INTEGRATING CO-TRAINING AND RECOGNITION FOR TEXT DETECTION

Wen Wu, Datong Chen and Jie Yang
School of Computer Science, Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

Training a good text detector requires a large amount of labeled data, which can be very expensive to obtain. Co-training has been shown to be a powerful semi-supervised learning tool for solving many problems using a large amount of unlabeled data. However, augmented data from a co-training process could potentially degrade the performance of classifiers due to added noises from unlabeled data. This paper makes two contributions by proposing a modified co-training scheme for text detection. First, to get cleaner augmented data, the new algorithm integrates some authority knowledge of unlabeled data into co-training. Text recognition output of each selected unlabeled image patch is used as the authority that is combined with classifier prediction to decide if the sample will be added to the augmented set. Second, instead of evenly combining predictions of two co-training classifiers, a weighted combination is learned and used to produce the final prediction. Contributions of the new algorithm have been evaluated on a standard text detection dataset.

1. INTRODUCTION

The explosive growth of Internet and widely usage of digital cameras have created tons of digital images. This has posed many new challenges for the image processing community. Understanding image is essential for indexing and retrieving these images. Text in an image can be very helpful for understanding the content of the image. Recently, there has been much work focusing on detecting text in images and videos [2, 4, 6, 9, 10, 12, 15, 16].

Detecting text from images is a challenging task because training a good text detector requires a lot of training data. However, there can be huge variations in text appearances. The varieties of text come from its variations in font, size, orientation, and position in an image. Text in an image can be blurred from motion when it is taken or occluded by other objects in the scene. Text can also be distorted by slant, tilt, and shape of objects on which they are printed. Also, it is very difficult to separate text foreground from its background in images using a fixed mask. That means it is difficult to get clean positive training data. Unlike labeling faces where we can easily remove face background using a mask, it is hard to mask background off text in an image unless you know exactly where and what text letters / characters are. The same text with different backgrounds can cause additional difficulty in training a good classifier.

Therefore, training a text detector needs a large amount of training data which are too expensive to be manually labeled. The goal of this work is to address the problem of lacking enough training data for supervised learning by exploiting a co-training scheme. Semi-supervised learning provides a way of utilizing the information available in a large amount of unlabeled data to train more robust classifiers than supervised learning which only uses labeled data. Co-training, one of semi-supervised learning methods, has been successfully applied to various domains, such as document classification [1, 13], video processing and vehicle detection from surveillance video [8], etc. One important criterion to apply co-training is that the features for the task can be partitioned into two compensate sets. In a text detection task, edge and color features can be separated from each other so they can be treated as an independent split of two compensate feature sets. In this paper, we apply a modified co-training algorithm to improve text detection from images. Our work has two differences from previous other co-training applications. First, recognition results of selected unlabeled image patches are combined with the predictions of the co-training classifiers during the data augment step. Previously, only predictions of co-training classifiers are used to label a few of samples per round. In contrast, when our method labels samples it uses recognition results as complementary information to reduce possibly added noises by unlabeled data. Second, a weighted combination of co-training classifiers is developed through a development set and used to predict the labels on the test set. In most of previous co-training works, two co-training classifiers are assumed to be equally important in testing. However, this assumption may not be true in some cases such as text detection and face recognition. Different features could contribute different amount of discriminative powers for classification.

Section 2 next gives an overview of the proposed approach. Section 3 describes feature selection for co-training in this task. Section 4 introduces a modified co-training algorithm combined with text recognition input. Section 5 shows experimental results and Section 6 concludes the paper.

2. PROBLEM DISCRIPTION

Figure 1 shows some images that contain texts, from which we can see some challenges in this problem. Texts are embedded in complex background / foreground; attributes of text in images are unpredictable, such as size, font, color and direction; lighting conditions are uncontrollable, etc.

These challenges lead us to focus on machine learning approaches instead of relying on empirical features only. To avoid the hard labeling task for text detection, we develop an approach which can take advantage of both co-training and text recognition feedbacks. Co-train two text detectors

on disparate feature sets and then use recognition results to obtain cleaner augmented data. The new approach works as follows. First, the algorithm extracts two feature sets: edge features and color features respectively for each image patch. Second, two margin-based classifiers are trained on these two feature sets using only labeled data. Next, the co-training algorithm comes to play with guidance of Optical Character Recognition (OCR) output. We add a data point into the positive set only when the classifier predicts it as "positive" and also OCR finds text(s) in it. Similarly, a sample is added to the negative set only when it is predicted as "negative" and OCR does not find any text(s) in it. Both conditions must be satisfied in order to move an unlabeled sample to either the positive or the negative set.

