Multi-feature fusion based spatial pyramid deep neural networks image classification

Qingyong Xu[1, 2]*, Shunliang Jiang[2], Wei Huang[2], Longzhen Duan[2], Shaoping Xu[2]

[1]*School of Economic and Management, Nanchang University, Nanchang 330031, China*

[2]*School of Information Engineering, Nanchang University, Nanchang 330031, China*

**Corresponding author e-mail: xyongle@ncu.edu.cn*

Received 1 October 2013, www.cmnt.lv

Abstract

The scalable and efficient multi-class classification algorithm is now a well-known hard problem. Traditional methods of computer vision and machine learning cannot match human performance on images classification tasks. This paper proposes a novel semi-supervised classifier called Spatial Pyramid Deep Neural Networks (SPDNN). SPDNN utilizes a new deep architecture to integrate the ability of neural networks and spatial pyramid model because deep neural networks do not considerable the spatial information. Feature fusion has been more and more important for image and video retrieval, indexing and annotation because of the lack of single feature. We use multiple feature fusion over any single feature instead of pixels of images. The features include color feature, shape feature and texture feature. The performance of experiment shows that the algorithm improved the state-of-the-art image classification.

Keywords: multi-feature fusion, spatial pyramid deep neural networks, image classification

1 Introduction

In the last decades, the availability of digital images produced by scientific, educational, medical, industrial and other applications has increased dramatically. Images have become one of the main sources of information representation in human life. Thus, images retrieval and images classification has become a challenging task. In order to reach the goals, some pattern recognition techniques have been proposed and become a research hotshot. Deep learning methods as one of pattern recognition techniques have become the focus of the study in image processing and computer vision. Recent advances in deep learning methods have led to a widespread enthusiasm among pattern recognition and machine learning researchers [1, 2]. Deep learning move machine learning towards the discovery of multiply levels of representation.

Deep learning is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using deep model architectures composed of multiple non-linear transformations. For deep model, it can extract more sophisticated and invariant feature from original raw input signals. Lower layers aim at extracting simple features, which are clamped into higher layers [7]. Generally speaking, deep architectures can be exponentially more efficient than shallow ones. For shallow architectures, it need more nodes in order to increase performance and leads to more time to train. For deep architectures, it increases deeper layers other than number of nodes. Those make it more efficient than shallow architecture. So the depth of architecture may be more important from the point of view of statistical efficiency [7].

The concept of neural networks started in the late-1800s as an effort to describe how the human mind performed. These ideas started being applied to computational models with Turing's B-type machines and the perceptron. A deep neural network (DNN) as one of deep learning is defined [4, 5] to be an artificial neural network with multiple hidden layers of units between the input and output layers. DNNs can model complex non-linear relationships. The extra layers enable composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network [4, 8]. The main purpose of DNNs is to extract generally useful features from unlabeled data, to detect and remove input redundancies, and to preserve only essential aspects of the data in robust and discriminative representations [9]. DNNs fully unfold their potential when they are big and deep [10].

In recent years, spatial pyramid model has been extremely popular in image classification. Spatial Pyramid is a widely used method for embedding both global and local spatial information into a feature, and it shows good performance in terms of generic image recognition and classification [11]. For spatial pyramid model, the image is divided into a sequence of increasingly finer grids on each pyramid level. Then the features are extracted from every grid cell and are concatenated to form one huge feature vector. Spatial information is usually embedded in the feature extraction process.

Since the emergence of extensive multimedia data, because of the lack of single feature, feature fusion has been more and more important for image and video retrieval, indexing and annotation. Existing feature fusion techniques simply concatenate a pair of different features or use canonical correlation analysis based methods for joint dimensionality reduction in the feature space.

In this paper we propose a novel semi-supervised classifier called spatial pyramid deep neural networks (SPDNN). The SPDNN utilizes a new deep architecture to integrate the

abstraction ability of deep neural nets (DNN) and discriminative ability of spatial pyramid model. For input data, we use multiple feature fusion over any single feature instead of pixels of images because of the lack of single feature and pixels.

