Urdu Caption Text Detection using Textural Features

Ali Mirza
Bahria University
Islamabad, Pakistan
alimirza@bahria.edu.pk

Marium Fayyaz
Bahria University
Islamabad, Pakistan
fayyaz_marium@yahoo.com

Zunera Seher
Bahria University
Islamabad, Pakistan
seherzunera@hotmail.com

Imran Siddiqi
Bahria University
Islamabad, Pakistan
imran.siddiqi@bahria.edu.pk

ABSTRACT

The amount of multimedia data has increased manifolds in the recent years. This calls for development of efficient retrieval techniques. Among various aspects of content based retrieval, textual content appearing in videos and images serves as a powerful semantic index. Development of such a retrieval system requires detection of text regions, recognition of detected text and generation of indices on keywords. Among these, the focus of the present study lies on detection of textual content from video frames. More specifically, we target the caption Urdu text appearing in News and entertainment channels. A series of image analysis operations is first carried out to identify candidate text blocks in the image. Features extracted from text and non-text regions using Gabor filters and Curvelet transform are fed to two classifiers namely artificial neural network and support vector machine. Evaluations on a database of 1000 video frames reported promising precision and recall.

KEYWORDS

Artificial Urdu Text; Text Detection; Textural Features; Gabor Filters; Curvelet Transform

ACM Reference Format:
Ali Mirza, Marium Fayyaz, Zunera Seher, and Imran Siddiqi. 2018. Urdu Caption Text Detection using Textural Features. In *Proceedings of Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPrai'18).* ACM, New York, NY, USA, 7 pages. https://doi.org/https://doi.org/10.1145/3177148.3180098

1 INTRODUCTION

Remarkable advancements in the amount of multimedia data in general and, digital videos in particular, have led to an increase in the demand of effective retrieval techniques. While the conventional tag-based search continues to be popular, intelligent content based search techniques are being introduced to meet the retrieval challenges offered by huge collections of online and offline video databases. In addition to the visual content and audio stream, an important component of these videos is the textual content which can be exploited for retrieval purposes and makes the subject of our present study.

Scene text and artificial text are generally two broad classes of text appearing in images and videos [30, 39]. Scene text refers to the textual content that is captured using camera. Typical examples are sign boards, advertisement panels & license plates etc. Artificial text or caption text (Figure 1), on the other hand, is purposely superimposed on the video, the news tickers or score cards or the names of anchors, for instance. This caption text, in general, is correlated with the content and is known to be more useful for retrieval applications [18]. Detecting and recognizing instances of artificial text can be employed to develop a complete semantic retrieval framework where videos can be indexed on textual keywords. This allows users to later query the video databases on given keywords and retrieve all videos where a particular keyword has been flashed. Such retrieval frameworks offer an interesting solution to media houses, regulatory bodies and security agencies allowing them to monitor and analyze the content appearing on different News, entertainment and sports channels.

Figure 1: Samples of Urdu Artificial Text in News Videos

The development of a textual content based retrieval system includes a number of components. The candidate text regions are first detected followed by fine localization of textual content. Text is then segmented from background for further processing. In case of videos involving text in multiple scripts, the script of detected text needs to be identified as well [25] so that each script can be processed by the respective recognition engine. Finally, keywords of interest can be matched against predefined vocabularies and videos can be indexed. Among these, the present study aims at the text detection part; more specifically, we focus on artificial Urdu text that appears in hundreds of television channel videos with millions of viewers all over the world.

MedPrai'18, March 27–28, 2018, Rabat, Morocco

ACM ISBN 978-1-4503-5290-1/18/03…$15.00
https://doi.org/https://doi.org/10.1145/3177148.3180098

This paper presents a system to detect occurrences of artificial Urdu text in video frames. The effectiveness of textural features based on Gabor filters and curvelet features is investigated to discriminate between text and non-text regions. Features extracted from text and non-text blocks are employed to train two classifiers, an artificial neural network (ANN) and support vector machine (SVM). Evaluations on a database of 1000 video frames reported promising precision and recall as discussed later in the paper.

