Digital Signal Processing ••• (••••) ••••••

Contents lists available at ScienceDirect

Digital Signal Processing

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

More learning with less labeling for face recognition

Barış Büyüktaş [a], Çiğdem Eroğlu Erdem [b], Tanju Erdem [a]

[a] Dept. of Electrical and Electronics Engineering, Özyeğin University, Istanbul, Turkey
[b] Dept. of Computer Engineering, Marmara University, Istanbul, Turkey

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Face recognition
Active learning
Self-paced learning
Minimum sparse reconstruction
Deep learning

ABSTRACT

In this paper, we propose an improved face recognition framework where the training is started with a small set of human annotated face images and then new images are incorporated into the training set with minimum human annotation effort. In order to minimize the human annotation effort for new images, the proposed framework combines three different strategies, namely self-paced learning (SPL), active learning (AL), and minimum sparse reconstruction (MSR). As in the recently proposed ASPL framework [1], SPL is used for automatic annotation of easy images, for which the classifiers are highly confident and AL is used to request the help of an expert for annotating difficult or low-confidence images. In this work, we propose to use MSR to subsample the low-confidence images based on diversity using minimum sparse reconstruction in order to further reduce the number of images that require human annotation. Thus, the proposed framework provides an improvement over the recently proposed ASPL framework [1] by employing MSR for eliminating "similar" images from the set selected by AL for human annotation. Experimental results on two large-scale datasets, namely CASIA-WebFace-Sub and CACD show that the proposed method called ASPL-MSR can achieve similar face recognition performance by using significantly less expert-annotated data as compared to the state-of-the-art. In particular, ASPL-MSR requires manual annotation of only 36.10% and 54.10% of the data in CACD and CASIA-WebFace-Sub datasets, respectively, to achieve the same face recognition performance as the case when the whole training data is used with ground truth labels. The experimental results indicate that the number of manually annotated samples have been reduced by nearly 4% and 2% on the two datasets as compared to ASPL [1].

1. Introduction

Face recognition has become prevalent in many application areas in the last decade, including security, health care, law enforcement, entertainment, education, and marketing. Face recognition is a difficult computer vision and pattern recognition problem due to variations in head pose, illumination, age, facial expressions, facial hair, and makeup [2]. Remarkable improvements have recently been achieved for face recognition using deep convolutional neural networks (DCNN) and large scale training datasets of images collected in unconstrained environments [3–6]. However, annotating millions of face images to build large scale datasets is very tedious, time-consuming, and costly [7–10]. Furthermore, in some applications, data may be available incrementally, i.e., a small number of face images may initially be given for each person, and new images may later become available which also need to be annotated [11–13]. Therefore, it is important to be able to reduce the human annotation effort for very large datasets and new incoming data while keeping the labeling errors at a minimum.

Recently, an active self-paced learning (ASPL) framework has been proposed for progressive and cost-effective face identification [1]. This framework starts with a small annotated dataset and trains a preliminary classifier on this small dataset. Then, at the learning stage, self-paced learning (SPL) and active learning (AL) techniques are combined to incrementally expand the training data by selecting from incoming unlabeled data and then updating the classifiers using the expanded training dataset. The purpose of SPL is to automatically annotate new unlabeled samples, which have the highest classification confidence using the current classifier [14–18]. These high-confidence samples are "easy" samples for the current classifier, and gradually more difficult samples are automatically labeled as the classifier becomes more robust with increasing data. This is similar to the learning process of humans, which starts with easier concepts and then gradually builds on them with more difficult concepts. The purpose of active learning (AL) is to require expert (human) annotation only for the most informative unlabeled images [19–24]. The most informative samples

E-mail address: baris.buyuktas@ozu.edu.tr (B. Büyüktaş).

https://doi.org/10.1016/j.dsp.2023.103915
1051-2004/© 2023 Published by Elsevier Inc.

are generally those that are the most uncertain or have low classification confidence based on the current state of the classifier. The low-confidence samples are considered as "difficult" samples for the current classifier. Although the AL and SPL approaches use conflicting criteria, they complement each other and have been used in ASPL to divide the annotation task between the human (using AL) and the computer (using SPL). The ASPL system requires less hand annotated data but achieves the same accuracy as compared to the case where the whole dataset is hand annotated for face recognition.

In the literature, there are studies which use AL and SPL for other applications. Wang et al. [25] combined AL and SPL by using a new uncertainty quantification method for deep answer selection. In [26] a similar approach to [1] is used for synthetic aperture radar (SAR) automatic target recognition. The high-confidence samples selected with SPL are labeled automatically and low-confidence samples selected by AL are labeled manually. The algorithm is based on a dual-loss function between the newly labeled samples and the original ones to update the model parameters. Yin et al. [27] uses self-paced multi-criteria active learning for image classification. The algorithm utilizes clustering to reduce the negative effects of hard samples. In [28] two weights are used for unlabeled samples, representing easiness and informativeness which are combined to select the most optimal samples for expert labeling.

In this work, we improve the ASPL framework [1] by introducing a minimum sparse reconstruction (MSR) [29] based image selection process to the AL part. Our improvement relies on the fact that the low-confidence unlabeled images may be similar to each other, a fact which has not been utilized in ASPL [1]. Instead of requiring expert annotation for all low-confidence images as in ASPL, we propose to use MSR to select a diverse subset representing all low-confidence images, and require human annotation only for this subset. In other words, in our method, images selected for human annotation will only consist of low-confidence samples which are sufficiently different from each other. Hence, the proposed ASPL-MSR method further reduces the human annotation effort as compared to ASPL, while achieving a similar face recognition performance.

An illustration of high and low-confidence samples is given in Fig. 1, which depicts two classes separated by a linear decision boundary. In this figure, images indicated by red feature vectors are high confidence samples being far from the decision boundary, whereas images indicated by green or yellow feature vectors are low confidence samples being closer to the decision boundary. The collection of yellow feature vectors correspond to the diverse subset representing all low-confidence images, as determined by MSR. As an example, for the subject on the left, the two images with sunglasses are both low-confidence samples, but since they are similar to each other, it is not necessary to label both of them manually. Selecting one of them for human annotation should be sufficient. The other image with the sunglasses is likely to be annotated automatically when the system is re-trained with newly annotated samples and the classifiers become more mature.

The rest of the paper is organized as follows. In Section 2, a brief overview of the proposed framework is given. In Section 3, background information about SPL and ASPL is provided. The details of the proposed active self-paced learning with minimum sparse reconstruction (ASPL-MSR) method are given in Section 4. Experimental results on two large-scale datasets are provided in Section 5 and conclusions are drawn in Section 7.

