A Novel Technique to Image Annotation using Neural Network

Pankaj Savita
TIT Bhopal

Deepshikha Patel
Professor TIT Bhopal

Amit Sinhal
Professor, TIT Bhopal

Abstract: Automatic annotation of digital pictures is a key technology for managing and retrieving images from large image collection. Traditional image semantics extraction and representation schemes were commonly divided into two categories, namely visual features and text annotations. However, visual feature scheme are difficult to extract and are often semantically inconsistent. On the other hand, the image semantics can be well represented by text annotations. It is also easier to retrieve images according to their annotations. Traditional image annotation techniques are time-consuming and requiring lots of human effort. In this paper we propose Neural Network based a novel approach to the problem of image annotation. These approaches are applied to the Image data set. Our main work is focused on the image annotation by using multilayer perceptron, which exhibits a clear-cut idea on application of multilayer perceptron with special features. MLP Algorithm helps us to discover the concealed relations between image data and annotation data, and annotate image according to such relations. By using this algorithm we can save more memory space, and in case of web applications, transferring of images and download should be fast. This paper reviews 50 image annotation systems using supervised machine learning Techniques to annotate images for image retrieval. Results obtained show that the multi layer perceptron Neural Network classifier outperforms conventional DST Technique.

General Term- Pattern Recognition

Keywords: Image Annotation, Neural Network, MLP, DST

1. INTRODUCTION

Nowadays, the number of digital images is growing with an incredible speed, which makes the image management very challenging for researchers. Automatic image annotation aims to develop methods that can predict the relevant keywords from an annotation vocabulary for a new image. The final goal of automatic image annotation is to assist image retrieval by supplying semantic keywords for search. This capability makes large image database management easy. The image annotation has been extensively researched for more than a decade. There are mainly two methods for automatic image annotation: Statistics models and Classification approaches. Statistics models annotate images by computing the joint probability between words and image features. Image Annotation is regarded as a type of multi-class image classification with a very large number of classes, as larger as the vocabulary size. Therefore, automatic image annotation can be considered as a multi-class object recognition problem which is an extremely challenging task and still remains an open problem in computer vision. In spite of many algorithms proposed with different motivations, the underlying questions are still not well solved-

1) Most of the automatic image annotation systems utilize a single set of features to train a single learning classifier. The problem is: A single feature set, which represents an image category well, may fail to represent other categories. For example, the semantic word “flower” and “tree” may be different in color, so color features may work best, but for “tree” and “grass”, which has similarity in color may be distinguished by texture features. Such kind of problem degrades the performance of the automatic image annotation system when, the number of categories increase. 2) For each image, we often have keywords assigned with the whole image. Here it is not known which regions of the image correspond to these keywords.In this paper, we propose a novel automatic image Annotation system which can tackle the problems mentioned above: 1) our algorithm combines different kinds of feature descriptors to boost the annotation.2) Segments each image into several regions, and establishes one-to-one correspondence between image region and annotation keyword.

2. LITERATURE SURVEY

Recently, a number of models have been proposed for image annotation. One of the first attempts at image annotation was reported by Mori et al. [1], who tiled images into grids of rectangular regions and applied a co-occurrence model to words and low-level features of such tiled image regions. Duygulu et al. [11], described images using a vocabulary of blobs. First, regions are created using a segmentation algorithm like normalized cuts [6]. For each region, features are computed and then blobs are generated by clustering the image features for these regions across images. Each image is generated by using a certain number of these blobs. Their Translation Model applies one of the classical statistical machine translation models to translate from the set of blobs forming an image to the set of keywords of an image.Tsai and Hung [2] reviewed 50 image annotation systems using supervised machine learning techniques to annotate images via mapping the low-level or visual features to high-level concepts or semanticsVailaya et al.[3] proposed a hierarchical classification scheme to first classify images into indoor or outdoor categories, then, outdoor images are further classified as city or landscape. Finally, landscape images are classified into sunset, forest, and mountain classes. In other words, three Bayes classifiers are used for the three-stage classification.Correlation LDA proposed by Blei and Jordan [4] extends the Latent Dirichlet Allocation (LDA) Model to words and images. This model assumes that a Dirichlet distribution can be used to generate a mixture of latent factors. This mixture of latent factors is then used to generate words and regions. Expectation-Maximization is used to estimate this model. In Datta et al. [5], the authors surveyed almost 300 key theoretical and empirical contributions related to image retrieval and automatic image annotation and their subfields. They also

