IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Turkish Text Detection System from Videos Using Machine Learning and Deep Learning Techniques

Jawad Rasheed
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0003-3761-1641

Akhtar Jamil
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0002-2592-1039

Hasibe Busra Dogru
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
hasibe.dogru@std.izu.edu.tr

***Abstract*—With the advancement in smart devices and high-speed internet, a continual increase in videos demands an efficient and automatic video indexing and retrieval system. To accomplish it, content-based video indexing is an optimal solution by detecting text in videos. In this study, we proposed a text detection system based on machine learning approaches. We compared conventional machine learning approaches with deep learning method. For deep learning, we implemented Convolutional Neural Network (CNN), while Logistic Regression (LR) and Support Vector Machine (SVM) are employed as conventional machine learning techniques to predict the outcome as text or non-text data. We evaluated the proposed systems on our own dataset obtained from various Turkish videos. LR obtained an overall accuracy of 95.0%, whereas SVM achieved 98.7% while CNN secured 99.8% accuracy. The experimental results show that CNN (deep learning approach) was more effective for our Turkish text dataset as compared to LR and SVM.**

Keywords—text detection, CNN, SVM, LR

I. INTRODUCTION

The progression of world towards 5G data communication system triggers more multimedia data formation on hourly bases. Besides lightning-fast internet, the evolution of smart devices drastically contributed in generation of unstructured and structured audios, videos and images data that eventually requires good indexing and retrieval systems. Business platforms like Cincopa, free video hosting websites like Vimeo, Dailymotion, and Youtube™ or social-media networks like Facebook provide opportunity to publish and share videos among closed groups or public. The user has the hectic responsibility to manually annotate the description of video while uploading it to these web-platforms. The manual annotation sometimes may lead to unaligned description with respect to content present in video. Likewise, it also restricts the searching efficacy, thus strongly demands content-based video indexing and retrieval scheme.

Apart from audio content along video, text appearing within video frames provides important information about visual content that can be adopted for automatic video indexing. These textual content is categorized as scene text or graphical text. Graphical text is the supplementary insertion of textual content at the time of editing, well-known as artificial text or caption text. On the other hand, scene text occurs naturally at the time of recording the video or capturing the image. The text present inside video frames or images varies in fonts, sizes, and style with different positing and alignments. Scene text detection requires more complex job due to less readability as text appears with inconsistent orientation, irregular contrast and uncertain resolution because of capturing conditions and constraints. However, as graphical text is artificially embedded, the editor generally uniformly aligns the text in horizontal or vertical orientation, thus makes detection and extraction relatively easier.

As textual content detection and extraction from videos, images, or documents is an old problem, researchers presented various methods and techniques. These proposed research approaches are normally classified into unsupervised methods like [1] and supervised techniques such as presented in [2]. Unsupervised methods are generally based on connected components approaches like Stroke Width Transformation (SWT) or Maximally Stable Extremal Regions (MSERs). Whereas, supervised models usually based on sliding window that extracts contextual features like color, texture and edges to train the classifier either based on conventional machine learning algorithms like LR, SVM, or deep learning such as CNN.

Scientists designed distinct approaches for text detection with slight variations based on unsupervised methods, like in [1] we performed morphological procedure with some heuristic and geometric constraints to detect and extract Turkish text from news and sports videos. Similarly, authors in [3] calculated local entropy to exploit statistical features of image by horizontal projection analysis and computed gradient difference to get horizontally aligned artificial text in Urdu videos. For English text detection, author in [4] used another unsupervised approach based on Laplacian operator by labeling the candidate text region obtained through maximum gradient difference using k-means clustering. Later Sobel edge maps helped in extracting the boundaries of text fields.

For multilingual text detection researchers suggested different supervised approach like in [2], we presented a CNN model (consists of 3 convolutional layers) as feature extraction approach for Turkish text detection and achieved an overall F-measure of 99.95% on our own dataset created from Turkish videos. Another author proposed a modified version of CNN in [5] called Multi-scale Spatial Partition Network (MSP-Net) to do block-level image classification as text or not-text. It spatially divided feature maps into various block sizes and finally classified those by three fully connected layers with 94.6% F-measure score on TextDis benchmark.

