ is convex with respect to $v \in [0, 1]$; (2) The optimal weight of each sample should be monotonically decreasing with respect to its corresponding loss; and (3) The optimal weight of each sample should be monotonically decreasing with respect to the pace parameter λ.

In this axiomic definition, Condition 2 indicates that the model inclines to select easy samples (with smaller errors) in favor of complex samples (with larger errors). Condition 3 states that when the model "age" λ gets larger, it embarks on incorporating more, probably complex, samples to train a "mature" model. The convexity in Condition 1 further ensures that the model can find good solutions.

Ψ is the so called curriculum region that encodes the information of predetermined curriculums. Its axiomic definition contains two conditions [14]: (1) It should be nonempty and convex; and (2) If x_i is ranking before x_j in curriculum (more important for the problem), the expectation $\int_{\Psi} v_i \, d\mathbf{v}$ should be larger than $\int_{\Psi} v_j \, d\mathbf{v}$. Condition 1 ensures the soundness for the calculation of this specific constraint, and Condition 2 indicates that samples to be learned earlier is supposed to have larger expected values. This constraint weakly implies a prior learning sequence of samples, where the expected value for the favored samples should be larger.

The SPL model (1) finely simulates the learning process of human education. Specifically, it builds an "instructor-student collaborative" paradigm, which on one hand utilizes prior knowledge provided by instructors as a guidance for curriculum designing (encoded by the curriculum constraint), and on the other hand leaves certain freedom to students to ameliorate the actual curriculum according to their learning pace (encoded by the self-paced regularizer). Such a model not only includes all previous SPL/CL methods as its special cases, but also provides a general guild line to extend a rational SPL implementation scheme against certain learning task. Based on this framework, multiple SPL variations have been recently proposed, like SPaR [16], SPLD [15], SPMF [35] and SPCL [14].

The SPL related strategies have also been recently attempted in a series of applications, such as specific-class segmentation learning [36], visual

Complementarity Between AL and SPL. It is interesting that the function of SPL is very complementary to that of AL. The SPL methods emphasize easy samples in learning, which correspond to the high-confidence intra-class samples, while AL inclines to pick up the most uncertain and informative samples for the learning task, which are always located in low-confidence area near classification boundaries. SPL is capable of easily attaining large amount of faithful pseudo-labeled samples with less requirement of human labors (by reranking technique [16]. We will introduce details in Section 4), while tends to underestimate the roles of those most informative ones intrinsically configuring the classification boundaries; on the contrary, AL inclines to get informative samples, while need more human labors to manually annotate these samples with more carefully annotation. We thus expect to effectively mix these two learning schemes to help incremental learning both improve the efficiency with less human labors (i.e., Cost Less) and achieve better accuracy and robustness of the learned classifier against noisy samples (i.e., Earn More). This constructs the basic motivation of our ASPL framework for face identification under large-scale scenarios.

3 FRAMEWORK OVERVIEW

In this section, we illustrate how our ASPL model works. As illustrated in Fig. 2, the main stages in our framework pipeline include: CNN pretraining for face representation, classifier updating, high-confidence sample pseudo-labeling in a self-paced fashion, low-confidence sample annotating by active users, and CNN fine-tuning.

CNN Pretraining: Before running the ASPL framework, we need to pretrain a CNN for feature extraction based on a pre-given face dataset. These images are extra selected without overlapping to all our experimental data. Since several public available CNN architectures [37], [38] have achieved remarkable success on visual recognition, our framework supports to directly employ these architectures and their pretrained model as initialized parameters. In our all experiments, AlexNet [37] is utilized. Given the extra selected of annotated samples, we further fine-tune the CNN for learning discriminative feature representation.

Initialization: At the beginning, we randomly select few images for each individual, extract feature representation for them by pretrained CNN, and manually annotate labels to them as the starting point.

Classifier Updating: In our ASPL framework, we use one-vs-all linear SVM as our classifier updating strategies. In the beginning, only a small part of samples are labeled, and we train an initial a classifier for every individual using these samples. As the framework gets mature, samples manually annotated by the AL and pseudo-labeled by the SPL are growing, we adopt them to retrain the classifiers.

High-Confidence Sample Pseudo-Labeling: We rank the unlabeled samples by their important weights via the current classifiers, e.g., using the classification prediction hinge loss, and then assign pseudo-labels to the top-ranked samples of high confidences. This step can be automatically implemented by our system.

Low-Confidence Sample Annotating: Based on certain AL criterion obtained under the current classifiers, rank all
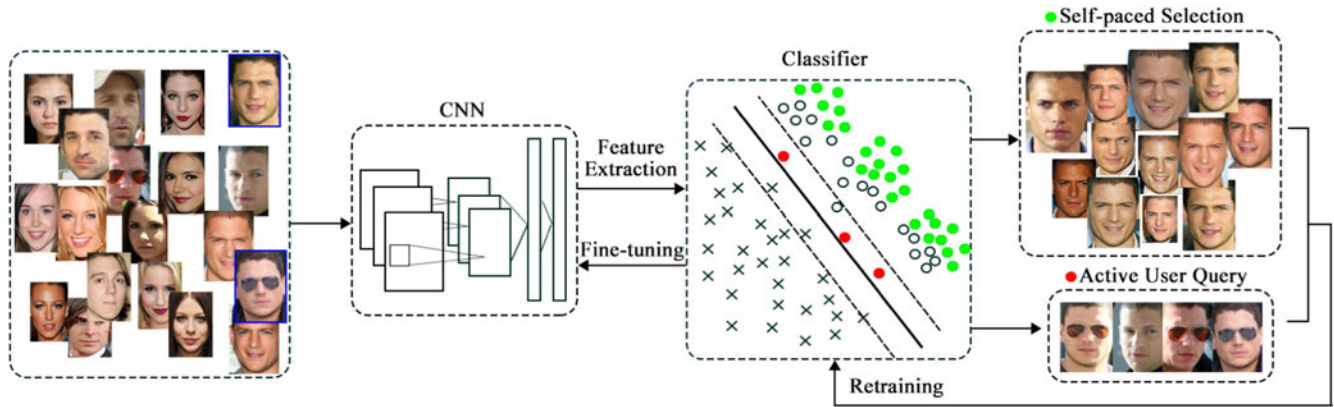
Fig. 2. Illustration of our proposed cost-effective framework. The pipeline includes stages of CNN and model initialization; classifier updating; high-confidence sample labeling by the SPL, low-confidence sample annotating by AL and CNN fine-tuning, where the arrows represent the workflow. The images highlighted by blue in the left panel represent the initially selected samples.

unlabeled samples, select those top-ranked ones (most informative and generally with low-confidence) from the unlabeled samples, and then manually annotate these samples by active users.

CNN Fine-Tuning: After several steps of the interaction, we make the neural nets fine-tuned by the backward propagation algorithm. All self-labeled samples by the SPL and manually annotated ones by the AL are added into the network, we utilize the softmax loss to optimize the CNN parameters via stochastic gradient decent approach.