2. Overview of the framework

In this section, we give an overview of our proposed face recognition framework. As illustrated in Fig. 2, the proposed framework

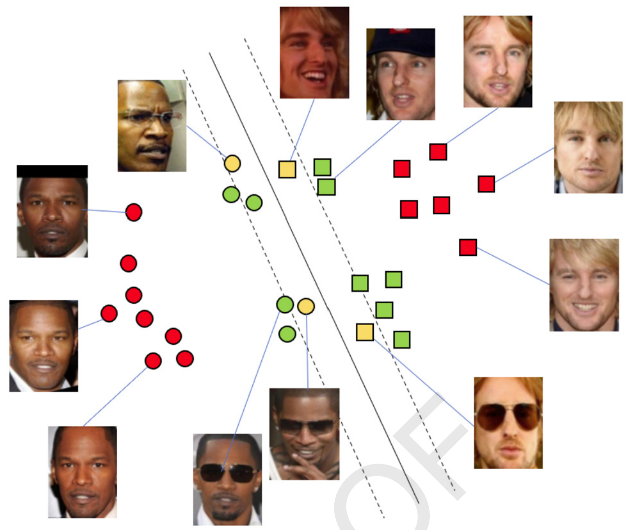

Fig. 1. Illustration of high-confidence and low-confidence samples for two classes. Red points represent high-confidence samples, which are located far from the decision boundary. Green and yellow points represent low-confidence samples, which are located close to the decision boundary. The yellow points form a diverse subset representing the low-confidence samples. Our proposed ASPL-MSR method requires the hand annotation of only the images represented by yellow points, while ASPL requires the hand annotation of images represented by green points, as well.

consists of four main steps: feature extraction, classification, annotation of unlabeled data, and updating the classifier and feature extractor. Each step is described briefly below.

Feature Extraction: We use learning-based features of images extracted using a deep convolutional neural network since it has been shown that they provide features with better discrimination properties for face recognition [30]. As in [1], we use the pre-trained AlexNet model [31] for extracting the features of images. We initially fine-tune AlexNet using 10% of the face dataset (CASIA-WebFace [32,1] or CACD [33,1]) containing randomly selected hand annotated samples for each person. Then, 20% percent of the rest of the dataset is reserved for testing and the remaining 80% is considered to be the unlabeled dataset.

Classification: We used a one-vs-rest linear SVMs for classification. Initially, an SVM is trained for each person with the features extracted using 10% of the data. These classifiers are later updated as the training data is augmented with incoming annotated data.

Annotation of unlabeled data: The current state of feature extractor and classifiers are used to classify the unlabeled images. The confidence level of each image is determined by the distance of its feature vector to the decision boundaries of one-vs-rest SVMs. The farther away a feature vector is from the decision boundary, the higher is its confidence level. If an image is classified with high-confidence (e.g., red points in Fig. 2), it is automatically annotated using SPL and added to the labeled dataset. On the other hand, if an image is classified with low-confidence (e.g., green and yellow points in Fig. 2), it is passed through a similarity detection step using MSR which generates a representative subset (e.g., yellow points in Fig. 2). Only this subset is sent to a human expert for manual annotation. Then, the training dataset is augmented with the automatically annotated and hand annotated samples of this step.

Updating the classifier and the feature extractor: The classifiers are updated using the augmented training dataset. After several classifier updates, the DCNN used for feature extraction is also fine-tuned. The whole process continues until all unlabeled images have been annotated.

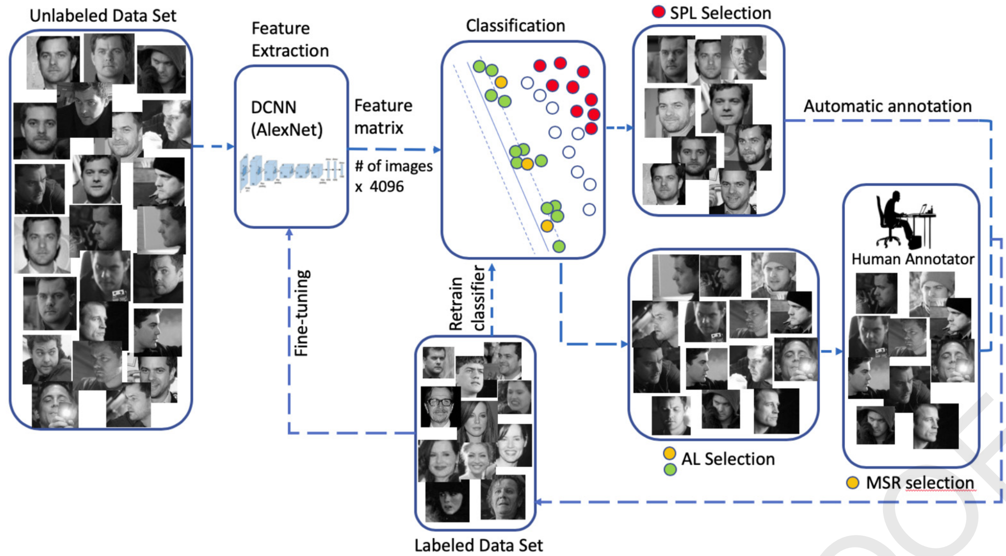

Fig. 2. Illustration of the proposed ASPL-MSR method. The main steps are feature extraction, classification, annotation of unlabeled data (using SPL and AL-MSR), and updating the classifier and feature extractor.

3. Background

In this section, we briefly review self-paced learning (SPL) and active self-paced learning (ASPL) methods in the literature to introduce the basic concepts and the mathematical notation.

3.1. Self-paced learning (SPL)

In education, it is customary to use a pre-defined curriculum which introduces easy concepts first followed by more difficult concepts. It has been shown that this curriculum approach, i.e., ordering of training samples in terms of their difficulty for classification, is also beneficial for machine learning tasks [34]. In self-paced learning, the curriculum is dynamically generated by the learner, as opposed to using a pre-defined (static) curriculum [35]. SPL has been formulated as an optimization problem by embedding the curriculum design as a regularization term into the learning objective function [36], where easy samples are considered to be classified with high-confidence.

Let us denote the training dataset as $D = \{(\mathbf{x}_i, y_i)\}$, $(i = 1, \ldots, n)$, where $\mathbf{x}_i$ is the d-dimensional feature representation of the i^{th} sample, and y_i is the corresponding ground-truth label. Let $L(y_i, h(\mathbf{x}_i, \mathbf{w}))$ denote the cost between the true label y_i and the estimated label $h(\mathbf{x}_i, \mathbf{w})$, and $\mathbf{w}$ denote the model parameters of the learner with the decision function h. In SPL, the goal is to jointly learn the model parameters $\mathbf{w}$, and the latent binary variables v_i, which indicate whether the i^{th} sample is easy or not. The function to be minimized is:

$$\min_{\mathbf{w},\mathbf{v}} \mathbb{E}(\mathbf{w}, \mathbf{v}; \lambda) = \sum_{i=1}^{n} v_i L(y_i, h(\mathbf{x}_i, \mathbf{w})) - \lambda \sum_{i=1}^{n} v_i, \tag{1}$$

where $\mathbf{v} = [v_1, \ldots, v_n]^T$ is a vector with elements 0 or 1, that is $\mathbf{v} \in [0, 1]^n$. The parameter λ is used for controlling the learning pace. The second term in (1) is a negative l_1-norm regularizer as $-\|\mathbf{v}\|_1 = -\sum_{i=1}^{n} v_i$ since $v_i \geq 0$. If λ is small, easy samples are the ones that give a small value of $L(.)$. Note that, samples are tied together in function $\mathbb{E}$ through the parameter $\mathbf{w}$, hence no single sample is easy, but a set of samples is considered to be easy if a parameter $\mathbf{w}$ can be learned such that the corresponding losses $L(.)$ are small. The problem in (1) can also be relaxed so that v_i is allowed to take any value in [0, 1].