The paper is structured as follows: In section 2 we will detail feature fusion. Sections 3 present the framework of our proposed spatial pyramid deep neural networks model. Section 4 asserts the validity of our method by the experiment using COREL 1000 data-set and section 5 draws the conclusions and points out future work.

2 Features fusion

In our lives, there are more and more technology and feature extracting methods to be proposed for image retrieval and image classification in order to increase the precise. In the beginning, the researchers mainly focus on the text based image retrieval(TBIR) using simple feature, then more and more researchers start to research the content based image retrieval(CBIR) because of the limitation of TBIR. If the features come from the entire image, we called it as global based image retrieval (GBIR) and otherwise we called region based image retrieval (RBIR). The features of GBIR are often low-level features from images. RBIR focuses on contents from regions of images not form entire image. Generally speaking, RBIR have better performance than GBIR and TBIR. The performance of the methods including CBIR, GBIR and RBID is depended on features extracting.

The features consist of colour feature, texture feature and shape feature.

The color feature is one of the most widely used visual features and is invariant to image size and orientation. But the color feature does not contain the spatial information. Some CBIR systems employ color to retrieve images such as QBIC system and Visual SEEK. Colour features consist of color moment, color histogram, the edge histogram, Gabor wavelet transform, partial binary image, GIST etc. Color moment is often used for color representation. It contains mean, variance and skewness.

Texture is another important characteristics for image retrieval. Image texture refers to surface patterns which show granular details of an image. Image texture gives us information about the spatial arrangement of color or intensities in an image or selected region of an image, and it can be used for image segmentation or classification. Texture includes edge detection, gray co-occurrence matrices (GLCM), autocorrelation features, Tamara [12] feature etc. Co-occurrence matrix is constructed based on the distance and orientation between pixels. GLCM is one of the most well-known and widely used texture features and is defined based on different combinations of pixel brightness values (i.e., grey levels). It considers the spatial relationship among pixels. Tamura is another important texture feature and consists of coarseness, contrast, directionality, linelikeness, regularity and roughness.

Shape feature is different from color feature and texture feature. Popular shape feature consists of edge histogram, Fourier descriptors, polygonal approximation, invariant moments, curvature scale space, etc. Invariant moments is proposed by Hu [13].

Obviously we cannot cover all features that have been proposed in our experiment. We select some features that can represent images. For color feature, we select the color moment in RGB and HSV color space because it is simple to adopt and effective for retrieval. It mainly describes the image color distribution. For texture feature, we use the GLCM and Tamara. For Shape feature, we use the invariant moments.

3 Spatial pyramid deep neural networks

In order to increase the performance of image retrieval, the machine learning methods are applied. Deep learning which is one of machine learning is proposed in recent years and is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using architectures composed of multiple non-linear transformations. And it is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text.

A deep neural network is one of the most important deep learning methods. For labeled training examples $(x^{(i)}; y^{(i)})$. Neural networks give a way of defining a complex, non-linear form of hypotheses $h_{W;b}(x)$, with parameters W; b that we can fit to our data. A neural network is put together by hooking together many of our simple neurons, so that the output of a neuron can be the input of another. For example, a small neural network as Figure 1.

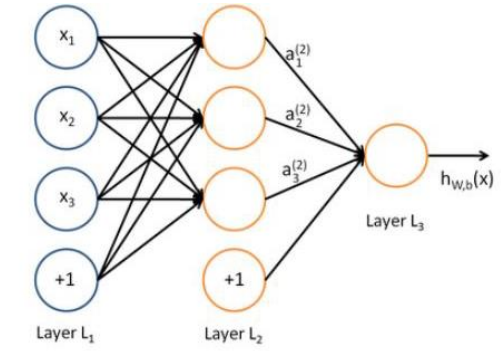

FIGURE 1 The Graph of an NN with hidden

A deep neural network (DNN) is defined [4, 5] to be an artificial neural network with multiple hidden layers of units between the input and output layers. Similar to shallow ANNs, DNNs can model complex non-linear relationships. The extra layers enable composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network. DNNs are typically designed as feed forward networks, but recent research has successfully applied the deep learning architecture to recurrent neural networks for applications such as language modelling [14].

