***Intersection of Language and Space***
*Department of Geography,*
**Chiang Mai University, Thailand**

# Text Localization, Script Identification and Recognition in Natural Scenes : Trends and Beyond

**Veronica Naosekpam**
*Research Fellow,*
*School of Digital, Technology, Innovation and Business*
***University of Staffordshire, United Kingdom***

# Outlines

- **Introduction**
- **Reviews:**
  - Scene text localization in images and videos
  - Script / language identification
  - Scene text recognition
- **Contributions**
  - Script identification
  - Text localization and motion tracking in videos
  - Datasets.
- **Scene text analysis and Geoinformatics**
- **Conclusion**

# Introduction

- New multimedia tools and devices.
- Digital data storage and access facility is **cheaper** and **freely** available.
- Increase in Social Media, News broadcast, Internet and digital content usage.


Fig 1 : Example of scene text image

**Need to automate Text Detection/Recognition for many useful purposes!**

- **Scene Text** is the text that appears in an image captured by a camera.
- Conveys semantic information.
- **Text detection, Language identification** and **Recognition** : Scene Text Understanding.

# Text Detection, Language Identification and Recognition


Fig 2 : Example of scene (a) Text detection (b) Language identification and (c) Text recognition

- **Scene Text Detection** is the localization of text regions present in an image.
- Represented by a rectangle, rotated rectangle, quadrilateral or multi-oriented polygon.
- **Language Identification** determine the language a particular text belongs to.
- **Scene Text Recognition (STR)** converts the detected text into machine readable format.

# Challenges with scene texts

- **Text diversity :** Irregular fonts, brightness, contrast.
- **Scene Complexity** : Cluttered backgrounds.
- **Distortion :** Non-uniform illumination, low resolution, motion blurring and partial occlusion.
- **Multi-lingual :** Presence of multiple languages.



Fig 3 : Challenges associated with scene texts understanding

# Challenges with scene texts in videos

- The quality of the image in videos is worst than the static image.
- Temporal information.
- Texts are from different modalities in case of lecture video.
- The instructor may write over the figures and equations thus, making the scene cluttered and decreases the visibility of the text.
- Occlusion may hurdle text tracking.

# Scene Text Detection / Localization

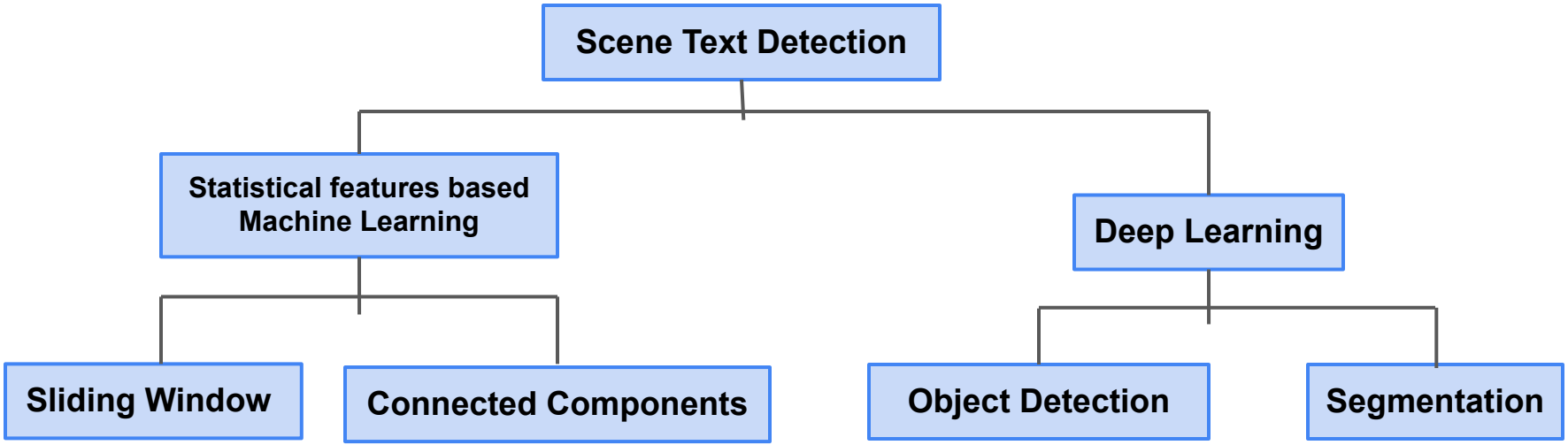



Fig 4 : Taxonomy of Scene Text Detection

# Sliding Window based Scene Text Detection

- Multi-scale sliding window scans the entire image, followed by candidate text regions extraction and use classifier to point out the presence of text in the sub-window or not.

