A Method for Extracting News Text Information from Converged Media Videos Based on SWT Algorithm

Lixiang Shi[1], Jing Liang[1,*], Qi Li[1], Rui Lv[1]
{355148443@qq.com, 15420948@qq.com, 782632722@qq.com, 409297066@qq.com}

[1] School of Computer Engineering, Chengdu Technological University,
No. 1, Section 2, Zhongxin Avenue, Pidu District, Chengdu 611730, Sichuan, China

Abstract. In converged media videos, only semantic images can be selected, resulting in low reliability of extracted information. Therefore, this paper proposes a method for extracting news text information from converged media videos based on the SWT algorithm. For news images, a hierarchical preprocessing framework is adopted, employing strategies such as spatiotemporal sampling dimensionality reduction, multi-channel separation and feature enhancement, and attention mechanism region weighting to select images with both semantic and spatial saliency. The news text region is determined based on the SWT algorithm, and an adaptive sliding window is used to suppress the influence of illumination. A maximum stable dynamic region criterion is proposed to detect spatiotemporally stable regions, and a triple feature encoding mechanism is designed. A multi-level feature fusion framework is proposed, generating positive and negative sample pairs to define the extraction loss function. Edge density is used to distinguish text from noise, and the extraction and classification losses are combined for optimization to achieve news text information extraction. Experimental results show that the proposed method reduces the loss value by 0.07 in each of the first 6 rounds, decreases to 0.32 in the $10th$ round, and finally stabilizes at 0.25. On three datasets, including NewsHub, the proposed method achieves an $F1$ score of up to 0.92 and an AP of 0.90, representing a 3.2%-3.8% improvement over the best comparison method. In feature space visualization, the average aggregation degree of similar texts reaches 91%, while the proportion of outliers drops to a minimum of 0.3%. This demonstrates the superior reliability of the extracted information, effectively addressing multimodal interference and possessing significant practical value.

Keywords: SWT algorithm, Converged media video, News text, Information extraction, Feature encoding, Loss function

1 Introduction

In the era of converged media, the carriers and forms of news dissemination have undergone tremendous changes. Traditional single media has been replaced by multiple forms, resulting in wider information dissemination and higher efficiency for users to receive information. However,

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365259

there are challenges in extracting news text information from converged media videos. Dynamic changes in video frames make text positioning difficult, complex backgrounds and low resolutions make character recognition blurry, and the mixing of multiple languages and professional terms affects the accuracy of semantic understanding. Under these circumstances, efficiently and accurately extracting news text information from massive amounts of video has become the key to improving the effectiveness of news dissemination.

In the field of text information extraction, the sentiment analysis three-level classification framework proposed by Mukasheva[1] provides a hierarchical analysis from text level, sentence level to object attribute level, and builds a theoretical model for structured data extraction. It has good interpretability in static scenarios such as news comments and social media texts. However, the text position shift, font size change and background interference between video frames will destroy its rule system and greatly reduce the extraction accuracy. Ivanovi S [2] explored the legality boundary of text data mining from the perspective of copyright law, and provided an important reference for the ethical norms of information extraction. However, his research focuses on legal compliance and does not involve the reliability optimization in technical implementation. News texts are diverse in form, such as subtitles, bullet comments, watermarks, etc. The copyright ownership and usage rights are complex. When technology developers pursue extraction efficiency, they often face the dual challenges of compliance risks and technical feasibility.

The image text extraction and translation system developed by Nagmoti S et al. [3]uses image enhancement and perspective transformation algorithms to improve the accuracy of static document recognition and performs well in scanned documents, photos and other scenarios. However, when processing video streams, due to the lack of time sequence information modeling ability, the dynamic text tracking and matching effect is poor, affecting the integrity of information extraction. The neural machine translation fusion key information method proposed by Hu S et al. [4] relies on the attention mechanism to strengthen semantic association and has made progress in cross-language text conversion. However, the model is mainly for pure text input and has poor compatibility with multimodal content. The extraction results often have semantic ambiguity or missing information.