Figure 1. Example images from ICDAR '03 datasets.

3. FEATURE SELECTION FOR CO-TRAINING

Co-training algorithm is used to train a pair of classifiers iteratively. The sufficient condition for the success of co-training is to pick two "conditional independent" feature sets which are both sufficient for correct classification [1, 7]. The choice of feature set is critical to the co-training algorithm. Ideally, we should select totally independent feature sets which can provide compensatory discriminative power to distinguish text and non-text.

Edge and color features are such two naturally conditionally independent feature sets for image processing. Text differs from other objects in the scene due to its two important characteristics. First, background and foreground colors of text in image are in highly contrast. Thus, a text image patch contains regular and low noisy edge features. Second, the color of the text letters in an image patch is quite uniform, which indicates that text image patches may have different color distributions from non-text patches. By computing the Canny edge every 5 degree for each window patch, we obtain a 72 dimension edge feature vector. Since text image patches have variable size, and normalize the edge vector. This will also pose constraints on the later classifier learning. We model text color distribution independent to its real RGB values due to the fact that text can appear in any pair of background and foreground colors. For each patch, we first project its color distribution into normalized RGB color space which actually has only two degree-of-freedom. We then use EM algorithm to compute the data likelihood by assuming the distribution is generated by a Gaussian model. Similarly, we compute the other four likelihood values by assuming two, three, four and five Gaussians. By this way, we can project the data point from normalized RGB space to a new 5-dim space in which each coordinate represents data likelihood of a generative model.

4. A MODIFIED CO-TRAING ALGORITHM FOR TEXT DETECTION

Next we describe how to learn a text detector from labeled and unlabeled data by combining the co-training scheme with text recognition input. We select Support Vector Machines (SVM) as the classifier in the work. We begin the discussion with a brief overview of SVM classifier.

4.1 Support Vector Machines

SVM is widely used to solve binary classification problem. The algorithm is to find a decision surface that maximally separates data points into two classes based on the Structural Risk Minimization principle [5]. The decision function is defined as $z_k f(W, X_k) \geq 1,\ \mathrm{k} = 1, \cdots, \mathrm{K},$ where $z_k = 1$ stands for positive and $z_k = -1$ for negative cases, W is the parameters to be estimated, X_k is the k-th instance vector, and K is the total number of the training samples. A linear SVM uses the decision function, $f(W, X_k) = W \cdot X_k - b$. The parameters, W, can be estimated by solving a quadratic programming problem. For other non-linear cases, a kernel function is chosen and it can have various forms, such as the polynomial kernel and Radial Basis Function (RBF) kernel. In this work, we choose SVM with a RBF kernel.

4.2 A Modified Co-Training Algorithm

The co-training algorithm trains two SVM classifiers over the edge and color feature sets from labeled and unlabeled training data. In our experiments two SVM classifiers are initialized using a limited number of manually labeled data. In co-training algorithm, each classifier iteratively chooses a small set of unlabeled samples per class as candidates to add to the labeled set per iteration with their predicted labels. There can be some errors in predictions since co-training classifiers are not perfect. We attempt to improve the quality of augmented data by introducing some additional authority information. In other words, the selection of the final unlabeled samples to add is also based on recognition results. We use commercial OCR software to scan every candidate samples. Only the samples, in which at least one English letter or more are found, are added into the positive training set (we are only interested in English text detection). On the other hand, only samples, in which OCR does not find any text, are added to the negative set. Each classifier retrains itself from newly augmented labeled

sets. Such a process iterates until no more text can be detected by two classifiers from unlabeled patches.

Another issue is that text recognition is computation expensive, the algorithm cannot afford running OCR on each of thousands of unlabeled data, but only on those most confidently predicted samples of each class per round.

