Open Access
Full Length Article

ORIGINAL ARTICLE

A Text Tracking Method using Maximally Stable Extremal Regions and Speeded up Robust Features

Samabia Tehsin[1], Sumaira Kausar

[1] Associate Professor, Department of Computer Science, Bahria University Islamabad
[2]Associate Professor, Department of Computer Science, Bahria University Islamabad

ABSTRACT

Content-based image retrieval is an active research area because of the vast applications of image and video collections. Text embedded in the video can provide a significant contribution to identifying the contents of the multimedia data and facilitate the process of video indexing, retrieval, and analysis. Text tracking is a vital part of the text extraction process. It can speed up the text extraction process and also improves the text localization accuracy for videos. This paper proposes a novel approach for text tracking in videos. This method experimentally proposes Maximally Stable Extremal Regions (MSER) and Speeded up Robust Features (SURF) descriptors for text tracking in videos. The proposed method improves the tracking accuracy and efficiency of video text objects. MSER is used for interest point detection as Extremal regions are affine invariant, so it is a suitable option for text tracking which can undergo scale, rotation, and translation changes. SURF is chosen as a feature descriptor because of its scale and rotation invariance, speed, and robustness. The proposed technique is testified on two datasets; the first is designed to test the tracking methodology as part of this research and the second is a publicly available Youtube Video Text (YVT) dataset. It provides Multiple Object Tracking Precision (MOTP) of 0.53 and Multiple Object Tracking Accuracy (MOTA) of 0.48. Succeeding analysis on diverse datasets endures verification of the fact that the projected technique demonstrates visible improvements to text tracking in videos.

Keywords: Text tracking, video retrieval, Caption Test, Document Analysis

Author`s Contribution
[1] Data analysis, interpretation, and manuscript writing, Active participation in data collection
[2] Conception, synthesis, planning of research, Interpretation, and discussion

Address of Correspondence
Samabia Tehsin
Email: Stehseen.buic@bahria.edu.pk

Article info.
Received: September 03, 2021
Accepted: December 21, 2021
Published: December 30, 2021

Cite this article: *Tehsin S. Kausar S. A test tracking method using maximally stable extremal regions and speeded up robust features. inf. commun. technol. robot. appl.2021; 12(2):39-47.*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

This document provides instructions for preparing manuscripts for publishing in The Scientific Journal of Riga Technical University. The document is also a sample of the layout for the manuscripts submitted for publication. In these times, there is an increased requirement for efficient retrieval and indexing of multimedia data due to immense applications and the growth of images and video libraries. Plenty of work is done to index and retrieve multimedia content. These methods use different features e.g. color, texture, silhouette, etc. But for text-based retrieval systems, text inserted in images and videos is a very significant choice. Graphical texts embedded in multimedia data communicate information regarding news bulletins, the title of movies, brand names in advertisements, match scores and summaries, date and time of an incident. All these pieces of information are very useful for the efficient retrieval and indexing of multimedia libraries.

Multimedia embedded text is broadly categorized into two groups, namely caption and scene text. Caption text is superimposed on an image or video during the editing phase. It is also referred to as artificial text or superimposed text in literature. Examples of caption text include match scores and summaries, news bulletins, and movie credits. Scene text appears in the real scene of the image or video frame e.g. road directions, billboards brands, nameplates, etc. Scene text is also known as graphics text.

Caption text more often provides good knowledge about the real contents of the multimedia and hence can be a very choice for multimedia retrieval.

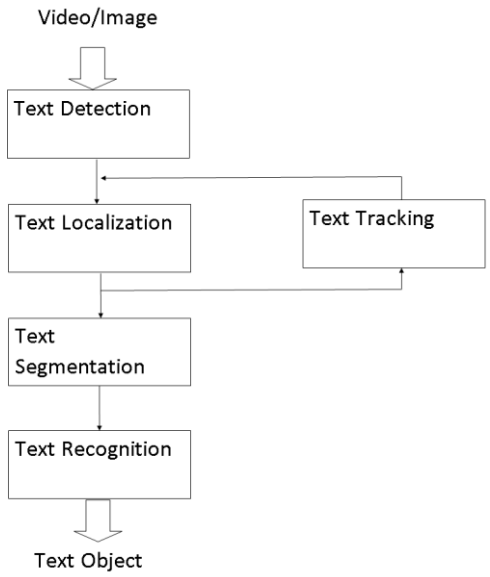


Figure 1. The architecture of text extraction and recognition process

The text extraction and recognition process include five stages that are text detection, text localization, text tracking, segmentation or binarization, and character recognition (See Fig. 1). Text detection and text localization phases aim to identify the boundaries of all the text instances and form bounding boxes around individual occurrences. Text tracking is used to identify all the instances of identical text that appear in adjacent frames of a video. Text tracking may use temporal information to eradicate incorrect text localization and improve the outcomes of text detection and localization phases. It also delivers the consolidated input to the binarization and recognition phase by eliminating the redundant text appearing in the consecutive frames.

