



The image annotation algorithm using convolutional features from intermediate layer of deep learning

Yuantao Chen¹ · Linwu Liu¹ · Jiajun Tao¹ · Xi Chen¹ · Runlong Xia² · Qian Zhang³ · Jie Xiong⁴ · Kai Yang³ · Jingbo Xie²

Received: 1 April 2020 / Revised: 7 September 2020 / Accepted: 16 September 2020 /

Published online: 28 September 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

The automatic image annotation is an effective computer operation that predicts the annotation of an unknown image by automatically learning potential relationships between the semantic concept space and the visual feature space in the annotation image dataset. Usually, the auto-labeling image includes the processing: learning processing and labeling processing. Existing image annotation methods that employ convolutional features of deep learning methods have a number of limitations, including complex training and high space/time expenses associated with the image annotation procedure. Accordingly, this paper proposes an innovative method in which the visual features of the image are presented by the intermediate layer features of deep learning, while semantic concepts are represented by mean vectors of positive samples. Firstly, the convolutional result is directly output in the form of low-level visual features through the mid-level of the pre-trained deep learning model, with the image being represented by sparse coding. Secondly, the positive mean vector method is used to construct visual feature vectors for each text vocabulary item, so that a visual feature vector database is created. Finally, the visual feature vector similarity between the testing image and all text vocabulary is calculated, and the vocabulary with the largest similarity used for annotation. Experiments on the datasets demonstrate the effectiveness of the proposed method; in terms of F1 score, the proposed method's performance on the Corel5k dataset and IAPR TC-12 dataset is superior to that of MBRM, JEC-AF, JEC-DF, and 2PKNN with end-to-end deep features.

Keywords Deep learning · Image annotation · Convolutional results · Positive mean vector · Eigenvector

✉ Yuantao Chen
chenyt@csust.edu.cn

1 Introduction

In only two decades, automatic image annotation has become been a research hotspot in the fields of image processing [3], image classification [4], image segmentation [5], computer vision and pattern recognition [19], among others [11]. The success of image annotation tasks mainly depends on the annotation model and visual feature vectors used, wherein the quality of the visual feature vector determines the upper limit of image annotation quality. In recent years, as image annotation models have become more and more mature, visual feature vectors have increasingly become the decisive factor for image annotation effects. Image annotation technology implements keywords that the semantic content of the image, thereby narrowing the gap between the underlying visual features of the image and the high-level semantic tags, improving the efficiency and accuracy of image retrieval, in image and video retrieval, scene understanding, and human-machine fields such as interaction have broad application prospects. However, due to the semantic gap, image annotation is still a challenging task, and it has always been a research hotspot in the field of computer vision.

Before 2016, image features in the research fields of computer vision and pattern recognition were artificial features designed by domain experts. The quality of these artificial features thus depended primarily on the experts' domain knowledge and experience. In complex computer vision applications, it is difficult to extract high-quality feature vectors from artificial features. In 2012, Alex et al. [1] built a deep learning model based on convolutional neural network (AlexNet) that won first place by an overwhelming margin in the ImageNet image sort competition. Since then, the era of deep learning was opened. Subsequently, researchers have proposed many excellent network models based on the AlexNet network [1], such as VGG-16 [30], GoogleNet [32], ResNet [17], etc. Before 2012, image features in the field of computer vision were referred to as artificial features; moreover, the features extracted via deep learning after 2012 are termed deep features (Johnson et al. 2015). Unlike artificial features, the latter is an end-to-end feature extraction processing procedure that does not require human involvement. For image feature extraction processing tasks, newer models can directly output high-quality deep features at the output-port after training the original image input at the input-port through a complex model. As it is based on high-quality end-to-end feature vectors, deep learning has made many breakthroughs in image classification and has been implemented in many research fields.

The reason why deep learning is able to make such breakthroughs depends largely on its complex network structure. In order to achieve better effects, contemporary deep learning network structure design is becoming more and more complex, while the number of layers (depth) is also steadily increasing. However, these millions of training model parameters require not only a large amount of training sample support, but also huge time overhead and high-end hardware configuration; these factors limit the application of deep learning. For example, the AlexNet network model [1] has a total of 61 million parameters, while the subsequently proposed VGG-16 model [30] has 138 million. Moreover, when sufficient training samples exist, the model training is adequate and complex deep learning can achieve the desired results; in the real world, however, it is difficult to provide sufficient training samples for most applications, which often leads to model overfitting, etc. and results in poor model training quality. Accordingly, scholars in this field have proposed some solutions to the above shortcomings, such as fine-tuning training based on pre-training models and applied to complex labeling algorithms [2, 7, 18, 20, 31, 34]. Although these methods have achieved good labeling results, they are still unable to extract high-quality deep features suitable for image annotation [27, 37].

In the remainder of this paper, section 2 introduces some important background knowledge by presenting the related works. Section 3 illustrates the improved deep feature extraction method which section 4 proposes the novel image annotation method. Through experiments on the Corel5k dataset and IAPR TC-12 dataset, we then compare and analyze our proposed method alongside several methods such as MBRM [16], JEC-AF [28], JEC-DF [33] and 2PKNN [29]. Finally, we summarize the present work and present some venues for future research.

2 Related works

Inspired by the successful application of deep learning methods in the field of image classification research, some researchers have attempted to apply deep learning methods to the automatic image annotation field. In 2014, Gong et al. [15] took the lead in applying deep learning to image annotation. Since then, more and more scholars have designed and conducted research into image annotation based on deep learning methods. In Table 1, the effect of image annotation models based on deep learning is compared with that of the traditional annotation models. The most popular and mainstream dataset used for experiments in this field is Corel5K dataset.

As can be seen from Table 1, compared with traditional image annotation methods, deep learning methods achieve improved performance, but not significantly so. In particular, when the network model is more complex (such as the VGG-16 network [30], the annotation performance actually decreases. The main reason for this is that small-scale image data cannot meet the training requirements of complex network structure models. Due to the over-fitting phenomenon, if there is insufficient support from the training dataset, these complex network models cannot achieve the ideal annotation effect; moreover, the more complex the network structure of the deep learning model, the worse the labeling performance will be under these circumstances. At the same time, this deep learning training method requires a huge time cost and high-end hardware configurations. In light of the above deficiencies, Li et al. [8] proposed a migration learning method. While applying migration learning to image annotation improved the effectiveness, the space-time overhead is still large and high-end hardware configurations are still required in the training process.

Although many deep learning models have a good theoretical foundation and a theoretically sound network structure, these complex network models cannot achieve the ideal annotation effect without sufficient support from the training dataset [9, 10, 12]. At the same time, the huge space-time overhead and high-end hardware configuration required for deep model training limits the application of these models. Therefore, scholars in the field have turned their research focus towards more complex annotation algorithms or better feature expression

Table 1 Comparison of the differences in effect between annotation models based on deep learning and traditional annotation models

Annotation Models	Feature Vectors	Precision	Recall	F1 Score	N ⁺
MBRM [16]	Artificial Features	24.00%	25.00%	24.88%	122
JEC-AF [28]	Multiple Artificial Features	27.00%	32.00%	29.28%	139
CAFFE [15]	Deep Learning Features	29.28%	33.00%	31.03%	146
AlexNet [1]	AlexNet End-to-end Features	25.80%	29.33%	27.45%	130
VGG-16 [7]	VGG-16 End-to-end Features	0.21%	1.92%	0.38%	5

quality; for example, merging deep features with other features for image annotation tasks [13, 14, 21, 22].

