Pattern Recognition 107 (2020) 107529

Contents lists available at ScienceDirect

Pattern Recognition

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

Two-stage knowledge transfer framework for image classification

Jianhang Zhou[a], Shaoning Zeng[a,b], Bob Zhang[a,*]

[a] PAMI Research Group, Department of Computer and Information Science, University of Macau, Taipa, Macau, China
[b] School of Computer Science and Engineering, Huizhou University, Guangdong 516007, China

ARTICLE INFO

Article history:
Received 8 December 2019
Revised 10 April 2020
Accepted 1 July 2020
Available online 2 July 2020

Keywords:
Image classification
Teacher-student model
Two-stage classification
Sparse representation

ABSTRACT

The two-stage strategy has been widely used in image classification. However, these methods barely take the classification criteria of the first stage into consideration in the second prediction stage. In this paper, we propose a novel Two-Stage Representation method (TSR), and convert it to a Single-Teacher Single-Student (STSS) problem in our two-stage knowledge transfer framework for image classification. Specifically, the first stage classifier is formulated as the teacher, which holds the 'gate value' to supervise the student classifier in the second stage. To transfer knowledge from the teacher classifier, we seek the nearest neighbours of the test sample to generate a set of candidate target classes in the first stage. Then, a student classifier learns from the samples belonging to these candidate classes in the second stage. Under the supervision of the teacher classifier, the teacher approves the student only if it obtains a higher score than the 'gate value'. In actuality, the proposed framework generates a stronger classifier by staging two weaker classifiers in a novel way. The experiments on several databases show that our proposed framework is effective, which outperforms multiple popular classification methods.

1. Introduction

Image classification is widely regarded as one of the most crucial research areas in computer vision. Usually, algorithms in image classification belong to one-step classification [1–4]. While one-step classification might not be credibly adequate, two-stage image classification has been successful in many tasks, i.e., face recognition [5,6] and object recognition [4,7]. Real-world recognition tasks often contain a number of complex data with noises and conditions. Under this scenario, the discriminative capability of one single classifier is likely to fail in picking the best result. Thus, it is understandable to use two-stage classification to perform a coarse-to-fine classification. Since a subset of the classes to determine the final result is used, the complexity of the distribution in the data will be reduced [8]. Besides this, according to the probability estimation, it is more effortless to choose multiple candidate classes containing the right class than to determine if one single class is the correct answer directly. For example, the Top-1 accuracy of ImageNet is even not comparable to the Top-5 accuracy in any condition. Therefore, two-stage classification, in various implementations, is more promising than others.

Two-stage method is a long-lasting popular technique in image classification [9–12]. For example, Xu et al. proposed a two-phase test sample representation (TPTSR) method [9], which used all training samples to represent a test sample so as to exploit its neighbors in the first stage, before representing each test sample via these neighbors in the second stage. The WSRC (weighted sparse representation for classification) [10] exploited weights of the representation to seek a sparser representation, which is a form of the sparse representation method implemented as a two-stage method. The RAMUSA [13] algorithm performed multi-task learning by using a multi-stage method, which is similar to the idea of two-stage methods. Also, the idea of two-stage classification is effective when applied to other problems, such as coarse-to-fine framework [12], and face recognition [14]. Besides these two-stage methods, we can see that all of these two-stage classification methods only paid attention to the two classifiers themselves, rather than information and knowledge sharing during the two-stages of classification.

By contrast, the Teacher-Student model has a much more specific role according to the definitions of the two classifiers. Most importantly, it concentrates on the knowledge transferring between the teacher and the student classifier. Recently, You et al. proposed g-SVM to solve the single-teacher multi-students problem [15]. Similarly, You et al. solved the multi-teacher single-student problem by averaging the soft outputs from multiple teacher deep neural networks [16]. To develop other applications further, Zheng et al. implemented a student-teacher strategy using GAN, which achieved practical results [17]. As far as we can see, the teacher-student model is a particular case of the two-

[*] Corresponding author.
E-mail addresses: mb85405@um.edu.mo (J. Zhou), zsn@outlook.com (S. Zeng), bobzhang@um.edu.mo (B. Zhang).

https://doi.org/10.1016/j.patcog.2020.107529
0031-3203/© 2020 Elsevier Ltd. All rights reserved.

stage classification, especially for the single-teacher single-student model. Nevertheless, no such findings are available to solve the single-teacher single-student problem through a score-based prediction mechanism.

Both the Two-Stage and Teacher-Student classification methods have various implementations. Among them, linear methods show promising performances, i.e., Sparse Representation (SR) and Support Vector Machine (SVM). As is widely known to all, SVM proposed in 1995 [18] is a powerful classifier. Presently, there are several variants and application scenarios. The sparse representation classifier (SRC) [1] shows promising performances and robustness in image classification as well by taking advantage of the sparsity and locality [19]. Inherited from the idea of representation-based classification, the collaborative representation-based classifier (CRC) [20] is designed, which has a competitive recognition performance with less computational complexity. There are several works in the fusion of SRC and CRC [2,21,22], trying to keep a balance between sparsity and collaboration.Noticeably, the kernel collaborative representation (KCR) framework [2] proposed by Wang et al. unified different linear representation-based methods and achieved promising performances by sufficiently taking the non-linear information into account. However, to the best of our knowledge, no such work combines them via a two-stage classification framework.

In this paper, we propose a novel knowledge transfer framework for image classification using a two-stage representation method and formulate the two-stage classification problem to a single-teacher single-student problem. We name it **T**wo-**S**tage image classification supervised by a **S**ingle **T**eacher **S**ingle **S**tudent model (TS-STSS) and utilize sparse representation and collaborative representation to implement the algorithm. In the first stage of classification, the teacher classifier performs classification and seeks the nearest classes to the test sample, which is denoted as the ‘candidate classes’. Then, a ‘candidate set’ containing the training samples is organized in terms of the ‘candidate classes’. In the second stage, we represent the test sample and perform classification on the candidate set. Next, we use the single-teacher single-student model to make a decision based on the scores from the student classifier and ‘gate value’ from the teacher classifier. The knowledge transferring occurs mainly in two aspects: (1) the construction of the ‘candidate set’ and (2) scores comparison in the decision-making stage. The proposed framework is explainable, we provided justification based on theory of probability, geometry, and theory of neighborliness to support the reasonability of this framework. Generally speaking, our proposed two-stage representation method is supervised by a teacher classifier in image classification. The contributions of our work can be summarized in three aspects as follows:

1) We propose a novel two-stage representation knowledge transfer method for image classification.
2) We formulate the decision-making problem between results of both stages to a single-teacher single-student problem, and solve it using a score-based mechanism.
3) We implement TS-STSS via L1-minimization (sparse representation) and L2-minimization (collaborative representation).
4) We provide justification of the reasonability of the proposed framework via three aspects (theory of probability, Geometry, and theory of neighborliness).

The remainder of this paper will be organized as follows. In Section 2, we introduce the related work of two-stage classification, sparse representation and teacher-student models. In Section 3, we first describe the two-stage test representation method and the single-teacher single-student strategy, before proposing our classification framework. In Section 4, experimental and comparison results on image datasets will be demonstrated to show the effectiveness and performance of our proposed method. Section 5 concludes this paper.

2. Related work

The concept of representation learning [23] is extensively studied. The main idea of representation learning is to acquire more informative features using different representation methods. The sparse representation based classifier (SRC) [1] performs robust face recognition by implementing sparse coding on images from different classes C with respect to the test sample $y \in \Re^{n \times 1}$. The closest subspace to y is found by calculating the distance between the test sample and each class. Rather than use the sparse coding, the collaborative representation based classifier (CRC) [20] uses L2-norm to encode training samples, which shows a competitive performance using less computational time. According to the experiments it is effective as well in object recognition. Based on the sparse representation and collaborative representation, other representation methods have been proposed (e.g., TPTSR [9], KCR-l_2 [2], WSRC [10], etc.). All these methods try to seek an optimal representation portfolio in order to achieve better effectiveness as well as the efficiency. KCR-L_2 [2] proved that the non-linear information is critical to the image classification task like face recognition. However, improvements from SCRC [24] shows neither sparse representation (SR) nor collaborative representation (CR) is able to represent knowledge fully by itself [25].

Due to the power of dark knowledge distillation [26], knowledge from a well-trained model can be transferred to a simpler model. Therefore, knowledge from a teacher model is able to be transferred to the student model, which also leads to a higher recognition performance for the student model. This student-teacher learning paradigm shows a positive result in improving the classification results. The Single-Teacher Multi-Student (STMS) model [15] uses a well-trained model to ‘teach’ $N-1$ binary student classifiers in the multi-class classification scenario. The prediction of each student classifier is obtained as follows:

$$\hat{h} = \{+1, \quad if \Gamma(x;\theta_T) \geq \Upsilon(x;\theta_S) - 1, \quad otherwise \tag{1}$$

where $\hat{h}$ is the predicted result, x is the test sample, θ_T and θ_S are parameters of the teacher classifier and student classifier, and $\Gamma(\cdot)$ and $\Upsilon(\cdot)$ are decision functions of the teacher and student classifiers. The output of $\Upsilon(\cdot)$ is termed as the ‘gate value ’, which contains the confidence of the teacher. Since the information has been transferred as prior knowledge from the teacher classifier, the student classifiers can achieve competitive classification accuracy with less computation time.