The paper is organized as follows. We first present a discussion on the related work in Section 2. Details of the proposed framework including feature extraction and classification steps are presented in Section 3. Section 4 summarizes the experimental settings and the realized results while conclusions are drawn in Section 5.

2 RELATED WORK

Over the recent years, a wide variety of approaches have been proposed for text detection, localization and extraction both in videos and still images. The problem still remains challenging because of complex backgrounds, various font sizes, low contrast and different text orientations. The subsequent sections discuss the well-known text detection methods proposed in the literature. These methods are discussed into two main categories, supervised and unsupervised methods.

In unsupervised approaches, image analysis techniques are applied and segmentation methods (edges, spatial grouping etc.) are used to differentiate text from rest of the image. Generally, unsupervised methods are classified into edge based (gradient based), connected component based (region based), texture based, and color based methods. Edge based methods [5, 12, 14, 20] exploit the high contrast between text and its background by finding the edges in an image. Regions of high edge density are then merged under some heuristics to filter out non-text regions. Typically, an edge detector (e.g. Sobel or Canny operator) is applied on the image to find the edges which is followed by smoothing and morphological operations. Connected component based methods [22–24, 28, 29] exploit the color/intensity of text pixels along with some geometrical heuristics to distinguish text from the background. These methods do not perform well in case of low contrast between text and the background.

Texture-based methods [4, 7, 16, 35] consider textual content in the image as a unique texture which distinguishes itself from the non-text regions. Texture features are generally computed from gray level images or by first transforming the image using filtering or applying frequency domain transformations. A number of studies [1, 8, 21] exploit different textural measures to detect and validate text regions in an image.

In [33, 37, 38] color based methods are used to distinguish text and non-text areas. These methods rely on the assumption that text pixels and the background contain separate color clusters and perform a color based segmentation to extract textual regions.

Supervised approaches for text detection [13, 34, 36, 40, 43] tend to be more sophisticated than the unsupervised techniques and to identify text and non-text blocks, machine learning methods are used. However, these machine learning based methods require significant training data to achieve acceptable classification rates. Among recent supervised techniques, Naive Bayes classifier is employed in [9] while an artificial neural network (trained with features based on local binary patterns (LBP)) is investigated in [3]. A set of mid-level primitives to capture the sub-structures of characters (termed as 'strokelets') is used for text detection in [4] while a character proposal network (CPN) for locating character proposals is used in [42].

In addition to the traditional learning algorithms, a number of recent contributions exploit deep learning techniques to detect (and recognize) textual content in images and videos. Convolutional neural networks (CNNs) have been widely employed to detect both artificial and scene text from videos [13, 15, 17, 31]. Such techniques outperform conventional methods but are computationally expensive and require huge amounts of training data.

3 PROPOSED SYSTEM

The proposed technique for detection of textual content relies on two main steps, detection of candidate regions using image analysis techniques and validation of detected regions using a trained model. Figure 2 presents an overview of the complete system while each of the processing steps is detailed in the following.

3.1 Detection of Candidate Text Regions

A series of image analysis operations is carried out on a video frame to identify the candidate text regions. The image is converted to gray-scale first which makes it independent of the color information (Figure 3-a). It is known that Urdu text is mainly composed of vertical strokes, consequently vertical edges in the image are enhanced using the Sobel operator (Figure 3-b). To remove the false and isolated edges, a horizontal window is employed to scan the image and central pixel is replaced with the average gradient in the window (Figure 3-c). This operation suppresses the isolated gradients and enhances them in areas of high density (potential text regions) [18].

To segment the candidate text regions, the average-gradient image is binarized (using global thresholding as shown in Figure 3-d) and run length smoothing algorithm is applied on the binarized image to merge neighboring components together to form larger components (Figure 3-e). Finally, a series of standard geometrical constraints based on aspect ratio and area are applied on the connected components to identify potential textual regions in the image (Figure 3-f) [19].

