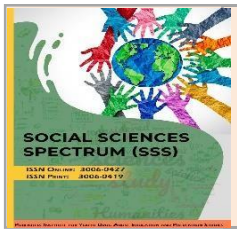

Social Sciences Spectrum

A Double-Blind, Peer-Reviewed, HEC recognized Y-category Research Journal

E-ISSN: 3006-0427 **P-ISSN:** 3006-0419
Volume 04, **Issue** 01, 2025
Web link:https://sss.org.pk/index.php/sss

Multimodal Feature Fusion Technique for Robust Classification Using Natural Scene Text Binary Images

Shahbaz Hassan Wasti
Department of Information Sciences, Division of Science and Technology, University of Education, Lahore, 54770, Pakistan
Email: shahbazwasti@ue.edu.pk

Ghulam Jillani Ansari
Assistant Professor Computer Science, Department of Information Sciences, Division of Science and Technology, University of Education, Lahore, 54770, Pakistan
Correspondence: ghulamjillani@ue.edu.pk

Sajid Ali
Associate Professor IT, Department of Information Sciences, Division of Science and Technology, University of Education, Lahore, 54770, Pakistan
Email: sajid.ali@ue.edu.pk

Article Information [YY-MM-DD]

Received 2025-02-11 **Accepted** 2025-03-29

Citation (APA):

Wasti, S, H., Ansari, G, J & Ali, S (2025). Multimodal feature fusion technique for robust classification using Natural Scene Text Binary Images. *Social Sciences Spectrum*, *4*(1), 799-810.
https://doi.org/10.71085/sss.04.01.447

Abstract

The scene text classification requires effective representation of diverse visual properties including geometric, appearance, contour and shape. This paper presents utilizing ensemble classification framework incorporating multimodal feature fusion for robust scene text images. Histogram of Oriented Gradient (HOG), Geometric, Local Ternary Pattern (LTP) and contour based features are extracted from segmented binary images to capture relevant complementary statistical and structural information. After that, these features are serially fused to form comprehensive multimodal feature fusion representation. The fused model is further optimized using feature selection process to retain highly discriminative features for classification. Multiple ensemble classifiers including k-Nearest Neighborhood (KNN), Decision Tree (DT) models and Support Vector Machines (SVM) variants are employed to determine optimal classification performance. The classification experiments are conducted on Street View Text dataset which is considered to very challenging with scenic variabilities. Among all, the L-SVM consistently outperforms when compared with competitive classifiers and trained on optimized multimodal feature. The proposed framework achieved substantial improvements gradually in precision, recall, F-measure, accuracy and AUC when compared to single and non-optimized baselines. The findings evidently ensure that ensemble learning trained on optimized multimodal feature fusion results discriminative and reliable solution natural scene text recognition under non-standard real-world imaging conditions.

Keywords: Multimodal Feature Fusion, LTP, HOG, Ensemble Classifiers, Non-optimized.

1. Introduction

Classification of scene text from natural images in an important role in bridging the gap between text recognition and detection. Once textual contents are retrieved accurately, then illustrating text and non-text cues and recognizing character patterns become challenging because of diverse visual properties in natural scenes. (Du et al., 2022; Harizi et al., 2022; Singh et al., 2021). Such diversity including complex background, font irregularities, arbitrary orientations, and varying illumination significantly complicate and challenge the classification process

Therefore, to cope with these variabilities, the researchers have increasingly confident on using handcrafted feature descriptors. These descriptors are capable to capture complementary aspects of text appearances including HOG effectively encode stroke and edge information, whereas LTP provide robustness against noise and illumination variations (Hong et al., 2020; Rassem & Khoo, 2014; Tan & Triggs, 2010). In addition, contour-based and geometric features capture structural and shape properties of text regions. Such features are useful for discriminating text from cluttered background (He et al., 2023; Kushwaha & Nain, 2020).