Different co-training classifiers play different roles in text detection task since they are trained on different feature sets. Therefore, it is natural to learn an optimal weighting combination of co-training classifiers from a development set. The idea is as follows. We randomly pick a subset from labeled training data and evaluate two classifiers on it. We then get the error rates of two classifiers. The weighting coefficients of two classifiers are then set as (*1- error rate*) and normalize them. Evaluate again and update the coefficients until they stabilize. Repeat the process by randomly selecting additional nine development sets with same size. Average coefficients from each set to get final weighting coefficients. Table 1 outlines the new algorithm.

Table 1. A modified co-training algorithm for text detection.

Input: 1). A labeled training set and an unlabeled set; 2). An OCR software [17].

Training Process: Iterate until there is no more text can be detected by two classifiers and OCR from unlabeled data:

1. Train an SVM classifier, *A*, using the edge features from labeled data;
2. Train an SVM classifier, *B*, using the color features from labeled data;
3. Select top *5%* unlabeled samples, about which both *A* and OCR are most confident that its class label is "text" or "non-text" and add them to the augmented sets of *B*;
4. Similarly, pick top *5%* instances about which *B* and OCR are most confident that their class labels are "text" or "non-text" and add them to the augmented sets of *A*;
5. Re-train *A* and *B* based on the augmented data.

Validation Process: Randomly select a development set and initialize the weighting coefficients of A and B as their *(1-error rate)*. Normalize the weighting coefficients. Select another development set and updates the weighting coefficients based on A and B's classification results on it. Iterate the process until coefficients stabilize.

Output: Two final SVM classifiers, *A'* and *B'*, that *A'* predicts labels for new samples using edge features and *B'* predicts labels using color features. The predictions of *A'* and *B'* are combined through the learned weighting coefficients and normalizing their class probability scores.

4.3 Sampling Negative Samples

Training data for text detection are image patches, so we need sample the training instances from each image by a sliding rectangle window (40×20) with *50%* overlap between adjacent patches at various scales. We found severe unbalanced distribution of positive and negative samples within sampled data. There are a much larger number of negative samples than that of positive ones. Here we apply the sub-sampling method to deal with this problem as in [8]. The basic idea is to use the importance sampling approach, by randomly selecting negative samples using a probability distribution related to the density estimation. Figure 2 shows some labeled positive and negative samples.

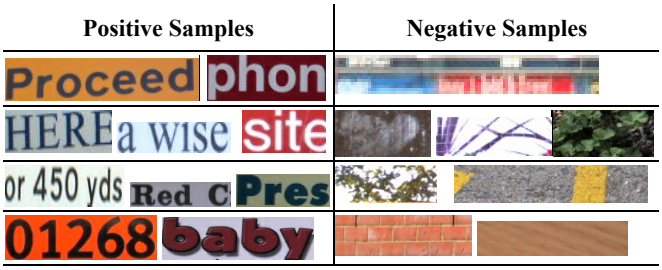

Figure 2. Labeled text image patches.

5. EXPERIMENTAL RESULTS

We randomly select *360* images, *205* for training and *155* for testing, from ICDAR'03 competition datasets [11]. In the *205* training images, *15%* images are randomly selected as the initial labeled training set and *10%* images are set as the size of development set. The rest images are used as unlabeled data. Figure 3 shows the ROC curves of two detectors that are computed on the testing data. Note that the bigger is area below the curve is the better performance. As we can see, co-training can dramatically reduce the false positive rate by over 50% when the detection rate is equal to 0.8. Figure 4 shows some detection results, among which almost all text segments in these examples were correctly detected except that the system missed "(*A12*)" in the upper-right image. Figure 5 shows some false positives. As you can see they are all "text-like" things. Table 2 compares co-training classifiers with supervised-learning classifiers using 5-fold cross validation on testing data. It shows that combining OCR and the weighting strategy improve text detection performance compared to the co-training classifier without OCR or other supervised learning classifiers. Since some state-of-the-art text detection methods [4, 11] did not use the same data set as us or used different experimental settings, so it is hard compare their results with ours.

6. CONCLUSION

We have proposed a modified co-training algorithm for text detection from images. There are two key elements of the new algorithm. First, text recognition is used in the co-training process to provide additional authority information about unlabeled data before augmenting data. Second, a weighted combination of two co-training classifiers is used to replace the even combination strategy. Experimental results have demonstrated the effectiveness of the proposed method on the text detection task. One future work is to

explore the feasibility of this modified co-training algorithm in other problem domains, e.g., web/email classification.