If $L(.)$ is convex in $\mathbf{w}$, (1) becomes a biconvex optimization problem with two sets of variables $\mathbf{w}$ and $\mathbf{v}$, so that when one set is fixed, the optimal value of the other set can be solved by using convex optimization. In particular alternative convex search (ACS) [37] can be used to solve (1). When $\mathbf{v}$ is fixed, the optimal parameter $\mathbf{w}^*$ can be solved, and when $\mathbf{w}$ is fixed, the global optimum $\mathbf{v}^* = [v_1^*, \ldots, v_n^*]^T$ can be calculated by [35]:

$$v_i = \begin{cases} 1, & \text{if } L(y_i, h(\mathbf{x}_i, \mathbf{w})) < \lambda \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The above alternative search strategy can be interpreted as follows. When updating $\mathbf{v}$ with fixed $\mathbf{w}$, samples with loss values less than a threshold λ are considered as easy (or high-confidence) and will be added to the training dataset. When updating $\mathbf{w}$ with $\mathbf{v}$ fixed, the classifier will be trained on the training dataset augmented with the newly selected easy samples. The parameter λ can also be considered as the "age" of the model. If λ is small (i.e., the model is "young") only samples with small losses will be included in the training set, which are easier to learn. As the model matures, λ may increase to include samples with larger losses.

3.2. Active self-paced learning for face recognition (ASPL)

In this approach, the classifiers are initially trained using a very small number of manually annotated samples. At each subsequent iteration, easy unlabeled samples are automatically labeled by SPL and difficult unlabeled samples are labeled by AL. In this section, we give a summary of the ASPL formulation; the reader is referred to [1] for further details.

Let $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, $(i = 1, \ldots, n)$, represent the training dataset where $\mathbf{x}_i$ is the d-dimensional feature representation of the i^{th} image, and $\mathbf{y}_i$ is the vector representing the ground truth label. A ground truth label is defined as $\mathbf{y}_i = \{y_i^{(j)} \in \{-1, 1\}\}_{j=1}^{m}$, where $y_i^{(j)}$ corresponds to the label of x_i for the j^{th} subject, as there are m subjects in total. If $\mathbf{x}_i$ belongs to the j^{th} subject, then $y_i^{(j)} = 1$, if not $y_i^{(j)} = -1$.

In ASPL, m linear one-vs-rest SVM classifiers are trained using the feature vectors extracted by the CNN. Inspired by [1], we define the function to be minimized as:

$$\min_{\{\mathbf{w},b,\mathbf{v},\mathbf{y}\}} \mathbb{E}(\mathbf{w}, b, \mathbf{y}, \mathbf{v}; \lambda, \Psi)$$

$$= \sum_{j=1}^{m} \left\| \mathbf{w}^{(j)} \right\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} L(y_i^{(j)}, h(\mathbf{x}_i, \mathbf{w}^{(j)}, b^{(j)})) - \mathscr{F}(\mathbf{v}^{(j)}; \lambda_j), \quad \text{s.t.}\, \mathbf{v} \in \Psi^{\lambda} \tag{3}$$

where $\mathbf{w} = \{w^{(j)}\}_{j=1}^{m} \subset \mathbb{R}^d$, $b = \{b^{(j)}\}_{j=1}^{m} \subset \mathbb{R}$ are the weight and bias parameters of the m linear decision functions, and C is a regularization parameter between the margin term (first term) and the loss term (second term), and it is set to 1 in the experiments. The last term in (3) $\mathscr{F}$ is the self-paced regularizer for the j^{th} classifier which determines how the weights $\mathbf{v}$ change in $[0, 1]$ and it can be the binary scheme as in (1) or linear, logarithmic can also be used [36]. The curriculum Ψ^{λ} is dynamic and determined by the samples selected by the SPL and AL steps as well as the parameter λ, which will be explained below. The loss function $L(.)$ is defined using the hinge loss of $\mathbf{x}_i$ in the j^{th} classifier as $(1 - y_i^j(\mathbf{w}^{(j)T} x_i + b^{(j)}))_+$.

4. Face recognition using active self-paced learning with minimum sparse reconstruction (ASPL-MSR)

In this section, we give the details of the proposed active self-paced learning with minimum sparse reconstruction (ASPL-MSR) method for face recognition. We explain each component of the proposed method, including automatic annotation of high-confidence samples using SPL and expert annotation using AL-MSR for low-confidence samples.

The system is initialized by extracting the feature vectors of all labeled and unlabeled samples using the pre-trained CNN as shown in Fig. 2.

Then, the following function is minimized, where we use the l_1 norm regularizer as the last term:

$$\min_{\{\mathbf{w},b,\mathbf{v},\mathbf{y}\}} \mathbb{E}(\mathbf{w}, b, \mathbf{y}, \mathbf{v}; \lambda, \Psi) = \sum_{j=1}^{m} \left\| \mathbf{w}^{(j)} \right\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} L(y_i^{(j)}, h(\mathbf{x}_i, \mathbf{w}^{(j)}, b^{(j)})) - \lambda \sum_{i=1}^{n} v_i, \quad \text{s.t.}\, \mathbf{v} \in \Psi^{\lambda} \tag{4}$$

where the parameters are as defined in section 3. The function (4) is minimized using an alternative search strategy as described in section 3.1. The first step of the alternative search strategy is *classifier updating*, which consists of training m one-vs-rest SVM classifiers, after fixing the current labeled (training) set. In the second step of the alternative search strategy, the m SVMs are fixed, and *high-confidence sample labeling* and *low-confidence sample labeling with MSR-based selection (AL-MSR)* steps are executed, as described below.

4.1. Classifier updating

In this step the parameters of one-vs-rest SVM classifiers $\{\mathbf{w}^{(j)}, b^{(j)}\}_{j=1}^{m}$ are updated while the labeled training dataset and the other parameters $\{\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}, \mathbf{v}, \Psi^{\lambda}\}$ are fixed. Therefore, (4) simplifies as follows:

$$\min_{\mathbf{w},b} \sum_{j=1}^{m} \left\| \mathbf{w}^{(j)} \right\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} L(y_i^{(j)}, h(\mathbf{x}_i, \mathbf{w}^{(j)}, b^{(j)})), \tag{5}$$

The function in (5) is minimized for each of the m SVM classifiers.