At present, the deep learning-related research on image annotation is mainly divided into three categories. Firstly, designing new network structures or improving existing models; for example, by modifying the loss function and the number of output categories and training using the object dataset, the original model can be made suitable for image annotation. Secondly, based on the pre-training model, they can modify the fully connected and output layers of a trained network model on a large dataset so that it fits the object dataset, while the other parameters and aspects of network structure remain unchanged (that is, they fine-tune the network based on the existing network weight). One example of this is the migration learning methods proposed by Chen et al. [6]. Thirdly, the pre-training model is directly used to complete feature extraction of the object dataset, after which other complex algorithms are used to complete the annotation, or the deep features are fused with other features for image annotation purposes. For example, Makadia et al. [28], who proposed the famous image annotation model CMRM [16] and MBRM [16], applied deep learning features to complex labeling algorithms such as JEC [28] and 2PKNN [33] and achieved a good annotation effect. This method's success is mainly due to the subsequent complex annotation model or multiple feature fusion. However, the problems of how to extract high-quality deep features suitable for image annotation and how to design an efficient annotation model when the dataset is small have still not been solved. In view of the above problems, this paper attempts to extract high-quality deep features under the conditions of limited resources and insufficient data, and accordingly proposes an innovative image annotation model.

3 Visual feature using image representation

3.1 Image themes modeling

In the field of image processing, images have the same underlying features, such as edges information, visual shapes, geometric changes, illumination changes, etc. These features can be applied to different tasks such as image classification [23, 35], object recognition [24, 36], automatic annotation [25, 26], etc., therefore, large-scale image training can be performed. The pre-trained network model is treated as a generic feature extractor that applies the extracted generic image features to new tasks.

At first, the themes model has mainly used in the researching field of text classification. In the processing of text modeling and analysis, it mainly involves three levels: corpus, document and word. The themes model classifies documents by analyzing the theme of words in corpus. In recent years, in the field of computer vision, researchers try to use the theme model to model and analyze image data. In the process of image subject modeling, image datasets, images and visual words corresponding to corpora, documents and words respectively. The basic attribute of image is unstructured pixel. According to different tasks, researchers had proposed different methods to capture different information of image, such as color, texture, shape and spatial relationship. However, single feature can't express the rich information of image. In order to capture enough information to model semantic concepts, it is necessary to integrate a variety of image features to complete the operation of image annotation. The visual word histogram can effectively integrate a variety of different image features, and the quantified features were called visual words.

Figure 1 is an intuitive representation of a visual word histogram. Visual histogram is also called bag of visual words. The first image was originally labeled “sky, mountains, sea, buffalo and grass”. The second image was originally labeled “sky, trees, grass and zebra”. In the processing of image processing, regular meshes are used to divide the image into equal sized blocks, each of which is represented by its own color and texture features. It can be seen from Fig. 1 that the visually similar areas are quantized into the same visual words, then the content of whole image can be described by counting the visual words. By comparing the visual words described by the two images, we can see that the visual histogram can capture the co-occurrence information of the image. The visual word #1 corresponds to the “sky” in the image. Both images obviously contain the area of “sky”. Therefore, the block count of visual word #1 in both images is more. The second row image uses not only the visual word #1 to describe the area of “sky”, but also the visual words #8 and #20. The first row image contains “sea” and “mountain” and is represented by visual words (#15, #16 and #18) and visual words (#2 and #4), respectively, while the second image does not contain visual information of “sea” and “mountain”, so the block count of these visual words is almost zero.

The x axis in Fig. 1 and Fig. 2 is the number of visual words. The basic idea of theme modeling for images is to associate visual words containing image semantic concepts with potential themes of image datasets, and main semantic themes can be modeled as multiple distributions on visual words. Figure 2 can show the distribution of the main semantic themes “sky”, “grass” and “sea” on the visual words (the visual words here are consistent with Fig. 1).

It can be seen from the three distributions in Fig. 2 that different probability distributions can be obtained by modeling the visual words with different regional characteristics. Figure 2(a) is to model the visual words related to “sky”, which corresponds to the higher probability that the visual words with more counts of “sky” areas in Fig. 1 will be displayed in multiple distributions. Figure 2(b and c) respectively correspond to the probability distribution of visual words of the semantic theme “grass” and “sea” in the image, although the second

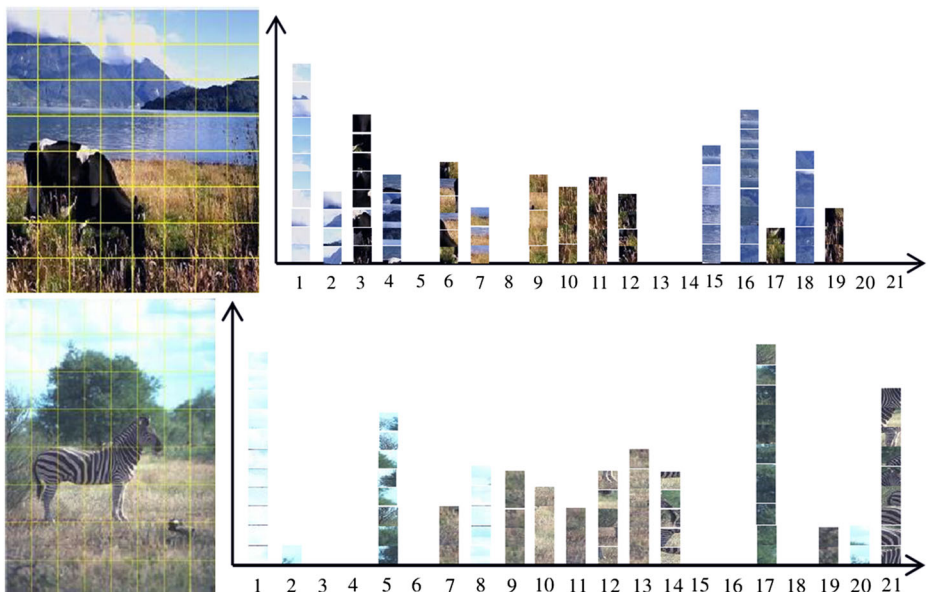


Fig. 1 Visual expression of word histogram

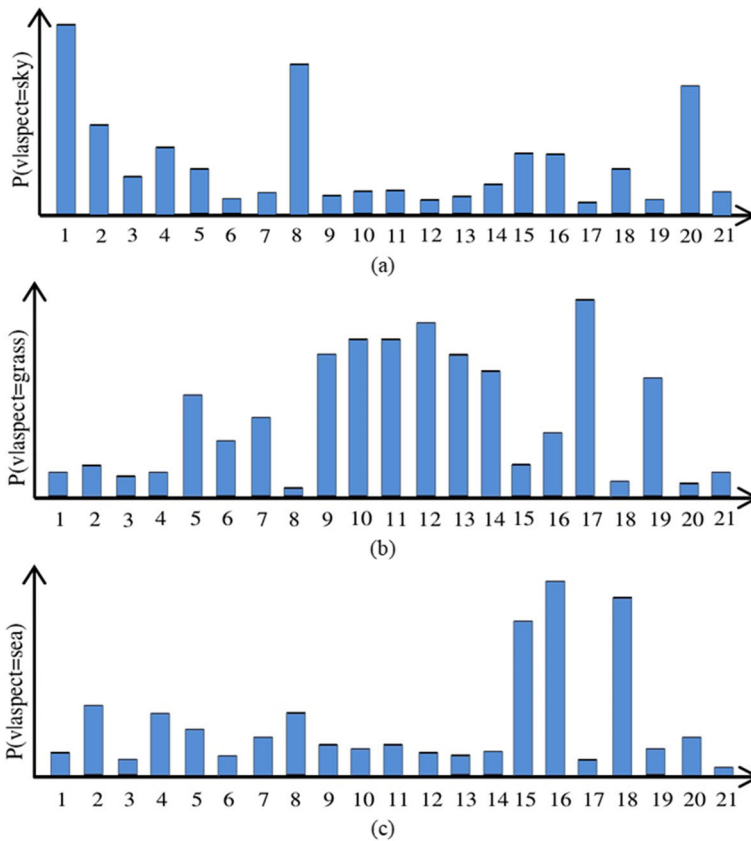


Fig. 2 Three main semantic themes: (a) sky, (b) grass, (c) sea of the distribution in visual words

image in Fig. 1 does not contain the visual information of “sea”. Similarly, the semantic themes “buffalo”, “mountain” and “tree” can be modeled in the same way. Therefore, the bag of visual word representation of different images can be modeled as a mixture of different themes, with only different weights assigned to each theme to represent its actual content. For example, the first image in Fig. 1 can be represented by a mixture of five themes: “sky, mountain, grass, sea and buffalo”. These mixed themes capture some common semantic information in the image dataset, so the image can be decomposed into a mixture of potential themes.