In order to break through the bottleneck of existing converged media video news text information extraction, a converged media video news text information extraction method based on SWT algorithm is proposed. This method uses the dynamic weight adjustment mechanism to enhance the model's perception of text stroke width, can accurately locate text regions in complex backgrounds, and also introduces a semantic enhancement module to integrate multimodal features and context to improve the semantic understanding accuracy of blurred and occluded text.

2 News Text Information Extraction Methods

2.1 Preprocessing News Images in Converged Media Videos

When disseminating news in converged media videos, news images are often embedded in the video stream as dynamic frame sequences. The content contains the core information of the news subject and may also convey key clues through implicit carriers such as backgrounds. This paper proposes a hierarchical preprocessing framework, which enhances the extraction of the core

semantics of news images through multi-stage feature screening and dynamic weight allocation [5] [6]. The overall process is shown in Fig. 1.

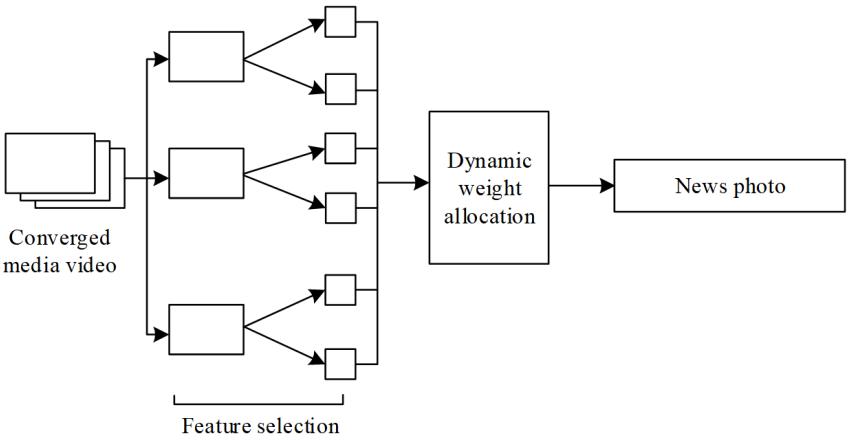

Fig. 1. News Image Layered Preprocessing Flow

To address the redundancy problem of video frame sequences, a spatiotemporal sampling strategy is first used to reduce the dimensionality of the original video [7]. Let the video stream be $V = [f_1, f_2, \ldots, f_N]$, where f_i represents the i frame image. Calculate the structural similarity index between adjacent frames:

$$\mathrm{SSIM}(f_i, f_{i+1}) = \frac{\left(2\mu_{f_i}\mu_{f_{i+1}} + C_1\right)\left(2\sigma_{f_i,f_{i+1}} + C_2\right)}{\left(\mu_{f_i}^2 + \mu_{f_{i+1}}^2 + C_1\right)\left(\sigma_{f_i}^2 + \sigma_{f_{i+1}}^2 + C_2\right)} \tag{1}$$

Where: μ_{f_i}, σ_{f_i} are the mean and standard deviation of the frame respectively, C_1, C_2 are the stability constant.When $\mathrm{SSIM}(f_i, f_{i+1}) > \tau$ (threshold $\tau = 0.95$), it is determined that the content of the two frames is highly similar, and only the key frame f_k is retained, thereby compressing the frame sequence to $V' = [f_{k_1}, f_{k_2}, \ldots, f_{k_M}]$, $\quad M \ll N$.

To improve the computational efficiency of subsequent processing, multi-channel separation and feature enhancement are performed on the retained key frames. First, the color frame f_k is decomposed into luminance channel I and chrominance channels UV, and apply adaptive contrast stretching only to channel I:

$$I'(x,y) = \alpha \frac{I(x,y) - I_{\min}}{I_{\max} - I_{\min}} + \beta \cdot \mathrm{SSIM}(f_i, f_{i+1}) \tag{2}$$

Where: α, β are the enhancement coefficient, $I_{\min}$, $I_{\max}$ are the minimum and maximum values of the channel.