4.2. High-confidence sample labeling

The idea of high-confidence labeling is to automatically label the unlabeled samples for which the classifier is highly confident. Hence, pseudo-labels $\mathbf{y}$ and the binary weights $\mathbf{v}$ indicating "easiness (confidence)" are estimated, while the SVM parameters and the labeled training dataset $\{\{\mathbf{w}^{(j)}, b^{(j)}\}_{j=1}^{m}, (\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}, \Psi^{\lambda}\}$ are fixed. Therefore (4) is simplified as follows:

$$\min_{\mathbf{v}} \sum_{j=1}^{m} C \sum_{i=1}^{n} v_i^{(j)} L(y_i^{(j)}, h(\mathbf{x}_i, \mathbf{w}^{(j)}, b^{(j)})) - \lambda \sum_{i=1}^{n} v_i, \tag{6}$$

This becomes a similar problem to (1) and the solution is similar to (2) where we use the "confidence" which is measured based on the distance from the decision boundary of the classifier (SVM). Let us denote the distance of $\mathbf{x}_i$ to the linear decision boundary of j^{th} classifier as $u_i^{(j)} = (\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)}) / \left\| \mathbf{w}^{(j)} \right\|_2$. Then,

$$v_i = \begin{cases} 1, & \text{if} \max_j u_i^{(j)} > \tau_{high},\ j = 1, \ldots, m \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

which states that the sample is high-confidence (or easy) if it is sufficiently far away from the decision boundary. Such samples are automatically labeled as belonging to the class $j^* = \text{argmax}_j\, u_i^{(j)}$.

4.3. Low-confidence sample labeling with MSR-based selection (AL-MSR)

An unlabeled sample is considered to be low-confidence if the classifier is not able to clearly identify the class label. A sample is estimated to be low-confidence if difference between the highest two signed distances to decision boundaries of m classifiers is small (i.e. less than a threshold τ_{low}).

Instead of manually labeling all low-confidence samples directly, which may have similar samples, we propose to select the most representative (or diverse) ones using minimum sparse reconstruction to reduce the human annotation effort and manually label only the diverse set.

Mei et al. [29] used MSR to find the most informative (key) frames in videos. We adopt it to select diverse images from the low-confident samples in our problem as described below.

Let us call the set of low-confident samples at the current iteration as the *low-confident set* with α elements and write their feature vectors in a matrix as $\mathbf{F} = [\mathbf{f}_1,\ \mathbf{f}_2 \ldots, \mathbf{f}_\alpha] \in \mathbb{R}^{d\times\alpha}$. MSR tries to find the optimal subset $\mathbf{F}_K = \left[\mathbf{f}_{k_1},\ \mathbf{f}_{k_2} \ldots, \mathbf{f}_{k_\beta}\right] \in \mathbb{R}^{d\times\beta}$, where $k_1,\ k_2 \ldots, k_\beta \in \{1, 2, ..., \alpha\}$ so that the original low-confident set $\mathbf{F}$ can be accurately reconstructed by $\mathbf{F}_K$ and the size of $\mathbf{F}_K$ is as small as possible. The objective function to be minimized is:

$$\min_{\mathbf{S}} \frac{1}{2} \left\| \mathbf{F} - \mathbf{F}_K \mathbf{A} \right\|_2 + \Lambda \left\| \mathbf{S} \right\|_0 \quad \text{s.t.}\ \mathbf{F}_K = \mathbf{FS}, \quad \mathbf{A} = g(\mathbf{F}, \mathbf{F}_K), \tag{8}$$

where $\mathbf{S}$ is a diagonal selection matrix that selects the diverse (or key) images from the low-confident set, where the diagonal elements are 1 only for columns to be selected, otherwise they are zero and $\|\mathbf{S}\|_0$ is the L_0 norm of the selection matrix. The number of nonzero elements in $\mathbf{S}$ is β, which is the total number of selected diverse images. In (8), the columns of the matrix $\mathbf{A}$ are the reconstruction coefficients of $\mathbf{F}$ given the matrix $\mathbf{F}_K$, $g(\mathbf{F}, \mathbf{F}_K)$ is the reconstruction function, and Λ is the weighting coefficient. The first term in (8) minimizes the least-squares reconstruction error (LSRE) and the second term minimizes the number of selected diverse images.

Assuming that β images have been selected and placed in columns of $\mathbf{F}_K$, the next image selected should be the one that gives the maximum LSRE, that the most dissimilar image to the ones in $\mathbf{F}_K$:

$$\mathbf{f}_{k_{\beta+1}} = \underset{\mathbf{f}_r \in \mathbf{F}/\mathbf{F}_K}{\operatorname{argmax}} \|\mathbf{f}_r - \mathbf{F}_K \mathbf{a}_r\|_2 \,, \tag{9}$$

where $\mathbf{f}_r$ represent the features vector of frame r; $\mathbf{a}_r$ represent the reconstruction coefficients of frame r and $\mathbf{F}/\mathbf{F}_K$ represent the images in the low-confident set, which have not been selected. This is equivalent to minimizing:

$$\mathbf{f}_{k_{\beta+1}} = \underset{\mathbf{f}_r \in \mathbf{F}/\mathbf{F}_K}{\operatorname{argmin}} \|\mathbf{F}_K \mathbf{a}_r\|_2 \,. \tag{10}$$

Since vectors with large magnitudes tend to have large LSRE, we can normalize with the magnitude and name as the percentage of reconstruction (POR):

$$\mathbf{f}_{k_{\beta+1}} = \underset{\mathbf{f}_r \in \mathbf{F}/\mathbf{F}_K}{\operatorname{argmin}} POR_r = \underset{\mathbf{f}_r \in \mathbf{F}/\mathbf{F}_K}{\operatorname{argmin}} \frac{\|\mathbf{F}_K \mathbf{a}_r\|_2}{\|\mathbf{f}_r\|_2} . \tag{11}$$

Reconstruction coefficients $\mathbf{a}_r$'s are calculated using orthogonal subspace projection (OSP), which projects all the images to the orthogonal space spanned by selected images $\mathbf{F}_K$:

$$\mathbf{a}_r = \left(\mathbf{F}_K^T \mathbf{F}_K\right)^{-1} \mathbf{F}_K^T \mathbf{f}_r = \mathbf{P}_K \mathbf{f}_r \tag{12}$$

The algorithm continues to add new images to $\mathbf{F}_K$ until the percentage of reconstruction error (POR) is below a pre-determined threshold T_p.

$$POR_r = \frac{\|\mathbf{F}_K \mathbf{a}_r\|_2}{\|\mathbf{f}_r\|_2} < T_p \tag{13}$$

The number of selected diverse images increases when the threshold T_p is increased. Note that a POR value of 1 means that the rth image in the low-confident set can be perfectly reconstructed as a linear combination of the images in $\mathbf{F}_K$. We summarized the MSR based image selection algorithm in Algorithm 1.

Algorithm 1: Minimum sparse reconstruction based diverse image selection on low-confident set (AL-MSR).