| Scheme | Technique | Comments |
|---|---|---|
| Lee *et al.* [1] | - Six types of features used.<br>- AdaBoost classifier with multi-scale sequential search. | - Brute force approach which results in slow detection speed. |
| Wang *et al.* [2] | - HOG [31] features extracted from the window.<br>- Random Ferns classifier. | - Works for horizontal texts. |

HOG : Histogram of Oriented Gradients

# Connected Component based Scene Text Detection

- Extracts components through clustering or edge detection and non-text regions are removed by using classifiers.
- Components grouping is usually done based on geometric properties : SWT, MSER.

| Scheme | Technique | Comments |
|---|---|---|
| Epstein *et al.* [3] | - SWT of the entire image.<br>- Letter candidate component grouping.<br>- Grouped letters to textline. | - Non-text components which easily confused with texts. |
| Koo *et. al* [5] | - Extract connected component.<br>- Candidate region generation.<br>- Word region normalization.<br>- Classification via multi-layer perceptron. | - Works for horizontal texts. |

SWT : Stroke Width Transform
MSER : Maximally Stable Extremal Region

# Object Detection based Scene Text Detection

- Forecast candidate bounding box by treating text as an object and apply object detection based algorithms.
- Detecting word instances require variable aspect ratio unlike the object.

| Scheme | Technique | Comments |
|---|---|---|
| Zhou *et al.* [8] | - U-shaped design [19] FCN is used to integrate features from different levels.<br>- The feature at each spatial location is used to regress the bounding box of the text instances directly.<br>- Post-processing includes thresholding and non-maximal suppression. | - Robust.<br>- Horizontal and multi-oriented English texts.<br>- The detection effect is not good for long text. |
| Ma *et al.* [9] | - Introduced Rotation Region Proposal Networks (RRPN), based on Faster-RCNN [22].<br>- RRoI. | - Multi-oriented , horizontal texts.<br>- Two staged text detector consumed more space. |

# Segmentation based Scene Text Detection

- Extract text blocks from the segmentation map and then obtain bounding boxes of the text by post-processing.
- Semantic segmentation and instance segmentation.

| Scheme | Technique | Comments |
|---|---|---|
| Yao *et al.* [29] | - Modified fully connected network.<br>- Three maps : text/non-text regions, character classes, and character linking orientations.<br>- Post-processing method. | - Works for multi-oriented texts.<br>- Fail to accurately separate the adjacent-word instances that tend to connect. |
| Wang *et al.* [30] | - Introduced PAN → Segmentation + post-processing<br>- Segmentation : FPEM + FFM<br>- Post-processing : Pixel Aggregation can precisely aggregate text pixels by predicted similarity vectors.<br><br>PAN : Pixel Aggregation Network<br>FPEM : Feature Pyramid Enhancement Module<br>FFM : Feature Fusion Module | - Multiple intermediate stages affect the final model and the processing speed is slow.<br>- Arbitrary shaped scene texts. |

# Detection Datasets & Metrics

- **Datasets:**
  - ICDAR03/05/11/13/15/17/19
  - COCO-Text
  - MSRA-TD500
  - Total-Text, CTW1500
- **Metrics:**
  - Precision/Recall/F1
  - DetEval
    - This evaluation metric takes into account for various cases that is, one-to-one, one-to-many and many-to-one
  - TIoU
    - Protocol for scene text detection that considers the tightness of the detection algorithm. In order to allow clear-cut attention on the detecting text contents, it involves annotation concepts: annotation does not cut the text instance, less background contents is present, and if the annotation does not match the text instance as expected, it should be as perfect as possible.
  - TedEval :
    - Calculates the accuracy of text detection algorithms via instance-level matching policy and character-level scoring policy.