Even though, single feature descriptor can provide useful text cues, but sometimes it leads to suboptimal performance when dealing with complex natural scenes. Therefore, the solution is multimodal feature fusion has emerged as an effective strategy to combine complementary information from different feature descriptors (Akhtar et al., 2023; Pandey et al., 2021). As a result, this strategy increased the dimensionality of fused descriptors is a great challenge for conventional classifiers and necessitate careful classifier design and feature optimization.

Numerous ensemble classifiers KNN, DT and SVM have been employed for scene text classification with varying degrees of success (Akhtar et al., 2023; Ansari et al., 2021). Among these, SVM variants demonstrated comparable generalization capabilities particularly when binary classification adopted for text and non-text regions involving high dimensional feature space. Furthermore, the classifier performance is highly sensitive depending upon the input features quality and presence of irrelevant cues.

Keeping this in mind, the foremost object of this paper is on multimodal feature fusion and classification. Hence, a comprehensive classification framework is proposed by serially fusing the selected feature descriptors to obtain optimized multimodal feature fusion. These obtained features from selected public datasets are trained using selected ensemble classifiers and the performance is systematically evaluated using standard parameters.

1.1 Research Contributions

The primary contributions of this paper can be summarized as follows:

1. A comprehensive multimodal feature space is determined from binary segmented image
2. The combined feature approach is systematically observed for its performance
3. Numerous ensemble classifiers are evaluated to assess classification robustness across different feature configurations
4. The efficacy of multimodal feature fusion is evaluated using measures including precision, recall, F-measure, accuracy, and AUC
5. For extensive experiments challenging public datasets are used and conclude that linear SVM consistently achieves competitive performance when compared with other competitive classifiers on optimized multimodal feature fusion

1.2 Paper Structure

The paper structure is as follows. **Section 2** reviews related work on multimodal feature fusion and scene text classification while **Section 3** highlights the proposed feature extraction, fusion, and optimization classification framework. **Section 4** depicts and discussed the comparative analyses and experimental results on selected state-of-the-art datasets. **Section 5** concludes the paper and discusses future research prospects.

2. Related Work

The central challenge in NST is scene text understanding because text regions exhibits severe variabilities in orientation, blur, illumination, font, scale and background clutter. In this regard the classification pipelines have evolved from handcrafted features to attention based deep architectures. This evolution persistently emphasis on generation and robustness across diverse and random environments (Cao et al., 2020).

In beginning, mostly journals framed their work on scene text classification as text and non-text regions using gradient and frequency driven descriptors. (Goto, 2008), published work using spatial frequency properties and refined Discrete Cosine Transform (DCT) features. The authors combine discriminative analysis and thresholding to enhance classification between non-text and text regions. This particular setup helps establish the importance of capturing repetitive and structured patterns typically character strokes. However, it also submits few limitations when background contains similar texture with high frequencies.

In another study, gradient based approach become widely adopted since the text strokes naturally poses consistent edge orientations. (Minetto et al., 2013), proposed and designed a gradient based descriptor named as T-HOG to efficiently handle single line text when contrast and brightness is locally adopted. The normalization approach is based on standard HOG pipelines. This setup reinforced and directs the role of stable gradient statistics for NST classification especially when SVMs are employed. In this case also, the pure gradient drive models can still reflects degraded performance when stroke edges are blurred, weak and contaminated by textured and background clutter (He et al., 2023; Kushwaha & Nain, 2020).

Few other studies demonstrated single feature descriptor can provide useful text cues, but sometimes it leads to suboptimal performance when dealing with complex natural scenes. Therefore, the solution is multimodal feature fusion has emerged as an effective strategy to combine complementary information from different feature descriptors (Akhtar et al., 2023; Pandey et al., 2021). As a result, this strategy increased the dimensionality of fused descriptors is a great challenge for conventional classifiers and necessitate careful classifier design and feature optimization.

In contrast to standalone text descriptors, probabilistic graphical modeling has also there to enhance classification generalizability and reliability. These methods enforce neighborhood and context consistency. (Wang et al., 2018), introduced CRF based framework for NST detection in order to estimates the confidence of text regions by evaluating model dependencies to reduce false positives. False positives often reflect form locally text patterns which plays important role in misclassification through structured dependency modelling.