Inputs: Low-confident set $\mathbf{F} \in \mathbb{R}^{d \times \alpha}$ with α images, where each feature vector is extracted using the fine-tuned CNN; Threshold T_p.
Output: The diverse image subset $\mathbf{F}_K \in \mathbb{R}^{d \times \beta}$, where $\beta < \alpha$.
1: Initialize the diverse image set using the first image in the low-confident set $\left[\mathbf{f}_{k_1}\right]$ and set $\beta = 1$.
2: Select the second diverse image as the one with the largest Euclidean distance from the first image $\mathbf{F}_K = \left[\mathbf{f}_{k_1}, \mathbf{f}_{k_2}\right]$ and set $\beta = 2$.
3: Calculate the POR for all images in set $\mathbf{F}/\mathbf{F}_K$ via Eqn. (13).
4: **while** $POR < T_p$ **do**
  5: Select the next diverse image via Eqn. (11).
  6: Update diverse image set $\mathbf{F}_K = \left[\mathbf{F}_K, \mathbf{f}_{k_{\beta+1}}\right]$.
  7: Calculate the POR for all images in set $\mathbf{F}/\mathbf{F}_K$ via Eqn. (13).
  8: Increase β by one.
end

A toy example for the MSR-based diverse image selection method is shown in Fig. 3. The images in the top row represent the images in the active set before selection, which contain similar images. Since MSR tries to reconstruct a subset with as few diverse images as possible, similar images are eliminated. In CASIA-WebFace-Sub and CACD datasets, there are indeed similar face images taken with short time differences. Instead of labeling all similar images manually, MSR selects a diverse subset for labeling. Later, unselected similar images are easily automatically labeled by SPL.

4.4. Overall algorithm for face recognition using ASPL-MSR

We combine self-paced learning, and active learning with minimum sparse reconstruction methods to build a cost-effective framework for face recognition by taking advantage of each approach. SPL selects the high-confidence images which are automatically labeled and AL selects the low-confidence images which are informative for training. MSR selects the best representative and diverse subset of the informative (low-confidence) images to reduce the number of images that need to be labeled manually.

The main steps for the proposed ASPL-MSR method are given in Algorithm 2. The main steps and parameters are similar to [1], except the AL-MSR based selection step (6:). MSR selects the most representative images for training from the low-confidence (active) set. Selected images are manually annotated, and unselected images are returned to the unlabeled set. SVMs are retrained after each SPL and AL-MSR annotation. After T rounds of SPL and AL-MSR updates, the CNN is fine-tuned with the labeled instances. This process continues until all images are labeled.

Algorithm 2: The proposed ASPL-MSR method for face recognition.

Inputs: A dataset: %10 annotated ($\{\mathbf{x}_i\}$), and %90 not annotated ($\{\mathbf{u}_j\}$); Model parameters of the pre-trained CNN (AlexNet); number of subjects to be recognized: m; T: number of iterations for updating classifiers.
Output: Optimal model parameters $\mathbf{w}^*$, b^* for m SVM classifiers.
while $\{\mathbf{u_j}\} \neq \emptyset$ **do**
  1: Use pre-trained CNN to extract feature vectors of images in $\{\mathbf{x}_i\}$ and $\{\mathbf{u}_j\}$
  for *iteration* $= 1 : T$ **do**
    2: Train m one-vs-all SVMs using the feature vectors of $\{\mathbf{x}_i\}$.
    3: Detect and automatically label high-confidence samples using SPL via Eqn. (7).
    4: Update $\{\mathbf{x}_i\}$ and $\{\mathbf{u}_j\}$.
    5: Train m one-vs-all SVMs using the feature vectors of $\{\mathbf{x}_i\}$.
    6: Detect and manually label diverse and low-confidence samples using AL-MSR (Algorithm 1).
    7: Update $\{\mathbf{x}_i\}$ and $\{\mathbf{u}_j\}$.
  end
  8: Update feature extractor: Fine-tune CNN using $\{\mathbf{x}_i\}$.
end
return $\mathbf{w}^*$, b^* for m SVMs.

5. Experimental results

In this section, we give the details of the experimental setup and provide results using self-paced learning, active learning, active self-paced learning, and the proposed active-self-paced learning with minimum sparse reconstruction (ASPL-MSR) methods on two different data sets, namely CASIA-WebFace-Sub and CACD.

5.1. Experimental setup

For both data sets, we use 80% of the images for training and 20% of the images for testing. Approximately 10% of the total training data is randomly selected as the initial set and used for fine-tuning the pre-trained CNN (AlexNet) model. The rest of the training data is treated as unlabeled, which will be annotated by the proposed method.

We fine-tune the last 2 fully connected (FC) layers of the pre-trained AlexNet using the initial set. Then, all images are passed through the network and 4096-dimensional feature vectors are obtained from the fc7 layer. We then train the m linear kernel SVMs using the feature vectors of the initial set. Low-confidence and high-confidence samples are specified based on the distance of their feature vectors to the one-vs-rest SVM decision boundaries as explained in the previous section.

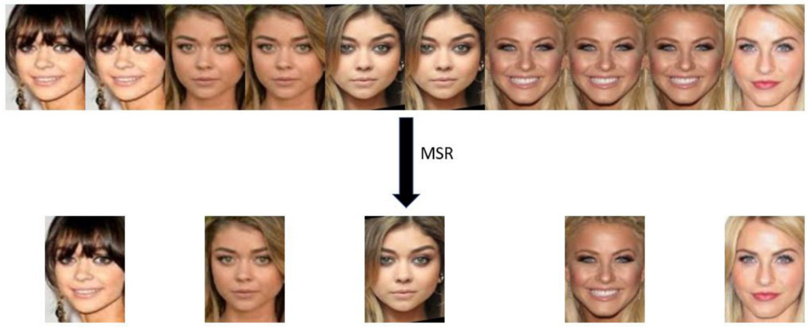

Fig. 3. The toy example of the MSR method. The images in the first row represent the candidates for selection, and the images in the second row represent the selected images with MSR.

The thresholds used in SPL, AL, and MSR methods are determined as follows:

- The parameter τ_{high} in SPL (7) used for automatic labeling is selected as 1, which is determined experimentally, and its value does not change during the iterations. Lower τ_{high} values cause more labeling error due to the incorrect annotation, and higher values result in less number of automatically selected images.
- The threshold for determining low-confidence images in AL is set to $\tau_{low} = 0.0075$ initially and it is doubled in each fine-tuning step. The network gets more accurate after each iteration, hence, the number of selected uncertain images decreases. Therefore, we increase the threshold value to keep the total number of selected uncertain images at a certain level.
- The parameter T in Algorithm 2 is determined experimentally to be 3, as in [1].
- DCNN is fine-tuned using stochastic gradient descent with momentum 0.9, learning rate 0.001, batch size 128, and weight decay 0.0001.