3.2 Image representation

Because a single feature usually can’t express image information well, this section has used bag of words model to integrate SIFT feature and HOG feature to represent image. The bag of words model was first used in the field of text. In this model, the words in the document are regarded as independent, and the bag of word has used to represent the document, and the bag is filled with independent and disordered words. The bag of words model had used a visual dictionary to quantify the words in the document, so a document can be represented by a word histogram. When the bag of words model has applied to the image processing field, the image can be regarded as a “bag of visual words”. Then, each image can be represented by the

histogram corresponding to the distribution of visual words. In the paper, the image has represented by the bag of words model in the following three steps, and the process of image processing with the bag of words model has been shown in Fig. 3.

- (1) Feature Extraction. Considering that the image is affected by many factors such as illumination, angle of view, occlusion and background information, a single feature can easily lead to the loss of image information. In the paper, SIFT features and HOG features with good properties were combined to represent the image.
- (2) Generate Visual Dictionary. As the main tool of quantifying visual words in image dataset, the main idea of visual dictionary is to divide the feature space of samples reasonably, and the feature vectors falling into the same partition area are regarded as the mapping range of same visual words. The classical methods of constructing visual dictionary are clustering algorithms, such as K-Means clustering and parallel K-Means clustering. In the paper, K-Means clustering method has been used to construct a visual dictionary. Firstly, some samples were selected from the dataset, and then the SIFT features and HOG features of these images are grouped into several categories, one cluster represents a visual word.
- (3) Histogram Statistics. After constructing the visual dictionary, we can quantize the features of the images in the dataset. The essence of feature quantization is to map the features of all images to the visual dictionary, and then make histogram statistics for the visual words of each image. In the process of feature quantization, the distance between two feature vectors has been calculated firstly, and then the clustering center closest to the feature vector is found in the visual dictionary, then the feature vector can be mapped to the visual word. By counting the number of visual words appearing in each image, the histogram of visual words corresponding to this image can be generated. By simply connecting the histograms of various features, the visual histogram of each image can be

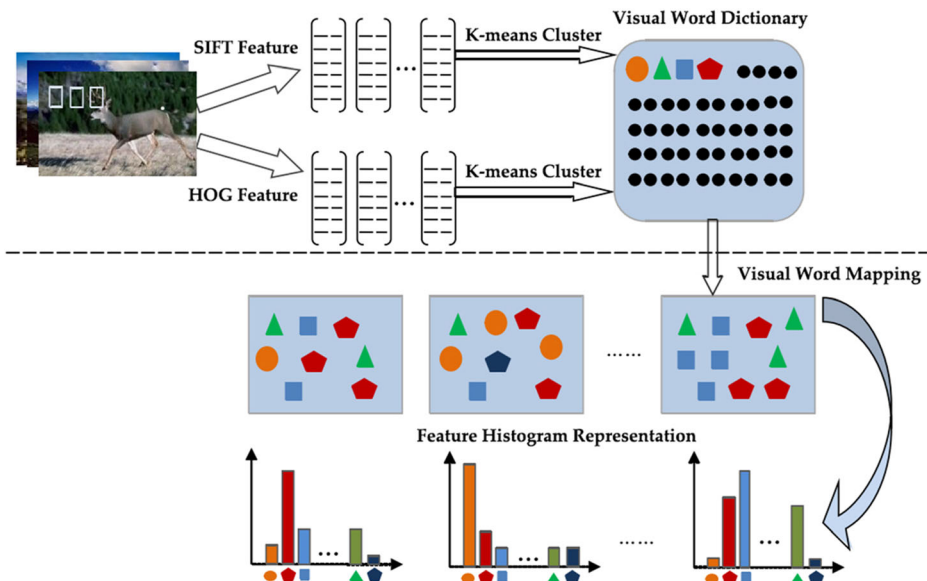


Fig. 3 Schematic diagram of image representation with Bag of Words model

obtained. For example, SIFT features and HOG features were used in the paper are clustered to obtain N_S and N_H visual words respectively, then an image d_i can be represented by $N_S + N_H$ dimensional histogram.

$$v(d_i) = \{n(d_i, v_1), \dots, n(d_i, v_j), \dots, n(d_i, v_{N_S}), n(d_i, v_{N_S+1}), \dots, n(d_i, v_{N_S+N_H})\} \quad (1)$$

Wherein, $n(d_i, v_j)$ represents the number of visual words v_j in image d_i .

4 Improved deep feature extraction method

At present, the end-to-end features extracted by the deep learning model can be regarded as the global features of the original image. While this improved method has achieved great success rates in the research field of image classification, it has not achieved significant results in the field of image annotation tasks. In deep learning-based image classification, the global feature that uses only the output layer at the end of the improved model, ignoring the output characteristics of the intermediate layer, is called the end-to-end model.

However, according to the theory of deep learning model, each layer has its own value when layering the abstraction of the image features; their receptive fields are different, and the extracted features are described in different scopes. The fully connected layer at the end of the network model depicts the global visual features of the image, while the intermediate layer depicts the regions or local features of the image. The perceptive field of the mid-level convolutional kernel of the deep learning model is small, but the number of layers is large. It is easier for the mid-level convolutional kernel to capture local or regional features; therefore, the mid-level features are better at characterizing objects in multiple-object or complex images. Moreover, directly extracting the mid-level features can avoid the high temporal and spatial overhead of the deep learning from the fully connected layer. In this paper, the intermediate convolutional layer features of deep learning are extracted, and the feature vectors of the images are generated by means of sparse coding theory.

The extraction process of generating visual features is shown in Fig. 4. The feature generation process is as follows:

- (1) First, extract the intermediate layer output features of the pre-trained deep learning model, expressed as $F \in \mathbf{R}^{(K \times W \times H)}$. K represents the number of feature maps, while W and H represent the width and height of feature map, respectively. Normalize the features and transform them into two-dimensional Eigen matrices, expressed as $F_{(W \times H, K)}$.
- (2) Gaussian normalization is carried out on the original features; moreover, Principal Component Analysis (PCA) is applied to reduce normalized data. The convolutional feature is represented by $F_{\text{new}(W \times H, n)}$; at this time, n represents the reduced dimension.
- (3) K-Means clustering of dimensionally reduced data is implemented to construct m visual ‘vocabularies’. In line with the visual bag-of-word principle, each image is represented as an m -dimensional bag-of-word vector.

- (4) The acquired clustering center points are used to encode the convolutional features in a Vector of Locally Aggregated Descriptor (VLAD). This is converted into the visual feature vector of the image, as shown in formula (2).

$$F_V = \left[\sum_{NN(f_i)=C_1} f_i - C_1, \dots, \sum_{NN(f_i)=C_m} f_i - C_m \right] \tag{2}$$

Here, f_i represents the feature of the graph block, $[C_1, C_2, \dots, C_m]$ represents the cluster center point, and $NN(f_i)$ is the cluster center closest to f_i .