This manuscript is an extension of our work [27], which has been accepted by the 30th British Machine Vision Conference (BMVC). In this paper, we present a more complete theoretical interpretation (in the view of theory of probability, geometry, and theory of neighborliness) compared to the original conference paper. What is more, we designed additional parameter experiments to enrich the analysis. We also tested our model using deep features from the test datasets extracted via ResNet [28]. Furthermore, one more dataset GT has been added in the evaluation.

3. The method

Our proposed method TS-STSS is a novel Two-Stage classification supervised by the Single-Teacher Single-Student model. In the first stage, a sparse representation based classifier via L1-minimization is learned as the teacher classifier, which computes the distances (or scores) to select a set of candidate classes and generates a ‘gate value’. Then, in the second stage, one single student classifier based on the faster L2-minimization, is trained using the samples of all candidate classes. Meanwhile, the scores of each class are generated. With the supervision via the ‘gate value’ of the teacher classifier, the student classifier in the second stage is capable of generating the final result. The detailed process of TS-STSS is depicted in the following sub-sections.

3.1. Two-stage (TS) representation

The representation procedure is performed in two stages. Specifically, in the first stage, all training samples are used to represent the test sample in a linear combination:

$$y = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_m x_m \tag{2}$$

where y is the test sample, and the θ_i is the coefficient of the ith instance in the linear combination, $x_i \in \mathbb{R}^{s\times 1}$is the s dimensional column vector of the ith instance, and m is the number of instances in the training set. For each class, we calculate its deviation with test sample by:

$$V_j = \left\| y - \sum_{m}^{i=1} \theta_{j,i} x_{j,i} \right\| \tag{3}$$

where V_j denotes the deviation of the jth class, $\theta_{j,i}$ and $x_{j,i}$ are the ith coefficient and ith sample of the jth class.

According to Eq. (3), we pick N nearest neighbors and insert their corresponding class label into the candidate classes set $C = \{c_1, c_2, \ldots, c_N\}$. We denote a sample set gathering samples from C as ‘candidate set’ $G = \{\gamma_{1,1}, \gamma_{1,2}, \ldots, \gamma_{1,k}, \ldots, \gamma_{N,1}, \gamma_{N,2}, \ldots, \gamma_{N,k}\}$.

In the second stage, each test sample will be represented by samples in G using a linear combination:

$$\tilde{y} = \lambda_1 \gamma_1 + \lambda_2 \gamma_2 + \ldots + \lambda_n g_n \tag{4}$$

where n denotes number of instances in G, and the $\hat{\gamma}_i$ is the coefficient of ith instance in the linear combination.

3.2. Single-teacher single-student (STSS) model

To take results from the first stage and second stage into consideration when classifying, we design a single-teacher single-student model and solve it using a score-based mechanism. We define the classifier in the first stage as the teacher classifier, and classifier in the second stage as the student classifier. Then, we calculate the ‘gate value’ of the teacher classifiers which will be used in the decision making procedure. Next, we calculate the scores corresponding to each class using the student classifier, and take the highest score as the final score of the student classifier. The ‘gate value’ indicates how confident the teacher classifier is, in regards to its classification output. For example, if the score of the student classifier is higher than the ‘gate value’, it means the student classifier learns better after absorbing the teacher‘s knowledge. In our strategy, we take the highest classification output value of the teacher classifier as the ‘gate value’, since it is the outputs with the most confidence. In this paper, We denote this solution as a single-teacher single-student model (STSS).

Firstly, we utilize a strong multi-class classifier T as a teacher classifier. Then, we apply a faster classifier as the student classifier ST. Next, we use the teacher classifier T to perform multi-class classification and obtain a score vector $S \in \mathbb{R}^{K\times 1}$ (K is the class number) for each class:

$$S_j = \delta\left(X_j\right) \tag{5}$$

where S_j is the j^{th} instance of score vector, X_j denotes the training set of the j^{th} class, and $\delta(\,\cdot\,)$ is a score evaluation function to evaluate the score of the sample vector. Classes with the highest score will be selected as the classification result by the teacher classifier, and its corresponding score will be taken as the ‘gate value’ g^*:

$$g^* = max(S) \tag{6}$$

The ‘gate value’ g^* can also be regarded as the confidence of the teacher (T). When the classification results of the student and the teacher are different, the final decision should be made between them. In this scenario, if the student (ST ’s) learned highest score is higher than the ‘gate value’ g^*, the ST ’s classification result will determine the final result. Otherwise, the classification result of T will be taken as the final result. Here we extend the binary decision function in Eq. (1) to predict the final result z as follows:

$$z = \left\{ \Psi(S_{ST}), \quad if S_{ST} > S_T \Psi(S_T), \quad otherwise \right. \tag{7}$$

where $\Psi(\,\cdot\,)$ denotes the function mapping of a score to its corresponding class label, and S_{ST}, S_T denote highest scores learned by a student ST and a teacher T, respectively.

It is obvious that the single-teacher single-student model uses the two-fold learning strategy, which motivates us to combine it with a two-stage representation method, as both of the two learning strategies require two steps to perform classification.

3.3. Two-stage knowledge transfer framework

Based on the two-stage representation method (TSR) and the single-teacher single-student model (STSS), we propose a two-stage image classification framework, named two-stage image classification supervised by a single-teacher single-student model (TS-STSS) for image classification. Fig. 1 depicts the general idea of our method intuitively. First of all, in the initial stage, we apply the sparse representation classifier (SRC) [1] as teacher classifier using all training samples for representation:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \|\theta\|_1 s.t. \|y - X\theta\|_2 < \varepsilon \tag{8}$$

where θ is the coefficient vector in the linear combination, and ε can be viewed as noise in y.

To transfer the knowledge from the teacher classifier, the candidate class set $C = \{C_1, C_2, \ldots C_m\}$ is built by appending the class associated with the M lowest deviations. Therefore we obtain the candidate set $G = \left\{X_{C_1}, X_{C_2}, \ldots X_{C_m}\right\}$. Next, we introduce the ‘gate value’ g^* to show the confidence of the teacher classifier. The larger g^* is, the more confident the teach is on its prediction. Since the probability of the prediction $p(z|x) \propto \left\|\bar{d} - d_{\min}\right\|_2$ we describe the ‘gate value’ calculation as follows:

$$H_T\left(y, \hat{\theta}_j, X_j\right) = \frac{1}{k} \sum_{i=1}^{k} \left\| y - \hat{\theta}_i X_i \right\|_2 - \left\| y - \hat{\theta}_j X_j \right\|_2 \tag{9}$$

$$g* = \max\left(H\left(y, \hat{\theta}_j, X_j\right)\right) \tag{10}$$

where $\Lambda_i = [0, \ldots, 0, \lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,k}, 0, \ldots, 0]$, X_i denotes the training set of the ith class, and n represents the number of all classes.

Following this, in the second stage, we apply the collaborative representation classifier (CRC) [20], who uses less computation time than SRC [20] as the student classifier using samples from the candidate set in the representation:

$$\hat{\lambda} = \operatorname{argmin}_{\lambda} \|y - G\lambda\|_2 s.t. \|\lambda\|_q \le \varepsilon \tag{11}$$

where λ is the coefficient vector in linear combination, and q can be 1 or 0. The solution of Eq. (11) is:

$$\hat{\lambda} = \left(G^T \cdot G + \lambda \cdot I\right)^{-1} G^T \tag{12}$$

To quantify the confidence of both the teacher and the student classifier, a score-based mechanism is designed. The score learned

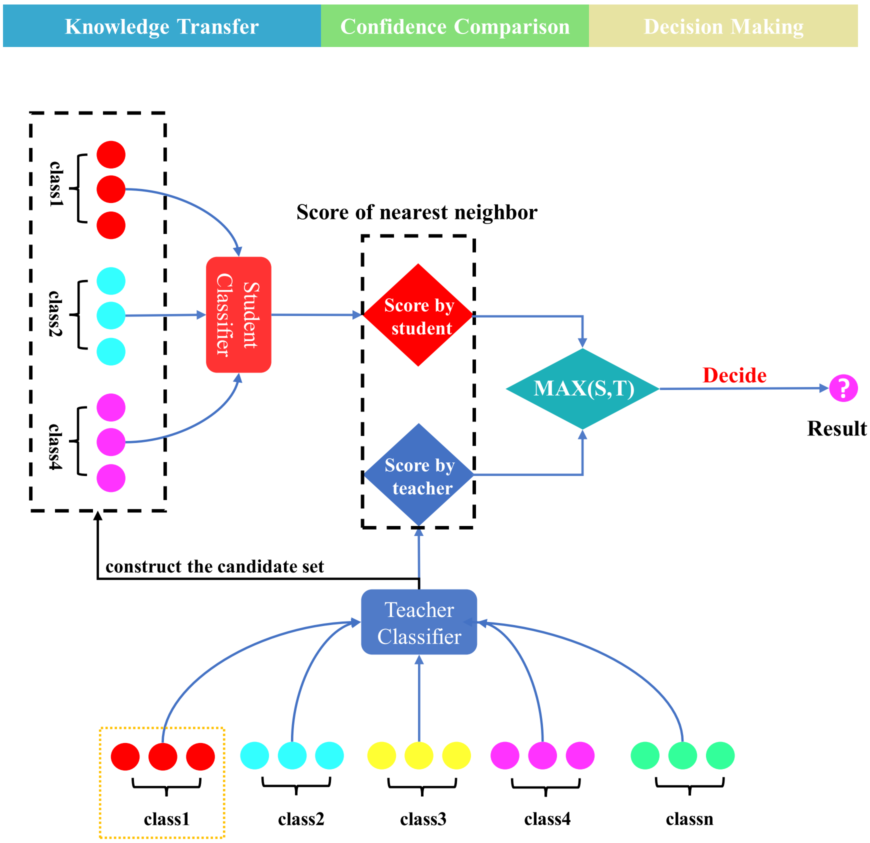


Fig. 1. Two-stage image classification supervised by a Single Teacher Single Student model.

by the student classifier can be described as follows:

$$H_s\left(y, \hat{\lambda}_j, X_j\right) = \frac{1}{k}\sum_{i=1}^{k} \|y - \lambda_i\gamma_i\|_2 - \left\|y - \lambda_j\gamma_j\right\|_2 \tag{13}$$

$$s^* = \max\left(H_s\left(y, \hat{\lambda}_j, X_j\right)\right) \tag{14}$$

where $\tilde{X}_i$ is the ith instance in G, and k is the number of instances in G.