5.2. Experimental results on the cross-age celebrity dataset (CACD)

Cross-Age Celebrity (CACD) dataset is a publicly available large-scale dataset, which contains more than 160,000 images from 2000 celebrities ranging in age from 16 to 32. Chen et al. [33] manually annotated a subset of 200 celebrities. To increase the number of annotated samples and to compare our results with [1], we used the augmented version of the subset, which was generated by Lin et al. [1]. They manually annotated 300 celebrities, hence, the subset contains 56,105 images from 500 celebrities. These images are in RGB format with a size of 150 × 200 pixels.

We present the experimental results on CACD dataset in Fig. 4, which shows the plots of accuracies of SPL, AL, AL-random, ASPL, and the proposed ASPL-MSR methods versus the percentage of manually annotated samples. The dashed black line at the top is the maximum accuracy (92.87%) achieved when we use the whole CACD dataset with the ground-truth labels.

Since the total number of manually annotated samples does not change during training with SPL, its maximum accuracy (51.78%) is shown with a flat blue line. SPL selects only high-confidence samples automatically, which are far away from the decision boundary. Hence, the accuracy of SVMs do not improve and a limited accuracy can be achieved.

AL (purple line) and ASPL (green line) reached the maximum accuracy with less percentage of manually labeled images, 45.93% and 40.15%, respectively. To observe the effects of AL in reducing the total number of annotated samples, the results of the AL and AL-random (cyan line) can be compared. AL-random selects the same number of images as AL in each iteration but in a random way. Its maximum accuracy (78.32%) is much lower than the highest accuracy. The proposed ASPL-MSR method achieves the maximum accuracy using the least number of human-annotated data (36.90%) as seen in Fig. 4, which indicated that human annotation effort can be significantly reduced.

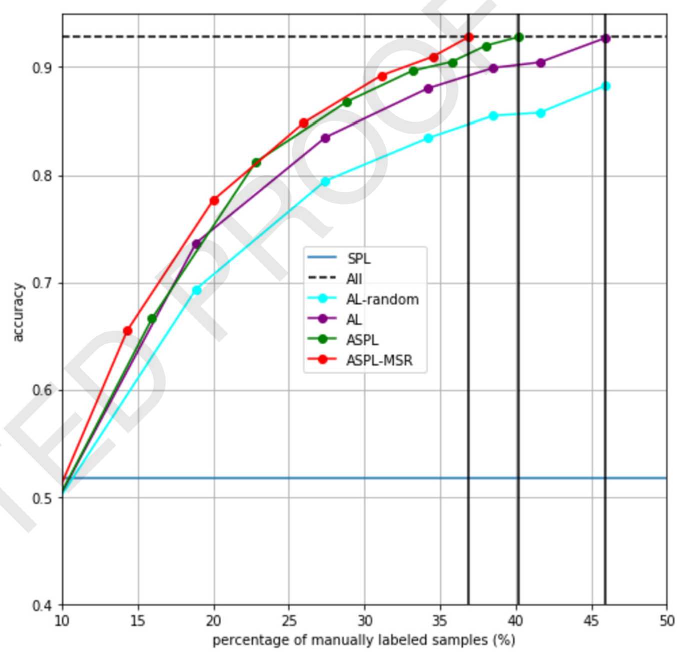

Fig. 4. Accuracies of the 5 different methods (SPL, AL, AL-random, ASPL, **proposed ASPL-MSR**) versus the percentage of manually labeled samples on the CACD dataset.

In Fig. 5, the percentage of manually labeled samples can be seen as the number of iterations of Algorithm 2 increase from 1 to 7. We can see that the proposed ASPL-MSR method requires the least number of manually annotated samples at all iterations.

In Table 1, we give a summary and comparison of the performance of different methods on the CACD dataset. We can see that the proposed ASPL-MSR method requires less manually annotated training samples than state-of-the-art methods.

5.3. Experimental results on the CASIA-WebFace-sub dataset

The CASIA-WebFace-Sub dataset used in the experiments contains 181,279 images from 923 individuals. Although the number of images per person ranged from 3 to 804 in the CASIA-WebFace database in [32], to reduce the imbalance in this distribution, individuals with fewer than 100 images were not used in the experiments as in [1]. All images are grayscale with a size of 100 × 100 pixels.

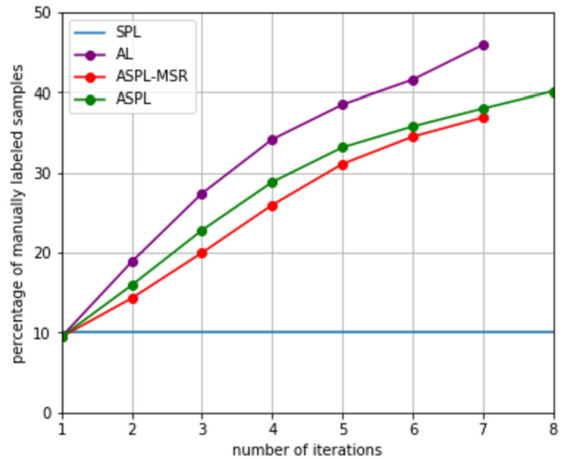

Fig. 5. Accuracy of the 4 different methods (SPL, AL, ASPL, **proposed ASPL-MSR**)with the increase of number of iterations on the CACD dataset. One iteration includes a fine-tuning of the CNN in Algorithm 2.

Table 1
Summary and comparison of the different methods on CACD dataset. The percentage of manually annotated data to achieve the maximum accuracy (92.87%) is given.

Method	Total Manually Annotated Samples (%)
AL	45.93
ASPL (Our implementation)	40.15
ASPL-MSR (Ours)	**36.90**
ASPL [1]	41.00

The plots of experimental results on the CASIA-WebFace-Sub dataset are given in Fig. 6, which shows the accuracies of SPL, AL, AL-random, ASPL, and the proposed ASPL-MSR methods versus the percentage of manually annotated samples.

The maximum accuracy that can be achieved when we use all of the training data with ground truth labels is 81.72%, which is shown by the dashed black line in Fig. 6. Since the total number of manually annotated samples does not change during training with SPL, its maximum accuracy (40.11%) is shown with a flat blue line. The accuracy of AL (purple line) almost reached the maximum level (81.52%) using manual labeling of 64.07% of the training set. It can be observed that the maximum accuracy was achieved when 55.83% of the entire dataset is annotated manually using ASPL, which is almost 10% lower than the percentage of total manually annotated images using the AL method. Using the proposed ASPL-MSR method (red line) the maximum accuracy of 81.59% is achieved using manual annotation of 54.10% of training data. This indicates a 1.73% improvement in reducing the percentage of manually annotated samples as compared to the ASPL approach.

In Fig. 7 the percentage of manually labeled samples can be seen as the number of iterations of Algorithm 2 increase from 1 to 7. We can see that the proposed ASPL-MSR method requires the least number of manually annotated samples at all iterations.

In Table 2, we give a summary and comparison of the performance of different methods on the CASIA-WebFace-Sub dataset. We can see that our ASPL-MSR method requires less manually annotated training samples than other state-of-the-art methods.