5 The proposed image annotation method

The vectors of artificial features are often used as the statistics of underlying vision. However, their visual patterns are not significant and their semantic level is inferior. Therefore, image classification and annotation models based on artificial feature vectors tends to be more abstract, with more complicated algorithms and higher space-time cost than other existing solutions.

Compared with traditional artificial features, the deep learning mid-level features have a significant visual mode and a high semantic level. Strong visual and semantic characterization can be achieved when using sparse coding theory such as visual lexicon. If visual feature vectors can be constructed for each text vocabulary, then the confidence problem of calculating lexical membership in images can be resolved. The traditional image annotation problem is converted to the similarity problem of calculating two visual feature vectors: namely, the text vocabulary visual feature vector and the image visual feature vector. The space-time overhead based on this text lexical visual feature vector annotation method is very small and independent of the training dataset size. These approaches are also more capable of handling large datasets than traditional algorithms.

The system structure diagram of the image annotation method proposed in this paper is presented in Fig. 5. In the training phase, the deep features of all training images are extracted and formed into VLAD vectors, which comprise the image visual feature gallery.

A positive sample mean vector method is used to construct a visual feature vector for each text vocabulary that can represent its most essential visual information, thus forming a positive case mean vector lexicon that contains all corresponding vocabulary features. In the annotation and testing phase, the feature vector of the testing image is extracted online and its VLAD vector is generated. The VLAD feature vector of the testing image is then used to calculate the

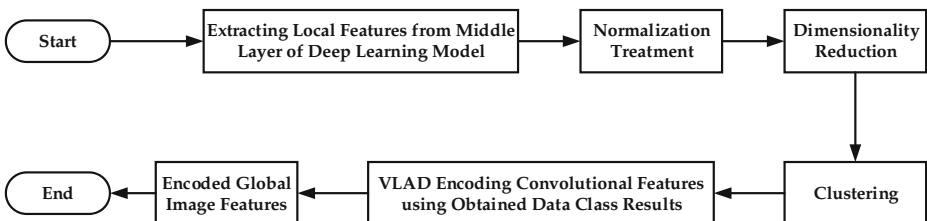


Fig. 4 Flowchart of generating visual feature

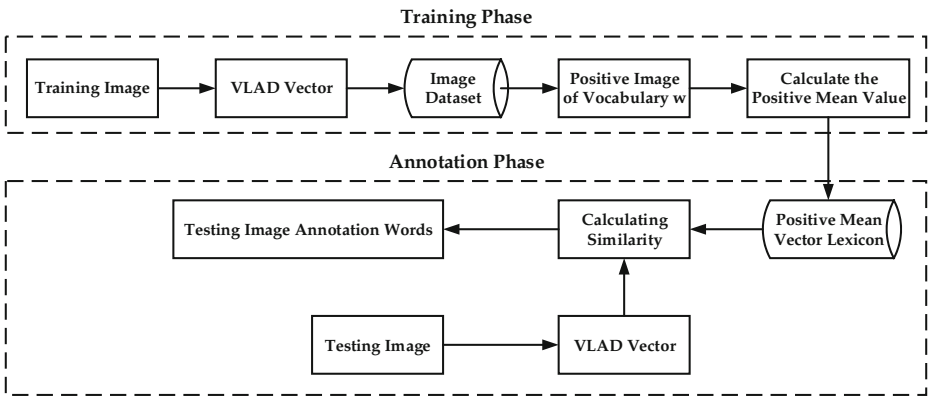


Fig. 5 The system architecture of the proposed image annotation approach

visual similarity with the positive case mean vector of each vocabulary in the positive case vector lexicon. Finally, the text semantic vocabulary corresponding to the largest feature vector of the similarity is selected as the vocabulary label of the test image. The following mainly describes the central concept behind and method used to construct the positive value vector of the positive examples.

In traditional visual dictionary representation methods, if there are M visual vocabularies, the visual dictionary representation method is equivalent to constructing an M -dimensional visual feature space; as each image is an object in the space, it can be represented linearly by M base feature vectors. M is the number of visual words. The value of M depends on the number of visual words. Unlike some existing annotation methods that fix the number of annotations for each image, the number of annotations for each image in the proposed method isn't a fixed value. Different images may have different numbers of annotations, which is more in line with the actual annotation situation. For images with simple semantics, the visual features of the image only contain the feature corresponding to a certain label, so the vector output by the model is basically close to the label each time, making the label corresponding to a higher number of times, while other labels appear less frequently. The threshold is filtered out, and the number of final annotations of the model is small. For complex images, the visual features of the image may include features corresponding to multiple tags. After random noise disturbance, each label in multiple labels has a greater probability of becoming The model outputs labels, so after passing multiple tests, the number of appearances of each label in the multiple labels will not be too small, and the final number of labels in the model will be large. From a semantic perspective, each image can be viewed as a combination of several textual concepts. If each textual concept can be represented as a feature vector in the visual feature space, then the visual feature vector of any image can be considered as the linear sum of several textual semantic vocabulary corresponding visual feature vectors, as shown in formula (3):

$$\mathbf{I} = a_1\mathbf{w}_1 + \dots + a_j\mathbf{w}_j + \dots + a_M\mathbf{w}_M \tag{3}$$

Here, the coefficient is expressed as a Boolean type: set to 1 if there is a corresponding vocabulary in the image, and zero otherwise. This represents the visual eigenvector of the vocabulary.

When both the image feature vector and the vocabulary information it contains are known, the visual vector of each text vocabulary can be obtained from the matrix knowledge. However, the following difficulties are encountered by the formulas:

(1) Ideally, the eigenvectors of all semantic objects will be linearly independent, and can be used as the basis vectors for the semantic space. In reality, however, there are related visual patterns between different concepts, meaning that this assumption is difficult to establish strictly. (2) The vocabulary distribution of most image datasets is not balanced, and the number of images corresponding to some low-frequency vocabularies is much lower than the vector dimension. (3) When the feature vector dimension is high, the space-time complexity of the solution also increases; therefore, it is difficult to solve by traditional matrix methods or machine learning methods.

The characteristics of regional and local adjustment and description for deep learning mid-level features are strong, distinguishable and have an inherent semantic characterization ability. This paper proposes a fast annotation method based on positive sample mean vectors. The basic idea is that while the problem is not directly solvable by the other program, for the characteristics of the deep learning mid-level feature, the feature vector of any text vocabulary can be approximated by the mean of all image feature vectors containing the vocabularies. Taking the vocabulary w_j as an example, if N images contain this item, then N images are represented by semantic concept eigenvectors, and formula (4) can be constructed as follows:

$$\begin{cases} \mathbf{I}_1 = a_{1,1}\mathbf{w}_1 + \dots + a_{1,j}\mathbf{w}_j + \dots + a_{1,M}\mathbf{w}_M \\ \mathbf{I}_2 = a_{2,1}\mathbf{w}_1 + \dots + a_{2,j}\mathbf{w}_j + \dots + a_{2,M}\mathbf{w}_M \\ \vdots \\ \mathbf{I}_N = a_{N,1}\mathbf{w}_1 + \dots + a_{N,j}\mathbf{w}_j + \dots + a_{N,M}\mathbf{w}_M \end{cases} \tag{4}$$

Here, the coefficient is expressed as Boolean. If there is a corresponding vocabulary in the image, otherwise 0), while w_j represents the visual feature vector of vocabulary w_j .