4 FORMULATION AND OPTIMIZATION

In this section we will discuss the formulation of our proposed framework, and also provide a theoretical interpretation of its entire pipeline from the perspective of optimization. In specific, we can theoretically justify that the entire pipeline of this framework finely accords with a solving process for an active self-paced learning optimization model. Such a theoretical understanding will help deliver more insightful understanding on the intrinsic mechanism underlying the ASPL system.

4.1 Active Self-Paced Learning

In the context of face identification, suppose that we have n facial photos which are taken from m subjects. Denote the training samples as $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n \subset R^d$, where $\mathbf{x}_i$ is the d-dimensional feature representation for the ith sample. We have m classifiers for recognizing each sample by the one-versus-all strategy.

Learned knowledge from data will be utilized to ameliorate our model after a period of pace increasing. Correspondingly, we denote the label set of $\mathbf{x}_i$ as $\mathbf{y}_i = \{y_i^{(j)} \in \{-1, 1\}\}_{j=1}^m$, where $y_i^{(j)}$ corresponds to the label of $\mathbf{x}_i$ for the jth subject. That is, if $y_i^{(j)} = 1$, this means that $\mathbf{x}_i$ is categorized as a face from the jth subject.

On our problem setting, we should give two necessary remarks. One is that in our investigated face identification problems, almost all data have not been labeled before our system running. Only very small amount of samples are annotated as the initialization. That is, most of $\{\mathbf{y}_i\}_{i=1}^n$ are unknown and needed to be completed in the learning process. In our system, a minority of them is manually annotated by the active users and a majority is pseudo-labeled in

a self-paced manner. The other remark is that the data $\{\mathbf{x}_i\}_{i=1}^n$ might possibly been inputted into the system in an incremental way. This means that the data scale might be consistently growing.

Via the proposed mechanism of combining SPL and AL, our proposed ASPL model can adaptively handle both manually annotated and pseudo-labeled samples, and still progressively fit the consistently growing unlabeled data in such an incremental manner. The ASPL is formulated as follows:

$$\min_{\{\mathbf{w},\mathbf{b},\mathbf{v},\mathbf{y}_i \in \{-1,1\}^m, i \notin \Omega^\lambda\}} \sum_{j=1}^m \frac{1}{2}\|\mathbf{w}^{(j)}\|_2^2$$
$$+ C \cdot L\big(\mathbf{w}^{(j)}, b^{(j)}, \mathcal{D}, \mathbf{y}^{(j)}, \mathbf{v}^{(j)}\big) + f\big(\mathbf{v}^{(j)}; \lambda_j\big)$$
$$s.t. \quad \mathbf{v} \in \mathbf{\Psi}^\lambda, \quad (2)$$

where $\mathbf{w} = \{\mathbf{w}^{(j)}\}_{j=1}^m \subset R^d$ and $\mathbf{b} = \{b^{(j)}\}_{j=1}^m \subset R$ represent the weight and bias parameters of the decision functions for all m classifiers. $C(C > 0)$ is the standard regularization parameter trading off the loss function and the margin, and we set $C = 1$ in our experiments. $\mathbf{v} = \{[v_1^{(j)}, v_2^{(j)}, \ldots, v_n^{(j)}]^T\}_{j=1}^m$ denotes the weight variables reflecting the training samples' importance, and λ_j is a parameter (i.e., the pace age) for controlling the learning pace of the jth classifier. $f\big(\mathbf{v}^{(j)}; \lambda_j\big)$ is the self-paced regularizer controlling the learning scheme. We denote the index collection of all currently active annotated samples as $\Omega^\lambda = \cup_{j=1}^m\{\Omega^{\lambda_j}\}$, where Ω^{λ_j} corresponds to the set of the jth subject with the pace age λ_j. Here Ω^λ is introduced as a constraint on $\mathbf{y}_i$. $\mathbf{\Psi}^\lambda = \cap_{i=1}^n\{\Psi_i^\lambda\}$ composes of the curriculum constraint of the model at the m classifiers' pace age $\lambda = \{\lambda_j\}_{j=1}^m$. In particular, we specify two alternative types of the curriculum constraint for each sample $\mathbf{x}_i$, as:

- $\Psi_i^\lambda = [0, 1]$ is for the pseudo-labeled sample, i.e., $i \notin \Omega^\lambda$. Then, its importance weights with respect to all the classifiers $\{v_i^{(j)}\}_{j=1}^m$ need to be learned in the SPL optimization.
- $\Psi_i^\lambda = \{1\}$ is for the sample annotated by the AL process, i.e., $\exists j \ s.t. \ i \in \Omega^{\lambda_j}$. Thus, its importance weights are deterministically set during the model training, i.e., $v_i^{(j)} = 1$.

Each type of the curriculums will be detailly interpreted in Section 2. Note that different from the previous

SPL settings, this curriculum Ψ_i^λ can be dynamically changed with respect to all the pace ages λ of m classifiers. This conducts the superiority of our model, as we discuss in the end of this section.

We then define the loss function $L\big(\mathbf{w}^{(j)}, b^{(j)}, \mathcal{D}, \mathbf{y}^{(j)}, \mathbf{v}^{(j)}\big)$ on $\mathbf{x}$ as:

$$
\begin{aligned}
& L\big(\mathbf{w}^{(j)}, b^{(j)}, \mathcal{D}, \mathbf{y}^{(j)}, \mathbf{v}^{(j)}\big) \\
&= \sum_{i=1}^{n} v_i^{(j)} l\big(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\big) \\
&= \sum_{i=1}^{n} v_i^{(j)} \big(1 - y_i^{(j)} (\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)})\big)_{+} \\
& \text{s.t.} \sum_{j=1}^{m} |y_i^{(j)} + 1| \le 2, y_i^{(j)} \in \{-1, 1\}, i \notin \Omega^\lambda,
\end{aligned}
\tag{3}
$$

where $\big(1 - y_i^{(j)}(\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)})\big)_{+}$ is the hinge loss of $\mathbf{x}_i$ in the jth classifier. The cost term corresponds to the summarized loss of all classifiers, and the constraint term only allows two kinds of feasible solutions: i) for any i, there exists $y_i^{(j)} = 1$ while for all other $y_i^{(k)} = -1$ for all $k \ne j$; ii) $y_i^{(j)} = -1$ for all $j = 1, 2, \ldots, m$ (i.e., background or an unknown person class). These samples $\mathbf{x}_i$ will be added into the unknown sample set U. It is easy to see that such constraint complies with real cases where a sample should be categorized into one pre-specified subject or not classified into any of the current subjects.

Referring to the known alternative search strategy, we can then solve this optimization problem. Specifically, the algorithm is designed by alternatively updating the classifier parameters $\mathbf{w}, \mathbf{b}$ via one-vs-all SVM, the sample importance weights $\mathbf{v}$ via the SPL, the pseudo-label $\mathbf{y}$ via reranking. Along with gradually increasing pace parameter λ, the optimization updates: i) the curriculum constraint Ψ^λ via AL and ii) the feature representation via CNN fine-tuning. In the following we introduce the details of these optimization steps, and give their physical interpretations. The correspondence of this algorithm to the practical implementation of the ASPL system will also be discussed in the end.