# Language Identification

- High inter-class similarity + High intra-class variability.
- Asian Language : Textual data may not be language wise uniform in the wild.

| Scheme | Technique | Comments |
|---|---|---|
| Gomez and Karatzas [14] | - Convolutional features + Naive-Bayes Nearest Neighbor (NBNN) classifier. | - Patch based.<br>- Release benchmark data (Mle2e) |
| Gomez and Karatzas [15] | - Ensemble of Conjoined Networks<br>- Loss function with the global classification error for a group of N patches. | - High inter class variable data combination (English, Kannada, Chinese, Hangul) |
| Chakraborty *et al.* [16] | - Input image information in 5 channels : R,G,B,RGB and grayscale passed to a CNN.<br>- Outcomes of models are combined using the classifier combination approaches based on sum rule and product rule. | - Mle2e dataset.<br>- In-house data English, Bangla, Hindi |

# Language Identification Datasets & Metrics

- **Datasets:**
  - MLe2e
  - SIW-10, SIW-13
- **Metrics:**
  - Precision/Recall/F1
  - Accuracy
  - Category-wise P/R/F1/Accuracy

# Scene Text Recognition : Machine Learning Based

- Scene texts are recognized based on a set of hand crafted features.
- Bottom-up approach that classified characters are linked up into words.

| Scheme | Technique | Comments |
|---|---|---|
| Neumann *et al*. [12] | - A set of handcrafted features, such hole area ratio, convex ratios used.<br>- Features are fed to an SVM classifier. | - Low recognition accuracy. |
| Campos *et al.* [13] | - Object categorization based on a bag-of-visual-words (BoW) representation.<br>- Shape context, geometric blur, SIFT, spin image, Maximum response of filters and patch descriptors<br>- Used KNN and SVM classification for decision making. | - Building models that are able to handle text recognition in the wild is difficult. |

SVM : Support Vector Machine
SIFT : Scale Invariant Features Transform
KNN : K-Nearest Neighbour

# Scene Text Recognition : Deep Learning Based

| Scheme | Technique | Comments |
|---|---|---|
| Wang *et al.* [6] | - A CNN-based feature extraction framework was for character recognition.<br>- NMS (Non Maximal Suppression) is applied to obtain the final word. | - Require each character localization.<br>- Challenging due to the complex background, irrelevant symbols, and the short distance between adjacent characters. |
| Jaderberg *et al.* [18] | - Proposed a synthetic dataset containing 90K English words.<br>- Train basic CNN on the synthetic data. | - Cannot recognize out-of-vocabulary words.<br>- Deformation of long word images. |
| Shi *et al.* [17] | - Combined attention-based sequence features and a rectification module.<br>- The text within the rectified image is recognized by a RNN (Recurrent Neural Network). | - Training rectification method without considering human-designed geometric ground truth is difficult. |

# Text Recognition Datasets & Metrics

- **Datasets:**
  - SynthText90K
  - IIIT5K
  - SVT
  - ICDAR 03/13/15
- **Metrics:**
  - Word recognition accuracy
    - It is defined as the ratio of the correctly recognized words to the total number of ground-truth words.
  - Word Error Rate: It is defined by the value left after subtracting the WRA from 1.

**Naosekpam, Veronica, and Nilkanta Sahu. "Text detection, recognition, and script identification in natural scene images: a Review."** ***International Journal of Multimedia Information Retrieval*** **11.3 (2022): 291-314.**

# Text Detection and Recognition in Videos

| Scheme | Technique | Comments |
|---|---|---|
| Wang *et al.* [20] | - End-to-end scene text recognition system from video based on multi-frame tracking.<br>- Two steps : 1) Text detection and recognition of each individual frame ; and<br>2) Multiple frame text tracking is performed by association of the results obtained in Step 1. | - Temporal information is employed. |
| Kartik *et al*. [21] | - Proposed the first proper dataset called LectureVideoDB for lecture videos text detection and recognition. | - The existing scene text recognition algorithms do not work well. |

# My research : Why multi-lingual Indian scene texts?

- Scene text understanding solutions focused on English with some works on Chinese [10] and Arabic scene texts [11].
- Work on multi-lingual started around the end of 2018  but majority is horizontal.
- Texts appearance in most of the datasets is horizontal and multi-oriented.
- Irregular shaped text : English and Chinese dataset.

# Contribution : Multi-lingual Scene Text Understanding

- Scene text understanding involves **text localization** and **script identification** in natural scene images.
  - **Text localization** : find the area of the potential text regions.
  - **Script / Language identification** : classify the language / script of that particular detected text.