An English and Chinese text detection method introduced in [6], named as Efficient and Accuracy Scene Text (EAST) detector, used Fully Convolutional Network (FCN) and Non-Maximum Suppression (NMS) to predict outcome using

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

multiple channels pixel-level text score maps. Authors tested EAST network on various dataset and obtained F-measure score of 80% on ICDAR2015 and 76.08 on COCO-Text. While a slight variation of one-step NMS with FCN in [7] performed convolutional feature extraction, multi-level feature fusion and multi-task learning and passed the resultant to Recalled-NMS in order to remove quadrilaterals within text spaces that also maintains low confidence text regions. It achieved F-measure of 81%, 74%, and 86% on ICDAR2015, MSRA-TD500 dataset and ICDAR2013 respectively.

Moreover, author in [8] presented six text line detection models based on standard VGG-16 network with some modified convolutional layers that learns from weakly annotated data to predict textual regions at multiple scales with maximum F-measure score of 86.9% on ICDAR2013 dataset. An end-to-end hybrid model named as CRNN in [9] used Deep CNN for feature extraction and Recurrent Neural Network (RNN) for prediction, and achieved 89% F-measure.

In the same way, author in [10] detected Arabic text by training the CNN classifier and Bidirectional Long Short Memory (BLSTM) network with text images of five different orientations. In [11], author employed Convolutional Auto Encoder (CAE) for feature extraction and SVM for candidate region classification as Arabic text or non-text with an F-measure score of 84%.

For Turkish text detection, mostly unsupervised techniques have been employed except in [2]. In [12] author employed connected component and histogram analysis to extract the sliding text, then recognized characters and words by Transformation Based Learning (TBL) that achieved 92% words recognition accuracy and 99% character recognition accuracy. Later author enhance [12] by incorporating Hidden Markov, n-gram language and Glyph models in [13] that did semi-supervised training and accomplished 1% WER while recognizing. In addition, in [14] authors proposed Turkish News videos retrieval system with semantic annotation by segmenting news based on silence periods and performed text detection by vertical and horizontal histogram analysis and connected component.

Literature review depicts that Turkish text detection and recognition is in fancy state, therefore, in this study, we compared three distinct learning-based methods for text detection on Turkish Text dataset used in [2]. These methods includes LR and SVM as conventional machine learning techniques while CNN as deep learning approach. Our aim is to analyze the performance and efficiency of deep learning and traditional machine learning approaches for horizontally aligned artificial text detection in Turkish Videos.

The paper enlists three more sections. The next section presents dataset (materials) used for this study, while section 3 explains learning-based methods performed to detect text. Section 4 summarizes the experimental results. Lastly, section 5 outlines concluding remarks.

TABLE I. DATASET

Labels	Dataset Type		
	Total Samples	*Training Set*	*Testing Set*
Non-text	54680	32808	21872
Text	61131	36679	24452
Total	115811	69487	46324

II. MATERIALS

As text detection for Turkish videos is in fancy state, no benchmark dataset is available publically as per our knowledge. Thus we prepared our own dataset of horizontally aligned Turkish graphical text obtained from various Turkish news, sports and business channels. For information about dataset compilation see Section 3 of [2].

The dataset consists of 115811 grayscale image samples, each of size 32 x 128. Among these, 61131 images belongs to text regions while 54680 instances are non-text images, as summarized in Table 1. The dataset is divided randomly into training and testing sets. 60% of dataset is used to train the models while 40% is used for evaluation purposes. Fig. 1 depicts few samples from dataset.

III. METHODS

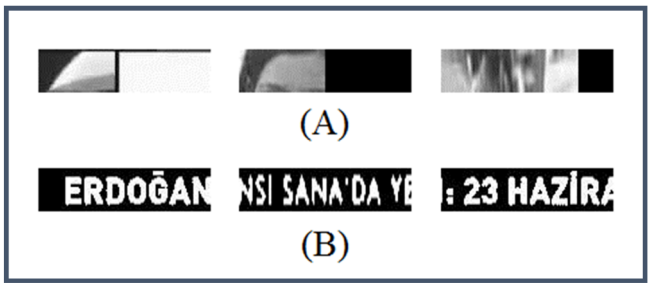

Fig. 1. Samples from dataset, each of size 32 x 128, (A) displays 3 non-text images samples, (B) shows 3 text images samples

A. Logistic Regression (LR)

Logistic Regression is a conventional machine learning algorithm used to classify categorical targets by performing regression analysis with help of sigmoid function as core operation as shown in Fig 2.