In the decision-making stage, we utilize the teacher classifier to supervise the student classifier. To make the supervision, we define the similarity score $L(s^*, s^*) = s^* - g^*$ to evaluate predictions from the teacher classifier and student classifier. A positive value of $L(H_s, g^*)$ indicates the student classifier has surpassed the teacher classifier, while a negative value indicates the teacher classifier still has a higher confidence than the student classifier. If has the classification results of the student and teacher are different from each other, we use the similarity score $L(H_{s^*}, g^*)$ to reach the decision-making:

$$\tilde{z} = \left\{\Psi(s^*), \quad ifL(s^*, g^*) > 0\Psi(g^*), \quad \text{otherwise}\right. \tag{15}$$

where $\Psi(\ \cdot\ )$ is the function to select the class label of the input score. For example, $\Psi(s^*)$ is the class label of the student classifier score. Given the test sample y, the classification is made as follows:

$$\Psi(y, \mu, \Lambda) = \underset{j}{\operatorname{argmax}} \frac{1}{k}\sum_{i=1}^{k} \| y - \mu_i\Lambda_i\|_2 - \| y - \mu_j\Lambda_j\|_2 \tag{16}$$

where μ_i is the coefficient of the ith training sample. We summarize our proposed TS-STSS framework is summarized in Algorithm 1.

3.4. Justification of the TS-STSS model

In this section, we analyze the proposed TS-STSS model, trying to interpret how the single-teacher single-student model improves recognition. To transfer the knowledge of the teacher classifier to the student, we utilize the two-stage representation. Here we provide the probabilistic interpretation of the two-stage representation. The distance d_i between each class and each test sample

```
Algorithm 1: TS-STSS classification framework.
Input: Training set $X$, test sample $y$
Output: Identity $I$
In the first stage, use SRC as the teacher classifier according to (8) to obtain a candidate set $C$ and classification result $R_t$:
$R_t = \underset{i}{\operatorname{argmax}} \|y - \theta_i X_i\|_2^2$
According to (10), calculate the 'gate value' $g*$.
Perform the second phase classification using CRC as the student according to (13) and obtain the result of student $R_s$:
$R_s = \underset{i}{\operatorname{argmax}} \|y - \lambda_i\Gamma_i\|_2^2$
Calculate score $s*$ learned by the student according to (14).
if $R_t \neq R_s$ then
    $I = \begin{cases} \underset{j}{\operatorname{argmax}} \frac{1}{k}\sum_{i=1}^{k} \|y - \lambda_i\Gamma_i\|_2 - \left\|y - \lambda_j\Gamma_j\right\|_2, & ifs^* > g^* \\ \underset{j}{\operatorname{argmax}} \frac{1}{L}\sum_{i=1}^{L} \|y - \theta_i X_i\|_2 - \left\|y - \theta_j X_j\right\|_2, & otherwise \end{cases}$
else
    $I = R_t$
return I.
```

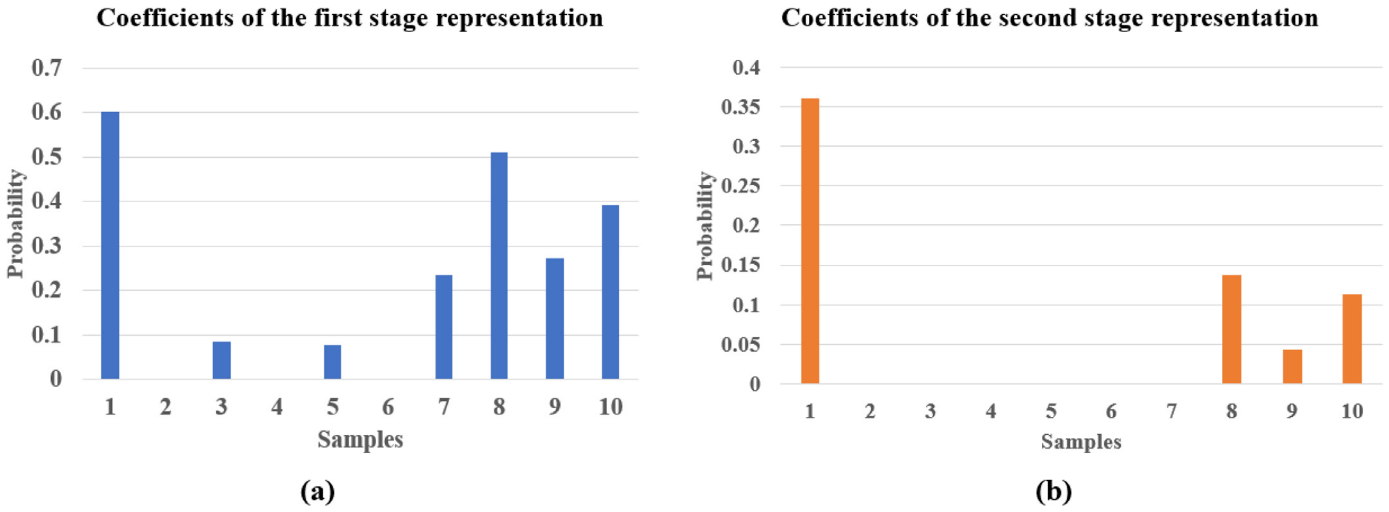


Fig. 2. Coefficients of the two-stage representation. (a) The first-stage representation, (b) The second-stage representation..

is able to determine the posterior probability $p(c_i|x) \propto d_{\max} - d_i$, which represents the probability of a test sample x belonging to c_i. The probability distribution of each class in the first representation and the second representation are displayed in the following Fig. 2. As shown in Fig. 2, the probability distribution in the second-stage representation is sparser than the first-stage representation, which makes it more discriminative for the correct class.

The knowledge transferring between the sparse representation and the collaborative representation are highly related, which is why our proposed framework works. In the first stage, the sparse representation makes the sample selection to generate a much sparser training set. In terms of [22,24], collaborative representation tends to achieve better results on the sparse training samples. Here we regard the collaborative representation based on sparse samples as a knowledge transfer procedure. In this scenario, the knowledge is composed of the selection from sparse representation and is transferred to the collaborative representation, which is a more efficient and competitive image representation method. In other words, the collaborative representation in the second stage makes the classification based on the selection of sparse representation, which enforces the classifier to take full advantage of the knowledge transferred from the sparse representation.

According to Fig. 1, the final decision is determined by comparing the score from the student classifier H_T and teacher classifier H_S, which is the significance of the single-teacher single-student model. The 'gate value' and score of the teacher and student classifiers are generated in terms of the classification outputs, which are regarded as the confidence of the teacher and student classifier. In fact, this improves the robustness of the model, which helps to improve the recognition rate to some extend. During classification, higher confidence determines the final result. There are two situations: (1) $H_T \leq H_S$, and and (2) $H_T \geq H_S$. In situation (1), the confidence of the teacher is higher than the confidence of the student, meaning the final result will follow the teacher's result. In situation (2), where the confidence from the student is higher than the teacher's, which means the student classifier learns a more discriminative representation and "corrects" the mistake of the teacher. We can also view it as the student learning 'more' under the supervision of the teacher. In this way, the robustness of the model is improved since the student classifier has the ability to correct the mistake from the teacher classifier.

The geometric illustration of TS-STSS is described in Fig. 3 in one subspace. Fig. 3(a) shows the geometric illustration of sparse representation, which is taken as the first-stage representation in the TS-STSS model. $V = y - \bar{y}$ represents the approximation error between the test sample y and its projection $\bar{y}$ in this subspace. V_i represents the deviation of one certain class. Correspondingly, τ_i is the i^{th} component in the linear combination to represent the sample $\bar{y}$ in this subspace. In other words, τ_i is considered as the class-specific representation of sparse representation and collaborative representation. The circle in the center describes a set of possible solutions of the class-specific representation, which means there may exist other representation T_j whose deviation $V_j = V_i$. As different representations have the same deviation, the misprediction will occur. Compared with Fig. 3(a1), which illustrates the collaboration representation (the second stage), the sparse representation uses fewer components to seek a minimal deviation within all classes of the samples. Although collaborative representation requires relatively more components, it has less computational complexity (comparatively speaking) via the closed-form solution of the coefficients [20]. Meanwhile, after the selection from the first stage (sparse representation), the second stage (collaborative representation) tends to achieve the result using fewer components, which makes it easier to achieve the correct representation. This explains why two-stage representation contributes to the classification. In the decision-making stage, the comparison is made between the "gate value" of the sparse representation and the score of the collaborative representation, which is geometrically illustrated in Fig. 3(b) and (b1), respectively. Rather than using V_i to determine the classification result, our proposed method uses the comparison result of two values, to further avoid the minimal deviation with incorrect representation, which improves the robustness. It is clear that the enhanced robustness helps the effectiveness of the model [24].