Initialization. As introduced in the framework, we initialize our system running by using pre-trained CNN to extract feature representations of all samples $\{\mathbf{x}_i\}_{i=1}^{n}$. Set an initial m classifiers' pace parameter set $\lambda = \{\lambda_j\}_{j=1}^{m}$. Initialize the curriculum constraint Ψ^λ with currently user annotated samples Ω^λ and corresponding $\{\mathbf{y}^{(j)}\}_{j=1}^{m}$ and $\mathbf{v}$.

Classifier Updating. This step aims to update the classifier parameters $\{\mathbf{w}^{(j)}, b^{(j)}\}_{j=1}^{m}$ by one-vs-all SVM. Fixing $\{\{\mathbf{x}_i\}_{i=1}^{n}, \mathbf{v}, \{\mathbf{y}_i\}_{i=1}^{n}, \Psi^\lambda\}$, the original ASPL model Eqn. (2) can be simplified into the following form:

$$
\min_{\mathbf{w}, \mathbf{b}} \sum_{j=1}^{m} \frac{1}{2} \|\mathbf{w}^{(j)}\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} l\big(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\big),
$$

which can be equivalently reformulated as solving the following independent sub-optimization problems for each classifier $j = 1, 2, \ldots, m$:

$$
\min_{\mathbf{w}^{(j)}, b^{(j)}} \frac{1}{2} \|\mathbf{w}^{(j)}\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} l\big(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\big).
\tag{4}
$$

This is a standard one-vs-all SVM model with weights by taking one-class sample as positive while all others as negative. Specifically, when the weights $v_i^{(j)}$ are only of values $\{0, 1\}$, it corresponds to a simplified SVM model under sampled instances with $v_i^{(j)} = 1$; otherwise when v_i^{j} sets values from $[0, 1]$, it corresponds to the weighted SVM model. And both of them can be readily solved by many off-the-shelf efficient solvers. Thus, this step can be interpreted as implementing one-vs-all SVM over instances manually annotated from the AL and self-annotated from the SPL.

High-Confidence Sample Labeling. This step aims to assign pseudo-labels $\mathbf{y}$ and corresponding important weights $\mathbf{v}$ to the top-ranked samples of high confidences.

We start by employing the SPL to rank the unlabeled samples according to their importance weights $\mathbf{v}$. Under fixed $\{\mathbf{w}, \mathbf{b}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{y}_i\}_{i=1}^{n}, \Psi^\lambda\}$, our ASPL model in Eqn. (2) can be simplified to optimize $\mathbf{v}$ as:

$$
\begin{aligned}
\min_{\mathbf{v} \in [0,1]} \sum_{j=1}^{m} C \sum_{i=1}^{n} v_i^{(j)} l\big(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\big) + f\big(\mathbf{v}^{(j)}; \lambda_j\big), \\
s.t. \quad \mathbf{v} \in \Psi^\lambda.
\end{aligned}
\tag{5}
$$

This problem then degenerates to a standard SPL problem as in Eqn. (1). Since both the self-paced regularizer $f(\mathbf{v}^{(j)}; \lambda_j)$ and the curriculum constraint Ψ^λ is convex (with respect to $\mathbf{v}$), various existing convex optimization techniques, like the gradient-based or interior-point methods, can be used for solving it. Note that we have multiple choices for the self-paced regularizer, as those built in [15], [16]. All of them comply with three axiomic conditions required for a self-paced regularizer, as defined in Section 2.

Based on the second axiomatic condition for self-paced regularizer, any of the above $f(\mathbf{v}^{(j)}; \lambda_j)$ inclines to conduct larger weights on high-confidence (i.e., easy) samples with less loss values while vice versa, which evidently facilitates the model with the "learning from easy to hard" insight. In all our experiments, we utilize the linear soft weighting regularizer due to its relatively easy implementation and well adaptability to complex scenarios. This regularizer penalizes the sample weights linearly in terms of the loss. Specifically, we have

$$
f(\mathbf{v}^{(j)}, \lambda_j) = \lambda_j \left(\frac{1}{2} \|\mathbf{v}^{(j)}\|_2^2 - \sum_{i=1}^{n} v_i^{(j)} \right),
\tag{6}
$$

where $\lambda_j > 0$. Eqn. (6) is convex with respect to $\mathbf{v}^{(j)}$, and we can thus search for its global optimum by computing the partial gradient equals. Considering $v_i^{(j)} \in [0, 1]$, we deduce the analytical solution for the linear soft weighting, as,

$$
v_i^{(j)} = \begin{cases} -\dfrac{C\ell_{ij}}{\lambda_j} + 1, & C\ell_{ij} < \lambda_j \\ 0, & \text{otherwise,} \end{cases}
\tag{7}
$$

where $\ell_{ij} = l\big(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\big)$ is the loss of $\mathbf{x}_i$ in the jth classifier. Note that the deducing way to Eqn. (7) is similar with in [16], but our resulting solution is different since our ASPL model in Eqn. (2) is new.

After obtaining the weight $\mathbf{v}$ for all unlabeled samples ($i \notin \Omega^\lambda$) according to the optimized $\mathbf{v}^{(j)}$ in a descending order. Then we consider the samples with larger important

weight than others are high confidences. We form these samples into high-confidence sample set $\mathcal{S}$ and assign them pseudo-labels: Fixing $\{\mathbf{w}, \mathbf{b}, \{\mathbf{x}_i\}_{i=1}^n, \mathbf{\Psi}^\lambda, \mathbf{v}\}$, we optimize $\mathbf{y}_i$ of Eqn. (2) which corresponds to solve:

$$\min_{\mathbf{y}_i \in \{-1,1\}^m, i \in \mathcal{S}} \sum_{i=1}^n \sum_{j=1}^m v_i^{(j)} \ell_{ij} \tag{8}$$
$$\text{s.t.,} \sum_{j=1}^m |y_i^{(j)} + 1| \le 2.$$

where $\mathbf{v}_i$ is fixed and can be treated as constant. When $\mathbf{x}_i$ belongs to a certain person class, Eqn. (8) has an optimum, which can be exactly extracted by the Theorem 1. The proof is specified in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2017.2652459.

Denote those js that satisfy $\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)} \ne 0$ and $v_i^{(j)} \in (0, 1]$ as a set M and set all $y_i^{(j)} = -1$ for others in default.[2] The solution of Eqn. (8) for $y_i^{(j)}$, $j \in M$ can be obtained by the following theorem.

Theorem 1.

a) If $\forall j \in M$, $\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)} < 0$, Eqn. (8) has a solution:

$$y_i^{(j)} = -1, \quad j = 1, \dots, m;$$

b) When $\forall j \in M$ except $j = j^*$, $\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)} < 0$, i.e., $v_i^{(j^*)}\ell_{ij^*} > 0$, then Eqn. (8) has a solution:

$$y_i^{(j)} = \begin{cases} -1, & j \ne j^*; \\ 1, & j = j^*; \end{cases}$$

c) Otherwise, Eqn. (8) has a solution:

$$y_i^{(j)} = \begin{cases} -1, & j \ne j^*; \\ 1, & j = j^*, \end{cases}$$

where

$$j^* = \arg\min_{1 \le j \le m} v_i^{(j)}\left(\ell_{ij} - \left(1 + (\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)})\right)_+\right). \tag{9}$$

Actually, only those high-confidence samples with positive weights, as calculated in the last updating step for $\mathbf{v}$, are meaningful for the solution. This implies the physical interpretation for this optimization step: we iteratively find the high-confidence samples based on the current classifier, and then enforce pseudo-labels $\mathbf{y}_i$ on those top-ranked high-confidence ones ($i \in \mathcal{S}$). This is exactly the mechanism underlying a reranking technique [16].