The detected regions, in addition to the textual content, also contain non-text regions exhibiting text-like properties. Conseuqently, we validate the detected regions using a machine learning approach. Features extracted from video frames comprising text and non-text blocks are used to train classifiers to discriminate between the two classes as discussed in the following.

3.2 Feature Extraction

The texture features are employed in our study include Gabor filters and curvelets and are detailed as follows.

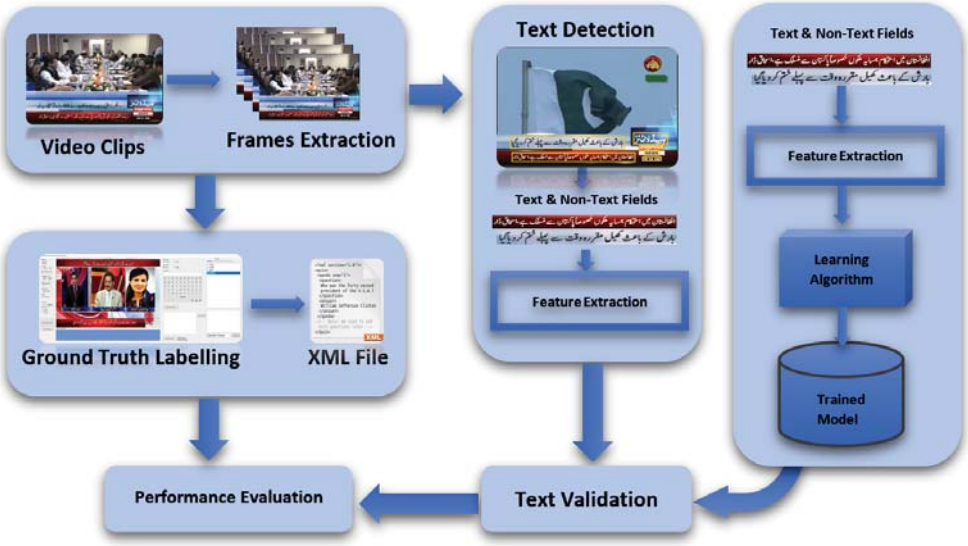

Figure 2: Proposed Framework

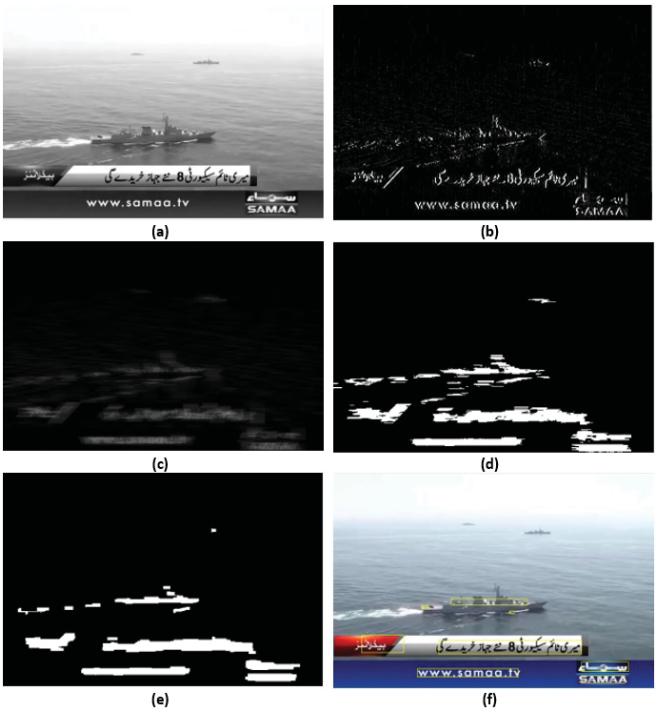

Figure 3: Process of Text Detection and Localization (a): Gray scale image (b): Gradient image (c): Average gradients (d): Binarized gradients (e): RLSA alogrithm (f): Geometrical constraints

3.2.1 Gabor Filters. One of the widely used and popular textural feature based filters are Gabor filters. Gabor filter shares similarities with the visual cortex of mammalian cells. Mammals are able to use bandpass and orientation selectivity as main characteristics of their visual cortex cells which make them respond to specific spatial frequency and direction.