By introducing the properties of the L_1-optimizer and the theory of neighborliness, we can use the concept of sparsity and locality to explain our knowledge transfer framework. According to [19], the role of the L_1-optimizer in the pattern classification is to achieve both sparsity and closeness, which is meaningful to the classification purpose. Specifically, the sparsity is able to depict the local characterization, and the closeness taking global similarity into consideration. In other words, the L_1-optimizer compromises of sparsity with closeness via the L_1-optimizer, which makes a good balance between them. In our framework, we utilize the sparse representation as the teacher classifier in order to simultaneously exploit the local characterization and global similarity. In addition, the sparse representation as a teacher, generating the possible homo-class candidates for the student classifier, is trans-

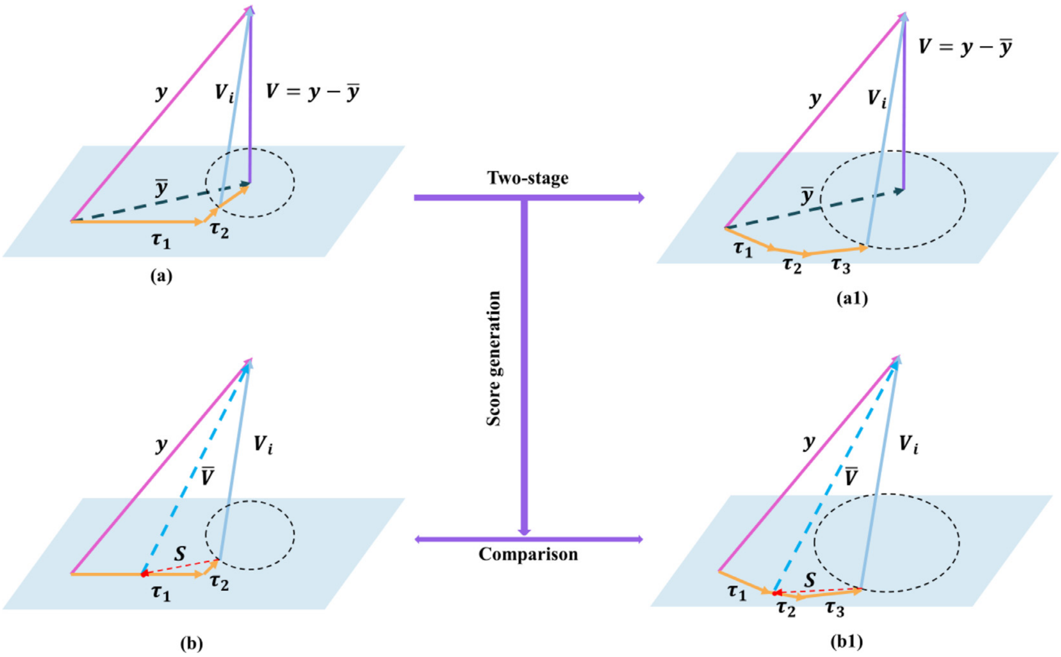


Fig. 3. Geometric illustration of the TS-STSS model.

ferring the local characterization knowledge and global similarity knowledge to the student classifier. As for the collaboration representation as the student classifier, it absorbs knowledge from the teacher and works out the scores of each class by rectifying the possibility of the test samples belonging to each hyperplane spanned by these support training samples. Here, the knowledge transferring is performed in a two-fold aspect: (1) homo-class candidates generated by the teacher classifier are sent to the student classifier in order to transfer local characterization knowledge (refer to the local knowledge); (2) the teacher's scores are compared to the student's scores in order to transfer global similarity knowledge (refer to the global knowledge). Fig. 4 shows these two folds.

4. Experiments

To verify the effectiveness of our proposed method in different classification tasks, we performed several experiments on six datasets. Specifically, we tested our method on the GT, FEI, MUCT, and YouTube facial datasets, the COIL-100 object dataset and the MNIST handwriting dataset, respectively. We used 10 classifiers for comparisons (SRC [1], CRC [20], K-SVD [25], SVM [18], KNN [29], TPSTR [9], NCRC [30], ProCRC [3], S*CRC [24], STMS [15]). Among them, STMS, ProCRC, S*CRC, and NCRC are newly proposed classifiers. The recognition rate was evaluated using the hand-out method, and we set different configurations for the different datasets (including the parameter λ of CRC and the number of candidate classes). On GT, FEI, MUCT, and COIL-100, we increased the number of training samples in each class for every iteration and took the remaining samples of each class as test sample. Then, we calculated the average accuracy and maximum accuracy respectively. On MNIST and YouTube, we directly used the pre-divided training set and testing set. Later on, we test our method using deep representation generated from popular pre-trained deep learning models (e.g., VGG16[31], ResNet[28]). The experiments were executed using MATLAB R2018b on a PC with one 3.40GHz CPU and 16.0 GB RAMs.

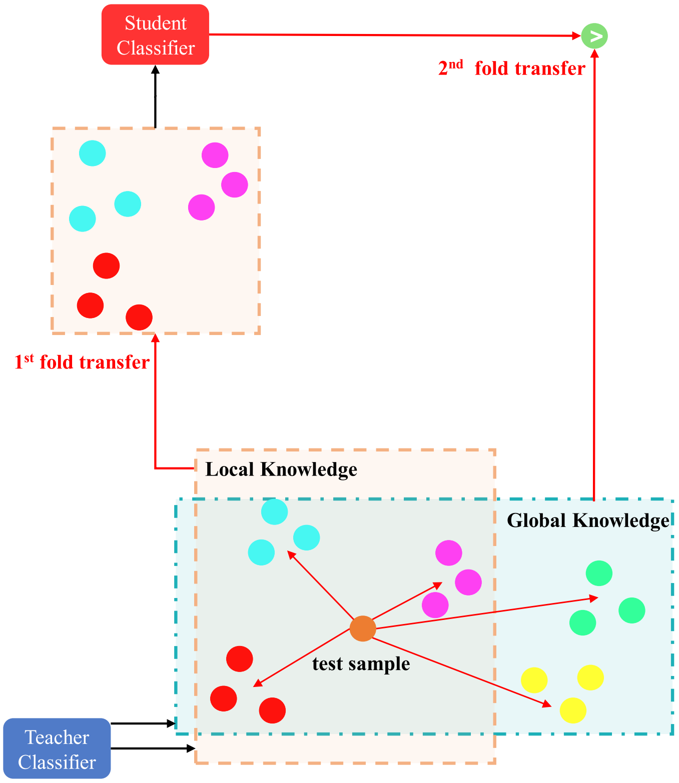


Fig. 4. Two folds of the knowledge transfer.

4.1. Dataset description

As shown in Fig. 5, we used six image datasets in total to evaluate our proposed method, including GT [32], FEI [33], MUCT [34], YouTubeFace [35],COIL-100 [36], and MNIST handwriting digits [37], respectively. The details of each dataset are summarized in Table 1.

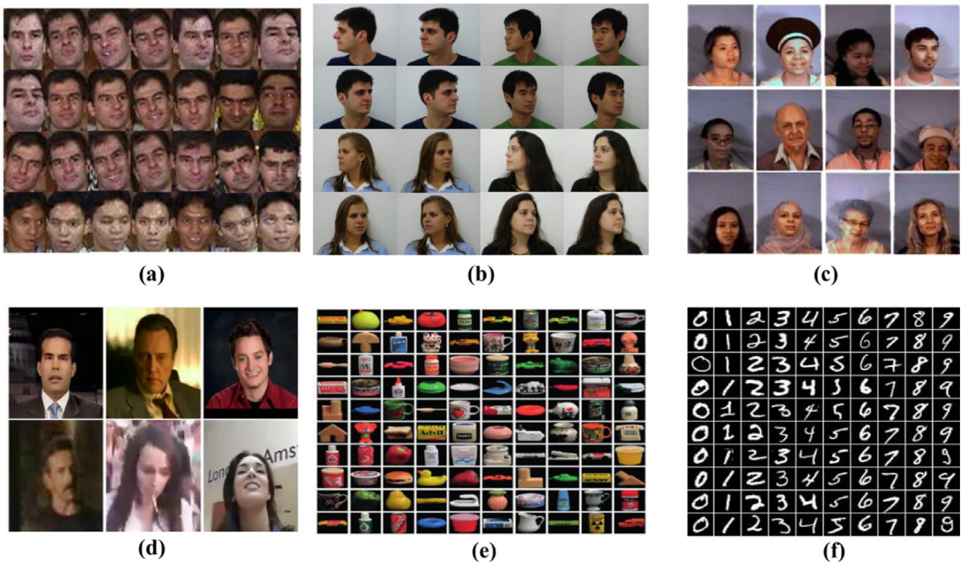


Fig. 5. Image Datasets: (a) GT, (b) FEI, (c) MUCT, (d) YouTubeFace, (e) COIL-100, and (f) MNIST.