Enormous research efforts are made in the field of text detection and localization. However, very limited approaches to text tracks are presented in the literature [1, 2, 3, 4]. The proposed methodology focuses on the text tracking part of the text extraction system. This phase can play a vital role in eliminating false alarms and can speed up the text extraction process in videos. In the proposed methodology, Maximally Stable Extremal Regions (MSER) along with the Speeded-Up Robust Features (SURF) features are used for tracking the occurrences of the same text in consecutive frames. The rest of the paper is structured as follows. Section 2 highlights important developments in the field, and Section 3 presents the proposed text tracking approach. Section 4 explains the dataset and metric used. It also shows the result of the presented methodology. Section 5 concludes the research findings.

LITERATURE REVIEW

A variety of approaches to text extraction is presented in the literature [5-15]. State-of-the-art techniques are demonstrated in comprehensive surveys [16-25]. But, the majority of the work ignores the temporal information in videos [26-31]. Very limited work focused on the text tracking aspect in the text extraction process. Conventional methods for the text recognition of video mainly focus on recognizing the text in every single frame independently [32, 36].

Lienhart [37] proposed a complex-valued multilayer feed-forward network to detect text in images and videos, but it has a relatively high computational complexity due to the multi-resolution approach. Li et al. [38] used the Sum of Squared Difference to track the text. It can track the text with a pure translational motion, but it cannot deal with severe background changes. Jiangbo Xu et al [39] presented a text extraction method in DCT compressed domain. Potential text blocks are identified by DCT texture energy. An adaptive temporal constraint method is also suggested to improve the text detection results. Gallavata et al. [40] used motion vectors of MPEG videos for text tracking. Qian et al. [41] proposed a methodology of text tracking for compressed video. Horizontal and

vertical projection profiles and Mean Absolute Difference (MAD) is applied to track the rolling and static text respectively.

W. Huang et al. [42] used temporal information obtained by dividing a video frame into sub-blocks and calculating the inter-frame motion vector for each sub-block. This work focuses only on text credits of uniform velocity. It fails to deal with scaling and randomly moving text. Tanaka et al. [43] suggested using the cumulative histogram to find the correspondence of detected text instances with the text instances of consecutive frames.

W. Zhen et al. [44] and LI et al. [45] exploit the temporal redundancy of neighboring frames but cannot deal with moving text. Some wearable camera applications are also proposed that provide the methodology for text tracking in real-time [46, 47]. Most of the presented work fails to deal with complex background text occurrences in videos. Furthermore, these approaches can only apply to stationary text or can deal with simple motions like simple scrolling credits. The proposed methodology presents the text tracking mechanism, which can deal with complex text movements such as zooming in and out and having complex backgrounds. It can also deal with motion blur, rotation, and scale changes.

RESEARCH METHODOLOGY

Temporal redundancy is a very significant feature of videos that is not present in image data. Text in videos persists for a certain duration to make it readable. Text normally remains on screen for more than one second. NTSC standard television videos refresh rate is app. 30 frames per second (FPS). It means a text should appear in at least thirty consecutive frames. This feature can be subjugated to minimize the false alarms in the detection and localization phase.

Animated effects are commonly applied to caption text in videos. Text appearing in videos may change in scale or/and location due to zooming in or out and translation. However, that change is very gradual to make it smooth for the viewers. It means the two consecutive frames having the same text should have the same characters and similar size and location, if not the same. Therefore, the text tracking phase aims to trace all the occurrences of the same text appearing in different frames of the video.

Text Detection and localization are applied before the tracking phase. The region-based methodology by Tehsin et. al. [48] is used for detecting and localizing text in the individual frame. This technique is chosen for its computational efficiency and good detection results. Tracking results are highly dependent upon the detection results and the chosen detection method focuses on the caption text, therefore the presented tracking method can deal with the caption text tracking only. Tracking of text events in a video is achieved in three steps. First interest points are extracted using the Maximally Stable Extremal Regions (MSER). Interest points are extracted for every detected text object in a frame. The second step is to define the feature descriptor for every interest point extracted; using the Speeded Up Robust Feature (SURF) descriptor and the last step is to match the descriptor of a detected text object with all the text objects in the adjacent frames. Combining the MSER and SURF provides the tracking tool that can deal with motion blur, scale change, and rotation variance.

Interest Point Detection

Interest points are detected using MSER [49]. Extremal regions are affine invariant, so it is a suitable option for text tracking which can undergo scale, rotation, and translation changes. MSER is used for interest point detection instead of corner points. Corner points are effective in high-resolution images but are not appropriate for low contrast low-resolution videos [50]. MSER is equally effective for low-resolution videos.

