EURASIP Journal on Image and Video Processing

RESEARCH Open Access

RGB-NIR image categorization with prior knowledge transfer

Xishuai Peng[1], Yuanxiang Li[1*], Xian Wei[1], Jianhua Luo[1] and Yi Lu Murphey[2]

Abstract

Recent development on image categorization, especially scene categorization, shows that the combination of standard visible RGB image data and near-infrared (NIR) image data performs better than RGB-only image data. However, the size of RGB-NIR image collection is often limited due to the difficulty of acquisition. With limited data, it is difficult to extract effective features using the common deep learning networks. It is observed that humans are able to learn prior knowledge from other tasks or a good mentor, which is helpful to solve the learning problems with limited training samples. Inspired by this observation, we propose a novel training methodology for introducing the prior knowledge into a deep architecture, which allows us to bypass the burdensome labeling large quantity of image data to meet the big data requirements in deep learning. At first, transfer learning is adopted to learn single modal features from a large source database, such as ImageNet. Then, a knowledge distillation method is explored to fuse the RGB and NIR features. Finally, a global optimization method is employed to fine-tune the entire network. The experimental results on two RGB-NIR datasets demonstrate the effectiveness of our proposed approach in comparison with the state-of-the-art multi-modal image categorization methods.

Keywords: Multi-modal image categorization, Knowledge distillation, Transfer learning, Deep learning

1 Introduction

In the past several decades, numerous computer vision methods have been developed to process visible RGB images. Recent studies demonstrate that if we have a larger spectrum of radiation than RGB-only images, the better performance would be obtained in many computational visual tasks, such as saliency detection [1, 2], scene categorization [3, 4], and image segmentation [5, 6].

RGB-NIR image categorization is one of the most challenging tasks for two reasons. Firstly, the labeled RGB-NIR images are scarce due to the difficulty of acquisition and labeling. Some "shallow" features used in traditional methods, e.g., *Scale-Invariant Feature Transform* (SIFT) [7], *Gist* [8], and *census transform histogram* (CENTRIST) [9], may do not need any labeled data. However, with increased scene or object categories, these algorithms gradually show their limitation [10]. Because in terms of the biology of color vision, these features are corresponding to the lowest level of processes in the hierarchically organized visual cortex. Different from these "shallow" features, deep networks, especially convolutional neural networks (CNNs), are able to learn the representations of complex data with multiple levels of abstraction [11–14]. However, these systems suffer from the so-called overfitting problem if the training data are limited.

As one of the most fruitful techniques addressing this problem, transfer learning (TL) methods are proposed to transfer knowledge from some auxiliary source data. Many researches demonstrated that deep architectures with TL have great capacity of generating transferable features [13, 14]. For example, the research in [10] proposed deep convolutional activation feature (DeCAF), which trained on different datasets with specific levels of network fixed. The authors in [15] trained the lower-layer and higher-layer features separately in supervised and unsupervised manners; this training process made the model achieved lower classification error rate in the classification of uppercase-lowercase letters. The research in [16] reused the different layers of features in a deep CNN model and evaluated the generality and specificity on both simulated and natural images. The research in [17] demonstrated scene-centric datasets, e.g., *Place365*

*Correspondence: yuanxli@sjtu.edu.cn
[1] School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, China
Full list of author information is available at the end of the article

dataset, and tend to produce better features than object-centric datasets, e.g., *ImageNet* [18] *dataset*, for the scene categorization. These works achieved significant progress when the feature distribution of source domain is similar to target domain. However, it is still a challenging problem when the distributions of involved domains are greatly different, such as RGB and NIR images. In this paper, we incrementally apply the existing TL theory to RGB-NIR domains and propose a feasible and effective method for learning effective RGB-NIR features.

Secondly, the statistical properties behind RGB-NIR images are significantly different, and the relation between them maybe highly non-linear. As a traditional method, canonical correlation analysis (CCA) [19] aims to find transformations for maximizing correlations between modalities. Many researches have demonstrated it is an efficiency technique for multi-modal feature fusion. For example, the research in [20] used CCA to explore the latent relation behind multi-modal features, e.g., *co-occurrence matrix*, and *Zernike moments*. The authors in [21] presented a three-view CCA model that explicitly incorporated the high-level semantic information as the third view. The authors in [22] proposed a multi-view semantic alignment process to address the projection domain shift problem. These methods are comprised of stacking self-contained algorithmic components, i.e., *feature extraction*, *feature fusion*, and *classifier training*. Consequently, the learned features are fixed once build, which results in the process of feature extraction, and fusion cannot benefit from the label information.

Recent works have demonstrated that it is effective to learn multi-modal features and the following classifiers simultaneously. The authors in [23] proposed a large margin multi-modal feature extraction (LM3FE) framework for multi-tasks, i.e.,*feature selection*, *transformation*, and *classification*. In comparison with the "shallow" networks, the general method used in deep architecture is intermediate fusion, i.e., *a additional perception layer is used to fully connect the flattened multi-modal features and classifier* [24]. The other promising work [25] was to implement CCA with deep architecture, i.e., *Deep CCA* (DCCA). Continuing along this line of research, the authors in [26] combined the DCCA with autoencoders and proposed the deep canonically correlated autoencoders (DCCAE). DCCAE was proposed to learn multi-modal features that are capable of reconstructing inputs as well as maximizing the correlation between modalities. Based on this research, the authors in [27] proposed to separately learn the modal-related features, which are used for data reconstruction, and the modal-invariable features, which are used for maximizing correlation between modalities. These works demonstrated that the prior knowledge of maximizing correlation across modalities is effective for exploring the complementary information behind multi-modal data. However, these architectures are still data-hungry models, which have not considered the problem addressed in this paper, i.e., *train a deep network using very limited training data*.

In the proposed training process, it is used as a sub-task in the pre-training phase to maximize correlation between RGB-NIR features. This idea was inspired by the work presented in [28], in which the authors assumed that there is a training barrier involved in the nature of difficult tasks, e.g., *scene recognition*. The experiments they conducted demonstrated that it is effective to learn several intermediate easier tasks, which are decomposed from a difficult task, rather than directly learn the difficult task. In order to facilitate such easier tasks for RGB-NIR image classification, we proposed a feature fusion method based on knowledge distillation (FFKD) in this paper. The knowledge distillation algorithm [29] was firstly proposed by Hinton for model compression. It transfers the knowledge from a cumbersome model to a small model. To the best of our knowledge, few researches have been conducted to address the multi-modal feature fusion problem using knowledge distillation architecture. The features generated by FFKD can be considered as an approximation to the features generated by CCA and a weight regularization. Despite its simplicity, we found it effective in multi-modal image categorization task.