Table 1
Configurations of the image databases in the experiments.

Database	Classes	Samples	Size
GT	50	750	40 × 30
FEI	200	2800	24 × 96
MUCT	276	3755	640 × 480
YouTubeFace	1283	128,300	32 × 32
COIL-100	100	7200	32 × 32
MNIST	10	70,000	28 × 28

There are four facial datasets used in the experiments. The GT face dataset contains 750 images from 50 people. The resolution of each image is 30 × 40 pixels. The FEI dataset is a face dataset containing 2800 images from 200 people (14 images per person). The resolution of each original image is 640 × 480 pixels. The MUCT face database contains 3755 face images from 276 people. The resolution of each image is 640 × 480 pixels. All images were captured by a CCD camera and stored in 24-bit RGB format. In our experiments, we used the 24 × 96 pixels version. GT, FEI and MUCT are relatively small, therefore, we wish to demonstrate our proposed TS-STSS works well in small datasets.

The last face dataset is YouTubeFace, which is a large-scale dataset. It is designed for studying unconstrained face recognition problems in video. In this dataset, 3425 videos from 1595 different people were collected from the YouTube website and labeled according to the LFW image collection method [38]. The resolution of each image is 32 × 32 pixels. In our experiments, we chose 1283 classes with over 100 samples, and randomly selected 100 samples per class, 128,300 samples in total. Hence, we can evaluate the performance of TS-STSS for large-scale recognition.

The COIL-100 dataset (Columbia Object Image Library) is the object dataset, which collected 100 objects and contains 7200 images in total with a black background. Each object has 72 images captured in different degrees by a CCD color camera. The resolution of each image is 32 × 32 pixels.

The MNIST hand-writing digits dataset is a hand-writing digits database built by LeCun et al. [39], containing 60,000 examples for a training set and 10,000 examples for a test set. The resolution of each image is 20 × 20 pixels. Each image was acquired from the center of 28 × 28 pixels from the original image and processed by a normalization algorithm.

4.2. Face recognition

4.2.1. Choosing the parameter of CRC

We first choose the parameter λ of CRC which is utilized in the second stage. Fig. 6 depicts changes in the recognition performance with the increasing λ value (from 0.001 to 0.9). As shown in Fig. 6, the TS-STSS model obtained the highest recognition rate: 78% (GT), 90.17% (FEI), 91.06% (MUCT), and 92.16% (YouTubeFace) when $\lambda = 0.001$. Therefore, we select $\lambda = 0.001$ as the optimal parameter in these four face datasets.

4.2.2. Choosing the number of candidate classes

After choosing the optimal parameter λ of each dataset, we test the proposed TS-STSS model using varying number of classes in the second stage, to find out the optimal number of candidate classes. The performances in face recognition are shown in Fig. 7. The optimal number of candidate classes are $n = 10$ (GT), $n = 30$ (FEI), $n = 20$ (MUCT), and $n = 50$ (Youtube), whose recognition rates are 82%, 91.83%, 91.66% and 92.16%, respectively.

4.2.3. Recognition performance

We used the MUCT, FEI and YouTubeFace datasets to perform experiments on face recognition. From the results shown in Table 2, we notice that TS-STSS achieves the highest accuracy on the MUCT dataset, which is 91.66% (max accuracy), indicating its effectiveness on face recognition. Besides this, TS-STSS outperforms most of the classifiers in face recognition both in maximum and average accuracy (MUCT: 91.66%, 80.54%; FEI: 91.83%, 62.45%). As for the YouTubeFace dataset, the proposed method achieved a recognition rate of 92.16%, which has an improvement of 1.12% over the best result from [15]. Compared with S*CRC, the proposed TS-STSS shows 23.41% of average accuracy improvement on the MUCT dataset and 4.65% average accuracy improvement on the FEI dataset. Fig. 8(a) and (b) show the accuracies generated by SRC, CRC and TS-STSS on the MUCT and FEI datasets respectively. In Fig. 8(a), it is obvious that TS-STSS produced a better recognition performance no matter the training samples. In Fig. 8(b), the accuracy of TS-STSS and SRC is quite competitive in beginning, while

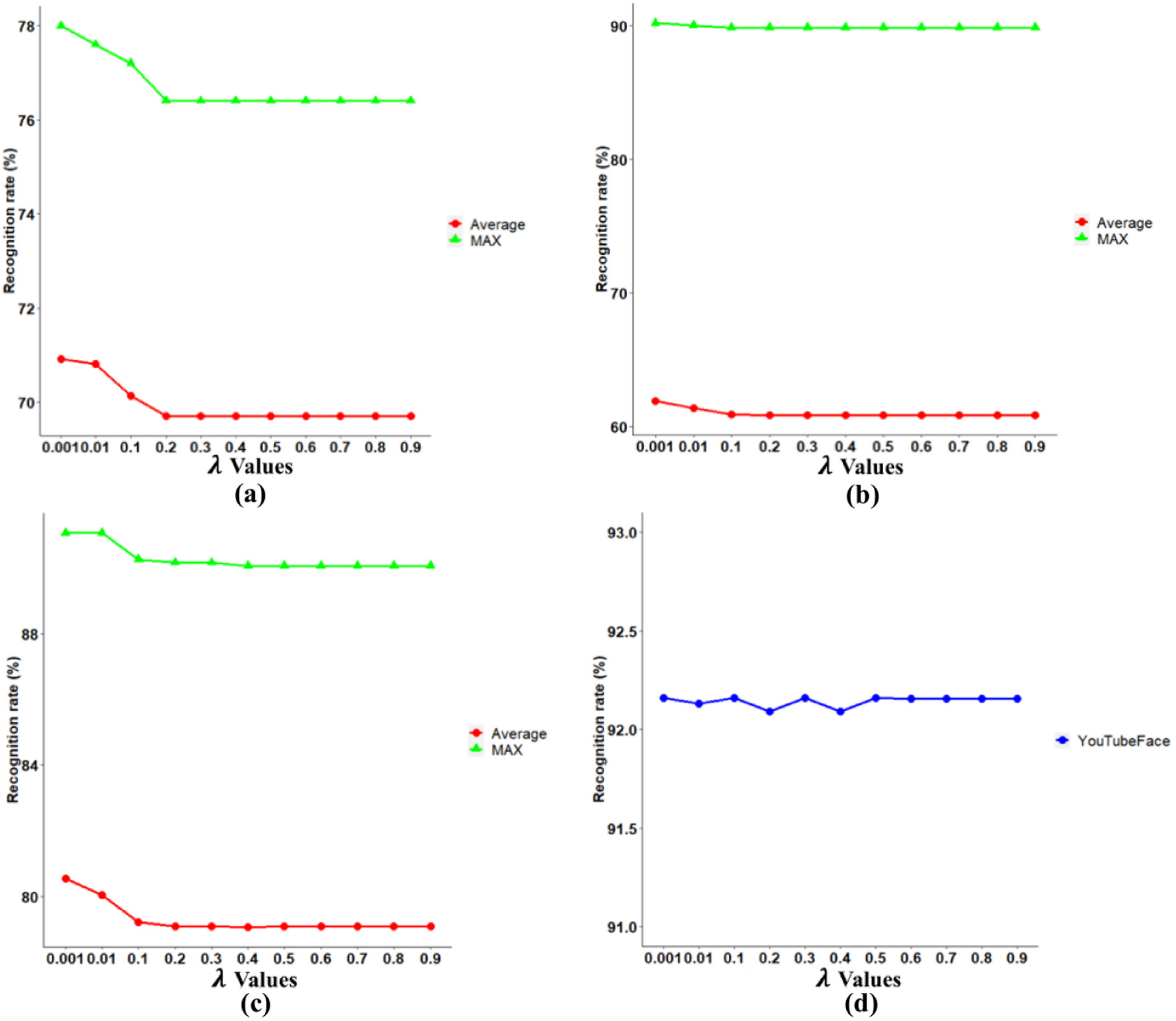


Fig. 6. Recognition performance on face datasets using different λ values. The line in green is the maximum recognition rate, the line in red is the average recognition rate. (a) GT dataset, (b) FEI dataset, (c) MUCT dataset, and (d) YouTubeFace dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Recognition rate comparisons to popular and state-of-the-art classifiers. (Note: Unit of data is %, and bold figures indicate the best results.).