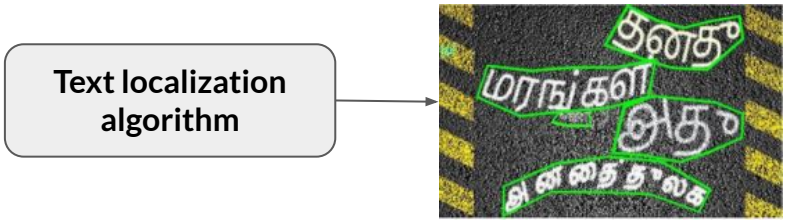



Fig. 5 :Text localization of curved text instances.

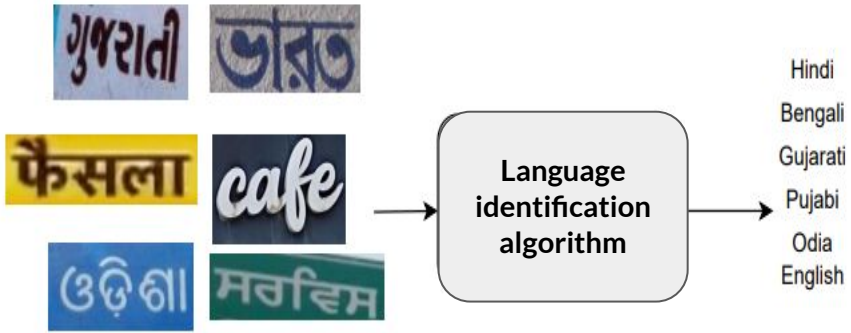
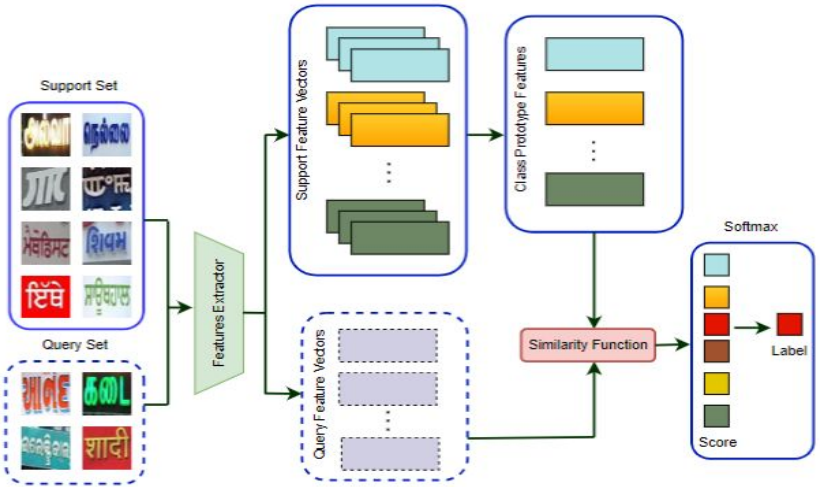



Fig. 6: Language identification of scene text.

# Contributions: Approaches

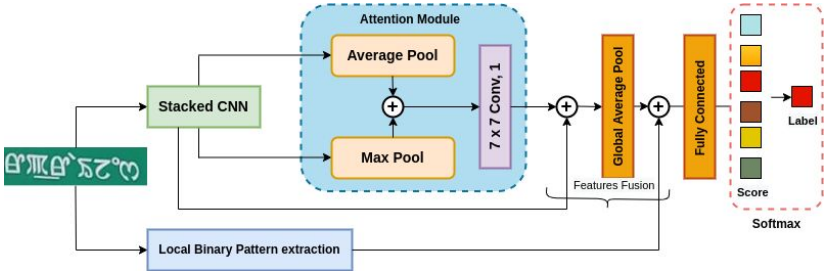
2. **Language identification** : Few-shots learning based.




- **Enhancement** : Fuzzy-based weighted distribution color histogram equalization [1]
- **Feature Extractor :** A 2-layer CNN with a multi-kernel spatial attention component maps input images into a metric space.

***Naosekpam, Veronica, and Nilkanta Sahu. "Few-shot learning for word-level scene text script identification." Computational Intelligence (2024).***

3. **Language identification :** Integration of statistical and spatial attention-infused deep features.




- Attention-assisted deep and statistical feature fusion network called WAFFNet
- Transfer learning: fewer trainable parameters as we leverage the pre-trained weights from the first 4 convolutional layers of the VGG16 [2].