To address the spatial saliency of target subjects in news images, a region-weighted strategy based on an attention mechanism is introduced.The target detector identifies n regions $[r_1, r_2, \ldots, r_n]$

in frame f_k, where each region r_j's spatial weight w_j is determined by the normalized distance between its center coordinate (x_j, y_j) and the frame center (x_c, y_c).

$$w_j = 1 - \sqrt{\frac{(x_j - x_c)^2 + (y_j - y_c)^2}{D_{\max} \cdot I'(x,y)}} \tag{3}$$

Where: $D_{\max}$is half the length of the frame diagonal.

This weight gives higher priority to regions closer to the center of the image. Finally, combining semantic saliency and spatial saliency, a comprehensive importance score s_j is constructed:

$$S_j = \lambda \cdot \mathrm{Sem}(r_j) + (1 - \lambda) w_j \tag{4}$$

Where: $\mathrm{Sem}(r_j)$ is constructed as the semantic confidence of the region r_j, and λ is the balance coefficient.

Through this formula, news images possessing both semantic keyness and spatial saliency can be dynamically selected, laying the foundation for subsequent accurate information extraction.

2.2 Determining News Text Regions in Converged Media Videos Based on SWT Algorithm

After completing the spatiotemporal preprocessing and multimodal alignment of the news images in the converged media video, it is necessary to accurately model the features of the news text region extracted from the dynamic frame sequence. Given that the video frame has time-varying characteristics, the window size is set to$W \times H$, and the local threshold of the center pixel $T(x,y)$is determined by the distribution of pixel intensity [8] in the window:

$$T(x,y) = \mu(x,y) + k \cdot \sigma(x,y) \tag{5}$$

Where :$\mu(x,y)$,$\sigma(x,y)$ are the mean and standard deviation of the pixels in the window, respectively,k is the dynamic adjustment coefficient.

By traversing the entire image through the sliding window, a threshold mapping map T_{map} is generated, which can effectively suppress the influence of local illumination fluctuations on text segmentation. Converged media video needs to process the dynamic changes between frames. This paper proposes the maximum stable dynamic region criterion [9], defining the dynamic change rate of the region T_{map} in the time series f_t:

$$\Delta R(t) = \frac{|A_R(t) - A_R(t-1)|}{A_R(t-1)} \cdot T(x,y) \tag{6}$$

Where: $A_R(t)$ is the area of the region R in the time series f_t.

When $\Delta R(t) < 0.1$ and continues for $\Delta t \geq 3$ frame, determine R as the stable dynamic region. Combining grayscale inversion operation, synchronously detecting positive and negative threshold regions, and finally merging them to obtain spatially and temporally stable news text regions.

When performing feature deconstruction on the detected text region, writing style, spatial layout, and semantic association must be taken into account. This paper designs a triple feature encoding mechanism:

1) Writing style features

The line lengths of text from the same news source exhibit a regular distribution, while the line lengths of noisy texts fluctuate randomly. Line length similarity is defined by clustering labeled samples using the SWT algorithm:

$$S(l_i, l_j) = \exp\left(-\frac{|l_i - l_j|}{\sigma_l}\right) \tag{7}$$

Where: l_i, l_j are the number of characters in two lines, σ_l is the standard deviation of the sample.

Select the line with the highest similarity to the target line m, calculate its mean line length μ_l and range δ_l, and construct the writing feature vector:

$$e_w = [\mu_l \cdot S(l_i, l_j), \mu_l - \delta_l, \mu_l + \delta_l] \tag{8}$$

2) Spatial layout features

The distribution of line numbers in news text reflects structural priority. Define the line number normalized coordinates:

$$e_p = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \tag{9}$$

Where: y_i is the vertical coordinate of the current line, $y_{\min}$ and $y_{\max}$ are the boundary of the text region.

Generate positional feature representation by mapping e_p to d_p-dimensional space through Gaussian kernel mapping.

3) Semantic association features

Semantic embedding e_s of line text is extracted using the SWT algorithm, and fuse writing and spatial features through an attention mechanism:

$$e_f = Attn[(e_w, e_p), e_s] \cdot W + b \tag{10}$$

Where: W,b are learnable parameters to realize dynamic weighting of multimodal features.