Method	Object		Handwriting	Face				
	COIL-100		MNIST	MUCT		FEI		YouTubeFace
	MAX	AVG	ACC	MAX	AVG	MAX	AVG	ACC
SRC	76.97	74.48	95.96	90.15	79.01	89.90	60.85	84.91
CRC	70.84	65.23	82.83	85.83	76.87	74.75	51.69	72.12
K-SVD	61.58	58.62	82.87	77.09	70.26	65.88	40.95	53.26
SVM	58.94	53.54	**98.60**	28.44	24.78	57.13	40.95	53.26
KNN	74.56	70.02	95.00	67.94	57.97	69.63	48.55	90.00
TPSTR	76.89	72.41	87.27	88.54	66.48	89.17	61.88	78.04
STMS	77.19	72.43	95.59	89.64	75.25	89.66	61.32	91.04
ProCRC	61.50	60.00	94.30	74.32	55.13	71.75	47.95	84.74
S*CRC	79.54	73.84	86.97	76.85	57.13	80.75	57.8	86.32
NCRC	69.45	65.96	61.53	77.78	57.63	85.33	53.64	87.31
TS-STSS	**81.32**	**75.88**	97.41	**91.66**	**80.54**	**91.83**	**62.45**	**92.16**

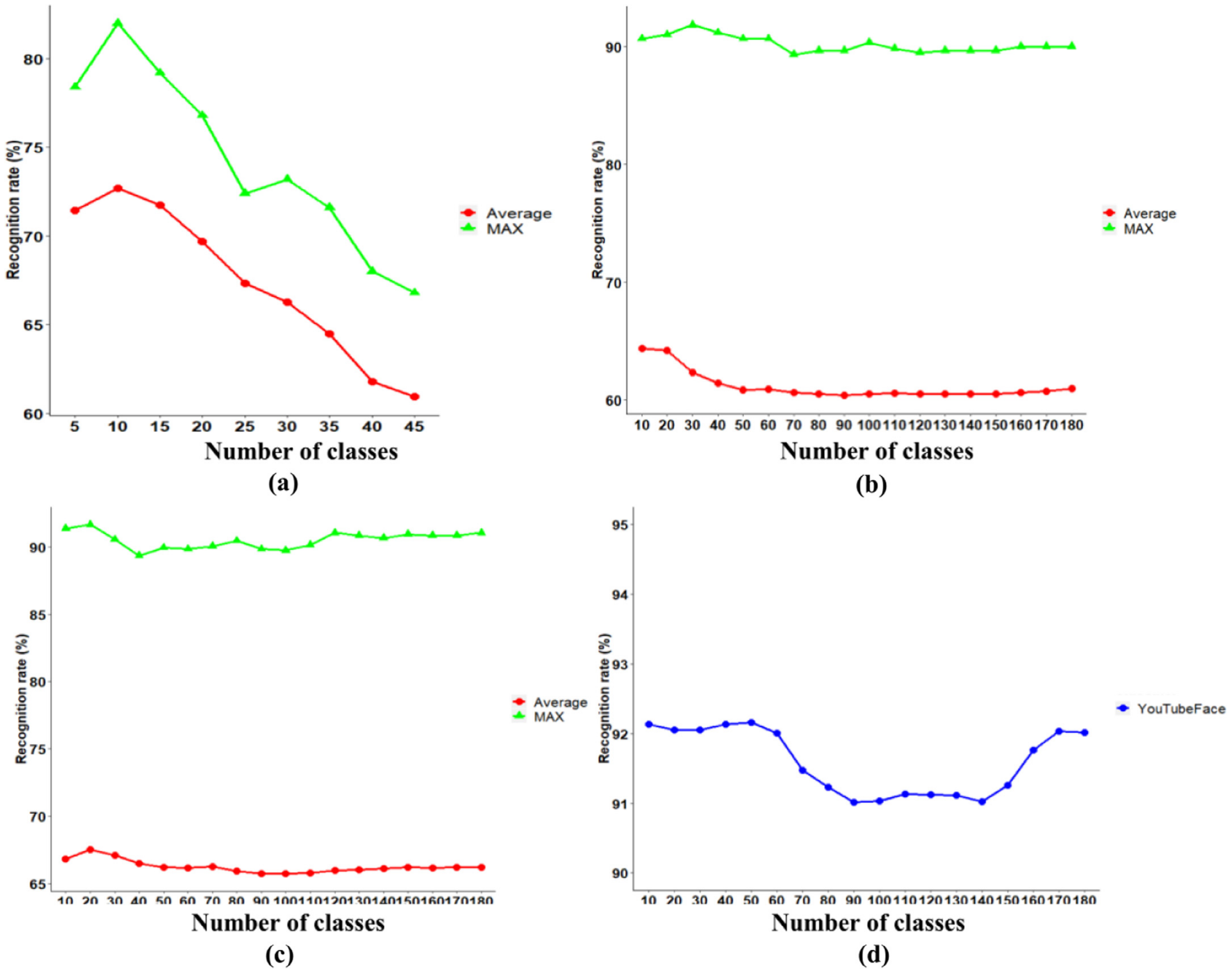


Fig. 7. Recognition performance on face datasets using difference number of classes. The line in green is the maximum recognition rate, the line in red is the average recognition rate. (a) GT dataset, (b) FEI dataset, (c) MUCT dataset, and (d) YouTubeFace dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

TS-STSS surpasses SRC after using seven training samples. Compared to ProCRC, S*CRC, and NCRC, the proposed TS-STSS has a better face recognition performance according to Table 2.

4.3. *Object recognition*

4.3.1. *Choosing the parameter of CRC*

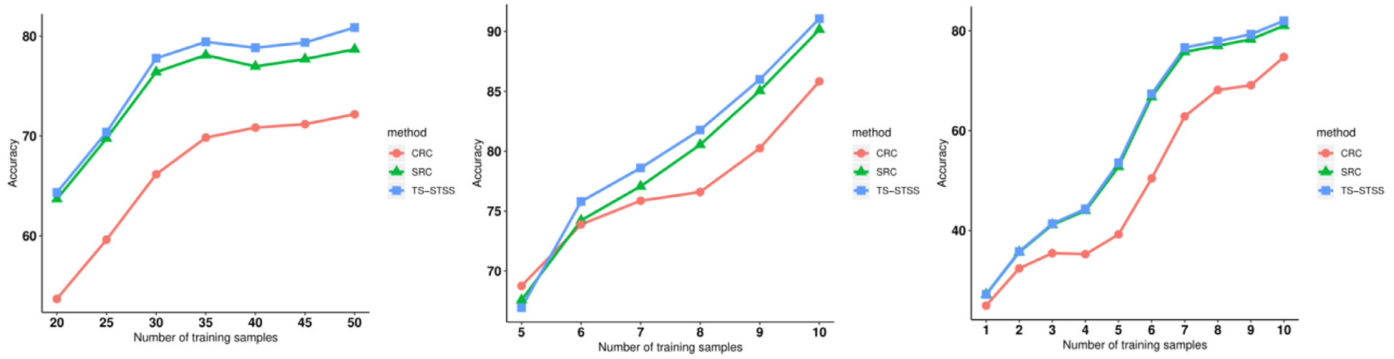
We choose the parameter λ of CRC which is utilized in the second stage. Fig. 9 depicts changes to the recognition performance with the increasing λ value (from 0.001 to 0.9). The x-axis represents different values of λ and the y-axis represents the recognition rate. As shown in Fig. 9, the TS-STSS model achieved the highest recognition rate (81.32%) when $\lambda = 0.001$. Therefore, we select $\lambda = 0.001$ as the optimal parameter in object recognition.

4.3.2. *Choosing the number of candidate classes*

After choosing the optimal parameter λ of CRC in the COIL-100 dataset, we test the proposed TS-STSS model using varying number of classes to find out the optimal number of candidate classes. The performance in object recognition is shown in Fig. 10. The optimal number of candidate classes of CRC in the COIL-100 dataset is $n = 10$, where the recognition rates is 81.32%.

4.3.3. *Recognition performance*

We used the COIL-100 dataset to perform experiments on object recognition. As can be seen in Table 2, the highest accuracy achieved by TS-STSS is 81.32%, which is higher than SRC (76.97%) and CRC (70.84%), respectively. Noticeably, the highest improvement for average accuracy compared with SRC and CRC is 4.35% and 10.48% correspondingly, showing that the proposed method has a significant effect on object recognition. Fig. 8(c) shows the accuracy generated by SRC, CRC and TS-STSS. We can observe directly that the gap between TS-STSS and CRC is larger when the size of the training set is increasing. As more training samples are used for representation, the difference between each class becomes larger. Therefore, the teacher classifier is able to supervise the student more accurately. Compared to the ProCRC (60% on average), S*CRC (73.84% on average) and NCRC (65.96% on average), our proposed TS-STSS (81.32% on average) shows a better performance than ProCRC, S*CRC and NCRC.



(a) **(b)** **(c)**

Fig. 8. Variation of the recognition rate on face datasets with increase of the number of training samples (a) MUCT, (b) FEI, (c) COIL-100.