Let $I_i^t : X \rightarrow [0,1]$ is a grayscale image defined on a domain $X \subset \mathbb{R}^2$. Here the image I_i^t is the i^{th} detected text object in the T^{th} frame. This image can be fully defined as a group of level sets. $\{x \in X : I_i^t(x) = l\}$ is a level set off I at l, where $l \in [0,1]$? A sub-level set can be characterized by an intensity level k and it can be defined as $\{x \in X : I_i^t(x) \leq k\}$. By varying the intensity level $0 \rightarrow 1$, we get the sequence of sub-level sets. This sequence can be visualized as a component tree. Let $R \subset X$ represent the region in the image surrounded by the boundary ∂R. R is said to be an extremal region if $\forall p \in R, q \in \partial R: I(p) > I(q) or I(p) < I(q)$. Let $R_1, \ldots., R_{k-1}, R_k, \ldots.$ is a sequence of nested extremal

regions, i.e. $R_{k-1} \subset R_k$. The stability of the region R_k can be defined as

$$\Phi(R_k) = \frac{\Delta(R_k)}{\left|\frac{d}{dk}\Delta(R_k)\right|} \quad (1)$$

Here $\Delta(R_k)$ is the area R_k. A stable region changes very slowly with the change in the threshold value k. A region R_k is maximally stable iff $\Delta(R_k)$ has a local maximum at k. The regions are fully defined by the extremal property of the intensity function of the region and its boundary. In general, an MSER is an extremal region whose area is left almost the same over a sequence of intensity levels. Extracted interest points for detected text objects are shown in Fig 2.

Figure 2: Detected Interest Points in the Text Objects

Feature Descriptor

In this phase, for every keypoint $p_{i_j}^{\iota}$, extracted in the previous step, the feature set $f_{i_j}^{\iota} \in \mathbb{R}^n$ is computed. Where $p_{i_j}^{\iota}$ is the j^{th} interesting point of the i^{th} text object in T^{th} the frame SURF [51] is used to define the feature set. SURF is chosen as a feature descriptor because of its scale and rotation invariance, speed, and robustness [51, 52]. It is studied that MSER is sensitive to image blur[53]. To compensate for this sensitivity, SURF is used as the feature descriptor that works very well in the presence of image blur [51].

SURF estimates the orientation of the interesting point by constructing a circular region around it. Haar wavelet responses, weighted with Gaussian centered at the key point, are used to estimate the orientation. This is used to make the descriptor rotation invariant. SURF descriptors are computed by constructing a square region around the key point. The square region is split into 4×4 windows $W(u,v)$. Four descriptors are extracted for each window $W(u,v)$ by shifting it by $(\delta x, \delta y)$. These descriptors are the intensity patterns of the particular window. The feature vector $f_{i_j}^{\iota}$ for the interesting point $p_{i_j}^{\iota}$ can be defined as

$$f_{i_j}^{T} = \bigcup_{w=1}^{W} \left(\sum dx_w, \sum dy_w, \sum |dx_w|, \sum |dy_w|\right) \quad (2)$$

Here, dx_w is the horizontal wavelet transform and dy_w is the vertical wavelet transform of the respective w window. These wavelet responses are summed up for each window that structures the first two entries of the feature vector $\sum dx$ and, $\sum dy$. Where, $|dx|$ and $|dy|$ are added in the feature vector to convey the information about the polarity of the intensity changes. W is the total number of windows in the square region and is equal to $U\ x\ V$. Here $U = V = 4$ and $W = 16$. So the length of the feature vector is 64 and the possible values $u\ v$ range from 1 to 4.

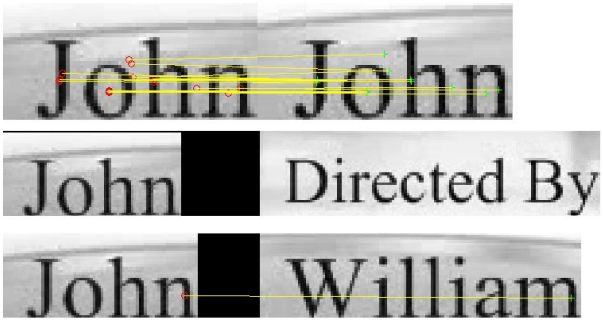


Figure 3: Matching results for different candidate text regions

Matching

Search space for feature matching is reduced by spatial constraints. Interest point is matched in neighborhood vicinity only because text event in neighboring frames may translate slightly to keep the smoothness of the video. Let (x_i^t, y_i^t) be the center of i^{th} a text object in the T^{th} frame and $(x_{i'}^{t'^{T+1}}, y_{i'}^{t'^{T+1}})$ be the center of i'^{th} the text object in the $T+1^{th}$ frame then before matching the interest points, it should obey the following spatial constraint:

$$\sqrt{(x_i^T - x_{i'}^{T+1})^2 + (y_i^T - y_{i'}^{T+1})^2} \leq \delta \quad (3)$$

Here δ is the threshold value and it is dependent upon the dimension of the text object and frame size

$$\delta \propto \frac{\min(L^T, H^T)}{\min(L_i^T, H_i^T)}$$

Where (L^t, H^t) are the length and height of T^{th} the frame and (L_i^t, H_i^t) are for the i^{th} text object of T^{th} the frame. After filtering through the spatial constraint, text objects are matched with the text objects of adjacent frames. Nearest Neighbor Ratio [54, 55] is used as the matching method. The Sum of squared distances is used as a feature matching metric. Fig 3 shows some matching results of the text object with the candidate text objects of the neighboring frame.