Through the above methods, news text regions can be stably extracted in complex dynamic scenes, providing a reliable foundation for subsequent information extraction.

2.3 Implementation of News Text Information Extraction from Converged Media Videos

In dynamic scenes of converged media videos, news text extraction must consider both spatiotemporal continuity and semantic integrity. This paper proposes a multi-level feature fusion framework based on the SWT algorithm, using defined text regions for context awareness to achieve efficient extraction in complex scenes.

To enhance feature discriminativeness, the SWT algorithm is used to generate positive and negative sample pairs [10]: positive sample pairs, word-word vectors (h_t^i, h_t^j) and sentence-word

vectors (h_s, h_t^i) of the same text Negative sample pairs, word-word vectors (h_t^i, h_t^{jk}) and sentence-sentence vectors (h_s, h_s') of different texts. Thus, the extraction loss function is defined:

$$L_c = -\log \frac{sim(h_t^i, h_t^j)/v}{sim(h_t^i, h_t^j)/v + sim(h_t^i, h_t^{jk})/v} \tag{11}$$

Where: *sim* is cosine similarity, v is temperature coefficient.

To distinguish between text and background noise, edge density ρ is defined:

$$\rho = \frac{\sum_G |G_x(x,y)||G_y(x,y)|}{L_c \cdot A_R \cdot \bar{I}_R} \tag{12}$$

Where: G_x, G_y are the gradients of the Sobel operator, $\bar{I}_R$ is the mean gray level of the region, A_R is the area of the region. When $\rho > 0.3$, candidate regions are retained.

Finally, the class probability is output through the Softmax classifier:

$$Q(y|x) = \frac{\rho \cdot e^{W+b}}{\sum e^{W+b}} \tag{13}$$

Where: e is the base of the natural logarithm.

Cross-entropy is used for classification loss optimization:

$$L_{CE} = -\sum \frac{sim(h_s, h_t^i)}{sim(h_s, h_s')} \cdot \log Q(y|x) \tag{14}$$

Combining extraction loss and classification loss, the overall objective function is:

$$L_t = L_{CE} - \vartheta L_c \tag{15}$$

Where: ϑ is the weight coefficient.

Through the above methods, accurate extraction and semantic understanding of news text can be achieved in video scenes, providing reliable support for converged media content analysis.

3 Experiments

3.1 Experimental Environment

This experiment focuses on the method of extracting news text information from converged media videos based on the SWT algorithm. A complete experimental platform was constructed from data acquisition to result presentation. Its architecture covers multiple stages including data collection, processing, querying, business logic, and display, as shown in Fig. 2.