In this paper, we consider the multi-modal data categorization problem in the context of images such that both standard visible RGB channels and near infrared (NIR) channels are available. The main contributions of this paper include the following: (1) we propose a novel feature fusion algorithm FFKD, which is based on CCA and knowledge distillation methods, and (2) we propose a novel training methodology that combined TL and FFKD methods. This allows us to bypass the burdensome labeling large image data to meet the big data requirements in deep learning, in the way of introducing the prior knowledge from auxiliary source data and classical models. The paper is organized as follows: Section 2 introduces the details of our proposed framework. Section 3 demonstrates the experimental results. Then the conclusion is given in Section 4.

2 Methods

Several studies have demonstrated that random initialization of deep neural network can yield rather poor results, while specifically targeted initialization can have a drastic impact [17, 24]. Based on these observations, we can deduce that the effective global minima obtained using sufficiently large data also could be obtained using less amount of data with proper guidance. Motivated by such hypothesis, two types of pre-training methods, transfer pre-training and distillation pre-training, are proposed to provide good initialization for application that have small amount of training data. Transfer pre-training transfers features learned from massive labeled source data to unlabeled target data. Though this technique has been popularly used in machine learning community, our work provides an effective way to transfer features from RGB data to NIR data. Distillation pre-training guides the network to fuse RGB-NIR features with the knowledge learned from a traditional model, which should be selected as related to the final task, e.g., *object or scene categorization in this paper*.

As described in Fig. 1, the proposed architecture consists of three functional layers, *the feature extraction layer*, *the feature fusion layer*, and *the classifier*. The feature extraction layer contains two off-the-shelf deep models for RGB and NIR data, which is introduced in Section 2.2. The feature fusion layer is a single linear perceptron, which is discussed in Section 2.3. The Softmax classifier is used to explore the label information to jointly fine-tune the feature extraction and fusion layers, which is presented in Section 2.4. The three functional layers are separately pre-trained before the global fine-tuning; the proposed training process can be described as follows: firstly, the two off-the-shelf deep models were pre-trained on a large labeled dataset, i.e., *ImageNet*, and then fine-tuned on our own RGB and NIR data respectively. After that, the feature fusion layer, which half connected with the RGB features and half connected with the NIR features, merged RGB-NIR features into one feature vector. Finally, a Softmax classifier was trained on the fused RGB-NIR features.

2.1 Model formulation

Let c_s be the number of source classes with n_s instances $S = \{X_s, Y_s\}$ and c_t be the number of target classes with n_t instances $T = \{V_t, N_t, Y_t\}$. The dimension of image vector both in source and target domains is d, $X_s \subseteq R^{n_s \times d}$, $V_t \subseteq R^{n_t \times d}$, and $N_t \subseteq R^{n_t \times d}$. Y_s and Y_t are the label vectors for source and target instances, respectively. The source and target classes are disjoint in the problem we addressed, $Y_s \bigcap Y_t = \emptyset$.

Firstly, the transfer pre-training aims to transfer the knowledge learned from vast labeled source data to target data. The knowledge in deep architectures is generally presented as layer weights and could be considered as a mapping function. f^s denotes the mapping function learned from the source labeled images and is used to map source images into features. Since the training process on source domain is surprised by sufficient data, it is reasonable that f^s is capable of mapping the source data into a more discriminative feature space. Taking the similarity between source and target data into account, f^s should be a good initialization for f^t, a good enough mapping function for target data. Specifically, the mapping function for target RGB data f^v and the mapping function for NIR data f^n should be fine-tuned from f^s. We use F^v and F^n denote the features generated from the two mapping functions.

Secondly, there are two models included in the distillation pre-training phase. One is the student model, which is an one-layer perceptron, and the other is the teacher model, which contains CCA and KCPA modules. During this training phase, the teacher model learns two transformations W_v and W_n to maximize the correlations between RGB and NIR features. The student model also contains two transformations: the weights G_v connect RGB features to the feature fusion layer, and the weights G_n connect NIR features to the feature fusion layer. The objective of this training phase is to make the output of the student model as similar to the output of the teacher model; this

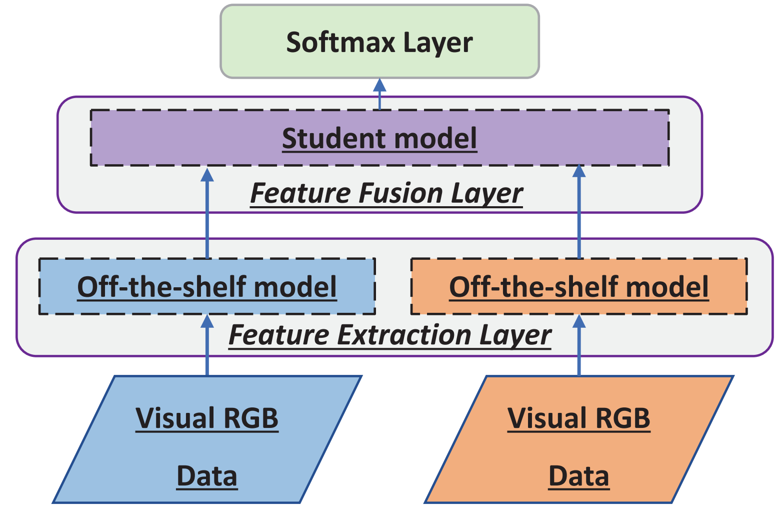

Fig. 1 The proposed architecture consists of three functional layers, *the feature extraction layer*, *the feature fusion layer*, and *the classifier*

can be easily obtained by back-propagation algorithm to minimize the metric between them.

After the two pre-training phases have been completed, a global fine-tuning process is used to learn a global mapping function $h : (V_t, N_t) \longrightarrow Y_t$ that maps the raw images into its predicted labels.

2.2 Transfer pre-training

CNNs have been used in most state-of-the-art deep architectures for image recognition, detection, and segmentation. CNN is a feed-forward architecture with three main types of layers: (1) 2D convolution layers, (2) 2D sub-sampling layers, and (3) 1D output layers. A convolution layer consists of several adjustable 2D filters. The output of each filter is a feature map that indicates the presence of a feature at a given pixel location. In practice, a non-linear activation [30, 31], e.g., *RELU* [30] and *Sigmoid* [32], is generally applied on the feature map to enhance its representation ability. Batch normalization (BN) [33] is used to normalize the feature map for addressing the internal covariate shift problem. The sub-sampling layer, e.g., *mean pooling layer* and *max pooling layer*, is used as a bottleneck to reduce the dimension of feature map. The fully connection layer maps the output of stacked convolution-pooling layers into the predicted labels.