These cortex cells are found in pairs with odd and even symmetry respectively. Various image processing applications are developed based on these similarities of Gabor filters and visual cortex. Fourier transformation of 2D Gabor function can be written as follows.

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-\frac{\acute{x}^2 + \gamma^2 \acute{y}^2}{2\sigma^2})cos(2\pi \frac{\acute{x}}{\lambda} + \psi)$$

Where:

$$\acute{x} = x \cos\theta + y \sin\theta$$

$$\acute{y} = -x \sin\theta - y \cos\theta$$

The above equation shows that complex sinusoids are used to define the product of a Gaussian function. Center frequency and bandwidth of above mentioned filters are controlled using standard deviation of two main components; Gaussian function and complex sinusoid frequency.

In many applications, Gabor filters bank is prepared using with different scales and orientations. With four scales and six orientations, a bank of Gabor filter can be seen in Figure 4 and the same is employed in our study. The visual representation of the Gabor filter bank used for feature extraction in our work is presented in Figure 5. Based on the combination of different orientations and scales, we get 24 filtered images. We calculate the mean and variance of each of the 24 images and place these values in two matrices. FFT is then applied on the mean and variance matrices to generate 48 dimensional feature vector.

3.2.2 Curvelet Features. Curvelet transform was first introduced in 2000 by Candes and Donoho [6] and is known to be an enhanced

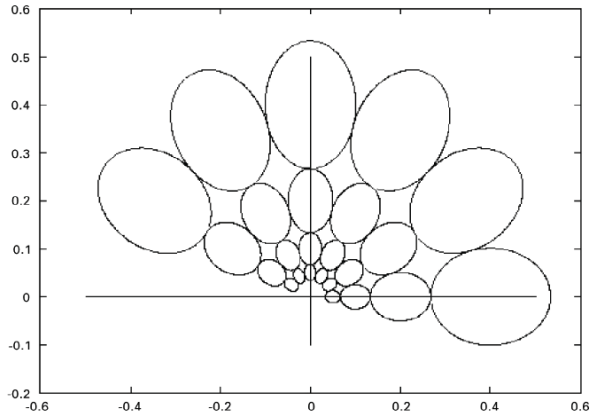

Figure 4: Gabor Filter Bank of different Scales and Orientations [32]

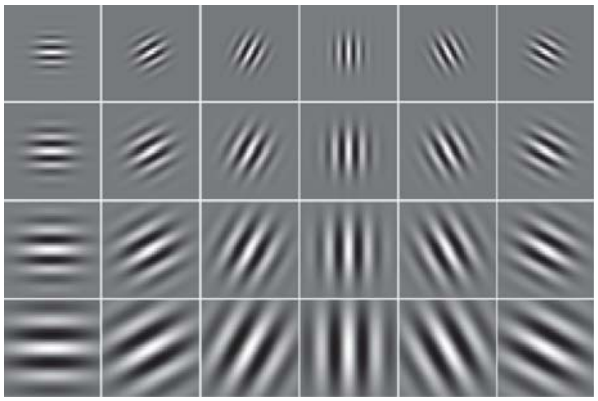

Figure 5: Bank Gabor Filters with 4 Scales and 6 Orientations

edge representation as compared to the wavelet transform. The initial version of curvelets was redesigned and then introduced as Fast Digital Curvelet Transform (FDCT). The wavelet transform is constrained in its orientations and curved singularity representations. Curvelet transform is a generalization of the wavelet transform in higher dimensions which aids in representing images at various scales and angles [10]. Figure 6 shows a sample Curvelet Transform applied to an Urdu word image.