Recently, deep and hybrid architectures have gaining researcher's interests, since they learn features to handle variabilities from complex background in a better way. (Kantipudi et al., 2021) published work in which they combine CNN features with Bi-LSTM sequence model for scene text classification. This integrated shift reflects from isolated component classification towards sequence consistent decision making process. Similarly, the authors proposed attention

mechanism to focus suppress background interference and computation on salient character regions. In another work, (Xie et al., 2019) analyzed Convolutional Attention Network (CAN) that employ CNN based can reduce or replace reliance on recurrent decoding from unconstrained text recognition and still showing robustness. No doubt, such approaches implicitly strengthen the quality of classification in terms of handling errors, noise and cluttered from learned information.

Interference and noise suppression is also an important and active phenomenon for researchers in order to control classifier from misclassification. (Tang et al., 2025), published recognition model, which explicitly target character noise interference and complex background using multi-level feature fusion. The setup become useful to enhance recognition accuracy across NST datasets. Consequently, recent applied techniques explored hybrid pipelines to segregate text and non-text regions for ensemble classifiers generalization. These strategies combine beneficial complementary details under practical constraints rather than relying on a single descriptor family.

As per related work, it is evidently concluding that two consistent approaches emerges i.e. firstly no single descriptor is uniformly robust under severe scenic variabilities. Hence, performance often improves when various cues are fused, integrated or combined. Secondly, reliability and generalizability is effectively deteriorating when redundancy control and robustness to clutter is compromised. These important factors motivate to proposed classification framework based on multimodal feature fusion for strong generalization and balanced discrimination considering abstruse NST conditions, Moreover, classifier performance is heavily influenced by presence of redundant information and feature representation quality. Therefore, these gaps allow us to explore optimized multimodal feature space and systematic evaluation of ensemble classifiers which is being addressed in this present work. The upcoming section highlights in proposed framework in detail.

3. Proposed Classification Model

This section presents the methodology used in building the classification model in detail. The motivation is to enhance reliability and generalizations of ensemble classifiers into text and non-text regions from binary segmented NST images.

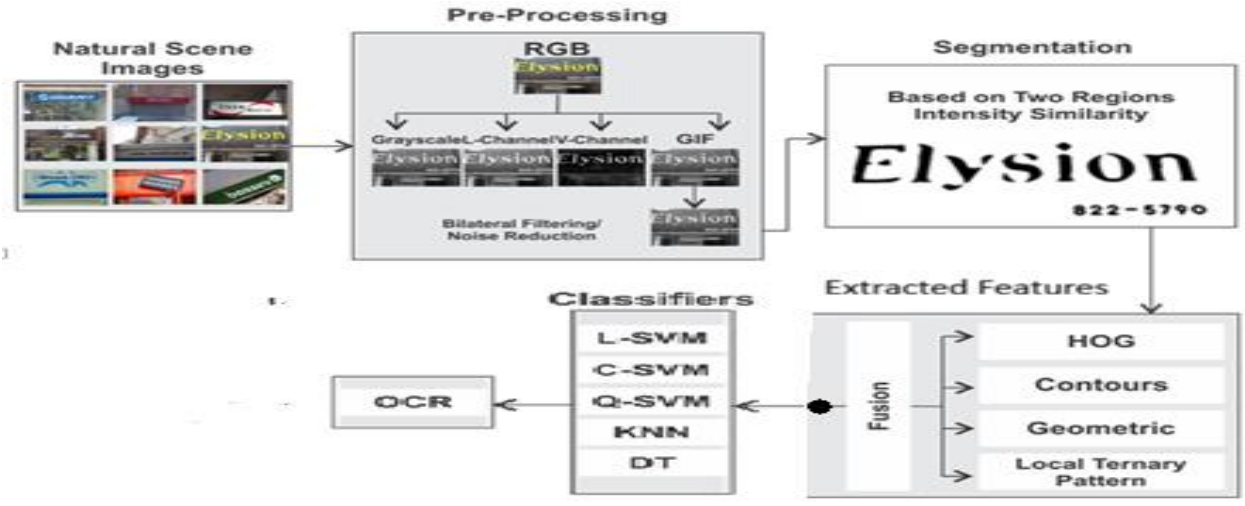

Figure 1: *Visual diagram of the intended classifications pipeline*

The features are extracted from prior binary segmented NST images as shown in the propose