Some recent researches demonstrated that CNNs pre-trained on large datasets contain general purpose features, which are transferable to many other tasks [10, 15, 17]. These works inspired us to fine-tune the off-the-shelf CNN-based architectures, e.g., *VGGNet* [34], *ResNet* [35], and *Inception Net* [12], on our own RGB and NIR data. To employ this fine-tuning process, we may need to select the fixed layer number. In general, the number of layers to freeze is determined by learning task. Some classical works [13, 16, 24] utilized the fine-tuning rule to decide the number of layers to freeze. Our experimental results are consistent with the latter, that is the less layers should be fixed to avoid the negative transfer if the dissimilarity between source and target datasets is large. Thus, we set $m = 4 > n = 2$ to be consistent with the rule, where m and n are the fixed layer number set for RGB and NIR fine-tuning process. The fine-tuning process could be spited as following three steps:

- Step 1: Train the off-the-shelf deep model on a large-scale database, i.e., *ImageNet*.
- Step 2: Remove the classification layer from the pre-trained off-the-shelf deep model and preserve the rest parts as the initial model of RGB and NIR data.
- Step 3: Fine-tune feature extractor on RGB data with first m layers frozen, while first n layers are frozen in fine-tuning NIR feature extractor.

After the fine-tuning process, the classification layer is removed, and we call the rest part as fine-tuned RGB and NIR feature extractors.

2.3 Distillation pre-training

There are two models included in the distillation pre-training phase, i.e., *the teacher model and the student model*. The feature vectors that generated from the feature extraction layer, i.e., $\mathbf{f^v}$ and $\mathbf{f^n}$, are fixed and used as inputs of this phase. Assemble each vectors $\mathbf{f^v}$ into the row of matrix F^v and similarly place $\mathbf{f^n}$ into matrix F^n. The process of distillation pre-training can be described as follows.

The proposed teacher model is an approach based on kernel principal components analysis (KPCA) [36] and CCA. KPCA is an extension of PCA using techniques of kernel methods. In this paper, the cosine kernel is selected to reduce the dimensions of F^v and F^n, and the generated low-dimension features are denoted as $F^{v'}$ and $F^{n'}$, respectively. Each row in $F^{v'}$ is denoted as $\mathbf{f^{v'}}$ similarly each row in $F^{n'}$ is denoted as $\mathbf{f^{n'}}$.

$$F^{v'} = \mathrm{KPCA}_{\mathrm{cosine}}\left\{F^v\right\} \tag{1}$$

and

$$F^{n'} = \mathrm{KPCA}_{\mathrm{cosine}}\left\{F^n\right\} \tag{2}$$

Then, CCA method is used to maximize the correlation between RGB and NIR features by learning two transformations W_v and W_n. CCA is a statistical method used to investigate relationships among two or more variable sets, each of them consists of at least two variables. The literature contains a number of sophisticated methods for multi-view learning, e.g., *CCA/Kernel CCA (KCCA)/DCCA* [25, 37, 38], *metric learning* [39], and *large-margin formulations* [40]. We found that the basic CCA formulation already gave very promising results for RGB-NIR image categorization without having to pay the price of increased complexity for learning and inference. The objective function of CCA can be formulated as follows:

$$\rho = \max_{\mathbf{w_v},\, \mathbf{w_n}} \frac{\left\langle F^{v'}\mathbf{w_v},\ F^{n'}\mathbf{w_n}\right\rangle}{\|F^{v'}\ \mathbf{w_v}\|\|F^{n'}\ \mathbf{w_n}\|} \tag{3}$$

Hence, it can be rewritten as:

$$\rho = \max_{\mathbf{w_v},\, \mathbf{w_n}} \frac{\mathbf{w_v'}\ C_{vn}\ \mathbf{w_n}}{\sqrt{\mathbf{w_v'}\ C_{vv}\ \mathbf{w_v}\ \mathbf{w_n'}\ C_{nn}\ \mathbf{w_n}}} \tag{4}$$

where C_{vv} and C_{nn} are the covariance matrices, and C_{vn} and C_{nv} are the between-sets covariance matrices of RGB and NIR features. Since the choice of re-scaling is arbitrary, the corresponding Lagrangian can be written as follows:

$$L(\lambda,\ \mathbf{w_v},\ \mathbf{w_n}) = \mathbf{w_v'}\ C_{vn}\ \mathbf{w_n} - \frac{\lambda_v}{2}\ \left(\mathbf{w_v'}\ C_{vv}\ \mathbf{w_v} - 1\right) - \frac{\lambda_n}{2}\left(\mathbf{w_n'}\ C_{nn}\ \mathbf{w_n} - 1\right) \tag{5}$$

Taking derivatives in respect to $\mathbf{w_v}$ and $\mathbf{w_n}$, then make them equal to zero, we have:

$$\lambda_n\ \mathbf{w_n'}\ C_{nn}\ \mathbf{w_n} - \lambda_v\ \mathbf{w_v'}\ C_{vv}\ \mathbf{w_v}\ =\ 0 \tag{6}$$

which together with the constraints implies that $\lambda_v - \lambda_n = 0$, let $\lambda = \lambda_v = \lambda_n$. Assuming C_{nn} is invertible, we have:

$$C_{vn}\ C_{nn}^{-1}\ C_{nv}\ \mathbf{w_v}\ =\ \lambda^2\ C_{xx}\ \mathbf{w_v} \tag{7}$$

As the covariance matrices C_{vv} is symmetric positive definite, it can be decomposed into a lower triangular matrix R_{vv} and its transpose R'_{vv} using a complete Cholesky decomposition [41]:

$$C_{vv}\ =\ R_{vv}\ \cdot\ R'_{vv} \tag{8}$$

Finally, we get $\mathbf{w_v}$ by solving the equation as follows:

$$R_{vv}^{-1}\ C_{vn}\ C_{nn}^{-1}\ C_{nv}\ R_{vv}^{-1'}\ R'_{vv}\ \mathbf{w_v}\ =\ \lambda^2\ R'_{vv}\ \mathbf{w_v} \tag{9}$$

After $\mathbf{w_v}$ is obtained, the $\mathbf{w_n}$ could be obtained using following equation:

$$\mathbf{w_n}\ =\ \frac{C_{nn}^{-1}\ C_{nv}\ \mathbf{w_v}}{\lambda} \tag{10}$$

At last, we assemble the top k projection vectors $\mathbf{w_v^i}$ in to columns of matrix W_v, and similarly place $\mathbf{w_n^i}$ into matrix W_n, then concatenate the feature transformed by W_n and W_v as the output of teacher model:

$$O_{\text{teacher}} = \text{concatenation}(\mathbf{f^v}'\ W_v, \mathbf{f^n}'\ W_n) \tag{11}$$