Curvelets are known to be effective as a feature descriptor for images containing textual occurrences as observed in [2, 11, 26]. In our study, we exploit the curvelet transform to discriminate between text and non-text regions. Pixels in the close proximity of one another give rise to edges, the strokes of text in our case. The 2D Fast Fourier Transform (FFT) of the curvelet transformed image is taken and the frequency plane obtained is partitioned into radial and angular divisions.

In our implementation, we applied the curvelet transform using 10 curvelets producing 10 values. Combining the Gabor and curvelet features generates a 58 (48+10) dimensional feature vector.

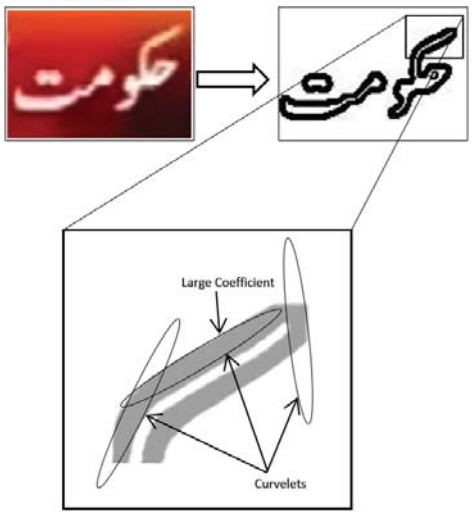

Figure 6: Curvelet Transform applied to an image

3.3 Classification

For classification, two state-of-the-art classifiers namely Support Vector Machine (SVM) and feed forward Artificial Neural Network (ANN) have been employed. Gabor and curvelet features are used to train these classifiers and results are reported on the individual as well as combined features as discussed in the following section.

4 EXPERIMENTS AND RESULTS

Experiments are performed on a custom developed dataset comprising 2000 video frames from various News and entertainment channels. A Software tool was developed for ground truth labeling of the video frames generating an XML file (Sample XML file shown in Figure:7) containing information on the location of each text line along with the actual textual content. A screenshot of the labeling tool is shown in Figure 8. Text and non-text regions of 1000 images were used to train the classifiers for the validation step while 1000 frames were employed as the evaluation set. For performance evaluation, area of text (in terms of number of pixels) is used to compute precision and recall. Let A_C be the text area detected by the system and A_G be the true area of text, then Precision and Recall are defined as follows.

$$Precision = \frac{A_C \cap A_G}{A_C}$$

$$Recall = \frac{A_C \cap A_G}{A_G}$$

Table 1 summarizes the precision and recall values realized using the Gabor and curvelet features (and their combination) with ANN and SVM classifiers. It can be seen from the realized results that Gabor filters report better precision and recall measures as compared to the curvelet features. This may be attributed to the relatively low dimensionality of curvelet features (10) as compared to that of Gabor features (48). The performance improves when

```
<?xml version="1.0" encoding="utf-8"?>
<VideoLabel>
  <FrameMetaData>
    <Video>Express_News_20170423_212136</Video>
    <Channel>Samaa News</Channel>
    <FrameNo>Express_News_20170423_212136_00301</FrameNo>
  </FrameMetaData>
  <TextFeeds TotalFeeds="6">
    <UrduFeeds TotalUrduFeeds="6">
      <TextLine ID="1" TextType="Artificial" X="347" Y="35" Width="203" Height="52" Text="بریکنگ نیوز" />
      <TextLine ID="2" TextType="Artificial" X="593" Y="170" Width="215" Height="67" Text="ڈاکٹر نے سیکیورٹی" />
      <TextLine ID="3" TextType="Artificial" X="581" Y="244" Width="228" Height="64" Text="گارڈ سے آبائی گھر" />
      <TextLine ID="4" TextType="Artificial" X="590" Y="326" Width="219" Height="51" Text="خالی کرنے کا کہا تھا" />
      <TextLine ID="5" TextType="Artificial" X="735" Y="515" Width="71" Height="23" Text="نیوز" />
      <TextLine ID="6" TextType="Artificial" X="95" Y="521" Width="576" Height="56" Text="دھماکا کے بعد علاقہ سیل، سرچ آپریشن جاری، آئی ایس پی آر" />
    </UrduFeeds>
    <EnglishFeeds TotalEnglishFeeds="0" />
  </TextFeeds>
</VideoLabel>
```