A DNN can be discriminatively trained with the standard back propagation algorithm. The weight updates can be done via stochastic gradient descent using the following Equation (1).

$$\triangle W_{ij}(t+1) = \triangle W_{ij}(t) + \mu \frac{\partial C}{\partial W_{ij}}, \quad (1)$$

where μ is the learning rate and C is the cost function. He choice of the cost function depends on factors such as the learning type (supervised, unsupervised, reinforcement, etc.)

and the activation function. But the deep neural networks do not considerable the spatial information.

A SPDNN is composed of one or more spatial pyramid layers with fully connected layers on top. It also uses tied weights and pooling layers. This architecture allows SPDNN to take advantage of the 2D structure of input data. In comparison with other deep architectures, convolutional neural networks are starting to show superior results in both image and speech applications. They can also be trained with standard back propagation.

4 Experiments and analysis

4.1 EXPERIMENTAL SETUP

In the following we give a detailed description of all the experiments we performed. We evaluate our architecture on various commonly used object recognition benchmarks and improve the state-of-the-art on all of them. The architecture of SPDNN is three layers used for the experiment. The description of the SPDNN is given as following: architecture of the pyramid is 1-4-16; architecture of the deep neural networks is 48-500-500-500-500-500-10(48 is the size of input data and 10 is the output of SPDNN). The architecture has five hidden layers with 500 hidden units and a fully connected output layer. All SPDNN are trained using on-line gradient descent. Initial weights are drawn from a uniform random distribution in the range [-0.05; 0.05] [15].

4.2 DATA-SET

The Corel image database contains a large amount of images containing various contents, ranging from animals and outdoor sports to natural scenes. It is often used for image retrieval system and image classification. There are two subsets. One is the min Corel image set with 1000 images and another is a bigger images set with 10000 images. In our experiment, the Corel 1000 images set is used in order to compare with other results of anthers. The Corel 1000 is a data-set have 1000 labeled high-resolution images with JPEG format belonging to 10 categories with 100 images each. The 10 categories are Africa, Beach, Buildings, Buses, Dinosaurs, Flowers, Elephants, Horses, Food and Mountains. The size of every image is 256×348 or 348×256. The images of every category are shown as Figure 2.

FIGURE 2 Examples of 10 class images

4.3 FEATURE EXTRACTING AND NORMALIZE

The feature extracted is one of the most important for image classification. Obviously we cannot cover all features that have been proposed in our experiment. However, we have tried to make the selection of features as representative and at the state-of-the-art as possible. Generally speaking, the features can be classified into three groups: colour features, texture features and shape features. Some good features is crucial for obtaining competitive performance in classify-cation. For color features, we select color moment proposed by stricker [16]. The colour moment include mean, variance and skewness. It does not need color space quantization and the dimension of feature vectors is low. It can be extracted from RGB and HSV space. For texture features, we select tamura, entropy and gray-level co-occurrence matrices (GLCM). Tamura consist of six texture features (coarseness, contrast, directionality, linelikeness, regularity, and roughness) corresponding to human visual perception: Image entropy is a quantity which is used to entropy measures the randomness of the distribution of intensity levels in bins. Co-occurrence matrix is constructed based on the distance and orientation between pixels. GLCM is one of the most well-known and widely used texture features and is defined different combinations of pixel brightness values (grey levels). It considers the spatial relationship among pixels. For shape features, Hu invariant moment is used. Hu derived these expressions from algebraic invariants applied to the moment generating function under a rotation transformation. They consist of groups of nonlinear centralised moment expressions. The result is a set of absolute orthogonal moment invariants, which can be used for scale, position, and rotation invariant pattern identification. Thus we can obtain 31 features (color:9, tamura:6, entropy:1, GLCM:8, Hu invariant moment:7) for each images. The data-set is 1000×31 array and every row represent an image. By this way, every row represented features of an images which contained the color feature, texture feature and shape feature. Then we used SPDNN to train the train data-set and test the testing data-set using the training results.