Table 2: Detection performance comparison.

Classifiers	Error rate
kNN (k = 5, w/o OCR)	0.65
SVM (RBF, w/o OCR)	0.53
Co-training (w/o OCR)	0.50
Co-training (w. OCR)	0.46
Weighted Co-training (w. OCR)	0.42

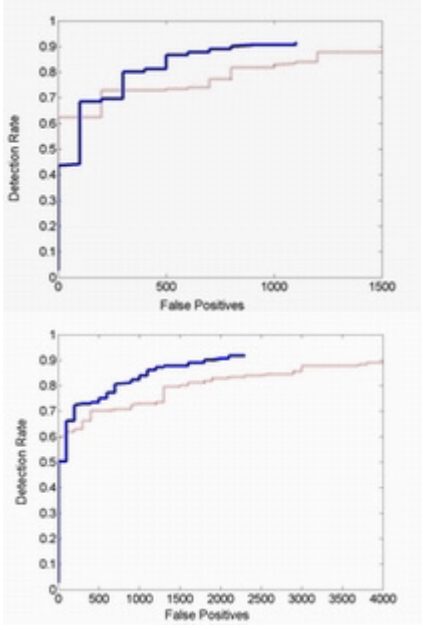

Figure 3. ROC curves. Solid blue line shows the classifier after co-training; dot red line shows the classifier before co-training. Upper using edge features. Lower using color features.

Figure 4. Detection results.

Figure 5. Some false positive images.

ACKNOWLEDGEMENTS

This research is partially supported by GM-CMU Research Lab, ARDA under contract number H98230-04-C-0406, and DAPAR under ASSIST program.

REFERENCES

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT:* Proceedings of the workshop on COLT, Morgan Kaufmann Publishers, 1998.

[2] D. Chen, J-M. Odobez, and H. Bourlard, "Text Detection and Recognition in Images and Video Frames", Pattern Recognition, 3(37) pp: 595-608, 2004

[3] X. Chen, J. Yang, J. Zhang and A. Waibel, Automatic detection and recognition of signs from natural scenes. IEEE Trans. on IP, 13, 1 (Jan. 2004), 87-99.

[4] X. Chen and A. Yuille. Detecting and reading text in natural scenes. In Proceedings of the CVPR 2004.

[5] C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20(3):273–297, September 1995.

[6] A.K. Jain and B. Yu. Automatic text location in images and video frames. Pattern Recognition, Dec. 1998, 2055-2076.

[7] C.W. Lee, K. Jung and H.J. Kim. Automatic text detection and removal in video sequences. Pattern Recognition Letters, 24, 15 (Nov. 2003), 2607-2623.

[8] A. Levin, P. Viola and Y. Freund. Unsupervised improvement of visual detectors using co-training. In Proceedings of the Intl. Conf. on Computer Vision, 2003.

[9] H. Li, D. Doermann and O. Kia. Automatic text detection and tracking in digital video. IEEE Trans. on Image Processing, 9, 1(Jan. 2000), 147-156.

[10] R. Lienhart and A. Wernicke. Localizing and segmenting text in images and videos. IEEE Trans. on CSVT, 12, 4, 256-268, 2000.

[11] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong and R. Young. "ICDAR 2003 Robust Reading Competitions", In 7th Intl. Conference on Document Analysis and Recognition, 2003.

[12] T. Sato, T. Kanade, Hughes, E.K., and Smith, M.A. Video OCR for digital news archives. In Proc. of the IEEE Int. Workshop on Content-Based Access of Image and Video Database, 52-60, 1998.

[13] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3):103–134, 2000.

[14] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In Computer Vision and Pattern Recognition, 2000.

[15] V. Wu, R. Manmatha and E. M. Riseman. TextFinder: an automatic system to detect and recognize text in images, *IEEE Trans. on PAMI*, 21, 11 (Nov. 1999), 1224-1229.

[16] W. Wu, X. Chen and J. Yang. Incremental Detection of Text on Road Signs from Video with Application to a Driving Assistant System. In the Proceedings of ACM Multimedia, 2004.

[17] http://www.scansoft.com/textbridge/