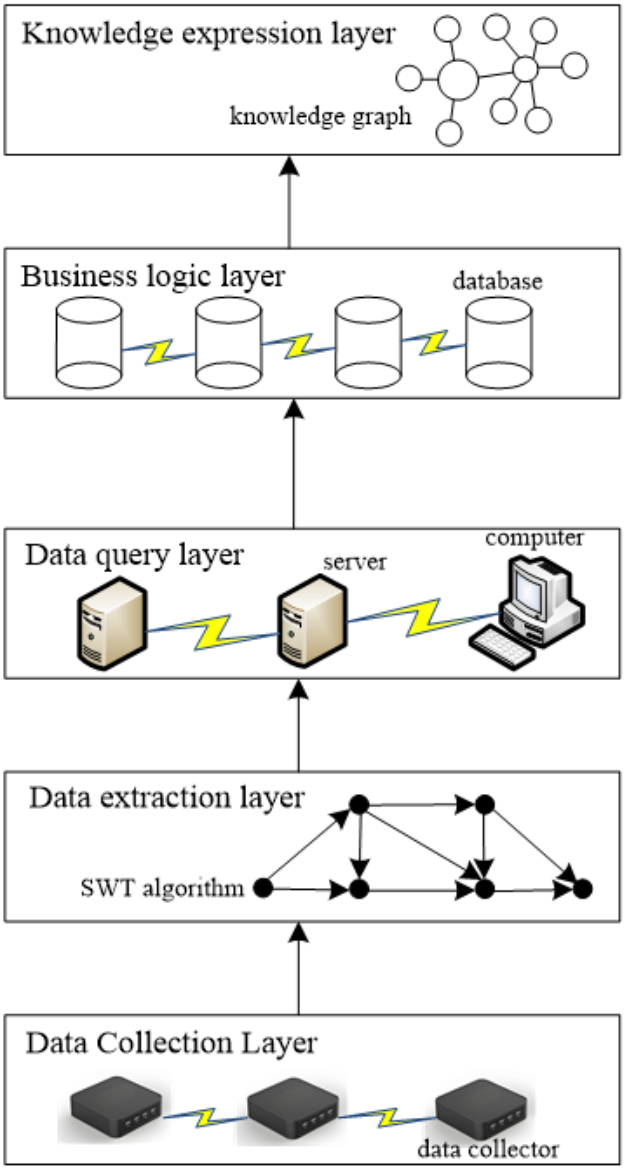

Fig. 2. Complete Experimental Platform Architecture

The experiment was conducted in a high-performance computing environment equipped with an Intel(R) Xeon(R) Platinum 8280 processor, 32GB of memory, and an NVIDIA Quadro RTX 6000 graphics card (24GB of video memory), providing powerful computing power for deep learning

model training. The experimental environment settings are shown in Table 1.

Table 1: EXPERIMENTAL ENVIRONMENT SETTINGS

Serial Number	Experimental Environment Component	Detailed Configuration Information
1	Operating System	Ubuntu 20.04 LTS 64-bit
2	Programming Language	Python 3.8
3	Deep Learning Framework	TensorFlow 2.6.0 (GPU version), PyTorch 1.9.0
4	Database	MongoDB 5.0, Neo4j 4.3
5	Front-end Framework	React 17.0.2
6	Visualization Library	ECharts 5.2.2
7	Driver	NVIDIA GPU driver version 470.57.02, CUDA Toolkit 11.4
8	Other Tools	OpenCV 4.5.3 (for image processing), Scikit-learn 0.24.2 (for machine learning assistance)

The SWT algorithm model is implemented using Python combined with TensorFlow, and PyTorch is used to assist in verification. The development is front-end and back-end separated. The back-end processes data based on the Python ecosystem and stores it in Neo4j. The front-end uses React to build pages and interacts through Ajax. Knowledge graph visualization is implemented using ECharts to ensure system stability.

3.2 Experimental Preparation

In the experimental preparation stage of the method of extracting news text information from converged media videos based on the SWT algorithm, three representative news datasets, namely NewsHub, DailyDigest, and MediaCorp, were selected to comprehensively evaluate the performance of the method. The experimental data were strictly divided into training set, validation set, and test set according to a 7:2:1 ratio to ensure the reliability of the evaluation results.

1) The NewsHub dataset collects five years of publicly reported news from several authoritative news websites. After screening and cleaning, it focuses on six core areas, including international and domestic. Each category contains about 12,000 data points, with an average text length of 22 characters, which can provide rich language samples for the model.

2) The DailyDigest dataset comes from a well-known news aggregation platform. It selects six dimensions: health, education, real estate, automobiles, culture, and society. Each category contains about 6,000 refined news articles with an average length of about 24 characters. It aims to examine the adaptability of the algorithm in different sub-fields.

3) The MediaCorp dataset is used as a supplement, covering a wider range of news categories, including 10 fields such as environment, art, fashion, and tourism. Each category has about 3,500 data points, with an average length of about 21 characters, further enriching the diversity of the experiment.

In the experimental preparation stage, the key parameters of the SWT algorithm were carefully set, as shown in Table 2.

The core text encoding uses GELU as the activation function to enhance nonlinear expression. The AdamW with weight decay is selected as the optimizer for training. After multiple rounds of experiments, the maximum input length of the fixed text is 64, which achieves the best balance between computational efficiency and information integrity.