After feature extracting, the features are normalized in order to keep as the unified scale because the scale of different feature is not same, so the normalization is needed. Normalization of the feature refers to adjusting values measured on different scales to a notionally common scale and brings the indicators into the same unit. The intention is that these normalized values allow the comparison of corresponding normalized values for different data-set in a way that eliminates the effects of certain gross influences. In our work, 0-1 normalization is use. This is also called unity-based normalization. The formulation as follows:

$$X_{\mathrm{i}} = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}}. \tag{2}$$

By this way, the feature scaling used to bring all values into the rang [0, 1].

4.4 PARAMETERS SETTING

After pre-processing, the parameters must be set for SPDNN. The architecture of SPDNN is three layers used for the experiment. The description of the SPDNN is given as following: architecture of the pyramid is 1-4-16; architecture of the deep neural networks is 31-500-500-500-500-500-10(31 is the size of input data and 10 is the output of SPDNN). The networks have one visible layer, five hidden layer and an out layer. Each hidden lever has 500 hidden units and the

output layer has 10 units. The sign function as feature mapping function is used. The sign functions as follows:

$$f(x) = g(Wx + b) = \frac{1}{1 + e^{-(wx+b)}} . \quad (3)$$

We do the experiment using the learning rate from 0.01 to 1, and the moment from 0.1 to 1. The experiment as Figure 3 shows that the learning rate has great effect on the results, and other elements have little influence on the result.

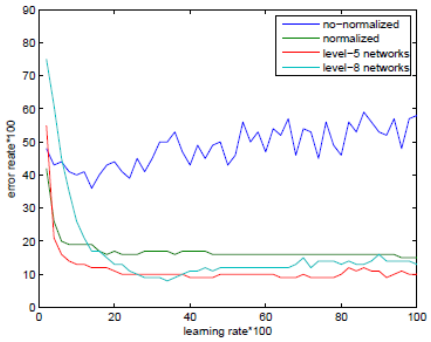

FIGURE 3 Learning rate figure

There are many hyper-parameters involved in training this deep learning method. The parameters in Table 1 which is determined by lots of experiments using different deep learning methods is used to test the effectiveness of our data-set and our experience.

TABLE 1 Parameters lists

Parameters	Value
number of hidden unit	90
learning rate	0.4
zero Masked Fraction	0.5
momentum	0.5
alpha	0.5
weight-decay	0.0001
number epochs	500
number epochs function	sigma

4.5 ANALYSIS

4.5.1 Data Grouping

In our experiment, there are 1000 images. The images was randomly into 10 groups, and each group has 100 samples. The results of groups is showed in Table 2.

TABLE 2 Data-set groups

Groups	Africa	Beach	Building	Buses	Dinosaurs	Flowers	Elephants	Horses	Food	Mountains	Total
1	10	5	10	11	12	9	10	14	9	10	100
2	11	14	6	7	14	9	8	12	10	9	100
3	8	8	10	12	10	8	5	17	11	11	100
4	10	14	11	16	5	12	9	7	4	12	100
5	11	10	16	12	4	11	7	10	7	12	100
6	9	11	11	7	7	11	10	11	7	16	100
7	13	10	9	9	6	10	14	11	10	8	100
8	8	12	11	11	10	10	14	7	10	7	100
9	12	6	9	10	18	11	10	4	12	8	100
10	8	10	7	5	14	9	13	7	20	7	100
total	100	100	100	100	100	100	100	100	100	100	1000

4.5.2 Experiment results

For our groups, we select 9 groups as training set and 1 group as testing set. We do 10 experiments, each with different test sets. 10 results are obtained. The average result of classification as Figure 4.

From the correct rate of every groups, we know that the best is 91% for tenth group. The worst is 78% for seventh group. The average correct rate is 84.2%. It is better than the-state-of-the-art.

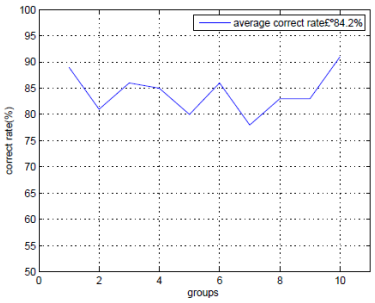

FIGURE 4 Correct rate for every group

4.5.3 Compared with single feature

Table 3 lists the classification results of several common features, including histogram, color straight direction, gray level co-occurrence matrix, color co-occurrence matrix and the results in this paper. It is seen that the average correct classification rate of single features are not more than 70%, and the results of multi feature fusion is achieved 84.2%. It has better performance than single feature.