discussed on challenge for systems development in the adaptation on existing image retrieval techniques. Metzler and Manmatha [7], segmented training images, connecting them and their annotations in an inference network, whereby an unseen image is annotated by instantiating the network with its regions and propagating belief through the network to nodes representing the words. Oliva and Torralba [8], showed that images can be described with basic scene labels such as 'street', 'buildings or 'highways', using a selection of relevant low-level global filters. They further showed how simple image statistics can be used to infer the presence and absence of objects in the Scene.Yavlinsky et al. [9] followed the scene-based approach and discussed the possibility of applying global features on image annotation. Their method uses kernel smoothing on nonparametric density estimation and shows that the global features can effectively be used to annotate images even with relatively simple global features. Liu. et. al. [10], They estimated the joint probability by the expectation over words in a pre-defined Lexicon. It involves two kinds of critical relations in image annotation. First is the word-to-image relation and the second is the word-to-word relation. In their experimental results, their Dual Cross-Media Relevance Model (DCMRM) outperformed the last two relevance models [12]-[13] for image retrieval. For image annotation, their model also outperformed previous models [11]-[13]. Monay and Gatica-Perez [14] introduced latent variables to link image features with words as a way to capture co-occurrence information. This is based on latent semantic analysis (LSA) which comes from natural language processing and analyses relationships between images and the terms that annotate them. The addition of a sounder probabilistic model to LSA resulted in the development of probabilistic latent semantic analysis (PLSA) [15]

3. IMAGE SEGMENTATION

In general, visual content of an image can be represented by either global or local features. Global features take all the pixels of an image into account. On the other hand, image segmentation into local contents is able to provide more detailed information of images. In general, there are two strategies for extracting local features. The first one is to partition a set of fixed sized blocks or tiles (see Fig. 1 for some examples) and the second for a number of variable shaped regions of interest. After performing block or region based segmentation, low-level features can be extracted from the tiles or regions for local feature representation.

4. IMAGE LOW-LEVEL FEATURES

In general, low-level features such as color, texture, shape, and spatial relationship are extracted to represent image features.

4.1. Color

Color is the most used visual feature for image retrieval due to the computational efficiency of its extraction. All colors can be represented as the combinations of the three colors, so-called primary colors: red (R), green (G), and blue (B). There are some other color spaces for representing the color feature, such as HSV, L*u*v*, YIQ, etc.

4.2. Texture

Texture is an important element of images for surface, object identification, and region distinctions. In addition to colors, texture is extracted to classify and recognize objects and scenes. Texture can be regular or random.

4.3 Shape

Shape is one of the most important features for describing the contents or objects of an image. Compared with color and texture features, shape features are usually