***Naosekpam, Veronica, and Nilkanta Sahu. "A Hybrid Scene Text Script Identification Network for regional Indian Languages." ACM Transactions on Asian and Low-Resource Language Information Processing (2024).***

# Contributions : Approaches

4. **Video scene text detection, motion tracking and prediction using tracking-by-detection paradigm.**

**Key Components :**

- Frame level detection using YOLOv4-Tiny.
- Text motion tracking and predicting the location of lost text instances.
  - Kalman Filter
  - Deep appearance features.
  - Munkres algorithm.
- Re-associating text with the corresponding detection after a long absence.
  - Derive the velocity state from prior frames.
  - Weight assignment based on its relevance
  - Calculate acceleration and predict current velocity.
  - Estimate frame's location using the predicted velocity.

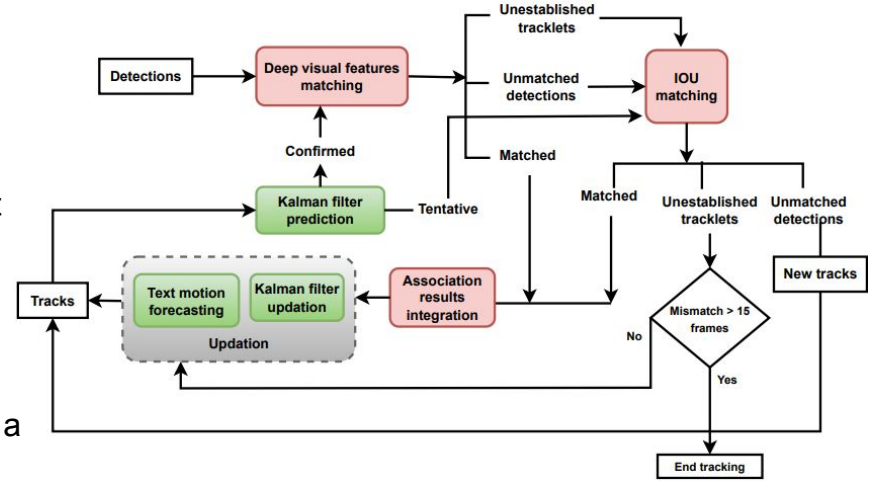



**Table** : Experimental results on ICDAR2015 Video.

| Scheme | Precision | Recall | F1-score |
|---|---|---|---|
| Li et al. [7] | **0.82** | 0.75 | 0.78 |
| Liu et al. [8] | 0.67 | 0.69 | 0.66 |
| Proposed | 0.75 | **0.83** | **0.78** |


***Naosekpam, Veronica, and Nilkanta Sahu. "Video text rediscovery: Predicting and tracking text across complex scenes." Computational Intelligence 40.3 (2024): e12686.***

# Contributions : Datasets

- **Indic Curved Synth Text (ICST)** dataset : Synthetic realistic curved scene text instances superimposed on natural scene background images for: Bengali, Tamil, and Telugu.
- **Indic-FSL2023** dataset : Includes some resource-constraint Indian languages for script identification.
- **EMBiL** dataset: An English-Manipuri Bi-lingual benchmark for scene text detection and language identification.

**Fig.** ICST dataset.


**Fig.** Indic-FSL2023 dataset.


**Fig.** EMBiL dataset.

# Text Detection-Recognition-Script Identification: Intersections of Language and Space

- Exploring how textual data in images defines and reflects our spatial environments
  - Bringing together AI-driven text analysis with geoinformatics and corpus linguistics
- **Bridging Computer Vision & Geoinformatics:**
  - **Corpus Linguistics**: Building and querying geographically anchored text corpora drawn from street-view imagery
  - **GeoDBMS Integration**: Storing detected text with spatial metadata for advanced GIS analyses.
  - **Street-View Text**: Mapping urban signage and public text as a reflection of socio-cultural landscapes
- **Deep Learning's Impact on Spatial Text Analysis:**
  - Leveraging neural networks to detect, recognize, and categorize text across large geospatial datasets.
  - Automated extraction of semantic information from imagery for urban planning, navigation, and heritage studies.