4.5.4 Compared with different methods

Table 4 lists the common image classification and related scholars are the classification result is at the COREL 1K gallery. It is seen that both in the average correct rate and maximum/minimum rate of correct classification, the SPDNN algorithm has better performance.

TABLE 3 Compared with Single Feature

Feature	Average correct rate (%)
Gray level histogram(size:16)	69.9
RGB histogram(size:16)	67.4
Index histogram(size:16)	57.2
Dominant colour descriptor(size:16)	48.6
Dominant Codebook(size:16)	38.1
Grey Level Co-occurrence Matrix	67.4
Color Co-occurrence Matrix	58.4
Gabor Wavelets	58.8
Scan pattern co-occurrence matrix	50.2
This paper	84.2

TABLE 4 Compared with different methods

Methods	Best	Worst	Average（%）
SIMPLIcity(2013) [26]	98.1/Dinosaurs	33.0/Building	46.7
Edge based(2013) [26]	95.0/Dinosaurs	25.0/Elephants	51.0
Fuzzy Club(2013) [26]	95.0/Dinosaurs	30.0/Elephants	55.9
DD-SVM(2004) [25]	99:7/Dinosaurs	----	81.5
CS_LBP(2012) [25]	96.2/Dinosaurs	31.4/ Mountains	59.1
LEPSEG(2012) [25]	96.0/Dinosaurs	37.2/ Mountains	65.2
LEPINV(2012) [25]	95. 5/Dinosaurs	34.9/ Beach	60.8
Hiremath's method (2007)[25]	95.0/Dinosaurs	30.4/Beach	54.9
Wang-yu-yang method (2010)[28]	95.0/Dinosaurs	30.0/Mountains	59.2
M.Babu Rao, Ch.Kavitha etc method(2013) [26]	99.0/Dinosaurs	55.0/Beach	75.13
Fazal Malik Baharum (2013) [27]	100.0/Dinosaurs	60.0/Food	82.0
This paper	100.0/Dinosaurs	64.0/Building	84.2

Due to the characteristics of the image itself, such as the difference between the differences of different objects, different in the image size, the foreground and background color of the size and not the same image, image classification correct rate of different category has certain difference. From the experimental results, ten types of images, classification of each class is the correct rate of each are not identical, the dinosaur a set of correct classification rate is highest, for 100%, all classified correctly; secondly is the flower and the automobile, the correct rate of classification was 99% and 98%; the correct ratio of less than 80% of the building, Africa, elephant and beach four class.

FIGURE 5 Some misclassification images

In real images, images belong to the same category sometimes have the obvious difference, and the images which belong to different categories and sometimes very similar. This is mainly because the channel between the image low-level features and high-level semantic. The semantics for the same class, both in the form are quite different, image semantic belong to different categories, may form is very similar, this will cause great difficulty for image classification. For example, the beach has 8 images is divided into the mountains of this group, mountains has 8 images is divided into beach in this group, the 16 images in Figure 5.

In Figure 5, the images of first two rows belong to beach but they are misclassified to mountains. The images of last two rows belong to mountains but they are misclassified to beach. As show in Figure 5, the image itself is not much difference and they are very similar.

5 Conclusion

Deep neural networks (DNN) as one of deep learning methods and Spatial pyramid are an active research topic in image processing and computer vision Based on DNN and spatial pyramid, we proposed a new methods called multi-future fusion spatial pyramid deep neural networks (SPDNN). SPDNN utilizes a new deep architecture to integrate the advantage of deep neural networks (DNN) and overcome the disadvantages of DNN without considering the spatial structure of image. Then it is successfully applied to visual data classification. For input data-set, images pixels are replaced by the based features for input of SPDNN. By this way, the size of input data vector can be reduce and keep the information of images. In our experiment, we stochastically classify the images database into 10 groups. The results show that SPDNN has better performance than the-state-of-the-art.