Figure 7: XML File of a Labeled Frame

Figure 8: Ground Truth Labeling Tool

both the features are combined. Comparing the performance of the two classifiers, SVM reports better results as compared to ANN. The best performance reads a recall of 0.89 and a precision of 0.72 with SVM classifier and a combination of both the features.

Table 1: Detection Performance on 1000 video frames

	Classifier			
	ANN		SVM	
Features	**Precision**	**Recall**	**Precision**	**Recall**
Gabor	0.66	0.78	0.68	0.83
Curvelets	0.59	0.72	0.61	0.77
Combined	0.69	0.85	0.72	0.89

While detection of text from video frames is a mature area of reserach, limited literature is available when it comes to Urdu text. We compare the performance of the proposed textural measures with our previous work that employed gray-level co-occurrence matrices (GLCM) for recognition of text and non-text chunks [19]. Furthermore, we also summarize the performance of well-known text detection systems that work on Arabic text as both Arabic and Urdu share many common characteristics. The comparison is presented in Table 2. The only work that reports quantified results on detection of Urdu text is presented in [19], a multi-lingual system that, in addition to other scripts, also considers Arabic and Urdu. On 200 frames with occurrences on Urdu text, the system reports a precision of 0.65 and a recall of 0.80. Among studies on Arabic video text, Zayene et al. [41] reports the best performance wih precision and recall values of 0.83 and 0.85 respectively on a large dataset of 1843 video frames.

Table 2: Evaluation of Text Detection Methods

Method	Frames	Script	Precision	Recall
Epshtein et al. [7]	425	Arabic	0.53	0.36
Zayene et al. [27]	425	Arabic	0.67	0.73
Jamil et al. [19]	200	Arabic	0.66	0.85
Zayene et al. [41]	1843	Arabic	0.83	0.85
Jamil et al. [19]	200	Urdu	0.65	0.80
Proposed Method	**1000**	**Urdu**	**0.72**	**0.89**

5 CONCLUSION AND FUTURE WORK

We presented an effective method for detection and validation of artificial Urdu text appearing in video images. A series of image

analysis techniques is first applied to identify the potential text regions. These regions are later validated using two classifiers trained to distinguish between text and non-text regions using textural measures. Experiments on a custom developed database reported promising precision and recall values. In our future study on this subject, we intend to develop a video OCR to recognize the detected textual content. This will lead to the development of a complete textual content based video indexing and retrieval system, especially targeting Urdu text. We also intend to extend the system to incorporate processing of audio streams so that indexing can be carried out on spoken (key) words as well.

ACKNOWLEDGMENT

This research is funded by IGNITE, National Technology Fund, Pakistan, under grant number ICTRDF/TR&D/2014/35.

REFERENCES

[1] Ansari Aasif and Muzammil H. Mohammed. 2015. Content based Video Retrieval Systems-Methods, Techniques, Trends and Challenges. *International Journal of Computer Applications 112.7* (2015).

[2] Majumdar Angshul. 2007. Bangla basic character recognition using digital curvelet transform. *Journal of Pattern Recognition Research* 2, 1 (2007), 17–26.

[3] Lluis Gomez-Bigorda Anguelos Nicolaou, Andrew D. Bagdanov and Dimosthenis Karatzas. 2016. Visual Script and Language Idenfication. *2016 12th IAPR Workshop on Document Analysis Systems* (2016).

[4] Xiang Bai, Cong Yao, and Wenyu Liu. 2016. Strokelets: A Learned Multi-Scale Mid-Level Representation for Scene Text Recognition. *IEEE TRANSACTIONS ON IMAGE PROCESSING* 25 (2016).