In order to transfer the knowledge from the teacher model to the proposed architecture, the feature fusion layer (student model) and FFKD is proposed. As illustrated in Fig. 2, the student model is an one-layer perceptron that half connects with the RGB features and half connects with the NIR features. FFKD is designed to make the output of student and teacher models as similar as possible. The object function of FFKD can be formulated as follows:

$$\text{Loss} = \|O_{\text{teacher}} - O_{\text{student}}\|_2 + \|G_n\|_2 + \|G_v\|_2 \tag{12}$$

where O_{teacher} and O_{student} denote the outputs of the teacher model and the student model, respectively. G_n and G_v are weights connected with NIR and RGB features. As shown in Fig. 2, the process of FFKD can be split into three steps:

- Step 1: Randomly select a batch of RGB and NIR data as the inputs of teacher and student models and get the outputs from both models.
- Step 2: Calculate the residual error using Eq. 12.
- Step 3: Update the weights of the student model by back propagating the residual error, until the residual error is small enough.

After the distillation pre-training, the feature fusion layer is well initialized. In spite of its simplicity, the experimental results show an evidence that the distillation method can improve the performance significantly.

2.4 Global fine-tuning

As an optimization technique, global fine-tuning is commonly used in deep learning. The works in [16] demonstrated that when layers were fine-tuned, the performance slightly improved with more layers; however, when layers were frozen (fixed, without fine-tuning), the performance degraded with more layers. This observation is very similar to human vision system; some evidences

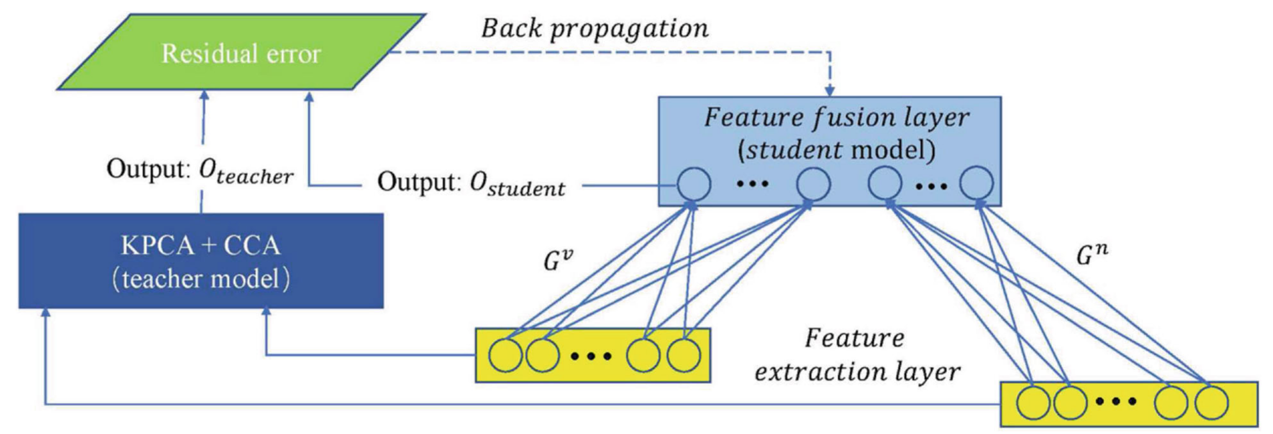

Fig. 2 The flowchart of FFKD method. The teacher model consists the KPCA and CCA modules; the student model is an one-layer perceptron. The FFKD methods is proposed to transfer the knowledge from the teacher to the student model

have suggested that humans construct high-level features for different visual tasks by altering the combination of different low-level features. Inspired by these works, we fine-tuned the whole network which was initialized by transfer (Section 2.2) and distillation (Section 2.3) pre-training. The process of global fine-tuning can be described as follows:

- Step 1: Randomly select a batch of training data as input to the pre-trained architecture.
- Step 2: Update the network by back propagating the classification error, with the first four layers of both RGB and NIR feature extractors frozen, until the error is small enough. Otherwise, go to step 1.

3 Results and discussion

3.1 Datasets and experimental setting

We evaluated the proposed method on both scene and object categorization dataset, i.e., *EPFL* [7] and *4-classes dataset* [42, 43]. The statistics of the two datasets are illustrated in Table 1.

The EPFL scene classification dataset (EPFL) contains 477 images distributed in 9 categories as follows: country (52), field (51), indoor (56), forest (53), mountain (55), old building (51), street (50), urban (58), and water (51). Although the number of images in this dataset is small, the image classes are challenging. To the authors' best knowledge, this is the only public benchmark dataset for scene categorization that provides both RGB and NIR channels for every image. Figure 3 shows several examples in order to demonstrate the complexity of this dataset.

The 4-classes object classification dataset (4-classes) contains 1464 images distributed in 4 categories: camouflage tanks (390), camouflage artillery (350), non-camouflage tanks (400), and vehicles (360). This is a unmanned aerial vehicle (UAV) aerial image dataset that includes both RGB and NIR channels for each image.

To facilitate the comparison, we follow the same experiment setup as in [7, 43] and [42], which is described as follows:

- EPFL: randomly selecting 99 images for testing (11 images per category) and the rest for training.
- 4-classes: randomly selecting 200 images for training and the rest for testing.

In all our experiments, we repeat 10 trials with a randomly selected training/test split and calculate the mean of the recognition rates (fraction of correct matches) over all 10 trails.

Table 1 Statistics of the two datasets

Dataset	Type	No. of example	No. of feature	No. of class
EPFL	Scene	477	89,401	9
4-classes	Object	1464	89,401	4

3.2 Comparison with other methods on object categorization

In order to demonstrate the effectiveness of the proposed method, we compared our method with seven classification methods on object database (4-classes), which are described as follows.

1. Linear discriminant analysis (LDA) [44]: LDA maximizes the ratio of the between-class scatter to the within class scatter then find out the projection of samples so that samples can be separated. We concatenated RGB and NIR image pixels as input to LDA.
2. Histogram of oriented gradients (HOG) [45]: HOG describes the distribution of gradient strength and gradient direction of object region in image; it is good at representing the appearance and shape of object. We used HOG descriptor as input of support vector machine (SVM)[46].
3. Speeded-up robust feature (SURF) [47]: SUPF is a variation of SIFT and it has good adaptability for the scale and rotation changes of objects. We used SURF descriptor as input of SVM.
4. Dictionary learning (DL) [48]: DL finds some basic elements for the image data then uses the linear combination of these basic elements to represent the image data approximately. We used DL features as input of SVM.
5. Kernel dictionary learning (KDL) [49]: KDL is proposed to learn over-complete dictionary from the training data set, and it is a non-linear dictionary learning method. We used KDL features as input of SVM.
6. DL+LDA: this method feeds DL feature to LDA as recommended in [42].
7. KDL+LDA: this method feeds KDL feature to LDA as recommended in [42].