This paper proposes that the visual feature vector of vocabulary w_j can be approximated by the sample mean vector containing the lexical positive case, as in formula (5):

$$\mathbf{w}_j \approx \mathbf{v}_j = \frac{1}{N} \sum_{I \in S_j} \mathbf{I} \tag{5}$$

Here, \mathbf{I} represents the feature vector of image I , while S_j represents a positive example image dataset of vocabulary w_j . N is the number of positive example images. Substituting formula (4) into formula (5), the positive sample mean vector \mathbf{v}_j can be expressed by formula (6), as follows:

$$\mathbf{v}_j = \frac{\sum_{i=1}^N a_{i,1}}{N} \mathbf{w}_1 + \dots + \frac{\sum_{i=1}^N a_{i,j}}{N} \mathbf{w}_j + \dots + \frac{\sum_{i=1}^N a_{i,M}}{N} \mathbf{w}_M \tag{6}$$

Here, N is the number of positive examples, while coefficient a is Boolean (a is 1 if there is a corresponding vocabulary in the image, otherwise a is 0) and w_j represents the visual feature vector of vocabulary w_j .

The larger the dataset size, the larger the size of the image subset containing the vocabulary, and the closer the \mathbf{v}_j calculated by Formula (5) is to the vector w_j of the vocabulary w_j ; this is because, when the image size increases, the coefficient w_j of the j^{th} vector $\frac{1}{N} \sum_{i=1}^N a_{i,j}$ is closer to 1, while the coefficient $\frac{1}{N} \sum_{i=1, x \neq j}^N a_{i,x}$ of the other vectors is closer to 0. Therefore, the larger the

dataset size, the closer the lexical visual feature vector \mathbf{v}_j constructed by the positive mean vector method is to the true eigenvector w_j of the vocabulary w_j . Therefore, the positive case mean vector of all keywords is generated according to Formula (4), and the text semantic concept is transformed to visual lexical vector representation, so that the visual feature vector library of the text vocabulary can be constructed.

The annotation process based on the positive case mean vector involves the visual feature vectors of all words in the lexicon being calculated to be similar to the visual feature vectors of the image to be labeled, while some words with the greatest visual similarity are taken as the annotation words of the image. The similarity distance uses L_2 distance, as in formula (7).

$$\text{sim}(I, w_j) = \text{Dist}_{L_2}(\mathbf{I}, \mathbf{w}_j) \quad (7)$$

This is the visual feature vector of the vocabulary, which is the visual feature vector of the image.

The training steps are as follows:

- (1) The image in the training dataset is divided into blocks of size 16×16 , and the SIFT features and HOG features of the image block are respectively extracted.
- (2) The different features of the image are processed by the bag of words model, and after the feature is quantized, the bag of words representation $v(d_i)$ of each image is formed.
- (3) The model is constructed to analyze the training samples represented by the word bag, and the topic distribution φ of each visual word and the potential topic distribution θ_d of each training image are obtained respectively.
- (4) The classification model is constructed using the obtained image potential topic distribution θ_d and the corresponding original annotation.

The annotation steps are as follows:

- (1) For the image to be labeled d_{new} in the training dataset, the word pocket representation $v(d_{new})$ of each image is obtained after step 1 and step 2 of the training process are performed.
- (2) The potential theme distribution θ_{new} of the image to be annotated is learned according to the visual word topic distribution φ obtained by the training process.
- (3) The visual theme distribution θ_{new} of the learned image to be labeled is input into the classifier of the training number, and the sequence of words to be labeled of the image to be labeled is obtained.
- (4) Selecting a plurality of to-be-labeled words with the highest probability to construct a set of annotated words of the image to be labeled.

6 The experimental results and analysis

6.1 Image annotation datasets and evaluation criterion

In order to accurately and objectively compare and evaluate the performance of the proposed method, the popular datasets Corel5k and IAPR TC-12 are selected; these two datasets are the most commonly used experimental datasets in the field of image annotation.

Corel5k is small, comprising 4500 training images and 500 testing images, along with a total of 260 semantic concepts. The overall size of the dataset is larger than the actual dataset size in many practical applications. Moreover, the IAPR TC-12 is larger in size than Corel5k, with a total of 19,623 images (17,663 training images and 1960 testing images) and 291 semantic concepts. The platform is Windows 10 with 64 bits. The hardware configuration is 3.60 GHz with i7–3790 CPU, NVIDIA GeForce GTX 1080 with GPU, and 8GB memory, and the software environment is Matlab2016a. At present, the general image annotation datasets generally have an imbalance of label distribution, that is, different semantic labels have a large variance in the frequency of image collection. For example, in general, the words “sky” and “tree” in the image appear much higher in the dataset than the words “canyon” and “whales”. For the multi-label image datasets Corel5K and IAPR TC-12, the label distribution imbalance problem is shown in Table 2. Approximately 75% of the labels in the Corel5K and IAPR TC-12 datasets appear less frequently than the average label frequency.

When training the model, because the label distribution of the training dataset is unbalanced, the network output value corresponding to the high frequency word is very different from the network output value corresponding to the low frequency word, that is, the learned model is more sensitive to the high frequency tag than the low frequency tag. Thereby, the system has high accuracy for high frequency labeling, while low frequency labeling performance is low. Due to the large number of low frequency tags, the accuracy of the labeling of low frequency tags has an important influence on the overall tagging performance of the model.

In order to improve the labeling performance of low-frequency tags, noise is added to the high-frequency tags, so that the model’s preference for high-frequency tags is appropriately weakened during the training process, which is equivalent to enhancing the low-frequency tags that were originally ignored. The network model improves the labeling performance of low frequency tags.

The performance evaluation index adopts the most extensive Precision, Recall, F1 (F1-Score) and N+ in the image annotation task research field. In order to objectively evaluate the performance of the deep learning intermediate convolutional layer features extracted in the paper. Suppose a keyword is w_i , then:

Table 2 Description of image annotation datasets

Parameter of Dataset	Corel5K	IAPR TC-12
Total number of images	4999	19,627
Total number of labels	260	291
Training Dataset	4500	17,665
Testing Dataset	499	1962
The size of image	192*168	480*360
Average number of labels per image	3.5	5.7
Label average number of images	58.6	347.7
Label minimum number of images	22	153
Label up to the number of images	1004	4999
Number of tags below average frequency	195	217
Proportion of low frequency tags / %	75.0	74.6

Table 3 Experimental result comparison between our method and others on Corel5k dataset

Annotation Model	Feature Vector	Precision	Recall	F1 Score	N+
MBRM [16]	Artificial Features	24.00%	25.00%	24.88%	122
JEC-AF [28]	Multiple Artificial Features	27.00%	32.00%	29.28%	139
JEC-DF [33]	VGG-16 End-to-End Features	31.00%	32.00%	32.00%	141
2PKNN [29]	VGG-16 End-to-End Features	33.00%	30.00%	32.00%	160
Proposed Method	VGG-16 Mid-Level Features	26.93%	41.43%	32.64%	161

$$P^i = \frac{\text{precision}(w_i)}{\text{prediction}(w_i)} \quad (8)$$

$$R^i = \frac{\text{precision}(w_i)}{\text{Ground}(w_i)} \quad (9)$$

Wherein, $\text{precision}(w_i)$ is the number of correctly labeled images in the returned results. $\text{prediction}(w_i)$ is the number of all images automatically labeled as the keyword in the returned results. $\text{Ground}(w_i)$ is the number of images actually labeled as the keyword in the testing dataset, and the average of the accuracy and recall rate of all keywords are recorded as P and R respectively, with the specific formulas as follows:

$$P = \frac{1}{M} \sum_{i=1}^M P^i \quad (10)$$

$$R = \frac{1}{M} \sum_{i=1}^M R^i \quad (11)$$

F1 comprehensively reflects the balance of accuracy and recall rate. The higher the value of F1, the better the performance of the model. The calculation method of F1 is as follows:

$$F_1 = \frac{2 * P * R}{P + R} \quad (12)$$

Table 4 Experimental result comparison between our method and others on IAPR TC-12 dataset

Annotation Model	Feature Vector	Precision	Recall	F1 Score	N+
MBRM [16]	Artificial Features	24.00%	23.00%	23.48%	223
JEC-AF [28]	Multiple Artificial Features	28.00%	29.00%	28.49%	250
JEC-DF [33]	VGG-16 End-to-End Features	28.00%	21.00%	24.00%	237
2PKNN [29]	VGG-16 End-to-End Features	38.00%	23.00%	29.00%	261
Proposed Method	VGG-16 Mid-Level Features	46.15%	32.80%	38.35%	258

In addition, the experiments also uses N^+ as the evaluation index, which reflects the number of keywords accurately predicted at least once in the dataset.

