Character Recognition in Images

Dhruva Adike
MBA Student IIM Calcutta
Kolkata, India

Abstract— **This paper tackles the problem of recognizing characters in images of natural scenes which is a challenging problem that has received significant amount of attention. Recently, many methods have been proposed to design better feature representatives and models for text detection and character recognition. In machine learning, we apply these recently developed large-scale algorithms for learning features automatically from unlabeled data and access the performance of various features based on these methods. Nearest neighbor and SVM classification allow us to construct highly effective classifiers for both detection and recognition, that are to be used in a high accuracy end- to-end system that can be far superior to that of commercial Optical Character Recognition systems.**

Keywords—Bag-of-visual-words; K-means; descriptor; SIFT

I. INTRODUCTION

Text detection and character recognition is a challenging recognition problem in scene images. Primary focus is on the recognition of individual characters. Font style, thickness, background color, foreground color, texture geometric distortions caused by camera, illumination and image resolution are to be taken-into account. These factors add on to give a problem on object recognition instead of OCR. These factors are the cause for which we cannot apply OCR techniques [1].

Reading text includes many problems like text localization, character segmentation, word segmentation, recognition, integration of language models, and context etc. For assessing the feasibility of posing the problem as an object recognition task, we consider the performance of various features based on a bag-of-visual-words representation. The results show that even the isolated character recognition task is very challenging. Furthermore, training data of few characters might occur very rarely in natural scenes. We therefore investigate whether surrogate training data, either in the form of hand-printed characters or font generated characters, can be used to bolster recognition in such a scenario [2].

II. RELATED WORK

The task of text detection and character recognition in natural scenes is related to problems considered in camera-based recognition and document analysis. Most of the work in this field is based on locating and rectifying the text areas (e.g. (Krempp, 2002), (Kumar, 2007), (Clark & Mirmehdi, 2002) and (Brown et al., 2007)), followed by the implementation of OCR techniques (Kise & Doermann, 2007) [2].

These approaches are therefore limited to scenarios where OCR works well. Furthermore, even the rectification step is not directly applicable to the problem, as it is based on the detection of printed document edges or assumes that the image is dominated by text. The methods for off-line recognition of hand printed characters (Pal et al., 2007), (Plamondon and Srihari, 2000) have successfully tackled the problem of intra-class variation due to differing writing styles. However, these approaches typically consider only a few appearance classes, not dealing with variations in foreground/background color and texture.

For natural scenes, few researchers have designed systems that integrate text detection, segmentation and recognition into a single framework to accommodate contextual relationships. For instance, (Tu et al., 2005) used insights from natural language processing and presented a Markov chain framework for parsing images. (Jin and Geman, 2006) introduced composition machines for constructing probabilistic hierarchical models which accommodate contextual relationships. This approach ensures re-usability of parts among multiple entities and non-Markovian distributions. (Weinman and Learned Miller, 2006) proposed a method that fuses language and image features information (such as bi-grams and letter case) in a single model and integrates dissimilarity information among character images.

Simpler recognition pipelines on the concept of classifying raw images have been widely explored for digits recognition (see (Zhang et al., 2006), (le Cun et al., 1998) and other works on the MNIST and USPS datasets). Another approach is based on modeling it as a shape matching problem (e.g. (Belongie et al., 2002)): several shape descriptors are detected and extracted and matching is computed between pairs of images [1].

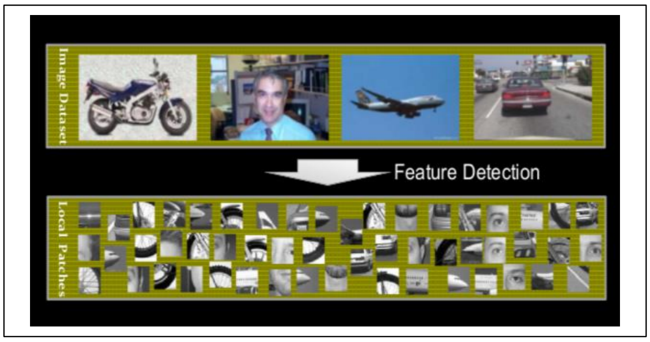
III. DATA SETS

Our focus is on recognizing the characters in natural scene. Towards this end, we compile a database of characters taken from images. However, collecting and annotating a large number of images for training can be expensive and time consuming. So, in order to provide complementary training data, we also acquire a database of hand-printed characters and other from characters generated by computer fonts. For English, we treat upper-case and lower-case characters separately and include digits to get a total of 62 classes.

Individual characters are manually segmented from these images. We experiment with rectangular bounding boxes segmentation. The hand-printed data set captures using a tablet PC with the pen thickness set to match the average thickness found in hand painted information boards [2].