The same id (identification number) is assigned to the matched text objects of consecutive frames and a new id is assigned to the text object having no match.

RESULT AND ANALYSIS

For our experiments, we used a dataset consisting of ten videos collected from different sources of news, sports, movies, and cartoon videos. These videos are having static and animated caption text. Animations include scrolling up and down, horizontal moving credits, and zooming in/out. Fig 4 shows some sample tracking results of the proposed technique on the dataset videos.

Figure 4: Sample results of the proposed method

Evaluating text tracking task is a Spatio-temporal task. Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA) is used for the evaluation of tracking task [56]. This evaluation metric penalizes fragmentation in spatial as well as temporal domains.

Let $G_i^{(t)}$ is the i^{th} ground truth object in the t^{th}frame and $D_i^{(t)}$ is the corresponding detected object, then MOTP can be defined as

$$\aleph = \frac{\sum_{i=1}^{N_m}\sum_{t=1}^{N_f}\left[\frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)} \cup D_i^{(t)}|}\right]}{\sum_{j=1}^{N_f} N_m^j} \quad (4)$$

Here, $N_m$are the mapped text objects in the entire track, N_f is the total number of frames, and N_m^t are the number of mapped objects in the t^{th}frame. MOTA is used to compute the accuracy of text tracking tasks. Three error ratios naming missed count(τ), false alarm (ψ), and several switches(ж), are used to calculate the MOTA$(\beth)$ of any video sequence.

$$\beth = 1 - \frac{\sum_{i=1}^{N_f}(\varepsilon_m(\tau_i) + \varepsilon_f(\psi_i) + log_e(\text{ж}))}{\sum_{i=1}^{N_f} N_G^i} \quad (5)$$

Where ε_m and ε_f are the cost functions for missed counts and false alarm penalties. Here the cost functions are considered as the associated weights. N_G^i is the number of ground truth boxes in the i^{th}frame. Evaluation results of the proposed methodology and its comparison with the Shi et. al. method [57] are presented in Table 1. Shi's method is a discrete cosine transformation (DCT) based method. Shi's method is chosen for comparison for its efficiency and accuracy.

Table 2 shows the proposed method for tracking the performance of stationary text, moving texts in horizontal and vertical directions, and scale-changing text. Results show that static text gives better results than animated once. The reason for misdetection and mistracking in the animated text is the variation in size and position of the text. Moving text changes its background and is hence difficult to detect and track.

Table 1. Performance Evolution of Proposed Methodology

Algorithm	MOTP	MOTA
Proposed	0.53	0.48
Shi [39]	0.41	0.37

Table 2. Performance Evaluation for Different Types of Videos Test		
Animation Type	**MOTP**	**MOTA**
Static	0.63	0.58
Scroll	0.49	0.46
Zoom In/Out	0.48	0.44

DISCUSSION

Lower MOTP and MOTA are mostly because of constraints in the detection phase. To only check the tracking phase capabilities, we assume the detection phase is error-free. MOTA is dependent upon text detection as well as text tracking results. To highlight the results of text tracking only, text detection results are simulated and represented as the blue plot in Fig 5. As the simulated detection results increases, corresponding MOTA results also improves. In the ideal scenario, when text detection has a 100% result, then MOTA reaches the accuracy of 94%. Fig 5 shows the tracking results with the simulated detection results.

Two factors play a vital role in the performance of text tracks, one is the complexity of the background and the second is the number of text objects to track at once. Fig 6(a) shows that the performance of text tracking degrades with the increase in the average number of texts per frame. Fig 6(b) shows the effect of average text objects on three error ratios. The proposed method suffers less from missed count than the Id switch and false alarm.