6.2 Parameter settings

In proposed method, the number of model themes determines the dimension of the representation vector in the image, so the number of themes will have a certain impact on the model performance and system efficiency. The number of subjects is too small, the training model can't fully express the internal association of image dataset, and the large number will lead to the decrease of system efficiency. In the experiment, 30, 50, 70, 90, 110, 130 and 150 subjects were selected to experiment on the Corel 5 K dataset. The experimental results are shown in Table 3. When the number of subjects is 90, the system performance is better than others. Therefore, 90 subjects were used as system parameters in the paper.

In addition, the size of the visual dictionary also affects the performance of the proposed model. Its size determines the degree of image information loss. Theoretically, the larger the number of visual dictionaries is, the less the lost features will be, and better experimental results will be obtained. However, a large number of experiments can show that for specific tasks, when the size of visual dictionary reaches a certain range, the performance growth of image annotation is not obvious. With the continuous increase of the visual dictionary, the opposite effect may appear, and also reduce the experimental efficiency. For specific tasks, when the size of visual dictionary is 2000, the experimental effect is the best. If the number of visual words of SIFT features and HOG features is 1000 respectively, an image can be represented by 2000 dimensional histogram.

6.3 The experimental results and comparative analysis

The experimental results are compared with the results of traditional annotation techniques using traditional artificial features (MBRM, JEC) and the application of deep learning features to JEC-AF, JEC-DF, 2PKNN and other complex annotation algorithms. Similar to the deep learning network model in Murthy et al. [29], the deep learning network model in this paper uses a complex VGG-16 network. In line with the network structure and convolutional kernel, two layers of data with Conv5 are selected as the local feature information of original image. The results of experiments on the smaller-scale dataset Corel5k and the larger-scale dataset IAPR TC-12 are shown in Tables 3 and 4, respectively.

The experimental results in Tables 3 and 4 demonstrate that the main performance indicators of the proposed method are not only superior to the annotation models MBRM and JEC using artificial features (whether on the smaller dataset Corel5k or the larger dataset IAPR TC-12), but are also better than the annotation model JEC and 2PKNN using deep learning end-to-end features. The experimental results in Table 3 show that the proposed method is slightly better than other methods on the smaller-scale Corel5k dataset. Moreover,

Table 5 The annotation performance of words with different frequency level on Corel5k dataset

Word Count	Average Number	Precision	Recall	F1 Score
High Frequency Word	144.09	44.55%	19%	51.14%
Medium Frequency Word	23.60	23.68%	25%	31.35%
Low Frequency Word	7.37	12.39%	17.64%	14.56%

Table 6 The annotation performance of words with different frequency level on IAPR TC-12 dataset

Word Counts	Average Number	Precision	Recall	F1 Score
High Frequency Words	812.61	50.19%	40.39%	44.76%
Middle Frequency Words	160.53	47.84%	34.53%	40.11%
Low Frequency Words	69.98	40.43%	23.49%	29.72%

the experimental results in Table 4 show that in the larger-scale dataset IAPR TC-12, all indicators except for N+ (which is slightly lower than 2PKNN) are obviously superior for our solution than other solutions. The comprehensive evaluation index F1 score is higher than that for MBRM, JEC-AF (Artificial Features), JEC-DF (Deep Features), and 2PKNN (Deep Features), with increases of 63%, 35%, 60% and 32% respectively.

The frequency of semantic vocabulary in each dataset is very uneven. The number of vocabulary occurrences of the highest and lowest frequencies in Corel5k dataset are 1004 and 1, respectively; for the IAPR TC-12 dataset, these images are 4998 and 43. In order to further analyze the relationship between word frequency and the constructed lexical feature vector, the labeling effect of different frequency vocabularies in each dataset is counted. The vocabulary is divided into three levels according to word frequency: namely, high-frequency words, medium-frequency words and low-frequency words. The performance statistics of different frequency vocabularies on the Corel5K and IAPR TC-12 datasets are presented in Tables 5 and 6 respectively.




We then analyze the performance of different frequency annotations on the two datasets to present the following three characteristics. Firstly, in the same dataset, the annotation performance increases as the word frequency increases. Secondly, in the smaller-scale Corel5k dataset, the number of low-frequency words is too small (average less than eight samples), while the corresponding annotation performance is significantly lower. Thirdly, as the dataset size increases, the vocabulary samples of various frequencies also increase, such that the performance gap between low-frequency and high-frequency words begins to narrow. This confirms the reasoning in Formula (4) and Formula (5) that the larger the size of the image dataset, the closer the constructed feature vector is to the feature vector of the object vocabulary, and therefore, the better the image annotation effect achieved by the proposed method.

Due to the high complexity of the various artificial feature models involved in 2PKNN and JEC, the time costing of these features is not provided in the existing related data, and the associated models are more complicated. The experimental link failed to complete the contrast experiment in the same experimental environments; therefore, Tables 5 and 6 describe the annotation performance of words with different frequency level on Corel5k and IAPR TC-12 dataset.

Table 7 Experimental result comparison between our method and others (ESP Game)

Annotation Model	Feature Vector	Precision	Recall	F1 Score	N+
MBRM [16]	Artificial Features	18%	19%	18.50%	209
JEC-AF [28]	Multiple Artificial Features	22%	25%	23.40%	224
JEC-DF [33]	VGG-16 End-to-end Features	26%	22%	23.83%	234
2PKNN [29]	VGG-16 End-to-end Features	40%	23%	29.20%	250
Proposed Method	VGG-16 Mid-Level Features	43.74%	33.08%	37.67%	258

Table 8 Predicted and ground-truth labels for Corel5K and IAPR TC-12 datasets.

Dataset	Image	Ground-Truth Labels	Proposed Method
Corel5K Dataset		Coral Fish Ocean	<i>People</i> Coral Fish Ocean <i>Reefs</i>
		Flowers Petals	<i>Leaf</i> Flowers <i>Plants</i> <i>Tulip</i> Petals
IAPR TC-12 Dataset		Man Table	<i>Bottle</i> <i>Clothes</i> <i>Hand</i> Man Table

In theory, however, the time and space complexities of these algorithms are larger than the method proposed in this paper. The proposed method does not require training the deep learning model. The feature extraction time of the Corel5k testing image dataset is 55 s, while for the traditional end-to-end deep learning fine-tuning method, the model training time is eight hours; moreover, the testing image data feature extraction time is 70 s. In the larger-scale IAPR TC-12 dataset, the image feature extraction time of this method is 330 s, while the traditional end-to-end deep learning fine-tuning method model training time is 10 h and the testing image data feature extraction time is 360 s. If the number of testing images is T , the number of training images is N , and the number of visual words included in the dataset is W , the time complexity of JEC and 2PKNN is $O(T \times N)$. The time complexity of the proposed method is $O(T \times W)$, because the number of training images in the dataset N is much larger than the number of vocabularies W (e.g., Corel5k: 4500 training images, 260 words; IAPR TC-12: 17,825 training images, 291 words). Therefore, the time costing of the method proposed in this paper is also significantly lower than that of JEC and 2PKNN, which is also much lower than labeling models such as MBRM.