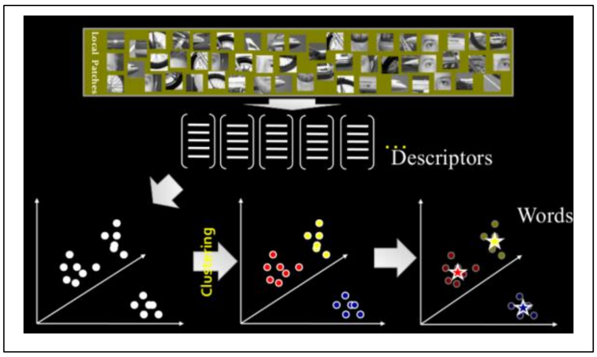
IV. FEATURE EXTRACTION AND REPRESENTATION

Bag-of-visual-words is a technique for representing image content for object category recognition. The idea is to constitute objects as histograms of feature counts. This representation quantizes the continuous high-dimensional space of scene image features to a manageable vocabulary of visual words. This is achieved, for instance, by grouping the low-level features collected from an image compile into a specified number of clusters using an unsupervised algorithm such as K-Means. One can then map each feature extracted from an image onto its closest visual word and represent the image by a histogram over the vocabulary of visual words.



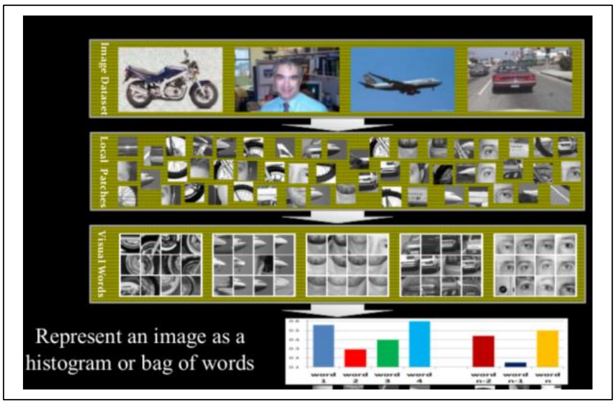
(source: Cordelia Schmid)

We evaluate six different types of local features. Not only did we try out shape and edge-based features, such as shape context, SIFT and geometric blur, but also features used for representing texture, such as filter responses, patches and spin images. We can treat an image as a document in the bag-of-words model. Bag-of-words model can also be defined as the histogram representation based on independent features. We need to define the meaning of words in images. For achieving this, we usually follow three steps which are feature detection, feature description and code-book generation.



(source: Cordelia Schmid)

The images are abstracted into several local patches after the feature detection. These patches are represented as numerical vectors by feature representation. These vectors are called as feature descriptors. Scale Invariant Feature Transform is a famous feature descriptor. A descriptor is known as a good descriptor if it can handle intensity, rotation, scale and affine variations to some extent.



Represent an image as a histogram or bag of words

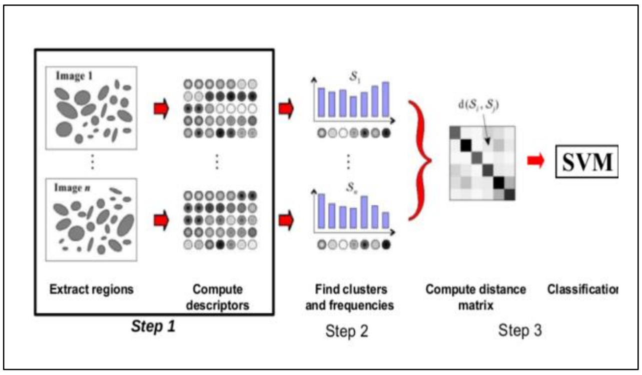
(source: Cordelia Schmid)

Shape Contexts (SC) is a descriptor for point sets and binary images. We sample points using the sobel edge detector. The descriptor is a log-polar histogram, which gives a theta cross n vector, where theta is the angular resolution and n is the radial resolution. Geometric Blur (GB) is a feature extractor with a sampling method similar to that of SC, but instead of histogram points, the region around an interest point is then blurred according to the distance from this point. For each region, the edge orientations are counted with different blur factor. This soothes the problem of hard quantization and allows its application to gray scale images. Scale Invariant Feature Transform are extracted on points located by the Harris Hessian-Laplace detector, which gives affine transform parameters. Feature descriptor is computed as a set of orientation histograms on 4×4 pixel neighborhoods. The orientation histograms formed are relative to the key-point orientation. The histograms contain 8 bins each, and each descriptor in-turn contains 4×4 array of 16 histograms around the key-point. This leads to feature vector with 128 elements.