classification model in **Figure 1**. The binary segmented images are already preprocessed and generated from images obtained from selected NST datasets. The prior preprocessing is important for enhancing classification accuracy and stable generalization. The process includes the combine implementation of Guided Image Filtering (GIF) and Bilateral Filtering (BF). Further to preserve edges and remove adaptive noise wiener filter is incorporated around each pixel of the NST image. The next stage of the pipeline i.e. extracted features and their fusion are discussed in detail in the subsequent sections.

3.1 Multimodal Feature Extraction and Fusion

Four types of features are extracted from prior binary segmented NST image as shown in Error! Reference source not found.. These features are i) appearance features extracted using HOG, ii) texture features fetched using LTP, iii) structural boundary information using contour descriptors, and finally iv) geometric features such as region shape and statistical attributes. Each feature contributes to a perfect, accurate, and precise classification. The whole feature extraction process is illustrated in **Figure 2**.

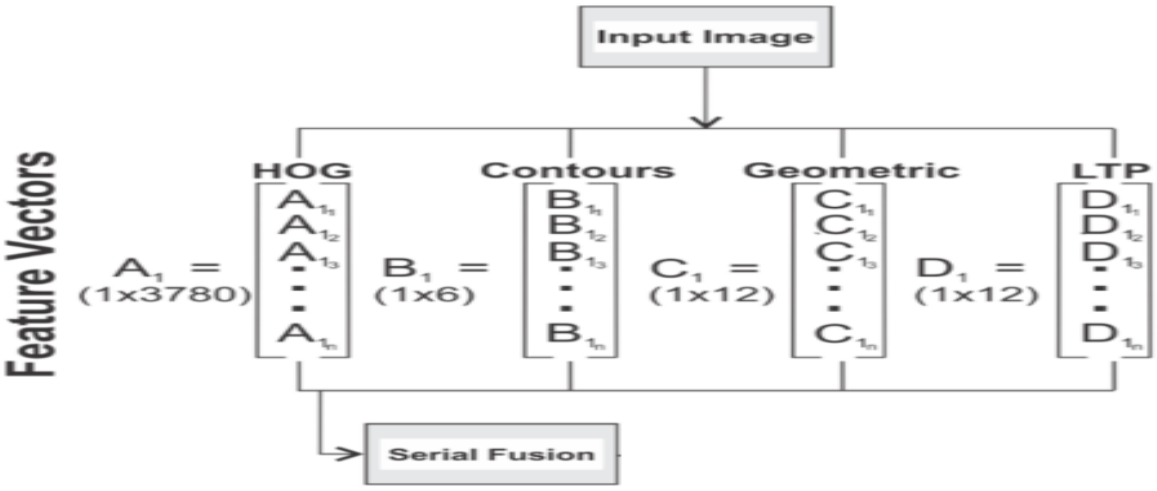

Figure 2 Feature extraction and fusion pipeline from input binary image

The multimodal feature representation can effectively improve the model performance. Following four algorithms explain the computation of each feature vector.

3.1.1 Algorithm 1 HOG feature vector

Input: Binary NST-segmented image *I*

Output: HOG feature vector $A_1 = [A_{1_1}, A_{1_2}, A_{1_3}, ..., A_{1_n}]$ length 1×3780.

Steps:

1. Divide input image *I* into fix number of blocks

2. HOG is calculated for each image block sub_i by computing the image gradient

3. All image blocks after calculations are then concatenated to form a single feature

Figure 3(b) illustrate the HOG appearance features.

3.1.2 Algorithm 2 Contour feature vector

This feature vector represents the edges and curves of the binary NST image. The detail of the

contour algorithm feature vector is given below:

Input: Binary NST-segmented image *I*

Output: Contour feature vector $B_1 = [B_{1_1}, B_{1_2}, B_{1_3}, ..., B_{1_n}]$ of length 1×6.