As shown in Table 2, the global descriptor, i.e., *HOG*, performed better than the local descriptor, i.e., *SURF*, for object categorization task. We also noticed that the model-driven methods, i.e., *HOG and SURF*, performed better than the data-driven methods, i.e., *DL and KDL*, when the training data is insufficient as the problem addressed in this paper. However, the proposed model, Inceptionv3-proposed model, achieved the best performance, though it is a data-driven method. It could be taken as an evidence that the proposed method leveraged the data-driven model by introducing the prior knowledge from auxiliary source data and other model-driven methods.

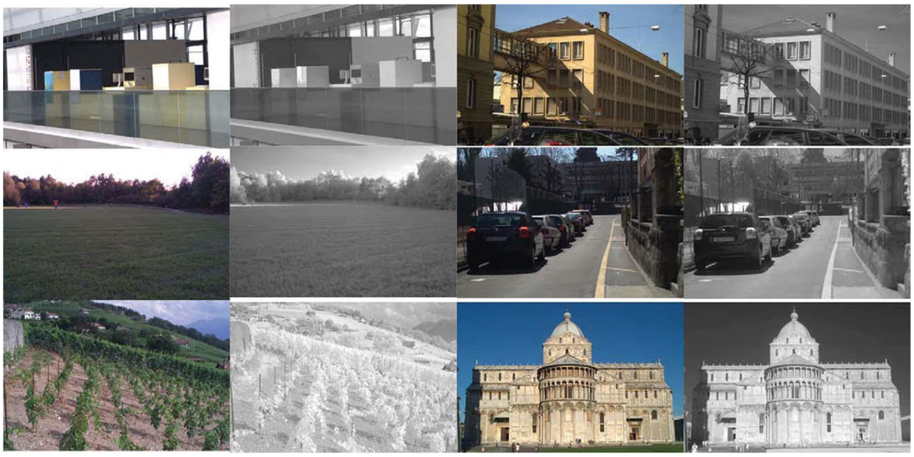

Fig. 3 Examples from EPFL database, each example includes visible image and its corresponding NIR image

3.3 Comparison with other methods on scene categorization

The scene categorization task is generally considered to be more difficult than the object categorization task due to it contains more abstract concept. In this section, we further compared our method with seven classification methods on scene database (EPFL). The seven classification methods are SVM using seven popular scene descriptors, which are described as follows.

1 Hierarchical model and X (HMAX) [50]: HMAX is inspired by a computational model of object recognition in cortex. We calculated the HMAX descriptors for each channel independently then concatenated them as one feature vector.
2 Multi-channel Gist (mGist) [7]: Gist descriptor was proposed in [51]; its coarse to fine processing was believed to be very similar to human vision. mGist was proposed to concatenate Gist descriptor from all channels.
3 Multi-channel SIFT (mSIFT) [52]: mSIFT descriptor concatenates SIFTs from all channels in the bag of visual words (BOV) framework [52].
4 Multi-spectral SIFT (msSIFT) [7]: msSIFT concatenates SIFTs from all channels with PCA post-processing.
5 Fisher Vector [53]: This method was proposed in [53] , which used local patch-based Fisher Vector image representation to encode both the texture and color information.
6 Multi-channel CENsus TRansform hISTogram (mCENTRIST) [3]: mCENTRIST was proposed to jointly encode the information within multi-channel images.
7 Concatenate CENTRIST (conCENTRIST) [3]: conCENTRIST was an intuitive way to extend CENTRIST [9] to multiple channels, it concatenates CENTRISTs from all channels.

Table 2 Comparison of performance on object categorization on 4-classes data

Method	Recognition rate (%)
LDA	95.5
HOG	95.91
SURF	94.25
DL	80.52
KDL	82.21
DL+LDA	94.61
KDL+LDA	97.79
Inceptionv3-proposed	99.3

As shown in Table 3, CENTRIST-based methods performed much better than the traditional descriptor, i.e., *mSIFT*, *msSIFT*, *mGist*, and *HMAX*. Among three CENTRIST-based methods, multi-channel joint information encoding made mCENTRIST better than its direct concatenation variant, conCENTRIST. The third best accuracy was yielded by Fisher Vector, which appended mSIFT with channel intensity statistical features. However, the Fisher Vector technique seemed to be the key to its performance improvement, which also suffered from the limited labeled data.

We also compared the proposed method against the reused or fine-tuned state-of-art deep architectures, i.e., *Inceptionv3*, *ResNetv1*, and *ResNetv2*. All the reused and fine-tuned models were pre-trained on ImageNet dataset;

Table 3 Comparison of performance on scene categorization on EPFL data

Method	Recognition rate (%)
HMAX	59.2
mGist	69.9
mSIFT	80.4
msSIFT	75.2
Fisher Vector	87.9
mCENTRIST	84.5
conCENTRIST	81.7
Inceptionv3-reused	72.3
Inceptionv3-finetuned	79.5
ResNetv1-reused	63.3
ResNetv1-finetuned	74.6
ResNetv2-reused	75.3
ResNetv2-finetuned	80.2
Inceptionv3-proposed	90.2
ResNetv2-proposed	90.8

the difference is that in the reused models, only the classifier was trained, but both the features and the classifier were retrained in the fine-tuned models. The fine-tuned deep models significantly improved the performance of reused deep models; however, the improvement was decreasing as the model complexity increases, i.e., *the number of parameters within three models are* ResNetv2 > Inceptionv3 > ResNetv1*; the performance improvements generated from fine-tuning for three models are* ResNetv2 < Inceptionv3 < ResNetv1.

We replaced the traditional fine-tuning approach with the proposed training approach for Inceptionv3 and ResNetv2 architectures, i.e., *Inceptionv3-proposed* and *ResNetv2-proposed*. The improvements are obvious on both architectures, which achieved the best and second-best accuracy. As suggested in [54], the relation behind the RGB and NIR data is highly non-linear, the more sophisticated solutions probably capture the highly complex relations. But in our experiment, we noticed that the Inceptionv3-proposed model achieved comparable performance to RestNetv2-proposed model, which usually should have much higher accuracy. We also tried to use non-linear activation in the feature fusion layer, i.e., *RELU, Sigmoid*, to enhance its non-linear representation ability. But the performance of Inceptionv3-proposed model and ResNetv2-proposed model was decreased to 89.6 and 88.3, respectively. This may be caused by the small number of our training samples, which is not enough to fully explore the advantage of too complicate models, such as ResNetv2.