Spin image is a two-dimensional histogram which encodes the distribution of image brightness values in the neighborhood of the particular reference point. The two dimensions of the histogram are d, distance from the center point, and i, the intensity value. We use 11 bins for distance and 5 for intensity value, resulting in 55-dimensional descriptors. The same interest point locations which are used for SIFT are used for spin images.

Maximum Response of filters (MR8) is a texture descriptor based on a set of 38 filters but only 8 responses. This filter is extracted densely, giving a large set of 8D vectors. Patch descriptor (PCH) is the simplest dense feature extraction method. For each position, the raw $n \times n$ pixel values are vectorized, generating an n power 2 descriptor.

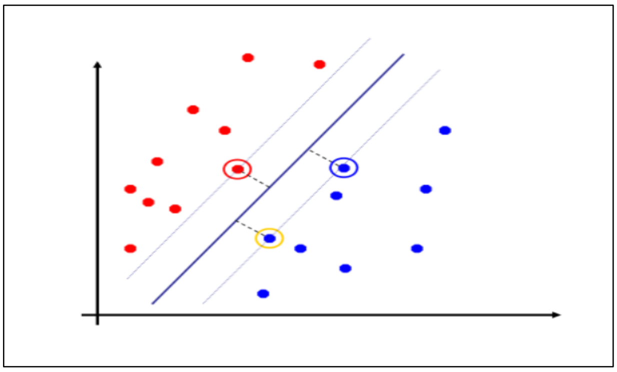
Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non- probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.



(source: Cordelia Schmid)

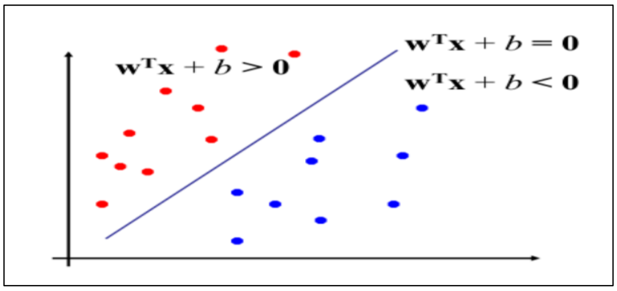
In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.



(source: Cordelia Schmid)

We need to maximize the margin according to intuition. This implies that we only consider the support vectors and ignore other training examples. Examples closest to the hyper plane are known as support vectors and hence this technique is known as support vector machines**.** Thus, in case of binary feature space we get a classifier similar to the above which separates the classes in feature space. This is why, this support vector machines are widely used in pattern recognition and object classification.



(source: Cordelia Schmid)

V CONCLUSIONS

In this paper, we tackled the problem of recognizing characters in images of natural scenes. We introduce a database of images and show that even commercial OCR systems are not well suited for reading text in such images. Working in an object categorization framework, we are able to improve character recognition accuracy. The best result on the English image database was obtained by the multiple kernel learning (MKL) method of when trained using 15 image samples per class. This could be improved further if we were not to be case sensitive. Nevertheless, significant improvements need to be made before an acceptable performance level can be reached.

Obtaining and annotating natural images for training purposes can be ex- pensive and time consuming. We therefore explore the possibility of training on hand-printed and synthetically generated font data. For equivalent size training sets, training on fonts using a NN classifier could actually be better than training on the natural images themselves. As regards features, the shape-based features and Shape Context, consistently outperformed SIFT as well as the appearance-based features. This is not surprising since the appearance of a character in natural images can vary a lot but the shape remains somewhat consistent.

VI LIMITATIONS AND RECENT DEVELOPMENTS

One main disadvantage of using bag-of-words model is that it ignores the spatial relationships among the patches, which are very important in image representation. Researchers have proposed several methods to incorporate the spatial information. For feature level improvements, correlation features can capture spatial co-occurrences of features. For generative models, relative positions of codewords are also taken-into account. The hierarchical shape and appearance model for human action introduces a new part layer between the mixture proportion and the bag-of-words features, which captures the spatial relationships among parts in the layer. For discriminative models, spatial pyramid match performs pyramid matching by partitioning the image into increasingly fine sub-regions and compute histograms of local features inside each sub-region.

The bag-of-words model has not been extensively tested yet for view point invariance and scale invariance, and the performance is unclear. The bag-of-words model for object segmentation and localization is not well understood.

REFERENCES

[1] Carl Case Sanjeev Satheesh Bipin Suresh Tao Wang David J. Wu Andrew Y. Ng Adam Coates, Blake Carpenter. Text detection and character recognition in scene images with unsupervised feature learning. Proceedings of the 2011 International Conference on Document Analysis and Recognition, 2011.

[2] T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. Proceedings of the International Conference on Computer Vision Theory and Applications, Lisban, Portugal, Reading, Massachusetts, 2009.