6. Discussion

The experimental results on two large scale face datasets show that the proposed method is successful in reducing the manual data annotation effort. This is achieved at the expense of an additional MSR step, which increases the number of computations. The

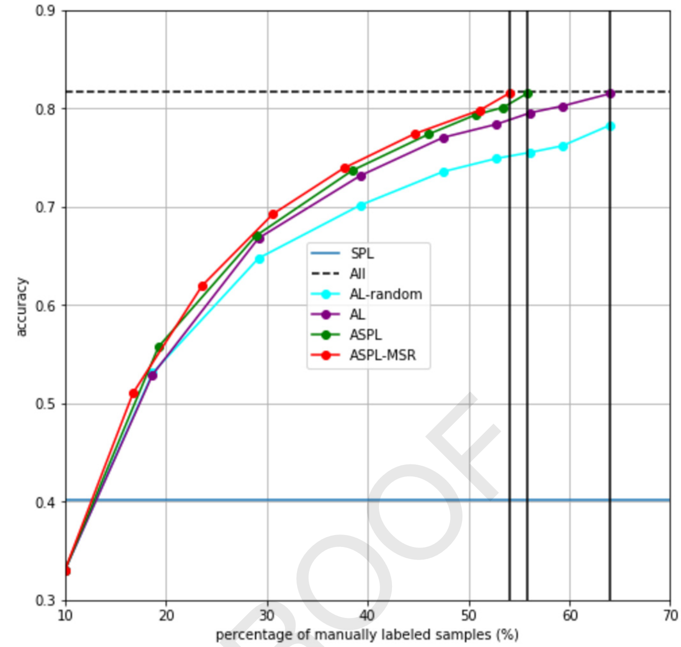

Fig. 6. Accuracies of the 5 different methods (SPL, AL, AL-random, ASPL, **proposed ASPL-MSR**) versus the percentage of manually labeled samples on the CASIA-WebFace-Sub dataset.

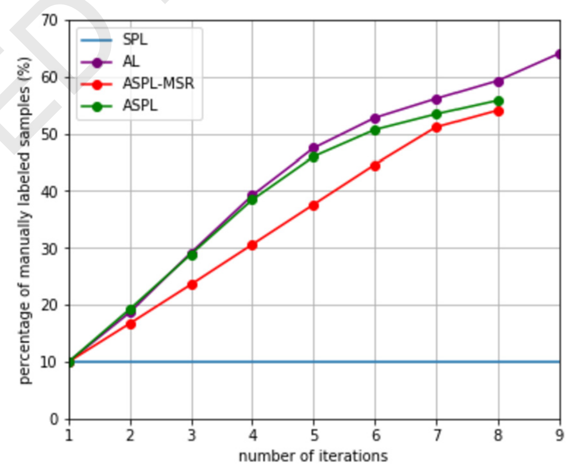

Fig. 7. Accuracy of the 4 different methods (SPL, AL, ASPL, **proposed ASPL-MSR**)versus the number of iterations on the CASIA-WebFace-Sub dataset. ASPL-MSR is the proposed method, which requires the least manually labeled data. One iteration includes a fine-tuning of the CNN in Algorithm 2.

Table 2
Summary and comparison of the different methods on CASIA-WebFace-Sub dataset. The percentage of manually annotated data to achieve the maximum accuracy is given.

Method	Total Annotated Samples (%)
AL	64.07
ASPL (Our implementation)	55.83
ASPL-MSR (Ours)	**54.10**
ASPL [1]	56.00

complexity of our method increases with the increase in the total number of images selected with AL-MSR. Especially, the matrix inversion in (12) is computationally costly if the size of the matrix $\mathbf{F}_K$ becomes large (i.e. the number of the selected images is large). Since the method calculates the POR value for all images at each iteration, if the number of images selected with AL-MSR is large, computational complexity will increase accordingly.

The proposed incremental SVM based method is a highly accurate multi-class classification method. However, pre-trained preprocessing may not always achieve the best results. The DCNN used may be replaced by other architectures in the literature, which may give higher accuracies. Since our goal was to experimentally validate the reduction in human annotation effort with the utilization of minimum sparse reconstruction together with ASPL, we used the same DCNN and experimental conditions as [1].

7. Conclusion

In this paper, we presented an improvement over the recently proposed active self-paced learning (ASPL) method [1] for face recognition in order to significantly reduce the human annotation effort. The proposed method relies on the use of minimum sparse reconstruction (MSR) based diverse subset selection for eliminating most of the low-confidence samples that ASPL would normally require to be manually annotated. Experimental results on the large scale CASIA-WebFace-Sub and CACD datasets indicate that the number of annotated samples is reduced significantly while obtaining an accuracy that is comparable to the case when the whole training data is used with ground truth labels. To the best of our knowledge, ours is the first work that uses MSR within the context of active learning in order to reduce the hand annotation effort.

The proposed ASPL-MSR method can be applied to other classification problems as future work. Since the system is modular, in order to improve the performance, other CNN structures can be used for feature extraction and different methods can be designed for the classification step to determine the low-confidence and high-confidence samples at each iteration. The method we introduced (ASPL-MSR) is a semi-supervised incremental method for face recognition. In the future, unsupervised methods could be explored and the face recognition accuracy could be compared with the proposed ASPL-MSR method.

CRediT authorship contribution statement

Barış Büyüktaş: Conceptualization, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Çiğdem Eroğlu Erdem:** Conceptualization, Formal analysis, Resources, Supervision, Writing – original draft, Writing – review & editing. **Tanju Erdem:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under project 116E088. The Titan V used for this research was donated by the NVIDIA Corporation.