Steps:

1. Calculation of the distance to the mass center using Euclidian
2. Calculation of the distance to the opposite mass center from contour pixel towards inner region
3. Aggregate the calculations to form the contour feature vector B_1.

3.1.3 Algorithm 3 Geometric feature vector

Geometric or silhouette features [38] is extracted from **Figure 3(c).**

Input: Binary NST-segmented image *I*

Output: Geometric feature vector $C_1 = [C_{1_1}, C_{1_2}, C_{1_3}, ..., C_{1_n}]$ of length 1×12.

Steps:

1. Geometric features area, solidity, perimeter, eccentricity, euler number, compactness and circularity are extracted
2. All features are computed.
3. All the computations are combined to geometric feature vector D_1.

3.1.4 Algorithm 4 Local Ternary Pattern (LTP) feature vector

Input: Binary NST-segmented image *I*

Output: LTP feature vector $D_1 = [D_{1_1}, D_{1_2}, D_{1_3}, ..., D_{1_n}]$ of length 1×12.

Steps:

1. Pixel-wise comparison of the neighboring pixels in image *I*
2. Pixels are assigned ternary codes -1,0,1 according to comparison results
3. Upper and lower binary patterns histograms are formed respectively.
4. All the histograms are concatenated to form the LTP feature vector D_1.

The **Figure 3:** *visually illustrates the use of LTP operator.*

36	55	109	0	1	1	0	1	1	0	0	0
65	30	18	1	0	-1	1	0	0	0	0	1
37	12	9	0	-1	-1	0	0	0	0	1	1
(a)			(b)			(c)			(d)		

(b) 3-valued coded (-1, 0, 1) pattern, when threshold $t = 7$, which generate interval

[23,37] considering the central pixel i_c . (c) Upper pattern binary form 01101000.

(d) Lower pattern binary form 10000111.

Figure 3: Illustration of LTP texture operator. (a) Gray level values

Finally, the fusion of these extracted features is formed as as $allFeatures = [N \times 3780, N \times 6, N \times 12, N \times 12]$, hence $allFeatures = [N \times 3810]$, where N is the total number of input images for which the proposed algorithm is trained and Experimented.

4. Results and Discussions

The section highlights the experimental results and implementation setup, datasets description and evaluation standards used in this work. The evaluation metrics used in this study including, precision, recall, F1-score, accuracy and AUC. While for ensemble classifiers settings for SVM, RBF kernel C=1.0, in KNN K=5, Euclidean Distance, and for the decision tree, Gini criteria is adopted. The final decision is obtained on the basis of a majority voting scheme to predict the class label. As the majority voting increased robustness and improved the stability against biasness of the classifiers. The methods are evaluated on public dataset Street View Text. The dataset is selected due its scenic variabilities and complexities in evaluating the proposed methods.

4.1 Implementation Setup

The proposed model is implemented in Google Collaborator a virtual resource to run Python based notebook. We used standard GPU-based configuration to implement the solution. The solution was developed in Python and related libraries. The datasets were distributed in 70% and 30% ratio for model training and testing purposes.

4.2 Classification Results

To evaluate the robustness of the proposed methods. Following, three experiments are carried out:

1. **Experiment A:** The results of this experiment are shown in shown in **Figure 4** and **Figure 5**. In this experiment results are obtained with and without feature optimization to reflect the impact of feature optimization in image processing field,
2. **Experiment B:** In this experiment feature fusion is applied without feature optimization. The result of this experiment are listed in **Table 1,**
3. **Experiment C:** In this experiment feature fusion is integrated with feature optimization to reflect the effects of feature fusion. **Table 2** listed the results with feature fusion and optimization.

Following classifiers are used in the above experiments:

5. Cubic-SVM (C-SVM),
6. Linear-SVM (L-SVM),
7. Quad-SVM (Q-SVM),
8. K-Nearest Neighbor (KNN),
9. Decision Tree (DT).

One of the prolific post-processing step is to remove the redundant features and preserve most salient features. To obtain this, a Non-Maximum Suppression (NMS) (Si et al., 2024) was applied to the multimodal feature fusion map to obtain optimal fused feature map.