2. $v_i^{(j)} = 0$ actually implies that the ith sample is with low-confidence to be annotated as the jth class, and thus it is natural to pseudo-label it as a negative sample for the jth class. $w^Tx + b = 0$ implies that a sample is located in the classification boundary of the jth class, and thus it is also a low-confidence j-class sample and thus we directly annotate it as negative. Actually, for these samples, pseudo-label them as positive or negative will not affect the value of the objective function of Eqn. (8). We tend to annotate these low-confidence samples as negative since due to the constraint of Eqn. (8) (at most one positive class one sample is allowed to be annotated), this will not influence selecting a more rational positive class for each sample.

The above optimization process can be understood as the self-learning manner of a student. The student tends to pick up most high-confident samples, which imply easier aspects and faithful knowledge underlying data, to learn, under the regularization of the pre-designed curriculum $\mathbf{\Psi}^\lambda$. Such regularization inclines to rectify his/her learning process so as to avoid him/her stuck into a unexpected overfitting point.

Low-Confidence Sample Annotating. After pseudo-labeling high-confidence samples in such a self-paced uncertainty modeling, we employ AL fashion to update the curriculum constraint $\mathbf{\Psi}^\lambda$ in the model by supplementing more informative curriculums based on human knowledge. The AL process aims to select most low-confidence unlabeled samples and to annotate them as either positive or negative by requesting user annotation. Our selection criteria are based on the classical uncertainty-based strategy [27], [28]. Specifically, given the current classifiers, we randomly collected a number of randomly unlabeled samples, which are usually located in low-confidence area near the classification boundaries.

1) *Annotated Sample Verifying:* Considering the user annotation may contain outliers (incorrectly annotated samples), we introduce a verification step to correct the wrongly annotated samples. Assuming that labeled samples with lower prediction scores from the current classifiers have higher probability of being incorrectly labeled, we propose to ask the active user to verify their annotations on these samples. Specifically, in this step we first employ the current classifiers to obtain the prediction scores of all the annotated samples. Then we re-rank them and select Top-L ones with lowest prediction scores and ask the user to verify these selected samples, i.e., double-checking them. We can set L as a small number ($L = 5$ in our experiments), since we do believe the chance of human making mistakes is low. In sum, we improve the robustness of the AL process by further validating Top-L most uncertain samples with the user. In this way, we can reduce the effects of accumulated human annotation errors and enable the classifier to be trained in a robust manner.

2) *Low-Confidence Definition:* When we utilize the current classifiers (m classifiers for discriminating m object categories) to predict the label of unlabeled samples, those predicted as more than two positive labels (i.e., predicted as the corresponding object category) actually represent these samples making the current classifiers ambiguous. We thus adopt them as so called "low-confident" samples and require active user to manually annotate them. Actually, in this step, other "low-confidence" criterion can be utilized. We employed this simple strategy just due to its intuitive rationality and efficiency.

After users perform manual annotation, we update the $\mathbf{\Psi}^\lambda$ by additionally incorporating those newly annotated sample set ϕ into the current curriculum $\mathbf{\Psi}^\lambda$. For each annotated sample, our AL process includes the following two operations: i) Set its curriculum constraint, i.e., $\{\Psi_i^\lambda\}_{i \in \phi} = \{1\}$; ii) Update its labels $\{\mathbf{y}_i\}_{i \in \phi}$ and add its index into the set of currently annotated samples Ω^λ. Such specified curriculum still complies with the axiomic conditions for the

curriculum constraint as defined in [14]. For those annotated samples, the corresponding $\Psi_i^\lambda = \{1\}$ with expectation value 1 over the whole set, while for others $\Psi_i^\lambda = [0, 1]$ with expectation value $1/2$. Thus the more informative samples still have a larger expectation than the others. Also, it is easy to see Ψ^λ is non-empty and convex. It thus complies traditional curriculum understanding.

New Class Handling. After the AL process, if active user annotates the selected unlabeled samples with u unseen person classes, new classifiers for these unseen classes are needed to be initialized without affecting the existed classifiers. Moreover, there is another difficulty that the samples of the new class are not enough for classifier training. Thanks to the proposed ASPL framework, we can employ the following four steps to address above mentioned issues.

1) For each of these new class samples, search all the unlabeled samples and pick out its K-nearest neighbors from the unseen class set U in the feature space;

2) Require active user to annotate these selected neighbors to enrich the positive samples for these new person classes;

3) Initialize and update $\{\mathbf{w}^{(j)}, b^{(j)}, \mathbf{v}^{(j)}, \mathbf{y}^{(j)}, \lambda_j\}_{j=m+1}^{m+u}$ for these new person classes according to above mentioned iteration process of {*initialization, classifier updating, high-confidence sample labeling, low-confidence sample annotating*}.

This step corresponds to the instructor's role in human education, which aims to guide a student to involve more informative curriculums in learning. Different from the previous fixed curriculum setting in SPL throughout the learning process, here the curriculum is dynamically updated based on the self-paced learned knowledge of the model. Such an improvement better simulates the general learning process of a good student. With the learned knowledge of a student increasing, his/her instructor should vary the curriculum settings imposed on him from more in the early stage to less in later. This learning manner evidently should conduct a better learning effect which can well adapt the personal information of the student.

Feature Representation Updating. After several of the SPL and AL updating iterations of $\{\mathbf{w}, \mathbf{b}, \{\mathbf{y}_i\}_{i=1}^n, \mathbf{v}, \Psi^\lambda\}$, we now aim to update the feature representation $\{\mathbf{x}_i\}_{i=1}^n$ through finetuning the pretrained CNN by inputting all manually labeled samples from the AL and self-annotated ones from the SPL. These samples tend to deliver data knowledge into the network and improve the representation of the training samples. A better feature representation is thus expected to be extracted from this ameliorated CNN.

This learning process simulates the updating of the knowledge structure of a human brain after a period of domain learning. Such updating tends to facilitate a person grasp more effective features to represent newly coming samples from certain domain and make him/her with a better learning performance. In our experiments, we generally conduct the CNN feature fine-tuning after around 50 rounds of the SPL and AL updating, and the learning rate is set as 0.001 for all layers.

Pace Parameter Updating. We utilize a heuristic strategy to update pace parameters $\{\lambda_j\}_{j=1}^m$ for m classifiers in our implementation.