[5] Sudipto Banerjee, Koustav Mullick, and Ujjwal Bhattacharya. 2013. A robust approach to extraction of texts from camera captured images. In *International Workshop on Camera-Based Document Analysis and Recognition*. Springer, 30–46.

[6] Emmanuel J Candes and David L Donoho. 2000. *Curvelets: A surprisingly effective nonadaptive representation for objects with edges*. Technical Report. DTIC Document.

[7] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. 2010. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2963–2970.

[8] Vigneshwari G. and A. Juliet. 2015. Optimized searching of video based on speech and video text content. *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on. IEEE* (2015).

[9] Lluis Gomez and Dimosthenis Karatzas. 2016. A fine-grained approach to scene text script identification. *2016 12th IAPR Workshop on Document Analysis Systems* (2016).

[10] Tanaya Guha and QM Jonathan Wu. 2010. *Curvelet based feature extraction*. INTECH Open Access Publisher.

[11] Joutel Guillaume, Eglin Véronique, Bres Stéphane, and Emptoz Hubert. 2007. Curvelets based feature extraction of handwritten shapes for ancient manuscripts classification. In *Electronic Imaging 2007*. International Society for Optics and Photonics, 65000D–65000D.

[12] D. S. Guru, S. Manjunath, P. Shivakumara, and C. L. Tan. 2010 USA. p. 501-506.. An Eigen Value Based Approach for Text Detection in Video. In *9th IAPR International Workshop on Document Analysis Systems*.

[13] Tong He, Weilin Huang, Yu Qiao, and Jian Yao. 2016. Text-Attentional Convolutional Neural Network for Scene Text Detection. *IEEE TRANSACTIONS ON IMAGE PROCESSING* 25 (2016).

[14] Rong Huang, Palaiahnakote Shivakumara, and Seiichi Uchida. 2013. Scene character detection by an edge-ray filter. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 462–466.

[15] Weilin Huang, Yu Qiao, and Xiaoou Tang. 2014. Robust scene text detection with convolution neural network induced mser trees. In *European Conference on Computer Vision*. Springer, 497–511.

[16] X Huang. 2011. A Novel Video Text Extraction Approach Based on Log-Gabor Filters. In *4th International Congress on Image and Signal Processing*.

[17] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision* 116, 1 (2016), 1–20.

[18] Akhtar Jamil. 2011. Edge-based Features for Localization of Artificial Urdu Text in Video Images. *Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE* (2011).

[19] Akhtar Jamil, Azra Batool, Zumra Malik, Ali Mirza, and Imran Siddiqi. 2016. Multilingual Artificial Text Extraction and Script Identification from Video Images. *International Journal of Advanced Computer Science & Applications* 1, 7 (2016), 529–539.

[20] A Jamil, I Siddiqi, F Arif, and A Raza. Jamil, A., et al. Edge-based Features for Localization of Artificial Urdu Text in Video Images. in International Conference on 2011.. Edge-based Features for Localization of Artificial Urdu Text in Video Images. In *International Conference on Document Analysis and Recognition*.

[21] et al. Khatri, Mohd Javed. 2015. Video OCR for Indexing and Retrieval. *International Journal of Computer Applications 118.2* (2015).

[22] Y.C. Kiran and L.N. C. 2012. Text extraction and verification from video based on SVM. *World Journal of Science and Technology* 2(5) (2012), 124–126.

[23] Hyung Il Koo and Duck Hoon Kim. 2013. Scene text detection via connected component clustering and nontext filtering. *IEEE transactions on image processing* 22, 6 (2013), 2296–2305.

[24] SeongHun Lee, Min Su Cho, Kyomin Jung, and Jin Hyung Kim. 2010. Scene text extraction with edge constraint and text collinearity. In *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 3983–3986.