4.2.1 Experiment A

The **Figure 4** shows the results obtained from **Experiment A** of L-SVM, C-SVM, Q-SVM, KNN and DT on SVT without feature optimization. According to the results it is evident that LTP features with L-SVM classifier obtained higher accuracy among all the classifiers, i.e., **59.4%**. Since the features are not optimized therefore, the accuracy of **Experiment A** are bit low for all the classifiers. While, the results shown in the **Figure 5** are after the feature optimization. The results of **Figure 5** clearly shows the improvement in the accuracy levels of all the classifiers. However, a similar pattern can be observed that L-SVM have achieved highest accuracy which is **69.2%** but this time with HOG features.

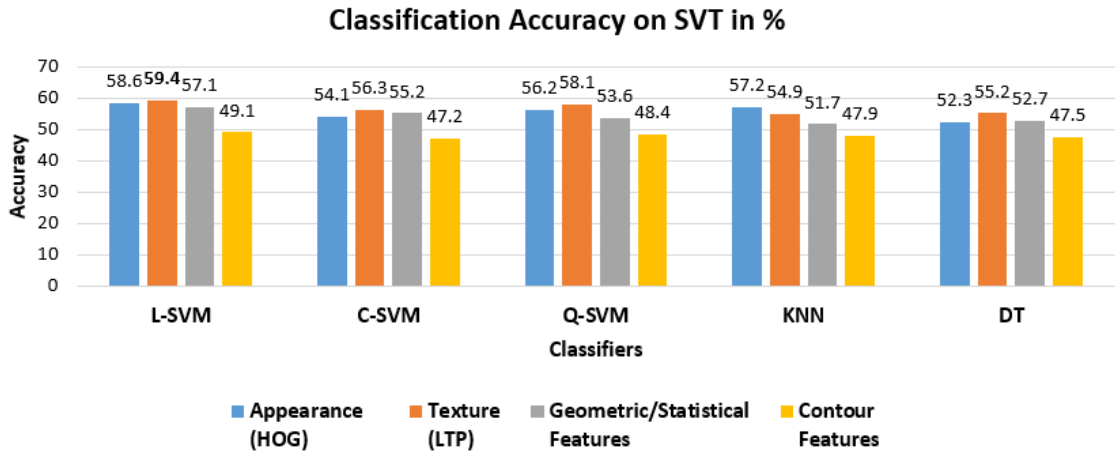

Figure 4: *Classification Accuracy on SVT in % (Experiment A)*

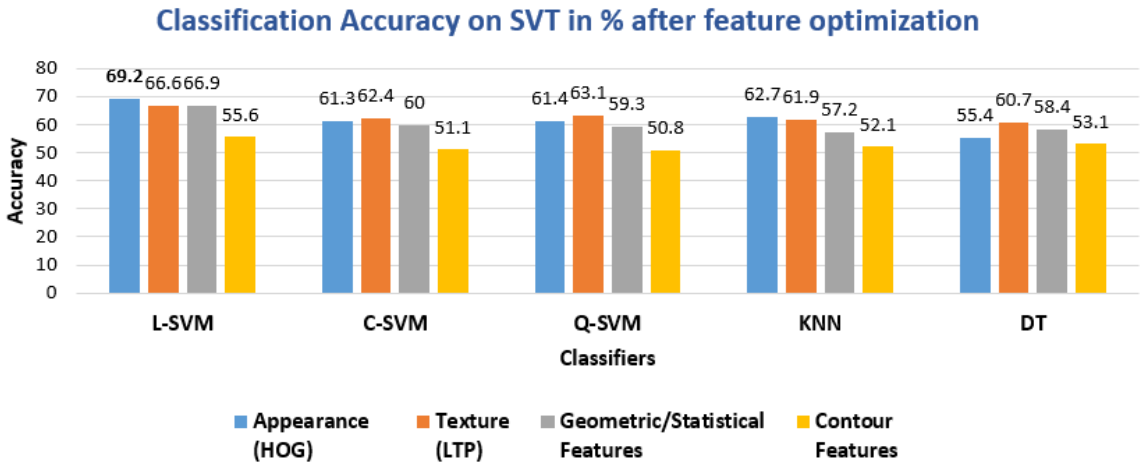

Figure 5: *Classification Accuracy on SVT in % with feature optimization (Experiment A)*