After multiple iterations of the ASPL, we specifically set the pace parameter λ_j for each individual classifier, and utilize a heuristic strategy in our implementation for parameter updating. For the tth iteration, we compute the pace parameter for optimizing Eqn. (2) by:

$$\lambda_j^t = \begin{cases} \lambda^0, & t = 0 \\ \lambda_j^{(t-1)} + \alpha * \eta_j^t, & 1 \leq t \leq \tau \\ \lambda_j^{(t-1)}, & t > \tau, \end{cases} \quad (10)$$

where η_j^t is the average accuracy of the jth classifier in the current iteration, and α is a parameter which controls the pace increasing rate. In our experiments, we empirically set $\{\lambda^0, \alpha\} = \{0.2, 0.08\}$. Note that the pace parameters λ should be stopped when all training samples are with $\mathbf{v} = \{1\}$. Thus, we introduce an empirical threshold τ constraining that λ is only updated in early iterations, i.e., $t \leq \tau$. τ is set as 12 in our experiments.

The entire algorithm can then be summarized into Algorithm 1. It is easy to see that this solving strategy for the ASPL model finely accords with the pipeline of our framework.

Algorithm 1. The Sketch of ASPL Framework

Input: Input dataset $\{\mathbf{x}_i\}_{i=1}^n$
Output: Model parameters $\mathbf{w}$, $\mathbf{b}$

1: Use pre-trained CNN to extract feature representations of $\{\mathbf{x}_i\}_{i=1}^n$. Initialize multiple annotated samples into the curriculum Ψ^λ and corresponding $\{\mathbf{y}_i\}_{i=1}^n$ and $\mathbf{v}$. Set an initial pace parameter $\lambda = \{\lambda^0\}^m$.
   while not converged do
2:   Update $\mathbf{w}$, $\mathbf{b}$ by one-vs-all SVM
3:   Update $\mathbf{v}$ by the SPL via Eqn. (7)
4:   Pseudo-label high-confidence samples $\{\mathbf{y}_i\}_{i \in \mathcal{S}}$ by the reranking via Eqn. (8)
5:   Update the unclear class set U
6:   Verify the annotated samples by AL.
7:   Update low-confidence samples $\{\mathbf{y}_i, \Psi_i^\lambda\}_{i \in \phi}$ by the AL
     if u unseen classes have labeled,
       Handle u new classes via the steps in Section 4.1
       Go to the step 2
     end if
8:   In every T iterations:

- Update $\{\mathbf{x}_i\}_{i=1}^n$ through fine-tuning CNN
- Update λ according to Eqn. (10)

9: **end while**
10: **return** $\mathbf{w}$, $\mathbf{b}$;

Convergence Discussion. As illustrated in Algorithm 1, the ASPL algorithm alternatively updates variables including: the classifier parameters w, b (by weighted SVM), the pseudo-labels y (closed-form solution by Theorem 1), the importance weight v (by SPL), and low-confidence sample annotations ϕ (by AL). For the first three parameters, these updates are calculated by a global optimum obtained from a sub-problem of the original model, and thus the objective function can be guaranteed to be decreased. However, just as other existing AL techniques, human efforts are involved in the loop of the AL stage, and thus the objective function cannot be guaranteed to be monotonically decreased in this step. However, just as shows in Section 5, as the learning processing, the model tends to be more and

more mature, and the labor of AL tends to be less and less in the later learning stage. Thus with gradually less involvement of the AL calculation in our algorithm, the monotonic decrease of the objective function in iteration tends to be promised, and thus our algorithm tends to be convergent.

4.2 Relationship with Other SPL/AL Models

It is easy to see that the proposed ASPL model extends the previous AL/SPL models and includes all of them as special cases. When we fix the curriculum and feature representations and only update other parameters, it degenerates to the traditional SPL models by rationally setting the self-paced regularizer. When we fix the SPL parameters, feature representations and do not involve pseudo-labels in learning, the model degenerates the a general AL learning regime. The amelioration to both SPL and AL is expected to bring benefits to both regimes. On one hand, introducing more high-confidence samples in the self-paced fashion is helpful to reduce the burden of user annotations, particularly when the classifier becomes reliable at later learning iterations. On the other hand, the low confidence samples selected by active user annotations tends to make our approach workable with less initial labeled samples than existing self-paced learning algorithms. All these benefits are comprehensively substantiated by our experiments.

5 EXPERIMENTS

In this section, we first introduce the datasets and implementation setting, and then discuss the experimental results and comparisons with other existing approaches.

5.1 Datasets and Setting

We adopt two public datasets in our experiments, the Cross-Age Celebrity Dataset (CACD) [39] and CASIA-WebFace-Sub dataset [40].

CACD is a large-scale and challenging dataset for evaluating face recognition and retrieval, and it contains a batch of images of 2,000 celebrities collected from Internet, which are varying in age, pose, illumination, and occlusion. And only a subset of 200 celebrities are manually annotated by Chen et al. [39]. For better convincing evaluation, we augment this subset by extra labeling 300 individuals and obtain a set of 56,138 images in total.

CASIA-WebFace dataset [40] is a large scale face recognition dataset with 10,575 subjects/persons and 494,414 images. CASIA-WebFace is extremely challenging for its images are all collected from Internet with different view points and light illumination under different scenes. Though the total person/subject number of CASIA-WebFace dataset is very large, the sample number for each person, varying from 3 to 804, is heavily unbalanced. For those persons who has very few samples (say below 100), the experiment analysis is not able to be performed. Hence, we select a subset of the CASIA-WebFace dataset by discarding its persons with less than 100 samples to form the CASIA-WebFace-Sub dataset. The CASIA-WebFace-Sub dataset has 181,901 images with 925 persons inside. The detailed information of above mentioned datasets is summarized in Table 1.

TABLE 1
The Summarization of Datasets We Used

Dataset	# images	# persons	# images/person
CACD	56,138	500	79~306
CASIA-WebFace-Sub	181,901	925	100~804

Experiment Setting. We detect the facial points using the method proposed in [41] and align the faces based on the eye locations. The experiments on both of the datasets are conducted as the following steps. We first randomly select 80 percent images of each individual to form the unlabeled training set, and the rest samples are used for testing, according to the setting in the existing active learning method [12]. Then, we randomly annotate n samples of each person in the training set to initialize the classifier. To get rid of the influence of randomness, we average the results over 5 times of execution with different sample selections. All of the experiments are conducted on a common desktop PC with i7 3.4 GHz CPU and a NVIDIA Titan X GPU.

On the two above mentioned datasets, we evaluate the performance of incremental face identification in two aspects: the recognition accuracy and user annotation amount in the incremental learning process. The recognition accuracy is defined as the rank-one rate for face identification. We compare our ASPL framework with several existing active learning algorithms and baseline methods under the same setting: i) Convex Programming based Active Learning (CPAL) [12]: Annotate a few samples in each step based on prediction uncertainty and sample diversity; ii) Confidence-based Active Learning via SVMs (CCAL) [28]: Select only one sample having lowest prediction confidence; iii) AL_RAND: Randomly select unlabeled samples to be annotated during the training phase. This method discards all active learning techniques and can be considered as the lower bound, and iv) AL_ALL: All unlabeled samples are annotated for training the classifier. This method can be regarded as the upper bound (best performance the classifier can achieve). For fair comparison, all of these methods utilize the same feature representation as ours in the beginning. As the training iteration increase, active user annotation is employed to those selected most informative and representative samples. Then, CNN fine-tuning is also exploited to improve the feature extractor for ASPL, CPAL, CCAL, AL_RAND, AL_ALL.