In addition, in order to verify the adaptability of the proposed method to domain migration, an experiment was completed on the ESP Game dataset; the ImageNet dataset used in the VGG-16 model pre-training in this paper is the natural scene domain image. The two most commonly used datasets in the image annotation field, Corel5k and IAPR TC-12, are mostly natural scene images. The ESP Game dataset has a total of 20,770 images, of which 18,689 are testing images and 2981 are testing images; the dataset also contains 268 semantic concepts. Results of experiments completed on the ESP Game dataset under the same experimental method are presented in Table 6. These results show that the annotation performance of the proposed method is better than that of other methods on the image datasets of different researching fields from the dataset of the pre-trained training model. This allows us to conclude that the proposed method has strong adaptability to domain migration (Table 7).

Table 8 can show an demo of the labeling results of some images after labeling using the proposed method in the paper. It can be seen from Table 8 that although some words predicted by the proposed method (expressed in italic bold font) are not the true labels of the testing

image, from the perspective of the semantic content of the image, these words can correctly express part of the image content. However, these predicted labeled words were judged as wrong labels in the process of calculating the precision and recall rates, which not only did not improve the recall rate, but led to a decrease in the precision rate. Combining the description of the datasets in Tables 3 and 4, the average number of annotations per image in the IAPR TC-12 dataset is 5.7, but in the Corel5K dataset there are only 3.5. This is due to the incomplete annotation of the datasets. The resulting of “weak annotation” is more obvious in the Corel5K dataset, which to a certain extent leads to the low precision of the Corel5K dataset.

7 Conclusions

Since the image is composed of an unstructured pixel array, to complete the automatic image annotation task, it is first necessary to extract effective visual features from these image pixels. Feature extraction can usually be performed on different scales such as global images or regional images. Currently, both methods were used in existing image annotation systems, but the more widely used was based on region features. Global features can be calculated directly from the entire image, while region-based methods need to be firstly processed using image segmentation techniques. Although deep learning has become a research hotspot in recent years, the data threshold and system configuration required to train these models are relatively high, which restricts the application of this approach. In this paper, based on the versatility of the visual model of the mid-level deep learning model, a method of extracting the convolutional characteristics of the mid-level is developed and studied. Based on this, an image annotation method based on positive examples is proposed. Compared with the traditional end-to-end deep features, which rely on large-scale dataset model training, the deep learning mid-level convolution feature extraction method used in the paper does not require a large-scale dataset training model, meaning that the deep feature data volume and hardware costs are reduced. The proposed annotation method has less space-time overhead and is more thus suitable for the processing and online labeling of large-scale datasets. In addition, since the final label vocabulary of the testing image depends mainly on the visual feature vector of the text vocabulary rather than the feature vector of the training image, the proposed method also helps to alleviate the problem of unbalanced training data categories.

Acknowledgments We are grateful to all students and teachers who participated in this study and all the colleagues working to realize this project.

Funding This work was supported in part by the National Natural Science Foundation of China under Grant 61972056, 61772454, 61402053, 61981340416, the Natural Science Foundation of Hunan Province of China under Grant 2020JJ4623, the Scientific Research Fund of Hunan Provincial Education Department under Grant 17A007, 19C0028, 19B005, the Changsha Science and Technology Planning under Grant KQ1703018, KQ1706064, KQ1703018–01, KQ1703018–04, the Junior Faculty Development Program Project of Changsha University of Science and Technology under Grant 2019QJCZ011, the “Double First-class” International Cooperation and Development Scientific Research Project of Changsha University of Science and Technology under Grant 2019IC34, the Practical Innovation and Entrepreneurship Ability Improvement Plan for Professional Degree Postgraduate of Changsha University of Science and Technology under Grant SJCX202072, the Postgraduate Training Innovation Base Construction Project of Hunan Province under Grant 2019–248-51, 2020–172-48.

Compliance with ethical standards

Conflict of interest No potential conflict of interest was reported by the authors.

References

1. Alex K, Ilya S, Hinton G E (2012) ImageNet classification with deep convolutional neural networks. In: proceedings of the 25th international conference on neural information processing systems, Lake Tahoe, Nevada, USA, 3–6 December 2012, pp 1106–1114
2. Budikova P, Batko M, Zezula P (2018) ConceptRank for search-based image annotation. *Multimed Tools Appl* 77(7):8847–8882
3. Chen YT, Wang J, Xia RL, Zhang Q, Cao ZH, Yang K (2019) The visual object tracking algorithm research based on adaptive combination kernel. *J Ambient Intell Humaniz Comput* 10(12):4855–4867
4. Chen YT, Wang J, Liu SJ, Chen X, Xiong J, Xie JB, Yang K (2019) Multiscale fast correlation filtering tracking algorithm based on a feature fusion model. *Concurr Comput*. <https://doi.org/10.1002/cpe.5533>
5. Chen YT, Zhang HP, Liu LW, Chen X, Zhang Q, Yang K, Xia RL, Xie JB (2020) Research on image inpainting algorithm of improved GAN based on two-discriminations networks. *Appl Intell*. <https://doi.org/10.1007/s10489-020-01971-2>
6. Chen YT, Xu WH, Zuo JW, Yang K (2019) The fire recognition algorithm using dynamic feature fusion and IV-SVM classifier. *Clust Comput* 22(10):S7665–S7675
7. Chen YT, Phonevilay V, Tao JJ, Chen X, Xia RL, Zhang Q, Yang K, Xiong J, Xie JB (2020) The face image super-resolution algorithm based on combined representation learning. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-020-09969-1>
8. Chen YT, Liu LW, Tao JJ, Xia RL, Zhang Q, Yang K, Xiong J, Chen X (2020) The improved image inpainting algorithm via encoder and similarity constraint. *Vis Comput*. <https://doi.org/10.1007/s00371-020-01932-3>
9. Chen YT, Tao JJ, Liu LW, Xiong J, Xia RL, Xie JB, Zhang Q, Yang K (2020) Research of improving semantic image segmentation based on a feature fusion model. *J Ambient Intell Humaniz Comput*. <https://doi.org/10.1007/s12652-020-02066-z>
10. Chen Y T, Tao J J, Zhang Q, Yang K, Chen X, Xiong J, Xia R L, Xie J B (2020) Saliency Detection via improved hierarchical principle component analysis method. *Wirel Commun Mob Comput*, vol. 2020, Article ID 8822777
11. Cheng QM, Zhang Q, Fu P, Tu CH, Li S (2018) A survey and analysis on automatic image annotation. *Pattern Recogn* 79(7):242–259
12. Diwakar M, Kumar M (2018) A review on CT image noise and its denoising. *Biomed Signal Process Control* 42:73–88
13. Diwakar M, Kumar M (2018) CT image denoising using NLM and correlation-based wavelet packet thresholding. *IET Image Process* 12(5):708–715
14. Diwakar M, Singh P (2020) CT image denoising using multivariate model and its method noise thresholding in non-subsampled shearlet domain. *Biomed Signal Process Control* 57:101754. <https://doi.org/10.1016/j.bspc.2019.101754>
15. Gong Y C, Jia Y Q, Leung T, Toshev A, Loffe S (2014) Deep convolutional ranking for multilabel image annotation. In: proceedings of international conference on learning representation, Banff, AB, Canada, 14–16 April 2014, <https://arxiv.org/abs/1312.4894v2>. Accessed 14 Apr 2014
16. Guillaumin M, Mensink T, Verbeek J, Schmid C (2009) TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. In: proceedings of IEEE international conference on computer vision, Kyoto, Japan, 27 September–4 October, 2009, pp 309–316
17. He K M, Zhang X Y, Ren S Q, Sun J (2016) Deep residual learning for image recognition. In: proceedings of IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016, pp 770–778
18. Jesus R, Abrantes AJ, Correia N (2011) Methods for automatic and assisted image annotation. *Multimed Tools Appl* 55(1):7–26
19. Ji Q, Zhang LY, Shu XB, Tang JH (2019) Image annotation refinement via 2P-KNN based group sparse reconstruction. *Multimed Tools Appl* 78(10):13213–13225
20. Johnson J, Ballan L, Li F F (2015) Love thy neighbors: image annotation by exploiting image metadata. In: proceedings of IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015, pp 4624–4632
21. Kumar M, Diwakar (2019) A new exponentially directional weighted function based CT image denoising using total variation. *Journal of King Saud University - Computer and Information Sciences*, 31(1), pp. 113–124