3.4 Experimental analysis

In order to further investigate the reason behind the performance improvement made by the proposed training process, we separately analyzed the performance improvement made by the transfer pre-training, distillation pre-training, and global fine-tuning on scene (EPFL) dataset. Since the Inceptionv3 had the best balance between performance and computational complexity, the model based on Inceptionv3 was selected for the following analysis.

As shown in Fig. 4, the reused Inceptionv3 model simply concatenated the RGB-NIR features and used the fixed features to train a softmax classifier. The transfer pre-training model fine-tuned the Inceptionv3 model on EPFL data using transfer pre-traing method (Section 2.2); then, the fine-tuned RGB and NIR features were concatenated together to train a Softmax classifier. The distillation pre-training model used the distillation pre-training method (Section 2.3) to fuse the features that generated from transfer pre-traing model; then, the fused RGB-NIR features were used to train a Softmax classifier. The global fine-tuning model used the global fine-tuning method (Section 2.4) to fine-tuning the whole network, which was initialized by transfer and distillation pre-training. By comparing these four models, we noticed that the three training processes, *transfer pre-training*, *distillation pre-training*, and *global fine-tuning*, gradually improved the model's performances. For example, distillation pre-training model started with a relative higher recognition rate, ended with a better performance, and reached convergence faster than the model without using it, i.e., *transfer pre-training model*. The same experimental results also could be observed when the transfer pre-training and global fine-tuning methods were used. These observations demonstrate the effectiveness of the proposed methods, especially when the training data are very limited.

In Fig. 5, we evaluated the performance of the RGB and NIR features that learned in different training phases. After each training phase, we additionally trained a Softmax classifier for each modality. The features generated from the last layer of feature extractor, i.e., *the inputs of feature fusion layer*, was used as input. Since in the phase of distillation pre-training the feature extractors were fixed, we did not show the evaluation results after this phase. But we compared the training process using distillation pre-training, i.e., *RGB-FFKD* and *NIR-FFKD*, against the training process without using it, i.e., *RGB-RI* and *NIR-RI*. The two training processes were evaluated on the same architecture, as illustrated in Fig. 1, but the random initialization was used to replace the FFKD method for initializing the feature fusion layer in RGB-RI and NIR-RI training process.

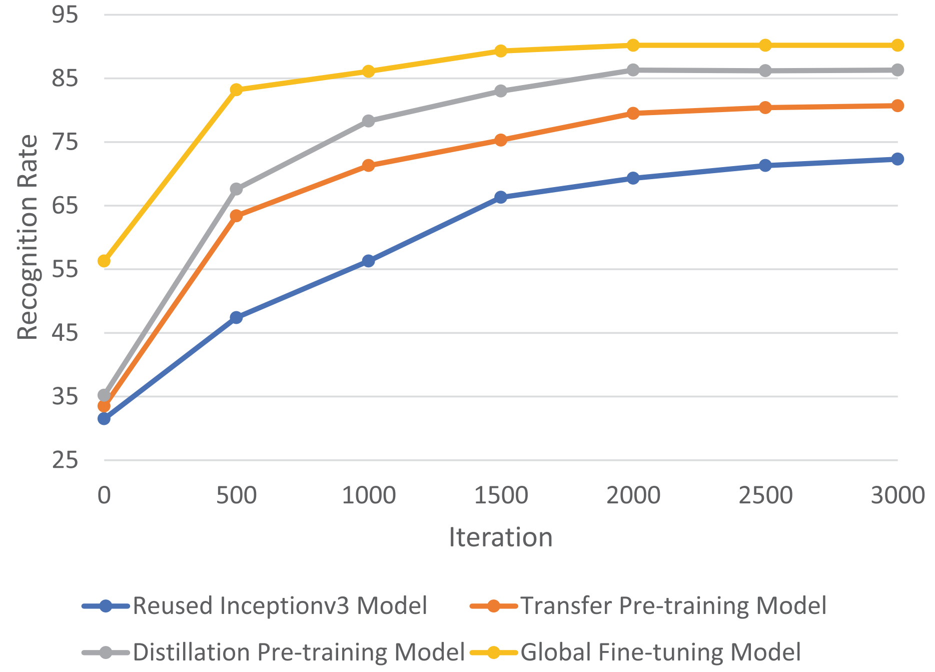

Fig. 4 Recognition rate of reused Inceptionv3 model, transfer pre-training model, distillation pre-training model, and global fine-tuning model

It is obvious that the reused Inceptionv3 has better performance on RGB data than NIR data due to the negative transfer problem. The transfer pre-training improved the model's performance on both RGB and NIR data. The feature extractors were fixed in the distillation pre-training phase; distillation pre-training only provided the feature fusion layer a meaningful initialization. But the analysis shows that the feature extractors, especially the NIR feature extractor, can benefit from the well-initialized fusion layer by the global fine-tuning. These observations

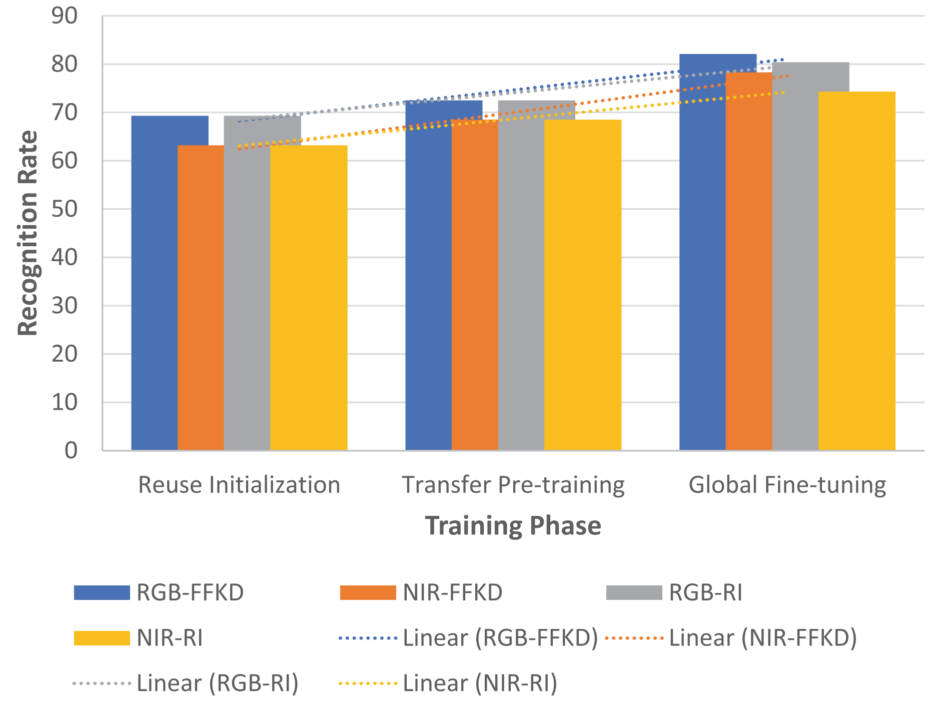

Fig. 5 The performance improvement of the RGB and NIR features along the training process

demonstrated that the distillation pre-training was not trivial; the well initialized fusion layer can help the model learn effective NIR features from RGB data and label information.