Table 2: Model Parameters and Their Range/Value Options

Parameter Category	Parameter Name	Range/Value Options
Model Architecture	Hidden Layer Dimension	768/1024/1280
	Number of Hidden Layers	8/12/16
Contrastive Learning	Temperature Coefficient	[0.05,0.1,0.15]
	Weight Balance Factor	0.2/0.3/0.4
Training Optimization	Learning Rate	[1e-5,2e-5,5e-5]
	Dropout Rate	0.3/0.4/0.5
Text Processing	Maximum Input Length	48/64/80

3.3 Experimental Result Analysis

Based on the above parameter configuration, fine-tuning training was carried out. The iterative changes of the loss function of the proposed method were recorded throughout the experiment, as shown in Fig. 3.

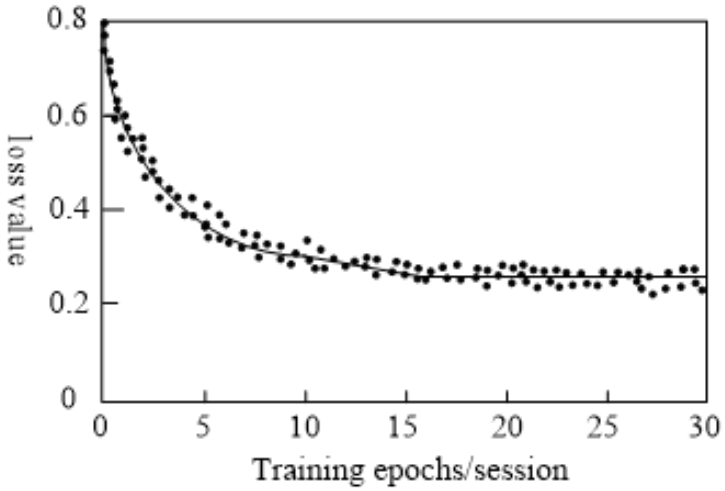

Fig. 3. Iterative Changes of the Loss Function of the Proposed Method

It can be clearly seen from the iteration curve in Figure 3 that the loss value decreases with the training rounds. In the first 6 rounds, the loss value decreases by 0.07 per round, and by the $10th$ round it decreases to 0.32, entering the fluctuation and convergence stage, and finally stabilizing at 0.25. This shows that the carefully tuned parameter combination can efficiently guide the model to capture the deep semantic features of news texts and performs well when dealing with long texts and complex sentence structures. This fully verifies that the SWT algorithm has practical value in the

converged media scenario.

To verify the effectiveness of the method in this paper, four representative methods from references [1], [2], [3], and [4] were selected as comparison benchmarks, and comparative tests were carried out on three datasets of different sizes. The $F1$ score and average accuracy (AP) of each method were evaluated simultaneously, and the specific experimental results are shown in Table 3.

Table 3: EXPERIMENTAL RESULTS OF F1 VALUE FOR EACH METHOD AND AVERAGE ACCURACY (AP)

	NewsHub		**DailyDigest**		**MediaCorp**	
Dataset	**F1 value**	**AP**	**F1 value**	**AP**	**F1 value**	**AP**
Reference[1]	0.82	0.79	0.78	0.76	0.75	0.73
Reference[2]	0.85	0.83	0.81	0.79	0.78	0.76
Reference[3]	0.87	0.85	0.83	0.81	0.80	0.78
Reference[4]	0.89	0.87	0.85	0.83	0.82	0.80
Proposed Method	0.92	0.90	0.88	0.86	0.85	0.83

According to Table 3, on the NewsHub dataset, the F1 score of the proposed method reached 0.92 and the AP was 0.90, which is 3.2% and 3.0% higher than the best benchmark in the literature [4], respectively. On the DailyDigest dataset, the $F1$ score and AP were 0.88 and 0.86, respectively, which are 3.5% and 3.6% higher. On the MediaCorp dataset, the two indicators were 0.85 and 0.83, which are 3.7% and 3.8% higher. As the data scale expands, the advantages of the proposed method become more obvious, with the $F1$ score increasing the most on NewsHub. The continuous improvement of AP indicates that the model improves accuracy while ensuring recall. The experiment shows that the SWT algorithm solves the multimodal interference problem of converged media video news text by means of dynamic weights and semantic enhancement mechanism, and provides a more reliable information extraction scheme.