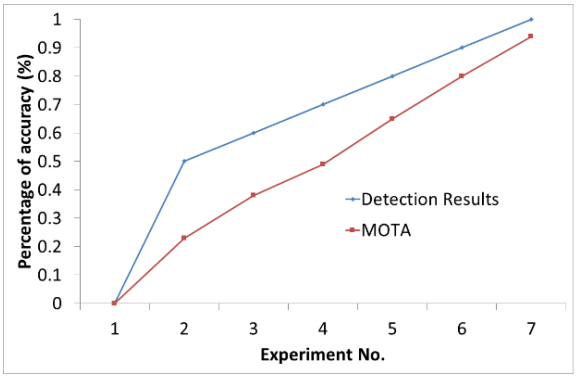


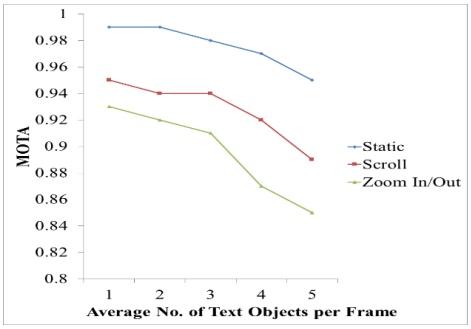
Figure 5: Comparison of tracking results for different simulated detection results

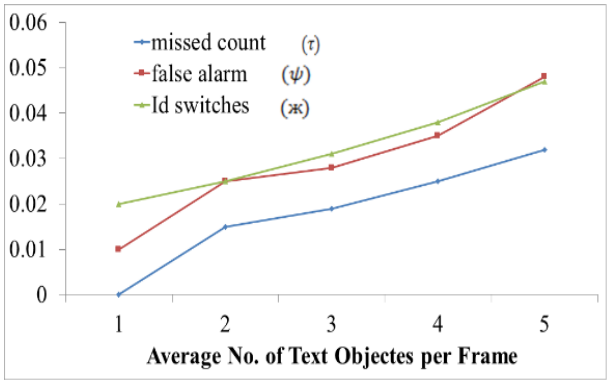
Another important factor in text tracking is the complex background. When the background is complex then it changes drastically with the motion of the text. Feature descriptor is based on the intensity map around the interesting point, so the proposed method mistrack some of the text objects that have a complex background (See Fig 6(c)). The complexity of background can be defined as the very frequent intensity changes in spatial and temporal domains.

Fig 7 shows the confusion matrix for the proposed method tracking results. It is observed that mostly mistracking results in the Id switching. The probability of mismatching a text object with a false object is less. In Fig 7, 'a' and 'b' are labels of the arbitrary text objects $a \neq b$.

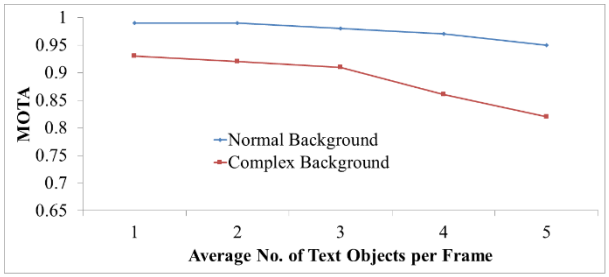
For the feature descriptor, we have experimented SURF with 64 features and with 128 features. Experiments propose the SURF with 64 features. There is a slight advantage in the accuracy of tracking using SURF with 128 features but it almost doubles the matching time. Experiments are conducted on a 2.1 GHz processor with 4.00 GB of RAM. The average time of processing a frame using 64 features is 0.0040 sec, while it takes 0.0076 sec for SURF-128. So we have chosen SURF with 64 features. Table 3 shows tracking results for both options. SURF-128 can be chosen for the systems, where speed can be compromised for accuracy.

The proposed tracking algorithm is also tested on the publicly available YouTube Video Text (YVT) dataset [1]. The evaluation metric for this dataset is the Average Tracking Accuracy (ATA) VACE metric [58]. ATA offers a Spatio-temporal evaluation that can quantify the correctly detected and tracked words while penalizing for fragmentation, false negatives, and false positives. Table 4 shows the performance of the proposed method on the caption text subset of the YVT dataset and its comparison with other methods.





(b)



(c)

Figure 6. (a) Effect of average text objects on MOTA (b) Effect of average text objects on Error Ratios (c) Effect of background complexity on MOTA

DISCUSSION

This paper proposes a text tracking method for videos. MSER and SURF-based tracking methodology have been presented. A combination of MSER and SURF provides a very effective tool for text tracking in videos. Results of the proposed technique are shown using a diverse dataset. Text tracking can help in improving the detection results and can improve the computational efficiency of text recognition by providing consolidated text. Hence, the proposed method can be helpful for indexing and retrieval of video content.

This paper is tested on the caption text data and provided promising results. It can also be used for scene text tracking. Moreover, other feature descriptors can also experiment for the better tracking results

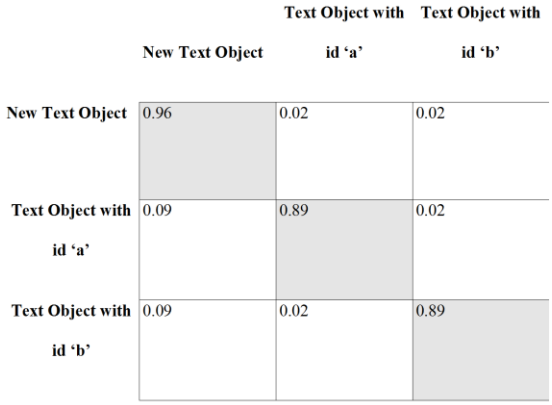


Figure 7: Confusion Matrix for text tracking

COMPARISON OF SURF-64 AND SURF-128 FOR TEXT TRACKING		
		MOTP
Animation Type	SURF-64	SURF-128
Static	0.63	0.63
Scroll	0.49	0.50
Zoom In/Out	0.48	0.51