References

[1] L. Lin, K. Wang, D. Meng, W. Zuo, L. Zhang, Active self-paced learning for cost-effective and progressive face identification, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 7–19.
[2] M. Taskiran, N. Kahraman, C.E. Erdem, Face recognition: past, present and future (a review), Digit. Signal Process. 106 (2020) 102809.
[3] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.
[4] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, 2015.
[5] C. Huang, Y. Li, C.L. Chen, X. Tang, Deep imbalanced learning for face recognition and attribute prediction, IEEE Trans. Pattern Anal. Mach. Intell. (2019).
[6] E. Zangeneh, M. Rahmati, Y. Mohsenzadeh, Low resolution face recognition using a two-branch deep convolutional neural network architecture, Expert Syst. Appl. 139 (2020) 112854.
[7] I. Kemelmacher-Shlizerman, S.M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 4873–4882.
[8] A. Bansal, A. Nanduri, C.D. Castillo, R. Ranjan, R. Chellappa, Umdfaces: an annotated face dataset for training deep networks, in: IEEE Int. Joint Conf. on Biometrics, IEEE, 2017, pp. 464–473.
[9] B. Maze, J. Adams, J.A. Duncan, N. Kalka, T. Miller, C. Otto, A.K. Jain, W.T. Niggel, J. Anderson, J. Cheney, et al., Iarpa janus benchmark-c: face dataset and protocol, in: Int. Conf. on Biometrics, IEEE, 2018, pp. 158–165.
[10] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: a dataset for recognising faces across pose and age, in: IEEE Int. Conf. on Automatic Face & Gesture Recognition, IEEE, 2018, pp. 67–74.
[11] M. Karasuyama, I. Takeuchi, Multiple incremental decremental learning of support vector machines, IEEE Trans. Neural Netw. 21 (2010) 1048–1059.
[12] S. Ozawa, S.L. Toh, S. Abe, S. Pang, N. Kasabov, Incremental learning of feature space and classifier for face recognition, Neural Netw. 18 (2005) 575–584.
[13] S. Madhavan, N. Kumar, Incremental methods in face recognition: a survey, Artif. Intell. Rev. 54 (2021) 253–303.
[14] J.S. Supancic, D. Ramanan, Self-paced learning for long-term tracking, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 2379–2386.
[15] B. Buyuktas, C. Eroglu Erdem, A.T. Erdem, Curriculum learning for face recognition, in: European Conf. on Signal Processing, 2020.
[16] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, A. Hauptmann, Self-paced learning with diversity, in: Advances in Neural Information Processing Systems, 2014, pp. 2078–2086.
[17] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, A.G. Hauptmann, Self-paced learning for matrix factorization, in: AAAI Conf. on Artificial Intelligence, 2015.
[18] H. Li, M. Gong, D. Meng, Q. Miao, Multi-objective self-paced learning, in: AAAI Conf. on Artificial Intelligence, 2016.
[19] B. Settles, M. Craven, S. Ray, Multiple-instance active learning, in: Advances in Neural Information Processing Systems, 2008, pp. 1289–1296.
[20] B. Settles, M. Craven, An analysis of active learning strategies for sequence labeling tasks, in: Conf. on Empirical Methods in Natural Language Processing, 2008, pp. 1070–1079.
[21] A. Culotta, A. McCallum, Reducing labeling effort for structured prediction tasks, in: AAAI Conf. on Artificial Intelligence, vol. 5, 2005, pp. 746–751.
[22] A. Holub, P. Perona, M.C. Burl, Entropy-based active learning for object recognition, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops, IEEE, 2008, pp. 1–8.
[23] A.J. Joshi, F. Porikli, N. Papanikolopoulos, Multi-class active learning for image classification, in: IEEE Conf. on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2372–2379.
[24] A.J. Joshi, F. Porikli, N.P. Papanikolopoulos, Scalable active learning for multiclass image classification, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 2259–2273.
[25] Q. Wang, W. Wu, Y. Qi, Z. Xin, Combination of active learning and self-paced learning for deep answer selection with bayesian neural network, in: European Conference on Artificial Intelligence, IOS Press, 2020, pp. 1587–1594.
[26] H. Ren, X. Yu, L. Bruzzone, Y. Zhang, L. Zou, X. Wang, A Bayesian approach to active self-paced deep learning for SAR automatic target recognition, IEEE Geosci. Remote Sens. Lett. 19 (2020) 1–5.
[27] T. Yin, N. Liu, H. Sun, Self-paced active learning for deep CNNs via effective loss function, Neurocomputing 424 (2021) 1–8.
[28] Y.-P. Tang, S.-J. Huang, Self-paced active learning: query the right thing at the right time, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 5117–5124.
[29] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, D.D. Feng, Video summarization via minimum sparse reconstruction, Pattern Recognit. 48 (2015) 522–533.
[30] Z. Lei, D. Yi, S.Z. Li, Learning stacked image descriptor for face recognition, IEEE Trans. Circuits Syst. Video Technol. 26 (2016) 1685–1696.
[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[32] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923, 2014.

[33] B.-C. Chen, C.-S. Chen, W.H. Hsu, Cross-age reference coding for age-invariant face recognition and retrieval, in: European Conf. on Computer Vision, Springer, 2014, pp. 768–783.
[34] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Int. Conf. on Machine Learning, ACM, 2009, pp. 41–48.
[35] M.P. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models, in: Advances in Neural Information Processing Systems, 2010, pp. 1189–1197.
[36] L. Jiang, D. Meng, Q. Zhao, S. Shan, A.G. Hauptmann, Self-paced curriculum learning, in: AAAI Conf. on Artificial Intelligence, 2015.
[37] J. Gorski, F. Pfeuffer, K. Klamroth, Biconvex sets and optimization with biconvex functions: a survey and extensions, Math. Methods Oper. Res. 66 (2007) 373–407.

Barış Büyüktaş received the B.Sc. degree in Electrical & Electronics Engineering from Ozyegin University in 2018. He started his M.Sc. as a full scholarship student at Ozyegin University in 2018. He was funded as a research assistant by the Scientific and Technological Research Council of Turkey (TÜBİTAK) and he worked as a graduate teaching assistant at Ozyegin University. Then, he worked as a research assistant at Bogazici University between August 2020 and December 2021. He has been doing Ph.D. at TU Berlin since December 2021. His research interests include image processing, computer vision, machine learning, remote sensing.

Cigdem Eroglu Erdem received the B.S. and M.Sc. degrees in Electrical and Electronics Engineering from Bilkent University, Ankara, Turkey in 1995 and 1997, respectively, with high honors. She received the Ph.D. degree in Electrical and Electronics Engineering from Bogazici University, Istanbul, Turkey, in 2002. From September 2000 to June 2001, she was a visiting researcher in the Department of Electrical and Computer Engineering, University of Rochester, NY, USA. Between 2003-2004, she was a postdoctoral fellow at the Faculty of Electrical Engineering at Delft University of Technology, the Netherlands, where she was also affiliated with the video processing group at Philips Research Laboratories, Eindhoven. Between 2002-2009, she was the director of research at Momentum Digital Media Technologies Inc., a technology SME located in Istanbul. Between 2009-2016, she was a faculty member in the Department of Electrical and Electronics Engineering at Bahcesehir University, Istanbul, Turkey. Since Sep. 2016, she is a faculty member in the Department of Computer Engineering at Marmara University, Istanbul, Turkey. Dr. Erdem's current research interests are in the areas of digital image and video processing, computer vision and pattern recognition with applications to biometrics, federated learning, and physiological signal measurement.

Arif Tanju Erdem received a B.S. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey in 1986. He received an M.S. degree in 1988 and a Ph.D. degree in 1990, both in electrical engineering, from University of Rochester, Rochester, New York, USA. He was a member of the Research Laboratories of Eastman Kodak Company, Rochester, New York, USA, from 1990 to 1998, where he worked on various aspects of image and video processing. He has co-founded Momentum A.S., a 3D digital media technologies company in 1998 in Istanbul, Turkey, and served as its CTO until 2009. He joined Ozyegin University in 2009, where he has served as Dean of School of Engineering and Director of Graduate School of Engineering during 2014-2018. He has been serving as Vice Rector for Academic Affairs at Ozyegin University since 2018.