To further verify the adaptability of the proposed method, samples from three major fields—technology, sports, and entertainment—were selected from the NewsHub dataset. Five methods were used to visualize the feature space of the test set, and the results are shown in Fig. 4.

As can be seen from Fig. 4, the literature [1] can distinguish news categories, but the same type of texts are discrete in the feature space. About 35% of the samples of the science and technology category are not effectively clustered, and 12% of the texts overlap with other categories. The feature extraction is insufficient in capturing semantic associations. Reference [2] showed that the aggregation degree of similar texts increased to 68%, but there were still 8% outliers, and the performance of long sports texts was unstable. Reference [3] showed that the clustering effect of science and technology texts reached 75%, while the entertainment texts had 15% of samples that were scattered due to the large number of professional terms. Reference [4] showed that the average aggregation degree of each category reached 82%, but there were 10% misclusterings in the colloquial entertainment texts. The method in this paper uses the SWT algorithm to increase the average

aggregation degree of each category to 91%, and the proportion of outliers in science and technology texts decreased to 0.3%, which verifies its optimization ability for complex semantic structures.

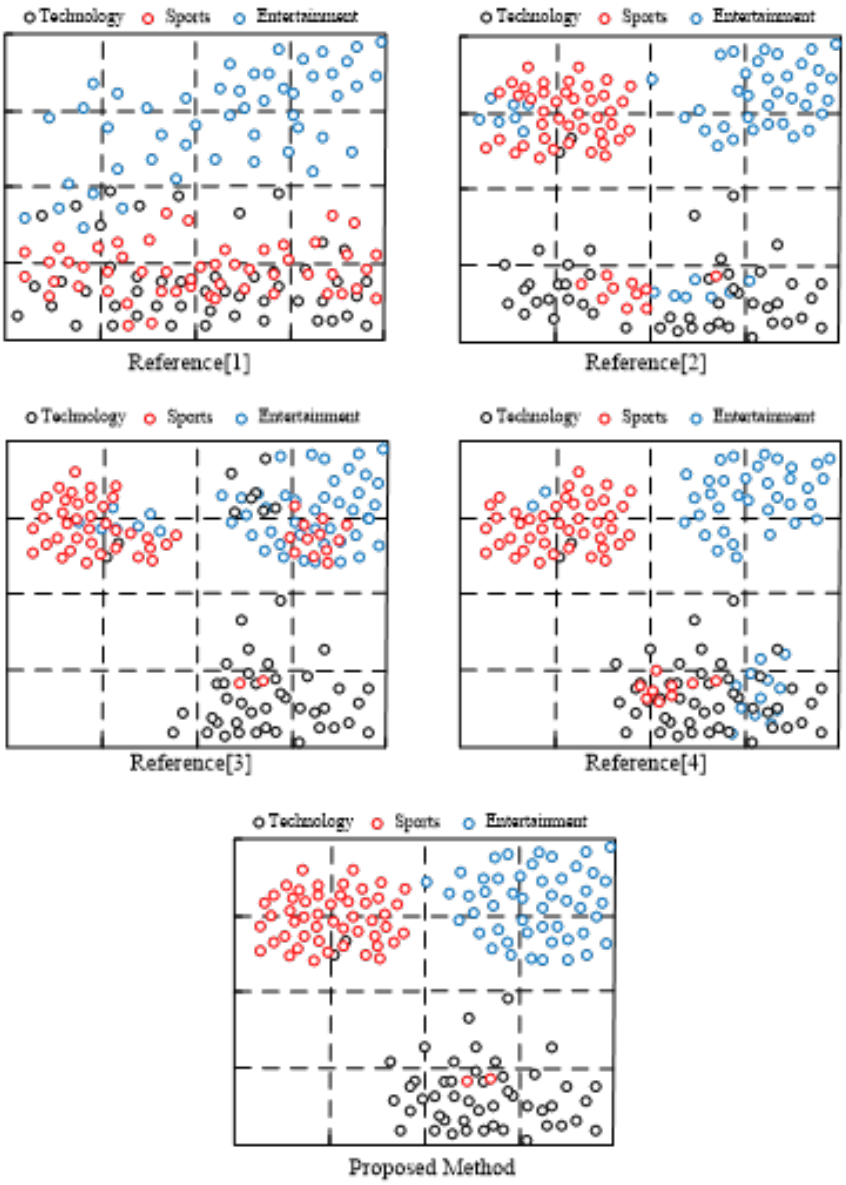

Fig. 4. Feature Space Visualization Results of Five Methods

4 Conclusions

Against the background of rapid media convergence and the growth of short-video news distribution, accurate extraction of textual information from video frames has become important for content understanding, archiving, retrieval, and intelligent analysis. However, practical scenarios are often affected by motion blur, illumination variation, background clutter, and semantic interference caused by multimodal content, which makes stable text extraction difficult. In response to these challenges, this study explored the method of extracting text information from converged media video news based on the SWT algorithm and achieved valuable results. The proposed dynamic text region detection method reduces the interference of inter-frame illumination changes and motion blur, and can accurately locate stable text regions, laying the foundation for subsequent feature extraction. Multi-granularity feature encoding