Details of CNN Implementation. The architecture of Alex-Net [37] is utilized in our all experiments. Thanks to the well pre-training, the CNN updating is only implemented few times during ASPL iteration in all our experiments, each only containing no more than 5 CNN updating steps. We generally conducted CNN steps after around 5 rounds of the SPL and AL updating, and the learning rate is set as 0.001 for all layers. Equal importance is imposed between the previous training examples and the newly labeled examples, and CNN is updated using the stochastic gradient decent methods with the momentum 0.9 and weight decay 0.0005.

5.2 Experimental Comparisons

The results on the two datasets are reported in Figs. 3a and 3b, respectively, where we can observe how the recognition
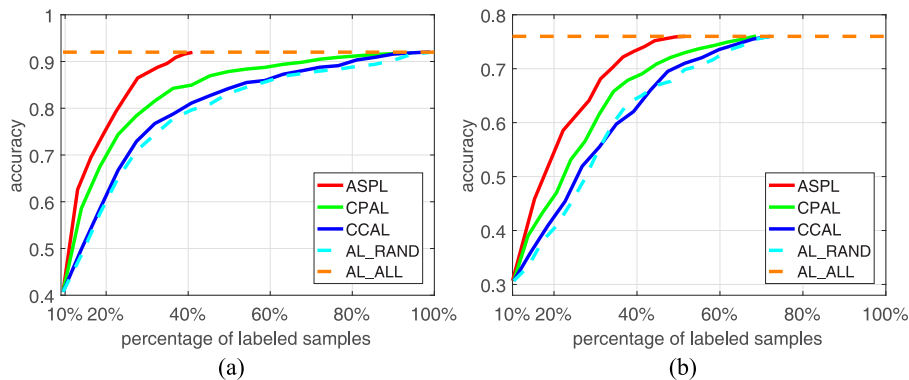
Fig. 3. Results on (a) CACD and (b) CASIA-WebFace-Sub datasets. The vertical axes represent the recognition accuracy and the horizontal axes represent the percentage of annotated samples of the whole set.

accuracy changes with increasingly incorporating more unlabeled samples. In CACD dataset, to achieve the same recognition accuracy, ASPL model requires few annotation of the unlabeled data. On the other hand, ASPL outperforms the competing methods in accuracy when the same amount annotations. ASPL can still have a superior performance as the iteration goes on. The similar results and phenomena can be discovered in CASIA-WebFace-Sub dataset. As one can see that, ASPL only requires about 40 and 45 percent annotations to achieve the-state-of-art performance on CACD and CASIA-WebFace-Sub dataset, respectively. While the compared methods AL_RAND, CCAL and CPAL all requires about 81 and 65 percent, respectively. Hence, our ASPL can performs as well as the AL_ALL with minimal annotations.

Note that the performances of RAND and CCAL are relatively close, and the similar results were reported in [12]. According to the explanation in [12], this comes from the fact that many samples have low prediction confidences and distribute not densely in the feature space. Thus, the randomizing sample selection achieves similar results compared to CCAL.

5.3 Component Analysis

To further analyze how different components contribute to performance, we implement several variants of our framework: i) ASPL (w/o FT): allowing both active and self-paced sample selection during learning while disabling the CNN fine-tuning, i.e., the feature extractor is kept the same as the iteration goes on for training; ii) ASPL (w/o SPL): discarding high-confidence sample pseudo-labeling via self-paced learning; iii) ASPL (w/o AL): ignoring low confidence samples for active user annotation; iv) AL_ALL: fine-tuning the CNN and train classifiers with all the labels of the training samples and v) AL_ALL (w/o FT): training classifiers with all the labels of the training samples without fine-tuning. Moreover, the full version of our proposed model is denoted as ASPL, which allows the convolutional nets to be fine-tuned during the training process. We further evaluate the ASPL variants in the following aspects.

Contribution of Different ASPL Components. Using AL_ALL and AL_ALL (w/o FT) as the baselines, we gradually add the AL, SPL and fine-tuning components to ASPL. These experiments are executed on the CASIA-Webface dataset. Fig. 4 illustrates the accuracy obtained using ASPL, ASPL (w/o FT), ASPL (w/o AL) and ASPL (w/o SPL). One

can observe that any of the three components is useful in improving the recognition accuracy. Especially, the additional SPL component can significantly improve the recognition accuracy and reduce the number of annotation samples by automatically exploiting the majority of high-confidence samples for feature learning.

We also observe that the CNN feature fine-tuning can dramatically improve the recognition accuracy in the early steps. This is mainly because the information gain (i.e., individual appearance diversity) deceases with progressively introducing new samples to the neural nets.

Analysis on Initial Samples. In SPL [18], classifier is first trained using the initial samples. With the current classifier, easy samples are preferred to be selected in the early training steps, and thus it is expected that the performance of SPL heavily relies on the initial samples. Fortunately, by incorporating with active learning, ASPL can evidently alleviate this problem. To verify this, we compare the performance of ASPL and SPL on 20 randomly selected individuals of CASIA-Webface-Sub dataset. The result is shown in Fig. 5. Given the same initialized feature representations, we also conduct the experiments to analyze the performance vs different initial portions to be handled by AL on this dataset. The results are illustrated in Fig. 6.

As one can see from Fig. 5, with different initial samples, ASPL reaches similar/stable results as the training continues, while SPL still varies a lot. This result indicates that the AL component is effective in handling the poor initialization. Fig. 6 illustrates that though poor performance is obtained at the beginning, the performance of our model increases
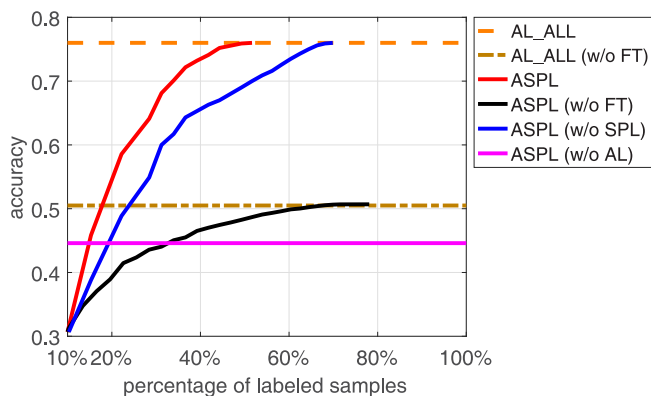

Fig. 4. Accuracies with the increase of annotated samples of different variants of our framework, using CASIA-WebFace-Sub dataset.
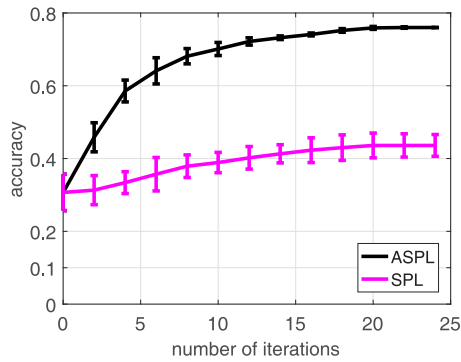
Fig. 5. The accuracy and standard deviation of ASPL and SPL on the CASIA-Webface-Sub dataset.
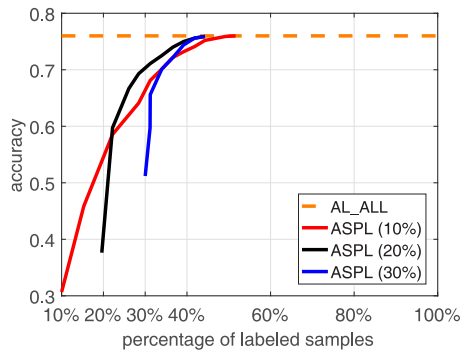

Fig. 6. The comparison of different number of initial samples and the further required annotation ported of the AL process on the CASIA-Web-face-Sub dataset. For fair comparison, these methods share the same feature representation as initialization.