22. Kumar M, Diwakar M (2018) CT image denoising using locally adaptive shrinkage rule in tetrolet domain. *Journal of King Saud University - Computer and Information Sciences* 30(1):41–50
23. Liao X, Li KD, Zhu XS, Liu KJR (2020) Robust detection of image operator chain with two-stream convolutional neural network. *IEEE Journal of Selected Topics in Signal Processing* 14(5):955–968
24. Liao X, Yu YB, Li B, Li ZP, Qin Z (2020) A new payload partition strategy in color image steganography. *IEEE Transactions on Circuits and Systems for Video Technology* 30(3):685–696
25. Liao X, Yin JJ, Chen ML, Qin Z (2020) Adaptive payload distribution in multiple images steganography based on image texture features. *IEEE Transactions on Dependable and Secure Computing*:1. <https://doi.org/10.1109/TDSC.2020.3004708>
26. Lu WP, Zhang X, Lu HM, Li FF (2020) Deep hierarchical encoding model for sentence semantic matching. *J Vis Commun Image Represent* 71:102794. <https://doi.org/10.1016/j.jvcir.2020.102794>
27. Luo YJ, Qin JH, Xiang XY, Tan Y, Liu Q, Xiang LY (2020) Coverless real-time image information hiding based on image block matching and dense convolutional network. *J Real-Time Image Proc* 17(1):125–135
28. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation. In *proceedings of European conference on computer vision, Marseille, France, 12-18 October 2008*: pp 316–329
29. Murthy V N, Maji S, Manmatha R (2015) Automatic image annotation using deep learning representations. In: *proceedings of ACM on international conference on multimedia retrieval, Shanghai, China, 23-26 June 2015*, pp 603–606
30. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *proceedings of the 3rd international conference on learning representation, San Diego, CA, USA, 7-9 may 2015*, <https://arxiv.org/abs/1409.1556>. Accessed 10 Apr 2015
31. Sun L, Ma CY, Chen YJ, Zheng YH, Shim HJ, Wu ZB, Jeon B (2019) Low rank component induced spatial-spectral kernel method for hyperspectral image classification. *IEEE Transactions on Circuits and Systems for Video Technology*:1. <https://doi.org/10.1109/TCSVT.2019.2946723>
32. Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S E, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: *proceedings of IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7-12 June 2015*, pp 1-9
33. Verma Y, Jawahar C V (2012) Image annotation using metric learning in semantic neighborhoods. In: *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012, 7574*, pp. 836–849
34. Xu HJ, Huang CQ, Huang XD, Huang MX (2019) Multi-modal multi-concept-based deep neural network for automatic image annotation. *Multimed Tools Appl* 78(21):30651–30675
35. Yu F, Liu L, Shen H, Zhang Z N, Huang Y Y, Cai S, Deng Z L, Wan Q Z (2020) Multistability analysis, coexisting multiple attractors and FPGA implementation of Yu-Wang four-wing chaotic system. *Math. Probl. Eng.*, vol. 2020, Article ID 7530976
36. Yu F, Liu L, Shen H, Zhang Z N, Huang Y Y, Shi C Q, Cai S, Wu X M, Du S C, Wan Q Z (2020) Dynamic analysis, Circuit design and Synchronization of a novel 6D memristive four-wing hyperchaotic system with multiple coexisting attractors. *Complexity*, vol. 2020, Article ID 5904607
37. Zhang JM, Xie ZP, Sun J, Zou X, Wang J (2020) A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection. *IEEE Access* 8:29742–29754

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Yuantao Chen received the B.S. degree in Computer Science and Technology from Jiangnan Petroleum Institute. He received the M.S. degree in Geodetection and Information Technology from Yangtze University. He received the Ph.D. degree in Control Science and Engineering from Nanjing University of Science and Technology in 2014. He is an associate professor at Changsha University of Science and Technology. His research interests include pattern recognition, image processing, etc. Email: chenyt@csust.edu.cn



Linwu Liu received his B.E. degree in Software Engineering from Hunan Institute of Technology. Currently, he is a postgraduate at Changsha University of Science and Technology. His research interests include image processing and pattern recognition. Email: liulinwu@stu.csust.edu.cn



Jiajun Tao received his B.E. degree in Software Engineering from Changsha University. Currently, he is a postgraduate at Changsha University of Science and Technology. His research interests include image processing and pattern recognition. Email: taojiajun@stu.csust.edu.cn



Xi Chen received the Master degree in Computer Science from Changsha University of Science and Technology in 2007. He is an associate professor at Changsha University of Science and Technology. His research interests include artificial intelligence, image processing, big data processing, etc. Email: chentianjun@163.com



Runlong Xia received the B.S. degree in Electronic Commerce from Hunan Normal University in 2010. He is a research assistant at Hunan Institute of Scientific and Technical Information. His research interests include news communication and public opinion analysis, etc. Email: xiarunlong@vip.qq.com



Qian Zhang received the B.S. degree in Electronic and Information Engineering from Xiangtan University in 2003. He is the department manager at Electronic Products Department of Hunan ZOOMLION Intelligent Technology Corporation Limited. His research interests include electronic engineering, intelligent control technology, etc. Email: zhangqian@zoomlion.com



Jie Xiong received the Ph.D. degree in Geodetection and Information Technology from China University of Geosciences in 2012. He is an associate professor at Yangtze University. His research interests include computer application technology, signal processing, etc. Email: xiongjie@yangtzeu.edu.cn



Kai Yang received the Master degree in Mechanical Engineering from Jilin University in 2014. He is an engineer at Technology Department of Hunan ZOOMLION Intelligent Technology Corporation Limited. His research interests include mechanical engineering, intelligent control technology, etc. Email: yangkai@zoomlion.com



Jingbo Xie was born in 1966. He received the B.S. degree in Thermal Power Machinery and Equipment from Northwestern Polytechnical University in 1988. He is an researcher at Hunan Institute of Scientific and Technical Information. His research interests include artificial intelligence, big data processing, etc. Email: xiejb@hnst.gov.cn

Affiliations

Yuanta Chen¹ · Linwu Liu¹ · Jiajun Tao¹ · Xi Chen¹ · Runlong Xia² · Qian Zhang³ · Jie Xiong⁴ · Kai Yang³ · Jingbo Xie²

¹ School of Computer and Communication Engineering & Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114 Hunan, China

² Hunan Institute of Scientific and Technical Information, Changsha 411105 Hunan, China

³ Department of Electronic Products, Hunan ZOOMLION Intelligent Technology Corporation Limited, Changsha 410005 Hunan, China

⁴ Electronics & Information School, Yangtze University, Jingzhou 434023, China