It squashes and maps the input values x to range between 0 and 1, then assigns weight w and bias (b_0 and b_1) coefficients to predict the output y as in (1).

$$y = \frac{e^{b0+b1*x}}{1+e^{b0+b1*x}} \quad (1)$$

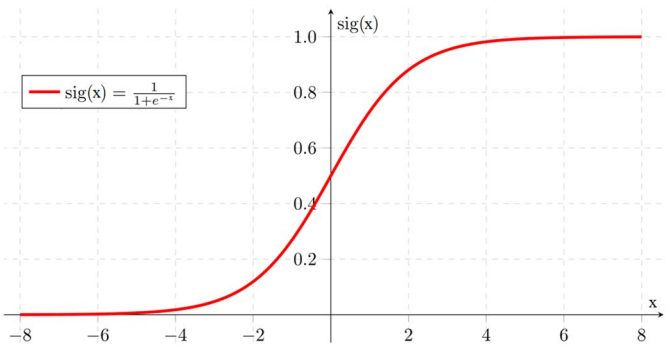

Fig. 2. LR - core function (sigmoid).

B. Support Vector Machine (SVM)

The SVM is a supervised machine learning arsenal preferred as discriminative classifier for various data

classification problems. SVMs are practiced for classification as well as regression tasks by finding the optimal separating hyperplane that distinctly classifies new instances. The hyperplane divides N-dimensional space (N is the number of features) into appropriate labels by finding group of data points (support vectors) that resides on the edge of class descriptors. SVM objective is to discover the best possible

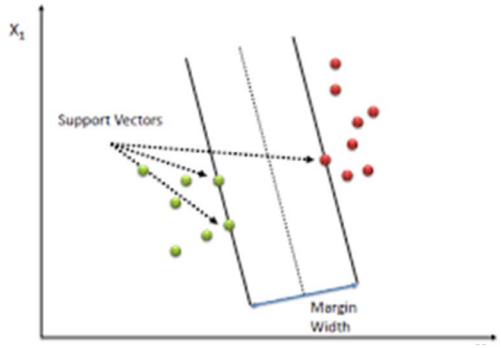

Fig. 4. Hyperplane with support vectors in SVM

hyperplane with maximum margin (distance between data points of classes) as illustrated in Fig. 3.

It requires less computation power as it only retains support vectors (data points that lies on border) while discarding rest of training instances. Initially, algorithm was formulated for data that is linearly separable, but with addition of kernel techniques [15], it can be exercised for non-linear cases by mapping non-linear to linear space. Different kernels have been evolved, but for our analysis, we used Radial Basis Function (RBF). RBF kernel is determined by (2) with as gamma (learnable parameter).

$$K(x_i + x_j) = \exp(-\gamma|x_i - x_j|^2) \qquad (2)$$

C. Convolutional Neural Network (CNN)

CNN is a simple deep learning based algorithm that converts the input data to useful representation to do complex image classification. It usually consists of convolutional layers, pooling layers and fully connected layers (known as dense layers) as depicted in Fig. 4.

Convolutional layers learn the feature representation of input image by assigning appropriate weights and biases to each neuron connecting with some neurons of preceding layer. The resultant is passed to activation function such as Rectified Linear Unit (ReLU). Pooling layer plays role to reduce features dimensionality. In the end, a fully connected layer combines all the learned features of previous layers to formulate the prediction in output layer (classification layer) with help of probabilities computed by softmax layer.

IV. Experimental Results

Each classifier outlined in previous section has few learnable parameters, which are fine-tuned to obtain best results, described in this section. The 115811 samples in Turkish text dataset are split randomly into training set consists of 69487 samples (60% of total dataset) and testing set containing 46324 instances (40% of dataset). Same number of training and testing samples were fed to all three proposed methods and evaluated performance in terms of precision (P), recall (R), F-measure (F), and overall accuracy (OA).

All the experiments were carried out on Jupyter Notebook with Python 3.6 environment along with Keras and Tensor flow packages, running on Nvidia GeForce 410M (512MB RAM) and 2.3GHz Intel Core™ i5 processor of 8GB RAM.

A. Logistic Regression

The LR model is trained with 'lbfgs' optimizer. It achieved an overall accuracy of 95.0% on testing dataset, with recall: 95.0%, precision: 95.0%, and F-measure: 95.0%. Table 2 shows the results obtained for LR classifier.

TABLE II. Classification Accuracy For LR

True Labels	Predicted Labels		Evaluation Results (%)		
	Text	*Non-text*	*P*	*R*	*F*
Non-text	1307	20565	95.3	94.0	94.7
Text	23443	1009	94.7	95.9	95.3
Overall			95.0	95.0	95.0

Fig. 5 shows the Receiver Operating Characteristics (ROC) curve as performance evaluation of LR on Turkish dataset.