4.2.2 Experiment B

The **Experiment B** is carried out with feature fusion but without feature optimization on SVT dataset. For consistency, the results are obtained for all the classifiers used in **Experiment A**. The feature fusion technique is preferred in this experiment over the handcrafted features. It is because handcrafted features are mainly dependent on statistical information about the images. **Tables 1** presents the results of all the classifiers on SVT dataset on PPR (Predictive Positive

Rate), TPR (True Positive Rate) or Recall, F1 (F1-Score or measure), FPR (False Positive Rate), FNR (False Negative Rate), ACC (Accuracy) and AUC (Area under ROC curve) evaluation metrics. The L-SVM showed a consistent, balanced and reliable performance on all the metrics. However, since the experiment is carried out without feature optimization, therefore, all the metrics are only in acceptable range for all the classifiers.

Table 1: *(%) Classification Results on SVT with Feature Fusion and Without Feature Optimization*

Classifiers	ACC	F1	AUC	PPR	TPR	FPR	FNR
DT	88.3	75.1	0.861	74.2	76.4	0.139	11.7
KNN	86.4	73.1	0.871	71.6	74.7	0.129	13.6
Q-SVM	87.7	79.0	0.882	78.7	79.4	0.118	12.3
C-SVM	88.3	79.2	0.891	79.4	79.1	0.109	11.7
L-SVM	**91.1**	**80.8**	**0.906**	**80.4**	**81.2**	**0.094**	**08.9**

PPR (Predictive Positive Rate), TPR (True Positive Rate) or Recall, F1 (F1-Score or measure), FPR (False Positive Rate), FNR (False Negative Rate), ACC (Accuracy) and AUC (Area under ROC curve)

4.2.3 Experiment C

The **Experiment C** results are obtained with feature fusion and optimization. Therefore, the results mentioned in **Table 2** shows a significant improvement among all the classifiers on SVT dataset on all metrics. These improvements are the evident that the NMS-based optimization has met with the expectations. From the result it can be easily observed that FPR and FNR have been reduced while the numbers of other metrics have improved significantly.

Table 2 : *(%)Classification Results on SVT with Feature Fusion and Feature Optimization*

Classifiers	ACC	F1	AUC	PPR	TPR	FPR	FNR
KNN	95.4	93.9	0.942	93.6	94.4	0.058	04.6
DT	96.1	91.8	0.959	92.8	90.9	0.041	03.9
Q-SVM	96.8	93.1	0.969	92.9	93.4	0.031	03.2
C-SVM	97.1	93.3	0.973	94.5	92.3	0.027	02.9
L-SVM	**98.6**	**95.4**	**0.981**	**95.7**	**95.2**	**0.019**	**01.4**

PPR (Predictive Positive Rate), TPR (True Positive Rate) or Recall, F1 (F1-Score or measure), FPR (False Positive Rate), FNR (False Negative Rate), ACC (Accuracy) and AUC (Area under ROC curve)

5.Conclusions and Future Work

This paper presents efficacy of classification pipeline consists of ensemble classification framework and multimodal feature fusion for scene text classification. The proposed fusion of texture, contour, appearance and geometric features effectively captures diverse visual variabilities in NST images. The broad level experimentations on challenging benchmark datasets significantly confirms that L-SVM gradually and consistently outperforming competing classifiers.

Despite great reduction in false positives and strong performance the framework relies on handcrafter feature descriptors that limit the proposed for more highly complex NST images in terms of scalability. Moreover, the framework focuses on character level classification rather than word level classification.