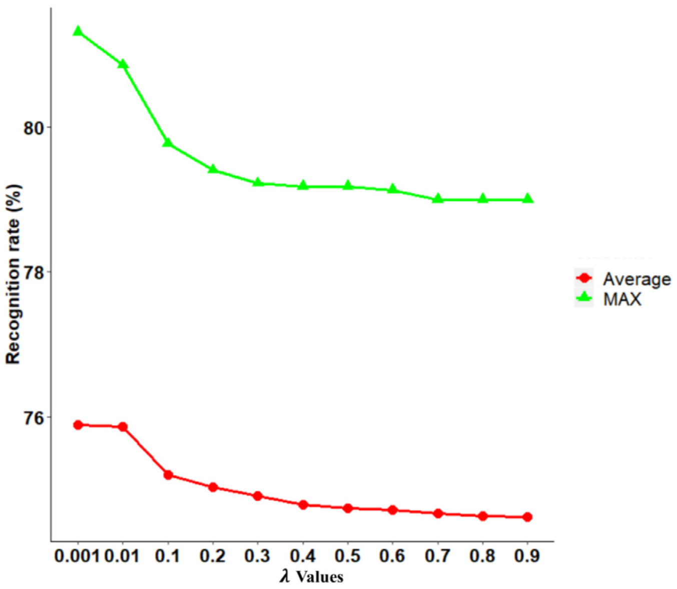


Fig. 9. Recognition performance on the object dataset applying different λ values. The line in green is the maximum recognition rate and the line in red is the average recognition rate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

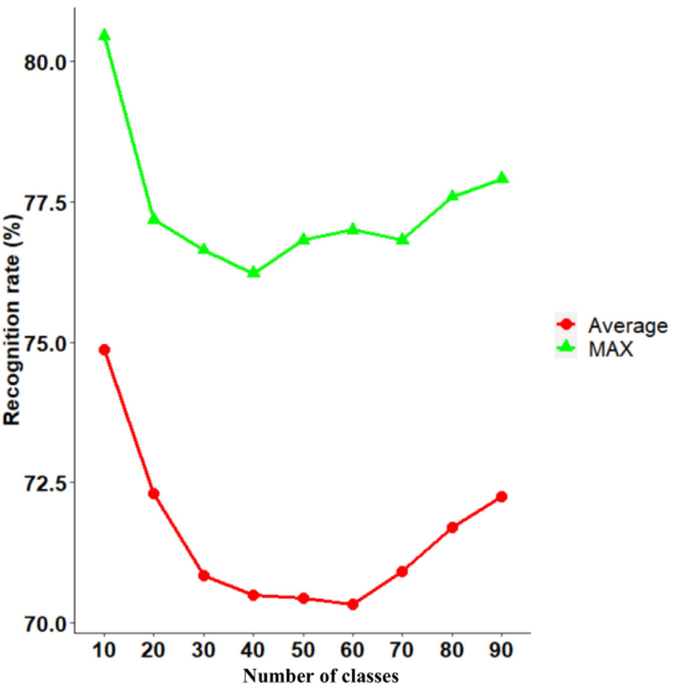


Fig. 10. Recognition performance on the object dataset with different number of classes. The line in green is the maximum recognition rate and the line in red is the average recognition rate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. *Hand-writing recognition*

4.4.1. *Choosing the parameter of CRC*

We choose the parameter λ of CRC which is utilized in the second stage. Fig. 11 depicts changes to the recognition performance with the increasing λ value (from 0.001 to 0.9). As shown in Fig. 11, the TS-STSS model achieved the highest recognition rate (97.10%) when $\lambda = 0.7$. Hence, we select $\lambda = 0.7$ as the optimal parameter in hand-writing recognition.

4.4.2. *Choosing the number of candidate classes*

After choosing the optimal parameter λ of the MNIST dataset, we test the proposed TS-STSS model using varying number of classes to find out the optimal number of candidate classes. The performance in hand-writing recognition is showing in Fig. 12. The optimal numbers of candidate classes in the MNIST dataset is $n = 2$, where the recognition rate is 97.41%.

4.4.3. *Recognition performance*

We applied the TS-STSS to perform experiments on hand-writing digits recognition, where the results are displayed in Table 2. Here, the accuracy of SRC and CRC is 95.96% and 82.83% respectively. We can see that TS-STSS improves the recognition rate with the highest accuracy of 97.41%, which is higher than SRC and CRC by 1.45% and 14.58%, respectively. Moreover, the proposed TS-STSS also achieves a better performance than TPSTR and STMS by 10.14% and 1.82%, correspondingly. Compared with other popular classifiers like K-SVD (82.87%), KNN (95.00%), and SVM (98.60%) TS-STSS is competitive as well. Our proposed TS-STSS is still higher than the recently proposed classifiers of ProCRC, S*CRC, and NCRC with a 3.11%, 10.44%, and 35.88% of improvement in average accuracy, respectively.

4.5. *Evaluation with the deep feature*

To achieve better recognition rates based on the TS-STSS model, we use the deep learning-based features of the different datasets extracted by ResNet [28]. The highest recognition rates achieved by TS-STSS with raw images and deep learning-based features are given in Table 3. We performed the evaluation in object recogni-

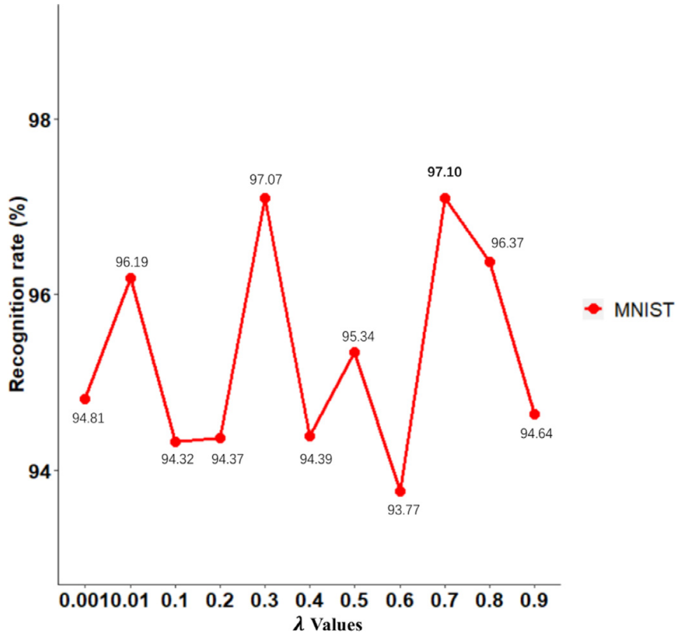


Fig. 11. Recognition performance on the hand-writing digits dataset with different number of classes.

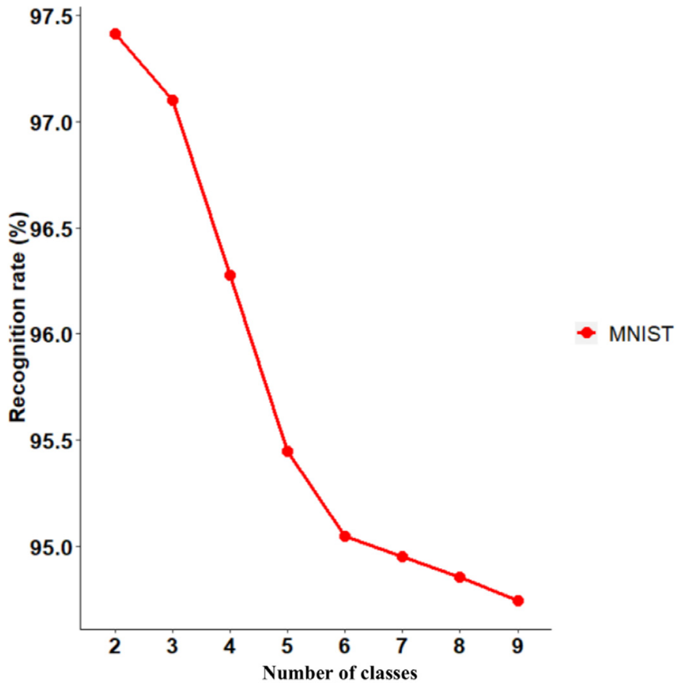


Fig. 12. Recognition performance on the hand-writing digit dataset of number employing different classes.

tion (COIL-100), hand-writing recognition (MNIST) and face recognition (GT) respectively. From the overall view of Table 3, the recognition rates in the different tasks become better by combining with deep learning-based features. By fusing deep learning-based features, our proposed classifier achieved 98.41% on object recognition, 98.15% on hand-writing digits recognition and 86% on face recognition. The improvements made in different tasks are 17.09% (object), 0.74% (hand-writing) and 4% (Face), respectively.

Table 3
Recognition rate of the TS-STSS model using deep features.

COIL100		MNIST		GT	
RAW	DEEP	RAW	DEEP	RAW	DEEP
81.32	**98.41**	97.41	**98.15**	82.00	**86.00**

4.6. Discussion

The experiments cover a variety of tasks and conditions in image classification, including face, object and hand-writing digits recognition, as well as different dataset sizes. The promising performance of our proposed method has been well proved. In addition, we can obtain the following inferences.

1) Enlarging the training set is helpful in the TS-STSS, since the implementation is based on sparse representation. As shown in Fig. 8, the FEI, MUCT, and COIL-100 datasets have no pre-split training and test sets, hence different training samples were utilized to train the classifiers. The accuracy keeps increasing as the training set becomes larger.
2) The supervision of the teacher classifier is the key to improve the Two-Stage classification. As shown in Table 2, TS-STSS outperforms both TPSTR [9] and STMS [15]. This confirms our expectation that applying a consistent scoring criteria in two stages is beneficial to classification, while the conventional two-stage classifiers lack this.
3) Compared to other linear methods, TS-STSS is very promising. Table 2 shows the recognition results of other popular linear classifiers, where TS-STSS consistently produces the highest accuracy in most cases. Besides, better performances are achieved in different circumstances with the help of deep features, especially in object recognition. Furthermore, TS-STSS introduces only one additional parameter, the number of candidate classes *k*, which can be set to the optimal value depends on the different datasets.
4) Compared with SRC and CRC, TS-STSS outperforms both of them on all experiments, indicating that our proposed framework successfully integrates two weaker classifiers to form a stronger classifier in image classification.
5) Compared with the state-of-the-art methods, including TPSTR, STMS, ProCRC, S*CRC, and NCRC, TS-STSS shows a better performance on different classification tasks. Classifiers like TPSTR, ProCRC, S*CRC, and NCRC are extensions of the SRC or CRC, but they only show a better performance in some specific areas. For example, the S*CRC performs better than SRC on object recognition, while it achieves a poor performance on face recognition and handwriting recognition. NCRC performs better than CRC, while it obtains a worse recognition rate than SRC in these tasks. In contrast, TS-STSS surpasses the original SRC, CRC, NCRC, and ProCRC in all tasks. TS-STSS also shows a better recognition rate than the STMS model especially in hand-writing digits recognition, which is a teacher-student based classifier. These indicate that the proposed TS-STSS method is superiority in image classification compared with other state-of-the-art methods.