Table 4. Results on YVT Dataset			
Methods	Proposed	Plex+T **Error! Reference source not found.**	DR+T [1]
ATA	**0.56**	**0.28**	**0.32**

REFERENCES

1. Z. Ahmed and H. Singh, "Text Extraction and Clustering for Multimedia: A review on Techniques and Challenges," presented at the 2019 International Conference on Digitization (ICD), 2019/11, 2019. [Online]. Available: http://dx.doi.org/10.1109/icd47981.2019.9105905.
2. G. J. Ansari, J. H. Shah, M. Sharif, and S. u. Rehman, "A novel approach for scene text extraction from synthesized hazy natural images," Pattern Analysis and Applications, vol. 23, no. 3, pp. 1305-1322, 2019/10/24 2019, doi: 10.1007/s10044-019-00855-7
3. G. J. Ansari, J. H. Shah, M. Yasmin, M. Sharif, and S. L. Fernandes, "A novel machine learning approach for scene text extraction," Future Generation Computer Systems, vol. 87, pp. 328-340, 2018/10 2018, doi: 10.1016/j.future.2018.04.074.

4. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346-359, 2008/06 2008, doi: 10.1016/j.cviu.2007.09.014.
5. K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," EURASIP Journal on Image and Video Processing, vol. 2008, pp. 1-10, 2008, doi: 10.1155/2008/246309.
6. L. Bhaskar and R. Ranjith, "Robust Text Extraction in Images for Personal Event Planner," presented at the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020/07, 2020. [Online]. Available: http://dx.doi.org/10.1109/icccnt49239.2020.9225417.
7. A. K. Bhunia, G. Kumar, P. P. Roy, R. Balasubramanian, and U. Pal, "Text recognition in scene image and video frame using Color Channel selection," Multimedia Tools and Applications, vol. 77, no. 7, pp. 8551-8578, 2017/05/05 2017, doi: 10.1007/s11042-017-4750-6.
8. D. Crandall, S. Antani, and R. Kasturi, "Extraction of special effects caption-text events from digital video," International Journal on Document Analysis and Recognition, vol. 5, no. 2-3, pp. 138-157, 2003/04/01 2003, doi: 10.1007/s10032-002-0091-7.
9. R. Deepa and K. N. Lalwani, "Image Classification and Text Extraction using Machine Learning," presented at the 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2019/06, 2019. [Online]. Available: http://dx.doi.org/10.1109/iceca.2019.8821936.
10. B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," presented at the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010/06, 2010. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2010.5540041.
11. S. P. Faustina Joan and S. Valli, "A Survey on Text Information Extraction from Born-Digital and Scene Text Images," Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, vol. 89, no. 1, pp. 77-101, 2018/01/18 2018, doi: 10.1007/s40010-017-0478-y.
12. P.-E. Forssen and D. G. Lowe, "Shape Descriptors for Maximally Stable Extremal Regions," presented at the 2007 IEEE 11th International Conference on Computer Vision, 2007. [Online]. Available: http://dx.doi.org/10.1109/iccv.2007.4409025.
13. J. Gllavata, R. Ewerth, and B. Freisleben, "Tracking text in MPEG videos," presented at the Proceedings of the 12th annual ACM international conference on Multimedia - MULTIMEDIA '04, 2004. [Online]. Available: http://dx.doi.org/10.1145/1027527.1027581.
14. H. Goto and M. Tanaka, "Text-Tracking Wearable Camera System for the Blind," presented at the 2009 10th International Conference on Document Analysis and Recognition, 2009. [Online]. Available: http://dx.doi.org/10.1109/icdar.2009.102.
15. L. Huiping, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," IEEE Transactions on Image Processing, vol. 9, no. 1, pp. 147-156, 2000, doi: 10.1109/83.817607.