TABLE 2
The Performance Comparison of Whether Handling Unseen New Classes or Not on the CASIA-WebFace-Sub Dataset

# Class Number	300	600	925
ASPL (ALL)	88.3%	81.0%	76.0%
ASPL	88.3%	81.6%	76.0%

ASPL (ALL) Denotes the ASPL Version of no Unseen Classes

during the training process. In summary, our model is insensitive to the diversity and quantity of initial samples.

Performance with New Classes. To justify the effectiveness of our ASPL for handling unseen new classes, we conduct the following experiment on the CASIA-WebFace-Sub dataset: We compare the performance of incrementally giving some classes (our ASPL) and directly giving all person classes. Specifically, given all person classes, we initialize all the classifiers at the beginning of the training and optimize them without handling unseen new classes. We denote this variant as ASPL (ALL). The experimental result is illustrated in Table 2 and shows that our proposed ASPL can handle unseen new classes effectively without substantially performance drop or even with slightly better performance, compared with the all classes given version ASPL (ALL).

Annotation Required for Large Scale Dataset. To demonstrate that our ASPL can be adopted under large scale scenario, we analyze the training phase of ASPL on the large scale CASIA-WebFace-Sub dataset. As illustrated in Fig. 7, the x-axis denotes the number of training iterations and the
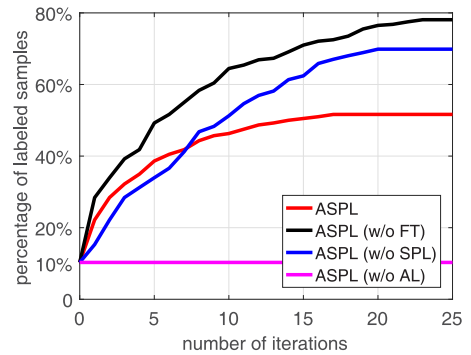

Fig. 7. The comparison of different number of initial samples and the further required annotation ported of the AL process on the CASIA-Web-Face-Sub dataset.

TABLE 3
The Error Rates of the Pseudo-Labels Assigned by SPL on High-Confidence Samples

# iteration	5	10	15	20	25
ASPL (w/o FT)	8.2%	6.9%	5.1%	5.0%	4.9%
ASPL	4.5%	4.1%	3.4%	3.3%	3.3%

y-axis denotes the amount of required user annotation. The curve in Fig. 7 demonstrates that our proposed ASPL model requires relatively larger annotations when the training iteration number is small. As the training continues, the amount required annotations began to be reduced due to the gradually mature model incrementally ameliorated in the learning process. This observation indicates that the burden of user annotations would be indeed relieved when the classifier becomes reliable at the later learning stage of the proposed ASPL method. Moreover, as illustrated in Table 3, with the increase of user annotations over time, ASPL can automatically assign more reliable pseudo-labels to the unlabeled samples selected in the self-paced way.

Robustness Analysis. We further analyze the robustness of ASPL when noisy images are deliberately included in two experiments. (i) Ex-1: a ($a = 0\%, 10\%, 30\%, 50\%$) noisy images are added to the initial samples for each individual. (ii) Ex-2: noise-free initials are used, but b ($b = 0\%, 10\%, 30\%, 50\%$) importers are deliberately annotated during the training process. These experiments are conducted on the CASIA-Webface-Sub dataset. To validate the effectiveness of the proposed annotated sample verifying step, we disable the verifying step and denote these modification as "Noise w/o VF".

Fig. 8a shows the result of Ex-1, where ASPL is initialized with different number of noisy images. In early steps of the iteration, noisy data have huge adverse effect on test accuracy. Along with the increase of iteration number, the genuine data gradually dominate the results. Fig. 8b illustrates the result of Ex-2, where noisy images are added to the labeled training set the 2th step of iteration. We can see that a sharp decline in the recognition accuracy. However, with the evolving of ASPL training, similar accuracy as compared with that got on the original clean data can be obtained when the number of iterations increases. As one can comparing "Noise (10/30/50 percent)" with "Noise (10/30/50 percent) w/o VF" from Fig. 8a, with the verifying
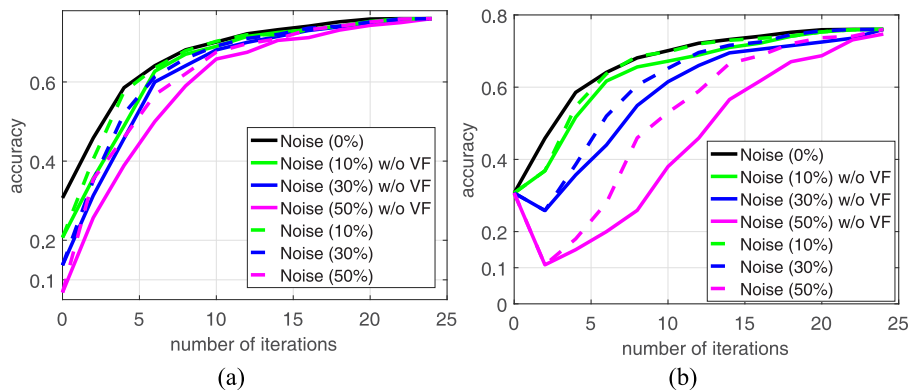
Fig. 8. Robust analysis of ASPL under two types of noisy samples. (a) Using different number of noisy samples as the initial annotation. (b) Adding different number of noisy samples at the 10th step (denoted by the black spots).

step, ASPL can recover from noisy images in a slightly fast way. This justifies the effectiveness of the proposed annotated sample verifying step.

6 CONCLUSIONS

In this paper, we have introduced, first, an effective framework to solve incremental face identification, which build classifiers by progressively annotating and selecting unlabeled samples in an active self-paced way, and second, a theoretical interpretation of the proposed framework pipeline from the perspective of optimization. Third, we evaluate our approach on challenging scenarios and show very promising results.

In the future, we will extend the system to support several video-based vision applications, which require large amount of user annotations. The proposed framework provides a rational realization to this task by automatically distinguishing high-confidence samples, which can be easily and faithfully recognized by computers in a self-paced way, and low-confidence ones, which can be discovered by requesting user annotation.
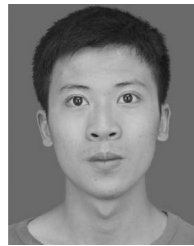
REFERENCES

[1] F. Celli, E. Bruni, and B. Lepri, "Automatic personality and interaction style recognition from facebook profile pictures," in *Proc. 22nd ACM Conf. Multimedia*, 2014, pp. 1101–1104.
[2] Z. Stone, T. Zickler, and T. Darrell, "Toward large-scale face recognition using social network context," *Proc. IEEE*, vol. 98, no. 8, pp. 1408–1415, Aug. 2010.
[3] Z. Lei, D. Yi, and S. Z. Li, "Learning stacked image descriptor for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1685–1696, 2016.
[4] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.