[25] Zumra Malik, Ali Mirza, Akram Bennour, Imran Siddiqi, and Chawki Djeddi. 2015. Video Script Identification using a Combination of Textural Features. In *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 61–67.

[26] Manuel Manju and Saidas SR. 2015. Handwritten Malayalam Character Recognition using Curvelet Transform and ANN. *International Journal of Computer Applications* 121, 6 (2015).

[27] Zayene Oussama, Hennebert Jean, Touj Sameh Masmoudi, Ingold Rolf, and Amara Najoua Essoukri Ben. 2015. A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 996–1000.

[28] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu. 2011. A hybrid approach to detect and localize texts in natural scene images. *IEEE Transactions on Image Processing* 20, 3 (2011), 800–813.

[29] T.Q Phan, P Shivakumara, and C.L Tan. 2009. A Laplacian Method for Video Text Detection. In *10th International Conference on Document Analysis and Recognition*.

[30] Ye Qiaoyang and David Doermann. 2015. Text detection and recognition in imagery: A survey. *IEEE Transactions On Pattern Analysis And Machine Intelligence* 37 (July 2015).

[31] Xiaohang Ren, Kai Chen, Xiaokang Yang, Yi Zhou, Jianhua He, and Jun Sun. 2015. A new unsupervised convolutional neural network model for chinese scene text detection. In *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 428–432.

[32] Srinivasan Selvan and Srinivasan Ramakrishnan. 2007. SVD-based modeling for image texture classification using wavelet transformation. *Image Processing, IEEE Transactions on* 16, 11 (2007), 2688–2696.

[33] P. Shivakumara, T.Q. Phan, and C.L. Tan. 2010. New Fourier-Statistical Features in RGB Space for Video Text Detection. *IEEE Trans. Circuits Syst. Video Techn* (2010), 1520–1532.

[34] P. Shivakumara, R.P. Sreedhar, Trung Quy Phan, Shijian Lu, and C.L. Tan. 2012. Multioriented Video Scene Text Detection Through Bayesian Classification and Boundary Growing. *IEEE Transactions On Circuits And Systems For Video Technology* 22(8) (2012).

[35] Weijuan Wen, , Xianglin Huang, Lifang Yang, and Zhao Yang. 2009. An Efficient Method for Text Location and Segmentation. In *WRI World Congress on Software Engineering (WCSE 09), Beijing, China. p. 3 - 7.*

[36] J. Ye, L.L. Huang, , and X. Hao. 2009. Neural Network Based Text Detection in Videos Using Local Binary Patterns. In *Chinese Conference on Pattern Recognition CCPR2009 p. 1-5.*

[37] Chucai Yi and Yingli Tian. 2012. Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification. *IEEE Transactions on Image Processing* 21, 9 (2012), 4256–4268.

[38] J Yi, Y Peng, and J Xiao. 2007. Color-based clustering for text detection and extraction in image. In *15th international conference on Multimedia, Germany.*

[39] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu. 2016. Text detection, tracking and recognition in video: a comprehensive survey. *IEEE Transactions on Image Processing* 25, 6 (2016), 2752–2773.

[40] X.-C.and X. Yin Yin and K. Huang. 2013. Robust Text Detection in Natural Scene Images. *CoRR abs/1301.2628.* (2013).

[41] Oussama Zayene, Mathias Seuret, Sameh M Touj, Jean Hennebert, Rolf Ingold, and Najoua E Ben Amara. 2016. Text detection in Arabic news video based on SWT operator and convolutional auto-encoders. In *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*. IEEE, 13–18.

[42] Shuye Zhang, Mude Lin, Tianshui Chen, Lianwen Jin, and Liang Lin. 2016. CHARACTER PROPOSAL NETWORK FOR ROBUST TEXT EXTRACTION. *ICASSP* (2016).

[43] W. Zhen and W. Zagiqiang. 2009. A comparative study of feature selection for SVM in video text detection. In *2nd International symposium on Computational Intelligence and Design p. 552 - 556.*