16. K. Jung, K. In Kim, and A. K. Jain, "Text information extraction in images and video: a survey," Pattern Recognition, vol. 37, no. 5, pp. 977-997, 2004/05 2004, doi: 10.1016/j.patcog.2003.10.012.
17. R. Kasturi et al., "Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 319-336, 2009/02 2009, doi: 10.1109/tpami.2008.57.
18. M. Leon, V. Vilaplana, A. Gasull, and F. Marques, "Region-Based Caption Text Extraction," in Lecture Notes in Electrical Engineering, ed: Springer New York, 2012, pp. 21-36.
19. J. Li, Z. Zhou, Z. Su, S. Huang, and L. Jin, "A New Parallel Detection-Recognition Approach for End-to-End Scene Text Extraction," presented at the 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019/09, 2019. [Online]. Available: http://dx.doi.org/10.1109/icdar.2019.00219.
20. L.-J. Li, J. Li, and L. Wang, "An integration text extraction approach in the video frame," presented at the 2010 International Conference on Machine Learning and Cybernetics, 2010/07, 2010. [Online]. Available: http://dx.doi.org/10.1109/icmlc.2010.5580492.
21. J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," International Journal of Document Analysis and Recognition (IJDAR), vol. 7, no. 2-3, pp. 84-104, 2005/07 2005, doi: 10.1007/s10032-004-0138-z.
22. R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 4, pp. 256-268, 2002/04 2002, doi: 10.1109/76.999203.
23. D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004/11 2004, doi: 10.1023/b:visi.0000029664.99615.94.
24. W. Lu, H. Sun, J. Chu, X. Huang, and J. Yu, "A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network," IEEE Access, vol. 6, pp. 40198-40211, 2018, doi: 10.1109/access.2018.2851942.
25. S. Mahajan and R. Rani, "Text Extraction from Indian and Non-Indian Natural Scene Images: A Review," presented at the 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018/12, 2018. [Online]. Available: http://dx.doi.org/10.1109/icsccc.2018.8703369.
26. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," Image and Vision Computing, vol. 22, no. 10, pp. 761-767, 2004/09 2004, doi: 10.1016/j.imavis.2004.02.006.
27. C. Merino-Gracia, K. Lenc, and M. Mirmehdi, "A Head-Mounted Device for Recognizing Text in Natural Scenes," in Camera-Based Document Analysis and Recognition, ed: Springer Berlin Heidelberg, 2012, pp. 29-41.
28. C. Merino-Gracia, K. Lenc, and M. Mirmehdi, "A Head-Mounted Device for Recognizing Text in Natural Scenes," in Camera-Based Document Analysis and Recognition, ed: Springer Berlin Heidelberg, 2012, pp. 29-41.
29. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp. 1615-1630, 2005/10 2005, doi: 10.1109/tpami.2005.188.
30. R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "T-HOG: An effective gradient-based descriptor for single-line text regions," Pattern Recognition, vol. 46, no. 3, pp. 1078-1090, 2013/03 2013, doi: 10.1016/j.patcog.2012.10.009.
31. R. Mittal and A. Garg, "Text extraction using OCR: A Systematic Review," presented at the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020/07, 2020. [Online]. Available: http://dx.doi.org/10.1109/icirca48905.2020.9183326.
32. S. Mujumdar, N. Gupta, A. Jain, and D. Burdick, "Simultaneous Optimisation of Image Quality Improvement and Text Content Extraction from Scanned Documents," presented at the 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019/09, 2019. [Online]. Available: http://dx.doi.org/10.1109/icdar.2019.00189.