5. Conclusion

In this paper, we propose a novel two stage knowledge transfer framework for image classification named Two-Stage image classification supervised by a Single-Teacher Single-Student model (TS-STSS). In the first stage, a candidate set of classes are chosen and the classification score vector is built using the L1-based SRC classifier (Teacher). Then, the L2-based CRC classifier (Student) represents the test sample using the candidate set in the second stage, under the supervision of the teacher classifier. In order to make a more precise score, we formulate it to the Single-Teacher Single-Student (STSS) problem. This image classification framework is able to combine two different weaker classifiers to form a stronger classifier. The reasonability of the knowledge transfer framework can be justified via theory of probability, geometry and theory of neighborliness. The experiments on six popular image datasets proved its effectiveness and promising capability in image classification, outperforming many other popular methods. Although it has not achieved the best performance in handwriting recognition, it still showed better results compared to all sparse representation-based methods in the experiments. Futhermore, with the help of deep features, the recognition rates of the proposed method on all three types of datasets have been improved to much higher values.

Currently, we have only implemented the proposed framework with linear sparse methods, SRC and CRC. It will be interesting to consider using the non-linear models as the teacher and student classifiers, i.e., KCR-l_2 [2]. Other linear classifiers, i.e., dictionary learning [40], SVM, KNN, etc. are potentially good choices as well. Despite our method utilizing image data, it should be helpful using deep features [22] in real-world applications. We will continue to observe and explore these options in the future.

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.

CRediT authorship contribution statement

Jianhang Zhou: Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Shaoning Zeng:** Conceptualization. **Bob Zhang:** Supervision.

Acknowledgments

The authors are grateful to the anonymous reviewers for all their constructive comments and suggestions. This work was partially supported by the University of Macau (File no. MYRG2019-00006-FST), partially supported by the Educational Science Planning Project of Guangdong (no. 2018GXJK183), and partially supported by the Fund of Science and Technology Plan Project in Huizhou (no. 2020SD0402030).

References

[1] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.
[2] D. Wang, H. Lu, M.-H. Yang, Kernel collaborative face recognition, Pattern Recognit. 48 (2015) 3025–3037.
[3] S. Cai, L. Zhang, W. Zuo, X. Feng, A probabilistic collaborative representation based approach for pattern classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2950–2959.
[4] H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, Z. Liu, Deep predictive coding network for object recognition, in: Proceedings of the 35th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, 80, PMLR, 2018, pp. 5266–5275.
[5] X. Dong, H. Zhang, J. Sun, W. Wan, A two-stage learning approach to face recognition, J. Vis. Commun. Image Represent 43 (2017) 21–29.
[6] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, X. Zhou, A two-phase weighted collaborative representation for 3D partial face recognition with single sample, Pattern Recognit. 52 (2016) 218–237.
[7] H. Liu, J. Qin, F. Sun, D. Guo, Extreme kernel sparse learning for tactile object recognition, IEEE Trans. Cybern. 47 (2017) 4509–4520.
[8] J. Li, J. Cao, K. Lu, Improve the two-phase test samples representation method for palmprint recognition, Optik 124 (24) (2013) 6651–6656.
[9] Y. Xu, D. Zhang, J. Yang, J.-Y. Yang, A two-phase test sample sparse representation method for use with face recognition, IEEE Trans. Circuits Syst. Video Technol. 21 (9) (2011) 1255–1262.
[10] Z. Fan, M. Ni, Q. Zhu, E. Liu, Weighted sparse representation for face recognition, Neurocomputing 151 (2015) 304–309.
[11] P. Zhang, Z.-Q. Zhao, J. Gao, X. dong Wu, Image set classification based on cooperative sparse representation, Pattern Recognit. 63 (2017) 206–217.
[12] S. Zeng, X. Yang, J. Gou, Using kernel sparse representation to perform coarse-to-fine recognition of face images, Optik 140 (2017) 528–535.
[13] L. Han, Y. Zhang, Multi-stage multi-task learning with reduced rank, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 1638–1644.
[14] Z. Liu, J. Pu, M. Xu, Y. Qiu, Face recognition via weighted two phase test sample sparse representation, Neural Process. Lett. 41 (1) (2015) 43–53.
[15] S. You, C. Xu, C. Xu, D. Tao, Learning with single-teacher multi-student, in: Thirty-Second AAAI Conference on Artificial Intelligence, AAAI Press, 2018, pp. 4390–4397.
[16] S. You, C. Xu, C. Xu, D. Tao, Learning from multiple teacher networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 1285–1294.
[17] Z. Xu, Y.-C. Hsu, J. Huang, Training student networks for acceleration with conditional adversarial networks., in: BMVC, 2018, p. 61.
[18] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[19] J. Yang, L. Zhang, Y. Xu, J. yu Yang, Beyond sparsity: the role of l1-optimizer in pattern classification, Pattern Recognit. 45 (2012) 1104–1118.
[20] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: which helps face recognition? in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 471–478.
[21] Z. Chen, W. Zuo, Q. Hu, L. Lin, Kernel sparse representation for time series classification, Inf. Sci. 292 (2015) 15–26.
[22] S. Zeng, B. Zhang, Y. Zhang, J. Gou, Collaboratively weighting deep and classic representation via l2 regularization for image classification, in: Proceedings of The 10th Asian Conference on Machine Learning, in: Proceedings of Machine Learning Research, 95, PMLR, 2018, pp. 502–517.
[23] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[24] S. Zeng, X. Yang, J. Gou, Multiplication fusion of sparse and collaborative representation for robust face recognition, Multimed. Tools Appl. 76 (20) (2017) 20889–20907.
[25] N. Akhtar, F. Shafait, A. Mian, Efficient classification with sparsity augmented collaborative representation, Pattern Recognit. 65 (2017) 136–145.
[26] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).
[27] J. Zhou, S. Zeng, B. Zhang, Two-stage image classification supervised by a single teacher single student model, in: Proceedings of the British Machine Vision Conference, 2019.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[29] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Stat. 46 (3) (1992) 175–185.
[30] J. Zhou, B. Zhang, Collaborative representation using non-negative samples for image classification, Sensors 19 (11) (2019) 2609.
[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[32] A.V. Nefian, Georgia tech face database.
[33] C.E. Thomaz, G.A. Giraldi, A new ranking method for principal components analysis and its application to face image analysis, Image Vis. Comput. 28 (6) (2010) 902–913.
[34] S. Milborrow, J. Morkel, F. Nicolls, The MUCT landmarked face database, Pattern Recognit. Assoc. South Africa 201 (2010) http://www.milbo.org/muct/.
[35] L. Wolf, T. Hassner, I. Maoz, Face Recognition in Unconstrained Videos with Matched Background Similarity, IEEE, 2011.
[36] S.A. Nene, S.K. Nayar, H. Murase, Object image library (coil-100) (1996).
[37] Y. LeCun, C. Cortes, MNIST handwritten digit database (2010).
[38] E. Learned-Miller, G.B. Huang, A. RoyChowdhury, H. Li, G. Hua, Labeled faces in the wild: a survey, in: Advances in Face Detection and Facial Image Analysis, Springer, 2016, pp. 189–248.
[39] L. Deng, The MNIST database of handwritten digit images for machine learning research [best of the web], IEEE Signal Process. Mag. 29 (6) (2012) 141–142.
[40] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 689–696.

Jianhang Zhou received the B.S. degree in Computer Science from Nanjing Forestry University in 2018, the M.S. degree in Computer Science from University of Macau in 2020. He is currently pursuing the Ph.D. degree in Computer Science in the Department of Computer and Information Science, Faculty of Science and Technology at the University of Macau. His research interest includes pattern recognition, machine learning, biometrics, deep learning for medical image processing, and computer vision.

Shaoning Zeng received the B.S. degree and M.S. degree from Beihang University (BUAA), Beijing, China, in 2004 and 2007, respectively. He is currently pursuing the Ph.D. degree in computer science in the Department of Computer and Information Science, Faculty of Science and Technology at the University of Macau. From 2009 to now, he is a Researcher and Lecturer in the School of Information Science and Technology at Huizhou University, China. His research interest includes computer vision, pattern recognition, machine learning and deep learning for multimedia and image processing applications. He has published over 10 scientific publications in these areas.

Bob Zhang received his B.A. in Computer Science from York University in 2006, a M.A.Sc. in Information Systems Security from Concordia University in 2007, and a Ph.D. in Electrical and Computer University from the University of Waterloo in 2011. After graduating from Waterloo he remained with the Center for Pattern Recognition and Machine Intelligence, and later worked as a Post-Doctoral Researcher in the Department of Electrical and Computer Engineering at Carnegie Mellon University. Currently, he is an Associate Professor in the Department of Computer and Information Science at the University of Macau, and is leading the team of Pattern Analysis and Machine Intelligence Group. His research interests focus on biometrics, pattern recognition, and image processing