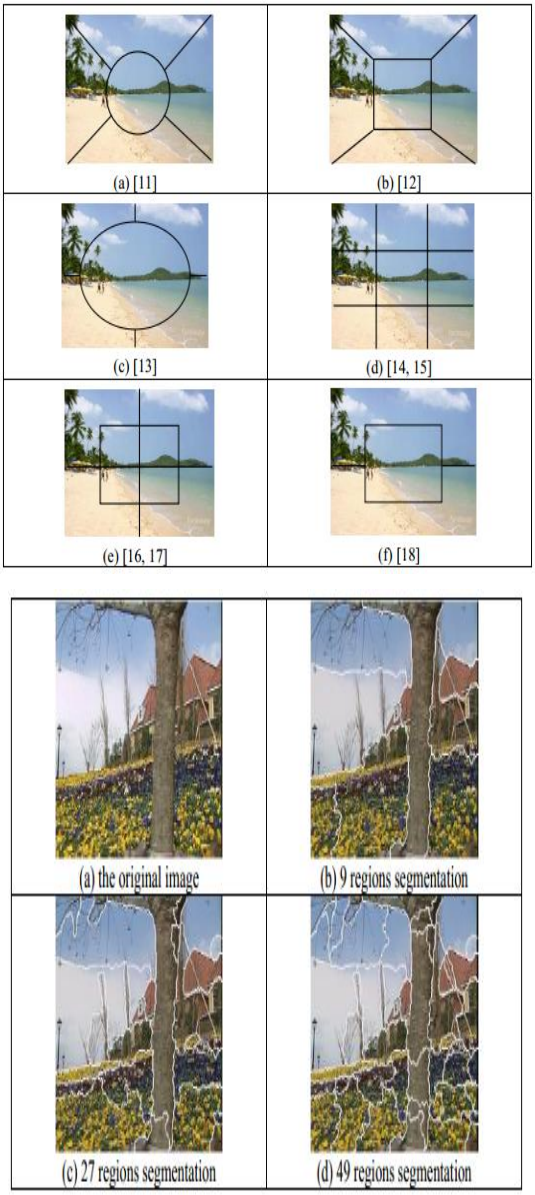

Fig.(1): Shape after segmentation

described after images have been segmented into regions . The shape representations can be divided into two categories, boundary-based (or edge detection) and region-based. The former uses only the outer boundary of the shape, while the later uses the entire shape region. However, to effectively extract shape features depends on segmentation methods.

4.4. Spatial Relationship

Objects and the spatial relationships (such as left of, inside, and above) among objects in an image are used to represent the image content. That is, an image can be divided into a number of sub-blocks and color, texture, or shape features are extracted from each of the sub-blocks, Then we can project them along the x and y axes, such as 'left/right', 'below/above' relationships between them.

5. ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (Neural Network) learn by experience, generalize from previous experiences to new ones, and can make decisions. A neural network can be thought of as a black box non-parametric classifier, that is, different from Naive Bayes. We do not need to make assumptions about the distribution densities. Neural networks are therefore more flexible. Neural network consists of an input layer including a set of sensory nodes as input nodes, one or more hidden layers of computation nodes, and an output layer of computation nodes. The input neurons are the feature values of an instance, and the output neurons, represent a discriminator between its class and all of the other classes. That is, each output value is a measure of the network's confidence and the class corresponding to the highest output value is returned as the prediction for an instance. Each interconnection has associated with it a scalar weight which is adjusted during the training phase.

6. PROPOSED TECHNIQUES FOR IMAGE ANNOTATION

6.1 Multilayer Perceptron Neural Network

Multilayer Perceptron algorithm is a widely used learning algorithm in Artificial Neural Networks. This section presents the architecture of a feed-forward neural network. The Feed-Forward Neural Network architecture is capable of approximating most problems with high accuracy and generalization ability. This algorithm is based on the error-correction learning rule. Error propagation consists of two passes through the different layers of the network, a forward pass, and a backward pass. In the forward pass the input vector is applied to the sensory nodes of the network and its effect propagates through the network layer by layer. Finally a set of output is produced as the actual response of the network. During the forward pass the synaptic weight of the networks are all fixed. During the backward pass the synaptic weights are all adjusted in accordance with an error-correction rule. The actual response of the network is subtracted from the desired response to produce an error signal. This error signal is then propagated backward through the network against the direction of synaptic conditions. The synaptic weights are adjusted to make the actual response of the network move closer to the desired response. In this research we have used a four layer feed-forward network having input, output, and two hidden layers. Fig. 2 shows an example of a four layer feed-forward network having input, output, and two hidden layers. Table 1 shows number of neuron, Activation function and learning rule for each layer.