B. Support Vector Machine

Classification parameters plays vital role in SVM, therefore we preferred RBF kernel due to its popularity and effectiveness in classification. We empirically calculated and fine-tuned the parameter C=12 and gamma=1.0E4. SVM

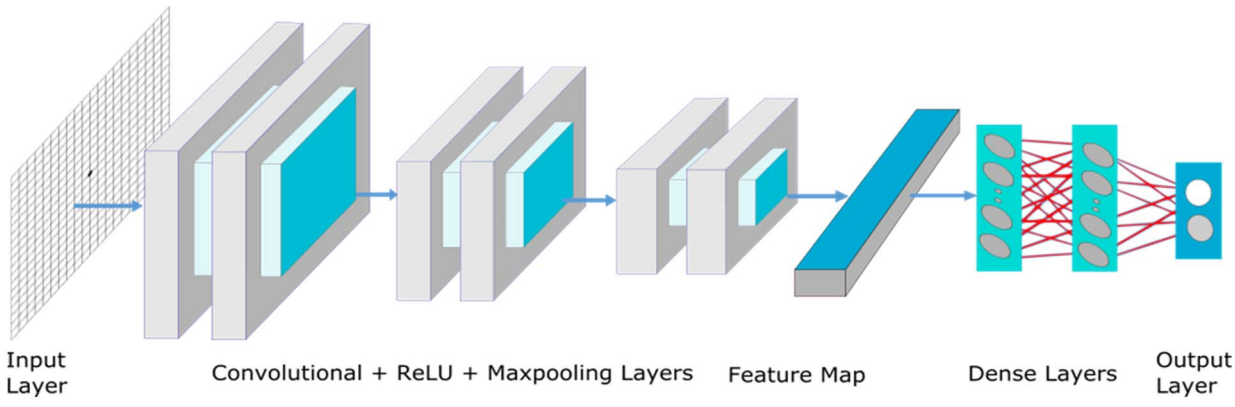

Fig. 3. Network architect of proposed CNN model.

produced promising results with an overall accuracy of 98.7%, recall: 98.7%, precision: 98.8%, and f-measure 98.7%, summarized in Table 3.

TABLE III. CLASSIFICATION ACCURACY FOR SVM

True Labels	Predicted Labels		Evaluation Results (%)		
	Text	*Non-text*	*P*	*R*	*F*
Non-text	399	21473	99.1	98.2	98.6
Text	24249	203	98.4	99.2	98.8
Overall			98.8	98.7	98.7

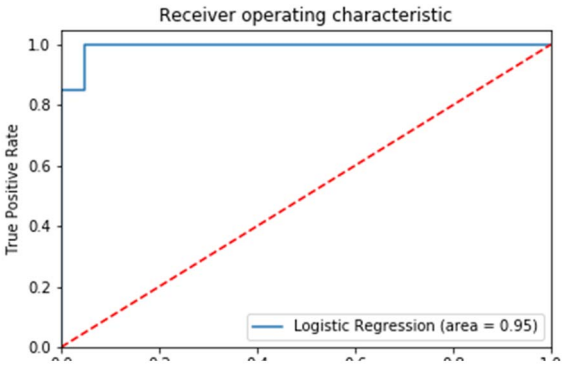

Fig. 6. ROC for LR model.

C. Convolutional Neural Network

For deep-learning approach, we proposed a CNN architect in Fig. 4. Image feature map is fed as input layer (size 32x128) to two consecutive convolutional layers, each with an output shape of 32x128x16, a kernel of size 3x3 and same size padding is turned-on. Followed by 2x2 max-pool layer, and dropout is fixed to 0.10. Next comes two more convolutional layers having a depth of 32 with 3x3 kernel and padding. Again 2x2 max-pool layer is placed for dimensionality reduction and dropout is set to 0.25. Two more convolutional layers with depth of 64 is placed followed by 2x2 max-pool layer and performed a dropout of 0.4. All convolutional layers have ReLU as activation function. The flattened features are fed to fully-connected layer with softmax activation function to predict the output either as text or non-text.

TABLE IV. CLASSIFICATION ACCURACY FOR CNN

True Labels	Predicted Labels		Evaluation Results (%)		
	Text	*Non-text*	*P*	*R*	*F*
Non-text	71	21801	99.8	99.7	99.8
Text	24424	28	99.7	99.9	99.8
Overall			99.7	99.8	99.8

The model is trained with Adam optimizer while learning-rate of 0.001. The proposed CNN model accomplished an overall accuracy of 99.8%, recall: 99.8%, precision: 99.7%, and F-measure: 99.8%. The results are tabulated in Table 4, and Fig. 6 shows the loss and accuracy curves for this trained model.