33. L. Neumann and J. Matas, "A Method for Text Localization and Recognition in Real-World Images," in Computer Vision – ACCV 2010, ed: Springer Berlin Heidelberg, 2011, pp. 770-783.
34. L. Neumann and J. Matas, "On Combining Multiple Segmentations in Scene Text Recognition," presented at the 2013 12th International Conference on Document Analysis and Recognition, 2013/08, 2013. [Online]. Available: http://dx.doi.org/10.1109/icdar.2013.110.
35. N. Phuc Xuan, K. Wang, and S. Belongie, "Videotext detection and recognition: Dataset and benchmark," presented at the IEEE Winter Conference on Applications of Computer Vision, 2014/03, 2014. [Online]. Available: http://dx.doi.org/10.1109/wacv.2014.6836024.
36. S. Puri and S. P. Singh, "Text recognition in bilingual machine printed image documents — Challenges and survey: A review on principal and crucial concerns of text extraction in bilingual printed images," presented at the 2016 10th International Conference on Intelligent Systems and Control (ISCO), 2016/01, 2016. [Online]. Available: http://dx.doi.org/10.1109/isco.2016.7727069.
37. X. Qian, G. Liu, H. Wang, and R. Su, "Text detection, localization, and tracking in compressed video," Signal Processing: Image Communication, vol. 22, no. 9, pp. 752-768, 2007/10 2007, doi: 10.1016/j.image.2007.06.005.
38. N. Rao, N. Panchal, and A. Shah, "Wind farm-power evacuation & grid integration," presented at the 2013 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia), 2013/11, 2013. [Online]. Available: http://dx.doi.org/10.1109/isgt-asia.2013.6698750.
39. A. Sain, A. K. Bhunia, P. P. Roy, and U. Pal, "Multi-oriented text detection and verification in video frames and scene images," Neurocomputing, vol. 275, pp. 1531-1549, 2018/01 2018, doi: 10.1016/j.neucom.2017.09.089.
40. B. Sheta, "Assessments Of Different Speeded Up Robust Features (SURF) Algorithm Resolution For Pose Estimation Of UAV," International Journal of Computer Science & Engineering Survey, vol. 3, no. 5, pp. 15-41, 2012/10/31 2012, doi: 10.5121/ijcses.2012.3502.
41. J. Shi, X. Luo, and J. Zhang, "DCT Feature-Based Text Capturing and Tracking," presented at the 2009 Chinese Conference on Pattern Recognition, 2009/11, 2009. [Online]. Available: http://dx.doi.org/10.1109/ccpr.2009.5344105.
42. P. Shivakumara, P. Trung Quy, and T. Chew Lim, "A Laplacian Approach to Multi-Oriented Text Detection in Video," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 2, pp. 412-419, 2011/02 2011, doi: 10.1109/tpami.2010.166.
43. H. Sidhwa, S. Kulshrestha, S. Malhotra, and S. Virmani, "Text Extraction from Bills and Invoices," presented at the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018/10, 2018. [Online]. Available: http://dx.doi.org/10.1109/icacccn.2018.8748309.
44. C. P. Sumathi, "A Survey On Various Approaches Of Text Extraction In Images," International Journal of Computer Science & Engineering Survey, vol. 3, no. 4, pp. 27-42, 2012/08/31 2012, doi: 10.5121/ijcses.2012.3403.
45. M. Tanaka and H. Goto, "Autonomous Text Capturing Robot Using Improved DCT Feature and Text Tracking," presented at the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2, 2007/09, 2007. [Online]. Available: http://dx.doi.org/10.1109/icdar.2007.4377101.
46. M. Tanaka and H. Goto, "Text-tracking wearable camera system for visually-impaired people," presented at the 2008 19th International Conference on Pattern Recognition, 2008/12, 2008. [Online]. Available: http://dx.doi.org/10.1109/icpr.2008.4761548.
47. S. Tehsin, A. Masood, S. Kausar, and Y. Javed, "Text localization and detection method for born-digital images," IETE Journal of Research, vol. 59, no. 4, p. 343, 2013, doi: 10.4103/0377-2063.118025.
48. C. Thorat, A. Bhat, P. Sawant, I. Bartakke, and S. Shirsath, "A Detailed Review on Text Extraction Using Optical Character Recognition," in ICT Analysis and Applications, ed: Springer Singapore, 2022, pp. 719-728.
49. K. Wang and S. Belongie, "Word Spotting in the Wild," in Computer Vision – ECCV 2010, ed: Springer Berlin Heidelberg, 2010, pp. 591-604.
50. Z. Wang and Y. Xiao, "An Approach for Video-Text Extraction Based on Text Traversing Line and Stroke Connectivity," presented at the 2010 International Conference on Biomedical Engineering and Computer Science, 2010/04, 2010. [Online]. Available: http://dx.doi.org/10.1109/icbecs.2010.5462313.
51. H. Weihua, P. Shivakumara, and T. Chew Lim, "Detecting moving text in video using temporal information," presented at the 2008 19th International Conference on Pattern Recognition, 2008/12, 2008. [Online]. Available: http://dx.doi.org/10.1109/icpr.2008.4761417.
52. J. Xu, X. Jiang, and Y. Wang, "Caption Text Extraction Using DCT Feature in MPEG Compressed Video," presented at the 2009 WRI World Congress on Computer Science and Information Engineering, 2009. [Online]. Available: http://dx.doi.org/10.1109/csie.2009.107.
53. J. Yang, S. C. Han, and J. Poon, "A survey on the extraction of causal relations from natural language text," Knowledge and Information Systems, 2022/03/12 2022, doi: 10.1007/s10115-022-01665-w.
54. C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," (in eng), IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, vol. 20, no. 9, pp. 2594-2605, 2011, doi: 10.1109/TIP.2011.2126586.
55. K. M. Yindumathi, S. S. Chaudhari, and R. Aparna, "Analysis of Image Classification for Text Extraction from Bills and Invoices," presented at the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020/07, 2020. [Online]. Available: http://dx.doi.org/10.1109/icccnt49239.2020.9225564.
56. X. Zhang, F. Sun, and G. Lei, "A combined algorithm for video text extraction," presented at the 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010/08, 2010. [Online]. Available: http://dx.doi.org/10.1109/fskd.2010.5569311.
57. M. Zhao, S. Li, and J. Kwok, "Text detection in images using sparse representation with discriminative dictionaries," Image and Vision Computing, vol. 28, no. 12, pp. 1590-1599, 2010/12 2010, doi: 10.1016/j.imavis.2010.04.002.
58. W. Zhen and W. Zhiqiang, "An Efficient Video Text Recognition System," presented at the 2010 Second International Conference on Intelligent Human-Machine Systems and Cybernetics, 2010/08, 2010. [Online]. Available: http://dx.doi.org/10.1109/ihmsc.2010.50.