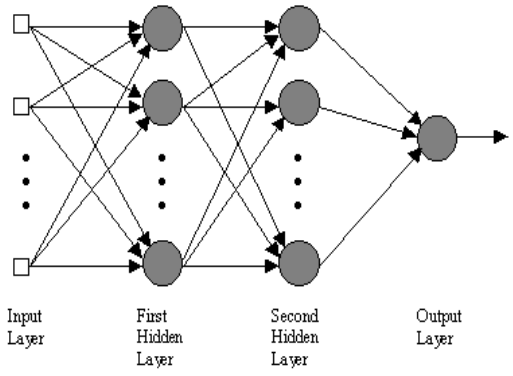

Fig (2):The four layer Neural network

6.2 Algorithm:

The algorithm for Perceptron Learning is based on the back-propagation rule discussed previously. This algorithm can be coded in any programming language. We are using VB.Net programming language. In this case we are assuming the use of sigmoid function in hidden layers. This is because it has a simple derivative. The learning rule for hidden layer is Levenberg-Marquardt.

Algorithm:

1. Initialize weights and threshold.

Set all weights and thresholds to small random values.

2. Present input and desired output

Present input $\mathbf{X}_p = x_0, x_1, x_2, \ldots, x_{n-1}$ and target output $\mathbf{T}_p = t_0, t_1, \ldots, t_{m-1}$ where n is the number of input nodes and m is the number of output nodes. Set w_0 to be ø, the bias, and x_0 to be always 1. For pattern association, $\mathbf{X}_p$ and $\mathbf{T}_p$ represent the patterns to be associated. For classification, $\mathbf{T}_p$ is set to zero except for one element set to 1 that corresponds to the class that X_p is in.

3. Calculate the actual output

Each layer calculates the following:

$$y_{pj} = f[w_0x_0 + w_1x_1 + ... + w_nx_n]$$

This is then passes to the next layer as an input. The final layer outputs values o_{pj}.

4. Adapts weights

Starting from the output we now work backwards.

$$w_{ij}(t+1) = w_{ij}(t) + \tilde{n}p_{pj}o_{pj}$$

Where ñ is a gain term and $þ_{pj}$ is an error term for pattern p on node j.

For output units

$$p_{pj} = ko_{pj}(1 - o_{pj})(t - o_{pj})$$

For hidden units

$$p_{pj} = ko_{pj}(1 - o_{pj})[(p_{p0}w_{j0} + p_{p1}w_{j1} + \ldots + p_{pk}w_{jk})]$$

Where the sum (in the [brackets]) is over the k nodes in the layer above node j.

7. PERFORMANCE EVALUATION

The proposed neural network Multi Layer Perceptron improves the classification accuracy. The construction of the proposed model is given in table:

Table 1

Input Neuron	50
Number of Hidden Layer Neuron	25
Output Neuron	2
Number of Hidden Layer	2
Transfer function of First Hidden Layer	Sigmoid
Learning Rule of First	Levenberg-Marquardt

Hidden Layer	
Transfer function of Second Hidden Layer	Sigmoid
Learning rule of Second Hidden Layer	Levenberg-Marquardt

The activation function in a neural network controls the amplitude of the output such that the range of output is between 0 and 1 or -1 to 1. Mathematically the internal activity of the neuron can be shown to be:

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

Where x_i is the input and w_{jk} is the weights. The output of the neuron, v_k would therefore be the outcome of some activation function on the value of v_k. The most common type of activation function used to construct the neural network is the sigmoid function.

A sigmoid activation function uses the sigmoid function to determine its activation. The sigmoid function is given as:

$$f(x) = \frac{1}{1+e^{-x}}$$

This function can range between 0 and 1, but it is also sometimes useful to use the -1 to 1 range. An example of the sigmoid function is the hyperbolic tangent function, where the range is -1 to 1.

$$\varphi(v) = \tanh\left(\frac{v}{2}\right) = \frac{1-\exp(-v)}{1+\exp(-v)}$$

7. RESULT

50 images are used in the experimental setup containing four class labels. The top 50 relevant attributes are selected using information gain. Fig.3 shows some of the images used in this work.