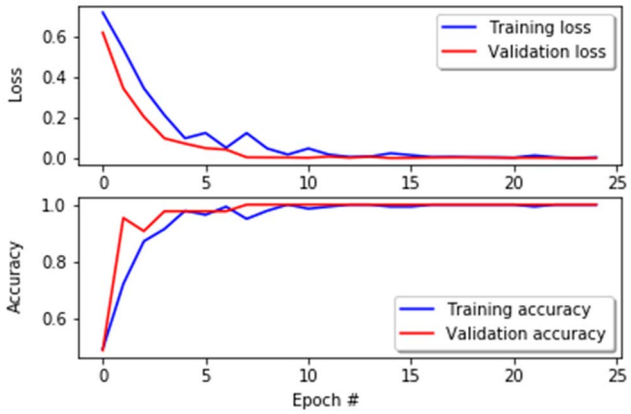

Fig. 5. Loss/Accuracy curves for CNN

V. CONCLUSION

In this study, we performed comparative analysis of deep learning and machine learning algorithms for detecting whether an image has text or not. Even though text detection is language independent, but we focused on Turkish language as it's still an under-stressed area. For this, a dataset compiled from Turkish videos were used to train LR and SVM models as machine learning methods while CNN model was trained as deep learning approach. Experimental results showed that CNN outperformed other two conventional methods on this dataset. CNN achieved an accuracy of 99.8%, while SVM secured 98.7% and LR produced 95.0% prediction accuracy on our dataset. Although methods obtained promising results, but cross-validation is worth exploring in our future work.

REFERENCES

[1] J. Rasheed, A. Jamil, A. Yahyaoui, and A. S. A. Madey, "Automatic Video Indexing and Retrieval System for Turkish Videos Experimental results showed that our proposed method," in *press of The 28th IEEE Conference on Signal Processing and Communications Applications*, 2020.

[2] J. Rasheed, A. Jamil, H. B. Dogru, S. Tilki, and M. Yesiltepe, "A Deep Learning-based Method for Turkish Text Detection from Videos," in *2019 11th IEEE International Conference on Electrical and Electronics Engineering (ELECO)*, 2019, pp. 935–939.

[3] A. Jamil, J. Rasheed, and B. Bayram, "Local statistical features for multilingual artificial text detection from video images," in *International Conference on Advance Technologies, Computer Engineering and Science (ICATCES)*, 2019, no. 2nd, pp. 256–260.

[4] T. Q. Phan, P. Shivakumara, and C. L. Tan, "A laplacian method for video text detection," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, no. January, pp. 66–70, 2009.

[5] X. Bai, B. Shi, C. Zhang, X. Cai, and L. Qi, "Text/non-text image classification in the wild with convolutional neural networks," *Pattern Recognit.*, vol. 66, no. December 2016, pp. 437–446, Jun. 2017.

[6] X. Zhou *et al.*, "EAST: An Efficient and Accurate Scene Text Detector," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 2642–2651.

[7] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep Direct Regression for Multi-oriented Scene Text Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 2017-Octob, pp. 745–753.

[8] S. Tian, S. Lu, and C. Li, "WeText: Scene Text Detection under Weak Supervision," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 2017-Octob, pp. 1501–1509.

[9] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[10] S. Bin Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, "Deep learning based isolated Arabic scene character recognition," in *2017 1st*

International Workshop on Arabic Script Analysis and Recognition (ASAR), 2017, pp. 46–51.

[11] O. Zayene, M. Seuret, S. M. Touj, J. Hennebert, R. Ingold, and N. E. B. Amara, "Text Detection in Arabic News Video Based on SWT Operator and Convolutional Auto-Encoders," *Proc. - 12th IAPR Int. Work. Doc. Anal. Syst. DAS 2016*, pp. 13–18, 2016.

[12] E. Dikici and M. Saraçlar, "Sliding Text Recognition in Broadcast News," *IEEE 16th Signal Process. Commun. Appl. Conf.*, pp. 8–11, 2008.

[13] T. Som, D. Can, and M. Saraçlar, "HMM-based sliding video text recognition for Turkish broadcast news," *24th Int. Symp. Comput. Inf. Sci. Isc.*, pp. 475–479, 2009.

[14] D. Küçük and A. Yazici, "Exploiting information extraction techniques for automatic semantic video indexing with an application to Turkish news videos," *Knowledge-Based Syst. Elsevier*, vol. 24, no. 6, pp. 844–857, 2011.

[15] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.