4 Conclusions

RGB-NIR image categorization is one of the most challenging computer vision tasks. Several studies have found that approaches based prior knowledge are promising. Continuing along this line of research, two pre-training approaches are proposed in this paper: (1) transfer pre-training, which transfers features from source data to alleviate the demand of labeled target data, and (2) distillation pre-training, which introduces an intermediate concept from a teacher model to guide student model to effectively fuse RGB-NIR features. Experimental results have demonstrated that the proposed approach gives better performance than existing methods.

However, there are still some research issues to be explored. For example, we employed the Inceptionv3 model pre-trained on Place365 to replace the Inceptionv3 model pre-trained on ImageNet. The experimental results showed the performance was increased to 92.3. This observation inspired us to wonder if we use the near-infrared database to train NIR features, will the model's performance increase as well, and if there are other concepts that can reduce the complexity more efficiently than maximizing multi-modal's correlation. These questions will be addressed, studied, and hopefully answered in our further research on this topic.

Abbreviations

BN: Batch normalization; BOV: Bag of visual words; CCA: Canonical correlation analysis; CENTRIST: Census transform histogram; CNN: Convolutional neural network; conCENTRIST: Concatenate census transform histogram; DCCA: Deep canonical correlation analysis; DCCAE: Deep canonically correlated autoencoders; DeCAF: Deep convolutional activation feature; DL: Dictionary learning; FFKD: Feature fusion method based on knowledge distillation; HMAX: Hierarchical model and X; HOG: Histogram of oriented gradients; KCCA: Kernel canonical correlation analysis; KDL: Kernel dictionary learning; KPCA: Kernel principal components analysis; LDA: Linear discriminant analysis; LM3FE: Large margin multi-modal multi-task feature extraction; mCENTRIST: Multi-channel census transform histogram; mGist: Multi-channel Gist; mSIFT: Multi-channel scale-invariant feature transform; msSIFT: Multi-spectral scale-invariant feature transform; NIR: Near-infrared; SIFT: Scale-invariant feature transform; SURF: Speeded-up robust feature; SVM: Support vector machine; TL: Transfer learning; UAV: Unmanned aerial vehicle

Acknowledgments

Not applicable.

Funding

This work was partially supported by the National Natural Science Foundation of China under Grant No. U1406404 and No. 91641129 and CAS pioneer Hundred Talents Program (Type C) with No.2017-122.

Availability of data and materials

The EPFL scene classification dataset supporting the conclusions of this article is available in the http://ivrlwww.epfl.ch/supplementary_material/cvpr11/nirscene1.zip.

Authors' contributions

XS contributed to the algorithmic implementation and paper writing. YL and YX contributed to paper revision. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, China. [2]Department of Electrical and Computer Engineering, University of Michigan-Dearborn, Dearborn, USA.

Received: 21 November 2017 Accepted: 29 November 2018
Published online: 27 December 2018

References

1. Q. Wang, P. Yan, Y. Yuan, X. Li, Multi-spectral saliency detection. Pattern Recogn. Lett. **34**(1), 34–41 (2013)
2. T. Shibata, M. Tanaka, M. Okutomi, Visible and near-infrared image fusion based on visually salient area selection. Proc. SPIE. **9404**, 9404–6 (2015). https://doi.org/10.1117/12.2077050
3. Y. Xiao, J. Wu, J. Yuan, mcentrist: A multi-channel feature generation mechanism for scene categorization. IEEE Trans. Image Process. **23**(2), 823–836 (2014)
4. Q. Zhang, G. Hua, W. Liu, Z. Liu, Z. Zhang, in *Computer Vision – ACCV 2014*. Can visual recognition benefit from auxiliary information in training? (Springer International Publishing, Singapore, 2015), pp. 65–80
5. N. Salamati, D. Larlus, G. Csurka, S. Süsstrunk, in *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Semantic image segmentation using visible and near-infrared channels (Springer Berlin Heidelberg, Berlin, 2012), pp. 461–471
6. C. Hung, J. Nieto, Z. Taylor, J. Underwood, S. Sukkarieh, in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Orchard fruit segmentation using multi-spectral feature learning (IEEE, Tokyo, 2013), pp. 5314–5320. https://doi.org/10.1109/IROS.2013.6697125
7. M. Brown, S. Süsstrunk, in *CVPR 2011*. Multi-spectral sift for scene category recognition (IEEE, Colorado Springs, 2011), pp. 177–184. https://doi.org/10.1109/CVPR.2011.5995637
8. A. Oliva, A. Torralba, Building the gist of a scene: the role of global image features in recognition. Prog. Brain Res. **155**, 23–36 (2006)
9. J. Wu, J. M. Rehg, Centrist: a visual descriptor for scene categorization. IEEE Trans. Pattern. Anal. Mach. Intell. **33**(8), 1489–1501 (2011)
10. J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, in *Proceedings of the 31st International Conference on Machine Learning*, ed. by E. P. Xing, T. Jebara. DeCAF: a deep convolutional activation feature for generic visual recognition (PMLR, Bejing, 2014), pp. 647–655
11. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ed. by L. Getoor, T. Scheffer. Multimodal deep learning (ACM, New York, 2011), pp. 689–696
12. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Rethinking the inception architecture for computer vision (IEEE, Las Vegas, 2016), pp. 2818–2826
13. J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: a survey. Knowl.-Based Syst. **80**(Supplement C), 14–23 (2015)
14. S. J. Pan, Q. Yang, A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
15. C. Kandaswamy, L. M. Silva, L. A. Alexandre, J. M. Santos, J. M. de Sá, in *Artificial Neural Networks and Machine Learning – ICANN 2014*. Improving deep neural network performance by reusing features trained with transductive transference (Springer International Publishing, Hamburg, 2014), pp. 265–272
16. J. Yosinski, J. Clune, Y. Bengio, H. Lipson, in *Advances in Neural Information Processing Systems*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D.