enhances the discriminative power of text features, and the model can distinguish text in different scenarios and mine news theme features. Edge density measurement and post-processing improve the accuracy and readability of text extraction and remove background noise. The study also found that dynamic region detection should consider both local and global stability, while physical features such as multi-edge density can help filter non-text regions. Overall, these findings enrich the theoretical system of video text extraction and provide a useful reference for subsequent research and engineering practice.

Acknowledgments

This work is supported in part by Sichuan Provincial Key Laboratory Open Research Project (Project No. 2024-ScL-MC&I-005).the Second Batch of Modern Industry School in Sichuan Province:“Industry School of Artificial Intelligence Media and Software”(project No.[2023]263).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Mukasheva TASKS AND METHODS OF TEXT SENTIMENT ANALYSIS. Scientific Journal of Astana IT University. 2021.

[2] Ivanovi S. Copyright law and text and data mining. Zbornik radova Pravnog fakulteta u Nišu. 2021;60(92):59-78.

[3] Nagmoti S, Bhoyar K, Raut S, Jamgade S, Mangrulkar N, Pathade A. IMAGE TEXT EXTRACTION AND ITS LANGUAGE TRANSLATION. Journal of Research in Engineering and Applied Sciences. 2021;(2).

[4] Hu S, Li X, Bai J, Lei H, Qian W, Hu S, et al. Neural Machine Translation by Fusing Key Information of Text. CMC-COMPUTERS MATERIALS amp; CONTINUA. 2022;74(2):2803-15.

[5] Liu S, Sun K, Fu L, Chen X, Zhang X, Lin Z, et al. SCRIBES: Web-Scale Script-Based Semi-Structured Data Extraction with Reinforcement Learning. 2025.

[6] Fromm H, Wambsgan T, Sllner M. Towards a Taxonomy of Text Mining Features. 2019.

[7] Garganas O. Digital Video Advertising: Breakthrough or Extension of TV Advertising in the New Digital Media Landscape? Journalism and Media. 2024;5(2):749-65.

[8] Fatourechi R, Momtazi S. Persian Text Summarization using Sparse Coding with Neural Text Representation. Iranian Journal of Information Processing and Management. 2021.

[9] Karimpour M, Slob E, Socco LV. Comparison of Straight and Curved-Ray Surface Wave Tomography at Near-Surface Scale: a 3D Numerical Example. 83rd EAGE Annual Conference & Exhibition. 2022:1-5.

[10] Al-Wesabi FN. Text Analysis-Based Watermarking Approach for Tampering Detection of English Text. CMES-COMPUTER MODELING IN ENGINEERING amp; SCIENCES. 2021;67(3):3701-19.