The further work will be explored from two aspects. Firstly, we will study how to determine the scale of deep architecture for various applications and the parameters are decide. Secondly, we will consider that how to improve the performance of region based image retrieval and classification use deep learning method in a large scale data-set.

Acknowledgments

This work was supported by Grants 61363046 & 41261091 & 61163203 approved by the National Natural Science Foundation of China.

References

[1] Markoff J 2012 Giant steps in teaching computers to think like us: neural nets mimic the ways human minds listen,see and execute *International Herald Tribune 24-25 (November)(2012)* 1-8
[2] Larochelle H 2007 An empirical evaluation of deep architectures on problems with many factors of variation *Proceedings of the 24th international conference on Machine learning ACM* 2007 473-80
[3] Lopes N, Ribeiro B 2013 Towards adaptive learning with improved convergence of deep belief networks on graphics processing units *Pattern Recognition* (47) 114-27
[4] Bengio Y 2009 Learning Deep Architectures for AI. Foundations and Trends in Machine Learning 1-127
[5] Schmidhuber J 2013 Deep Learning in Neural Networks: An Overview *arXiv preprint arXiv* 1404.7828
[6] Liou C-Y. 2013 Auto encoder for words *Neurocomputing* **139** 84-96
[7] Bengio Y, Courville A, Vincent P 2013 Unsupervised feature learning and deep learning: A review and new perspectives *IEEE Transaction Pattern Analysis and Machine Intelligence (PAMI)*
[8] Masci J 2011 Stacked convolutional auto-encoders for hierarchical

feature extraction.Artificial Neural Networks and Machine LearningCICANN *Springer Berlin Heidelberg* 52-9
[9] Ciresan D C 2011 Flexible, high performance convolutional neural networks for image classification. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (22)
[10] Harada T, Ushiku Y, Yamashita Y 2011 Discriminative spatial pyramid *Computer Vision and Pattern Recognition (CVPR) 2011 IEEE Conference on IEEE* 1617-24
[11] Tamura H, Shunji M, Takashi Y 1978 Textural features corresponding to visual perception *Systems, Man and Cybernetics IEEE Transactions on* 460-73
[12] Hu M-K 1962 Visual pattern recognition by moment invariants *Information Theory IRE Transactions on* 179-87
[13] Mikolov T 2010 Recurrent neural network based language model *Interspeech*
[14] Ciresan D, Meier U, Schmidhuber J 2012 Multi-column deep neural networks for image classification *Computer Vision and Pattern Recognition (CVPR)*
[15] Stricker M A, Orengo M 1995 Similarity of color images. IST/SPIE Symposium on Electronic Imaging: Science and Technology *International Society for Optics and Photonics*
[16] Rao M Babu, et al. 2013 A new feature set for content based image retrieval *Information Communication and Embedded Systems (ICICES), 2013 International Conference on IEEE*
[17] Hinton G E 2002 Training products of experts by minimizing contrastive divergence *Neural Computation* **14**(8) 1771-800
[18] Subrahmanyam Murala R P Maheshwari R Balasubramanian 2012 Directional local extrema patterns a new descriptor for content based image retrieval *Int J Multimed Info Retr* 191-203
[19] Hiremath P S, Pujari J 2007 Content based image retrieval using color, texture and shape features *Advanced Computing and Communications International Conference on IEEE*
[20] Wang X-Y, Yu Y-J, Yong H-Y 2010 An effective image retrieval scheme using color, texture and shape features *Journal of computer stabdards and interfaces*
[21] Fazal Malik, Baharum Baharudin 2013 Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain *Journal of King Saud University - Computer and Information Sciences* **25**(2) 207-18
[22] Zheng L, Wang S, Qi Tian 2013 Coupled Binary Embedding for Large-scale Image Retrieval *IEEE Transactions on Image Processing (TIP)* (8) 3368-80
[23] Zheng L, Wabg S, Liu Z, Qi Tian 2013 Packing and Padding: Coupled Multi-Index for Accurate Image Retrieval *In: CVPR*