[5] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3539–3545.
[6] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 787–796.
[7] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1489–1496.
[8] X. Wang, X. Guo, and S. Z. Li, "Adaptively unified semi-supervised dictionary learning with active points," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1787–1795.
[9] Y.-F. Li and Z.-H. Zhou, "Towards making unlabeled data never hurt," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 175–188, Jan. 2015.
[10] H. Zhao, et al., "A novel incremental principal component analysis and its application for face recognition," *IEEE Trans. Syst. Man Cybern.*, vol. 36, no. 4, pp. 873–886, Aug. 2006.
[11] T.-K. Kim, K.-Y. K. Wong, B. Stenger, J. Kittler, and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning set approximations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[12] E. Elhamifar, S. Guillermo, Y. Allen, and S. S. Shankar, "A convex optimization framework for active learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 209–216.
[13] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, 2016, Doi: 10.1109/TCSVT.2016.2589879.
[14] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2694–2700.
[15] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Proc. 27th Int. Conf. Advances Neural Inf. Process. Syst.*, 2014, pp. 2078–2086.
[16] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 547–556.
[17] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[18] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. 23rd Int. Conf. Advances Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.
[19] G. Hu, et al., "When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 384–392.
[20] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2010, pp. 253–256.
[21] K. Wang, L. Lin, W. Zuo, S. Gu, and L. Zhang, "Dictionary pair classifier driven convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 2138–2146.
[22] C. Gao, et al., "Interactive surveillance event detection through mid-level discriminative representation," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2014, Art. no. 305.
[23] L. I. Smith, *A Tutorial on Principal Components Analysis*, vol. 51. Ithaca, NY, USA: Cornell Univ., 2002, Art. no. 52.

[24] M. Karasuyama and I. Takeuchi, "Multiple incremental decremental learning of support vector machines," in *Proc. 22nd Int. Conf. Advances Neural Inf. Process. Syst.*, 2009, pp. 907–915.

[25] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Trans. Neural Netw.*, vol. 17, no. 6, pp. 1411–1423, Nov. 2006.

[26] S. Ozawa, S. L. Toh, S. Abe, S. Pang, and N. Kasabov, "Incremental learning of feature space and classifier for face recognition," *Neural Netw.*, vol. 18, no. 5/6, pp. 575–584, Jun. 2005.

[27] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1994, pp. 3–12.

[28] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, 2002.

[29] A. K. McCallumzy and K. Nigamy, "Employing em and pool-based active learning for text classification," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 350–358.

[30] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 2372–2379.

[31] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker, "Which faces to tag: Adding prior constraints into active learning," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 1058–1065.

[32] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with Gaussian processes for object categorization," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[33] X. Li and Y. Guo, "Adaptive active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 859–866.

[34] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proc. IEEE 20th Int. Conf. Mach. Learn.*, 2003, pp. 59–66.

[35] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann, "Self-paced learning for matrix factorization," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3196–3202.

[36] M. P. Kumar, H. Turki, D. Preston, and D. Koller, "Learning specific-class segmentation from diverse data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1800–1807.

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

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[39] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 768–783.

[40] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[41] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 532–539.
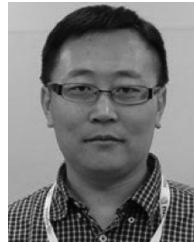
Liang Lin (M'08, SM'15) received the BS and PhD degrees from the Beijing Institute of Technology (BIT), Beijing, China, in 2003 and 2008, respectively, and was a joint PhD student with the Department of Statistics, University of California, Los Angeles (UCLA). He is a full professor of Sun Yat-Sen University. He is the Excellent Young Scientist of the National Natural Science Foundation of China. From 2008 to 2010, he was a post-doctoral fellow with UCLA. From 2014 to 2015, as a senior visiting scholar he was with The Hong Kong Polytechnic University and The Chinese University of Hong Kong. His research interests include computer vision, data analysis and mining, and intelligent robotic systems, etc. He has authored and co-authored on more than 100 papers in top-tier academic journals and conferences. He has been serving as an associate editor of the *IEEE Transactions Human-Machine Systems*. He was the recipient of the Best Paper Runners-Up Award in ACM NPAR 2010, Google Faculty Award in 2012, Best Student Paper Award in IEEE ICME 2014, and Hong Kong Scholars Award in 2014. More information can be found in the website of his group http://hcp.sysu.edu.cn

Keze Wang received the BS degree in software engineering from Sun Yat-Sen University, Guangzhou, China, in 2012. He is currently working toward the PhD degree in computer science and technology at Sun Yat-Sen University, advised by Professor Liang Lin. His current research interests include computer vision and machine learning.

Deyu Meng received the BSc, MSc, and PhD degrees from Xian Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. He is currently a professor in the Institute for Information and System Sciences, School of Mathematics and Statistics, Xian Jiaotong University. From 2012 to 2014, he took his two-year sabbatical leave with Carnegie Mellon University. His current research interests include self-paced learning, noise modeling, and tensor sparsity.

Wangmeng Zuo (M'09, SM'14) received the PhD degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. From July 2004 to December 2004, from November 2005 to August 2006, and from July 2007 to February 2008, he was a research assistant in the Department of Computing, Hong Kong Polytechnic University, Hong Kong. From August 2009 to February 2010, he was a visiting professor with Microsoft Research Asia. He is currently a professor in the School of Computer Science and Technology, Harbin Institute of Technology. His current research interests include image enhancement and restoration, visual tracking, weakly supervised learning, and image classification. He is an associate editor of the *IET Biometrics*. He is a senior member of the IEEE.

Lei Zhang (M'04, SM'14) received the BSc degree from the Shenyang Institute of Aeronautical Engineering, Shenyang, P.R. China, and the MSc and PhD degrees in control theory and engineering from Northwestern Polytechnical University, Xian, P.R. China, in 1998 and 2001, respectively. From 2001 to 2002, he was a research associate in the Department of Computing, Hong Kong Polytechnic University. From January 2003 to January 2006 he worked as a postdoctoral fellow in the Department of Electrical and Computer Engineering, McMaster University, Canada. In 2006, he joined the Department of Computing, Hong Kong Polytechnic University, as an assistant professor. Since July 2015, he has been a full professor in the same department. His research interests include computer vision, pattern recognition, image and video processing, and biometrics, etc. He has published more than 200 papers in those areas. As of 2016, his publications have been cited more than 20,000 times in the literature. He is an associate editor of the *IEEE Transactions on Image Processing*, the *SIAM Journal of Imaging Sciences* and the *Image and Vision Computing*, etc. He is a "Highly Cited Researcher" selected by Thomson Reuters. He is a senior member of the IEEE. More information can be found in his homepage http://www4.comp.polyu.edu.hk/ cslzhang/.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.