Lawrence, and K. Q. Weinberger. How transferable are features in deep neural networks? (Curran Associates, Inc., Montréal, 2014), pp. 3320–3328

17. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, in *Advances in Neural Information Processing Systems 27*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Learning deep features for scene recognition using places database (Curran Associates, Inc., Montréal, 2014), pp. 487–495
18. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, K. Andrej, K. Aditya, M. Bernstein, ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
19. H. Hotelling, Relations between two sets of variates. Biometrika. **28**(3), 321–377 (1936)
20. M. JIN, Y. HU, Feature-level fusion of infrared and visible images based on principal component analysis plus canonical correlation analysis. J. Shenyang Ligong Univ. **6**, 004 (2013)
21. Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics. Int. J. Comput. Vis. **106**(2), 210–233 (2014)
22. Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Transductive multi-view zero-shot learning. IEEE Trans. Pattern. Anal. Mach. Intell. **37**(11), 2332–2345 (2015)
23. Y. Luo, Y. Wen, D. Tao, J. Gui, C. Xu, Large margin multi-modal multi-task feature extraction for image classification. IEEE Trans. Image Process. **25**(1), 414–427 (2016)
24. D. Wu, L. Pigou, P. J. Kindermans, N. D. H. Le, L. Shao, J. Dambre, J. M. Odobez, Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Trans. Pattern. Anal. Mach. Intell. **38**(8), 1583–1597 (2016)
25. G. Andrew, R. Arora, J. Bilmes, K. Livescu, in *Proceedings of the 30th International Conference on Machine Learning, vol. 28*, ed. by S. Dasgupta, D. McAllester. Deep canonical correlation analysis (PMLR, Atlanta, 2013), pp. 1247–1255. http://proceedings.mlr.press/v28/andrew13.html
26. W. Wang, R. Arora, K. Livescu, J. Bilmes, in *International Conference on Machine Learning*, ed. by F. Bach, D. Blei. On deep multi-view representation learning (PMLR, Lille, 2015), pp. 1083–1092
27. A. Shahroudy, T. Ng, Y. Gong, G. Wang, Deep multimodal feature analysis for action recognition in RGB+D videos. IEEE Trans. Pattern. Anal. Mach. Intell. **40**(5), 1045–1058 (2018). https://doi.org/10.1109/TPAMI.2017.2691321
28. Ç. Gülçehre, Y. Bengio, Knowledge matters: importance of prior information for optimization. J. Mach. Learn. Res. **17**(8), 1–32 (2016)
29. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network. Stat. **1050**, 9–9 (2015)
30. V. Nair, G. E. Hinton, in *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*. Rectified linear units improve restricted Boltzmann machines (Omnipress, USA, 2010), pp. 807–814
31. P. Ramachandran, B. Zoph, Q.V. Le, searching for activation functions. CoRR. **abs/1710.05941** (2017). http://arxiv.org/abs/1710.05941
32. J. Han, C. Moraga, in *From Natural to Artificial Neural Computation*, ed. by J. Mira, F. Sandoval. The influence of the sigmoid function parameters on the speed of backpropagation learning (Springer, Berlin, 1995), pp. 195–201
33. S. Ioffe, C. Szegedy, in *Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 37*, ed. by F. Bach, D. Blei. Batch normalization: accelerating deep network training by reducing internal covariate shift (PMLR, Lille, 2015), pp. 448–456
34. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR. **abs/1409.1556** (2014). http://arxiv.org/abs/1409.1556
35. K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Deep residual learning for image recognition (IEEE, Las Vegas, 2016), pp. 770–778
36. B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**(5), 1299–1319 (1998)
37. P. Rai, H. Daume, in *Advances in Neural Information Processing Systems 22*, ed. by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta. Multi-label prediction via sparse infinite CCA (Curran Associates, Inc., Vancouver, 2009), pp. 1518–1526
38. A. Sharma, A. Kumar, H. Daume, D. W. Jacobs, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Generalized multiview analysis: a discriminative latent space (IEEE, Providence, 2012), pp. 2160–2167
39. N. Quadrianto, C. Lampert, in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ed. by L. Getoor, T. Scheffer. Learning multi-view neighborhood preserving projections (ACM, New York, 2011), pp. 425–432
40. N. Chen, J. Zhu, F. Sun, E. P. Xing, Large-margin predictive latent subspace learning for multiview data analysis. IEEE Trans. Pattern. Anal. Mach. Intell. **34**(12), 2365–2378 (2012)
41. E. Isaacson, H. B. Keller, *Analysis of numerical methods*. (Courier Corporation, New York, 1994)
42. J. Xu, Y. Li, X. Wei, X. Peng, Y. Lu, in *Signal Processing (ICSP), 2016 IEEE 13th International Conference On*. Object recognition with multi-source images based on kernel dictionary learning (IEEE, Chengdu, 2016), pp. 1100–1105
43. X. Peng, Y. Li, J. Luo, J. Xu, Y. Lu, in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. Multi-modal scene categorization using multi-tasks learning (IEEE, Chengdu, 2016), pp. 1106–1111
44. G. K. Demir, K. Ozmehmet, Online local learning algorithms for linear discriminant analysis. Pattern Recogn. Lett. **26**(4), 421–431 (2005)
45. N. Dalal, B. Triggs, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Histograms of oriented gradients for human detection (IEEE, San Diego, 2005), pp. 886–893. https://doi.org/10.1109/CVPR.2005.177
46. C. Cortes, V. Vapnik, Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
47. H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, speeded-up robust features (SURF). Comp. Vision Image Underst. **110**(3), 346–359 (2008)
48. R. Rubinstein, T. Peleg, M. Elad, Analysis k-SVD: a dictionary-learning algorithm for the analysis sparse model. IEEE Trans. Signal Proc. **61**(3), 661–677 (2013)
49. S. Bahrampour, N. M. Nasrabadi, A. Ray, K. W. Jenkins, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kernel task-driven dictionary learning for hyperspectral image classification (IEEE, Brisbane, 2015), pp. 1324–1328. https://doi.org/10.1109/ICASSP.2015.7178185
50. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. Nat. Neurosci. **2**(11), 1019–1025 (1999)
51. A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
52. K. V. D. Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern. Anal. Mach. Intell. **32**(9), 1582–1596 (2010)
53. D. L. Neda Salamati, G. Csurka, in *Proceedings of the British Machine Vision Conference*. Combining visible and near-infrared cues for image categorisation (BMVA Press, Scotland, 2011), pp. 49–14911. https://doi.org/10.5244/C.25.49
54. J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, I. B. Ayed, HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. CoRR. **abs/1804.02967** (2018). http://arxiv.org/abs/1804.02967