# Potential Research Directions

- **Cross-Lingual Spatial Text Mining**:
  - Unified models for detecting and recognizing text in dozens of co-occurring scripts across geographic regions
  - Zero- and few-shot learning to rapidly onboard emerging or under-resourced languages in map corpora.
- **Spatio-Temporal Text Dynamics**:
  - Tracking the evolution of public signage (e.g., commercial, political) via time series analysis of street-view imagery
  - Forecasting text changes for urban planning, resilience monitoring (e.g., pop-up ads vs permanent landmarks)
- **3D Geo-Text Integration**:
  - Projecting localized text onto 3D city models and point clouds for augmented reality wayfinding
  - Combining LiDAR and image text layers for immersive digital twins with semantic overlays
- **Interactive GeoDBMS & Corpus Linguistics**:
  - Query languages for spatial corpora (e.g., “find all bilingual shop signs within 500m of heritage site”)
  - Visualization tools to explore socio-linguistic landscapes via geotagged text occurrences
- **Multimodal Environment Understanding**::
  - Fusing environmental sensor data (air-quality, noise) with geotext to study public health or urban heat islands.
  - Integrating audio (speech), video, and text for comprehensive smart-city analytic.

# Conclusion

- Reviewed advances in text detection, recognition, script identification in natural scene images and video data.
- Highlighted key datasets, metrics, and challenges.
- Contributions.
- Bridging gap between Scene text analysis (Computer Vision) with geoinformatics

# References :


[1] S. Ghosh, B.B. Chaudhuri, in 2011 International Conference on Document Analysis and Recognition (IEEE, 2011), pp. 294–298

[2] Wang, K., Babenko, B., & Belongie, S. (2011, November). End-to-end scene text recognition. In *2011 International Conference on Computer Vision* (pp. 1457-1464). IEEE.

[3] Epshtein, Boris, Eyal Ofek, and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.

[4] Mathew, M., Jain, M., & Jawahar, C. V. (2017, November). Benchmarking scene text recognition in devanagari, telugu and malayalam. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 7, pp. 42-46). IEEE.

[5] Koo, Hyung Il, and Duck Hoon Kim. "Scene text detection via connected component clustering and nontext filtering." *IEEE transactions on image processing* 22, no. 6 (2013): 2296-2305.

[6] Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012, November). End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)* (pp. 3304-3308). IEEE.

[7] Huang, W., Qiao, Y., & Tang, X. (2014, September). Robust scene text detection with convolution neural network induced mser trees. In *European conference on computer vision* (pp. 497-511). Springer, Cham.

[8] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 5551-5560).

[9] Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., & Xue, X. (2018). Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, *20*(11), 3111-3122.

[10] Yuan, T. L., Zhu, Z., Xu, K., Li, C. J., & Hu, S. M. (2018). Chinese text in the wild. *arXiv preprint arXiv:1803.00085*.

[11] Ahmed, S. B., Naz, S., Razzak, M. I., & Yusof, R. B. (2019). A novel dataset for English-Arabic scene text recognition (EASTR)-42K and its evaluation using invariant feature extraction on detected extremal regions. *IEEE access*, *7*, 19801-19820.

[12] Neumann, L., & Matas, J. (2012, June). Real-time scene text localization and recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3538-3545). IEEE.

[13] De Campos, T. E., Babu, B. R., & Varma, M. (2009). Character recognition in natural images. *VISAPP (2)*, *7*.

[14] Gomez, L., & Karatzas, D. (2016, April). A fine-grained approach to scene text script identification. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)* (pp. 192-197). IEEE.

[15] Gomez, L., Nicolaou, A., & Karatzas, D. (2017). Improving patch-based scene text script identification with ensembles of conjoined networks. *Pattern Recognition*, *67*, 85-96.

[16] Chakraborty, N., Kundu, S., Paul, S., Mollah, A. F., Basu, S., & Sarkar, R. (2020). Language identification from multi-lingual scene text images: a CNN based classifier ensemble approach. *Journal of Ambient Intelligence and Humanized Computing*, 1-12.

[17] Shi, Baoguang, et al. "Robust scene text recognition with automatic rectification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[18] Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

[19] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[20] Wang, X., Jiang, Y., Yang, S., Zhu, X., Li, W., Fu, P., ... & Luo, Z. (2017, November). End-to-end scene text recognition in videos based on multi frame tracking. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1255-1260). IEEE.

[21] Dutta, K., Mathew, M., Krishnan, P., & Jawahar, C. V. (2018, August). Localizing and recognizing text in lecture videos. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 235-240). IEEE.

# Thank You