In future versions, integration of deep features will be incorporated for multilingual and word level NST understanding and recognition to further enhance the system performance.

References

Akhtar, Z., Lee, J. W., Attique Khan, M., Sharif, M., Ali Khan, S., & Riaz, N. (2023). Optical character recognition (OCR) using partial least square (PLS) based feature reduction: An application to artificial intelligence for biometric identification. *Journal of Enterprise Information Management*, *36*(3), 767–789.

Ansari, G. J., Shah, J. H., Farias, M. C. Q., Sharif, M., Qadeer, N., & Khan, H. U. (2021). An optimized feature selection technique in diversified natural scene text for classification using genetic algorithm. *IEEE Access*, *9*, 54923–54937.

Cao, D., Zhong, Y., Wang, L., He, Y., & Dang, J. (2020). Scene text detection in natural images: a review. *Symmetry*, *12*(12), 1956.

Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., & Jiang, Y.-G. (2022). Svtr: Scene text recognition with a single visual model. *ArXiv Preprint ArXiv:2205.00159*.

Goto, H. (2008). Redefining the DCT-based feature for scene text detection: Analysis and comparison of spatial frequency-based features. *International Journal of Document Analysis and Recognition (IJDAR)*, *11*(1), 1–8.

Harizi, R., Walha, R., Drira, F., & Zaied, M. (2022). Convolutional neural network with joint stepwise character/word modeling based system for scene text recognition. *Multimedia Tools and Applications*, 1–16.

He, Z., Li, Q., Zhao, X., Wang, J., Shen, H., Zhang, S., & Tan, J. (2023). ContourPose: Monocular 6-D Pose Estimation Method for Reflective Textureless Metal Parts. *IEEE Transactions on Robotics*.

Hong, D., Wu, X., Ghamisi, P., Chanussot, J., Yokoya, N., & Zhu, X. X. (2020). Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(6), 3791–3808.

Kantipudi, M. V. V. P., Kumar, S., & Kumar Jha, A. (2021). Scene text recognition based on bidirectional LSTM and deep neural network. *Computational Intelligence and Neuroscience*, *2021*(1), 2676780.

Kushwaha, R., & Nain, N. (2020). PUG-FB: Person-verification using geometric and Haralick features of footprint biometric. *Multimedia Tools and Applications*, *79*(3–4), 2671–2701.

Minetto, R., Thome, N., Cord, M., Leite, N. J., & Stolfi, J. (2013). T-HOG: An effective gradient-based descriptor for single line text regions. *Pattern Recognition*, *46*(3), 1078–1090.

Pandey, D., Pandey, B. K., & Wairya, S. (2021). Hybrid deep neural network with adaptive galactic swarm optimization for text extraction from scene images. *Soft Computing-A Fusion of Foundations, Methodologies \& Applications*, *25*(2).

Rassem, T. H., & Khoo, B. E. (2014). Completed Local Ternary Pattern for Rotation Invariant Texture Classification. *The Scientific World Journal*, *2014*(1), 373254. https://doi.org/https://doi.org/10.1155/2014/373254

Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., & Hassner, T. (2021). TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8798–8808. https://doi.org/10.1109/CVPR46437.2021.00869

Tan, X., & Triggs, B. (2010). Enhanced Local Texture Feature Sets for Face Recognition Under

Difficult Lighting Conditions. *IEEE Transactions on Image Processing*, *19*(6), 1635–1650. https://doi.org/10.1109/TIP.2010.2042645

Tang, S., Cao, Y., Liang, S., Jin, Z., & Lai, K. (2025). Scene Text Recognition That Eliminates Background and Character Noise Interference. *Applied Sciences*, *15*(7), 3545.

Wang, Y., Shi, C., Xiao, B., Wang, C., & Qi, C. (2018). CRF based text detection for natural scene images using convolutional neural network and context information. *Neurocomputing*, *295*, 46–58.

Xie, H., Fang, S., Zha, Z.-J., Yang, Y., Li, Y., & Zhang, Y. (2019). Convolutional attention networks for scene text recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*(1s), 1–17.