Fig.3 Sample images used in this work.

The results obtained from regular MLP Neural Network is shown in fig. 4.

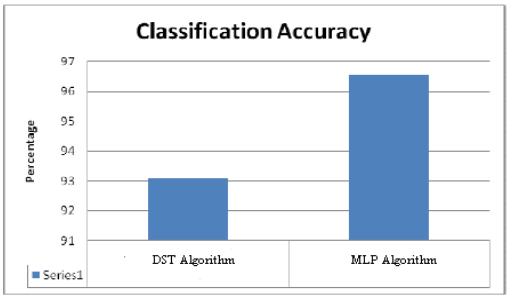

Fig.4Classification accuracy measured in percentage

8. CONCLUSION

In this paper it is proposed to extract features using Discrete Sine Transform (DST) and select the top 50 attributes based on class attribute using information gain. The extracted features are trained and compared with MLP Neural network classifier. The classification accuracy of the proposed method improved by a percentage of 3.45. Using less number of features in the proposed method decreases the overall processing time for a given query.

9. REFRENCES

[1]Y. Mori, H. Takahashi and R. Oka, "Image-to-word Transformation Based on Dividing and Vector Quantizing Images with Words," in MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.

[2]C. F. Tsai and C. Hung, "Automatically Annotating Images with Keywords: A Review of Image Annotation Systems," Recent Patents on Computer Science, Jan., 2008, vol 1, pp 55-68,

[3]Dirichl,A.Vailaya, M. Figueiredo, d H. Zhang, "Image classification for content-based indexing,"IEEE

Transactions on Image Processing, vol. 10, no. 1,2001,pp. 117–130.

[4]Blei, D.M.Jordan.,"LatentDirichlet Allocation", Journal of machine learning research, 2003, pp.993-1022.

[5]R.Datta, D. Joshi, J. Li and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys (CSUR), Apr. 2008, vol 40.

[6]J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," in IEEE Transactions of Pattern Analysis and Machine Intelligence, Aug. 2000, vol. 22, pp. 888–905.

[7]Metzler, D. and Manmatha, R. (2004). An inference network approach to image retrieval, Proceedings of the International Conference on Image and Video Retrieval, pp. 42-50.

[8]Oliva, A. and Torralba, A. (2001).Modeling the shape of the scene a holistic representation of the spatial envelope, International Journal of Computer Vision pp. 145-175.

[9]Yavlinsky, A., Schofield, E., and Ruger S. (2004). Automated image annotation using global features and robust nonparametric density estimation, Proceedings of the 4thInternational Conference on Image and Video Retrieval, Singapore, pp. 507-517.

[10]J. Liu, B. Wang, M. Li, Z. Li, W. Y. Ma, H. Lu and S. Ma, "Dual Cross-Media Relevance Model for Image Annotation," in Proceedings of the 15th International Conference on Multimedia 2007, p. 605 – 614.

[11]Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, and D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: European Conference on Computer Vision. (2002),p 97–112.

[12]V. Lavrenko, R. Manmatha and J. Jeon, "A Model for Learning the Semantics of Pictures," in Proceedings of Advance in Neutral Information Processing, 2003.

[13]S. Feng, R. Manmatha and V. Laverenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004,p. 1002-1009.

[14]Monay. F., Gatica-Perez, D.: On image auto annotation: constraining the latent space. In: Proceedings of the 12th international ACM Conference on Multimedia, Newyork, USA, ACM (2004),p.348-351

[15]Monay. F., Gatica-Perez, D.: Plsa –based image auto annotation with latent space models. In: Proceedings of the 11th international ACM Conference on Multimedia, Newyork, USA, ACM (2003),p. 275-278.