Automated Unauthorized Banner Detection System

Sang Hyun Park
Dept. of Artificial Intelligence
Ajou University
Suwon, Korea
hyun0666@ajou.ac.kr

Hyung Il Koo
Dept. of Electrical and Computer Engineering
Ajou University
Suwon, Korea
hikoo@ajou.ac.kr (corresponding author)

***Abstract*—Banners are widely used for public communication and advertising. However, due to their sheer volume, many unauthorized banners are posted, making manual monitoring impractical. In this study, we propose an automated system that detects banners from real-world video or image sources, such as fixed CCTVs and in-vehicle dashcams, and determines whether they are authorized. The system first detects and tracks banners using YOLOv8 for instance segmentation and ByteTrack for multi-object tracking. Then, it selects a frame in which the banner is most clearly visible and applies a general-purpose OCR engine to recognize the text. Finally, a large language model (LLM) analyzes the meaning of the text and classifies the banner as authorized or unauthorized based on context. Except for the segmentation component, the system relies solely on pre-trained general-purpose models without any task-specific fine-tuning. Experiments on real-world banner images show that simply combining existing models in a well-structured manner can yield strong performance. These results suggest that this approach holds significant practical potential.**

***Index Terms*—OCR, Unauthorized banner detection, Vision-language pipeline, Keyframe extraction, LLMs.**

I. Introduction

Banners remain one of the most widely used and effective forms of visual media for disseminating public information, promoting events, and advertising commercial services. Owing to their low cost, high visibility, and ease of production and installation, banners are commonly found in urban environments such as roadsides, building facades, pedestrian overpasses, and intersections. Within local communities, banners serve as a direct communication channel to the public and continue to complement digital media in offline contexts.

Despite their practicality, the widespread use of banners raises several concerns. When placed in unauthorized or inappropriate locations, banners can degrade the urban landscape, obstruct visibility, and disrupt public order. The rapid proliferation of unauthorized banners—those installed without permission or in violation of local regulations—has become a persistent social issue. Unregulated commercial advertisements and promotional materials frequently clutter public spaces, leading to safety risks, visual pollution, and increased administrative burdens.

Currently, the detection and removal of unauthorized banners are primarily carried out manually by on-site inspectors. Although rule-based approaches based on specific keywords or layout heuristics have been explored as simple automation techniques, these methods face significant limitations in real-world applications.

(a) Visible public institution name banner

(b) Occluded public institution name banner

(c) Cropped public institution name banner

Fig. 1. Limitation of Rule-Based Unauthorized Banner Detection. In (a), the name of the local government is clearly visible inside the red box. However, in (b) and (c), occlusion hides the name, making it unrecognizable.

A. Limitations of Rule-Based Unauthorized Banner Detection

As illustrated in Fig. 1, rule-based systems tend to rely heavily on the presence of explicit keywords, such as the names of city offices or public institutions. In Fig. 1(a), where the government name is clearly visible, the banner is correctly identified as authorized. However, in Fig. 1(b) and (c), the same identifiers are occluded or partially hidden, causing the system to misclassify them as unauthorized. These examples highlight a core limitation of rule-based systems: They perform lexical matching without contextual understanding of the banner content.

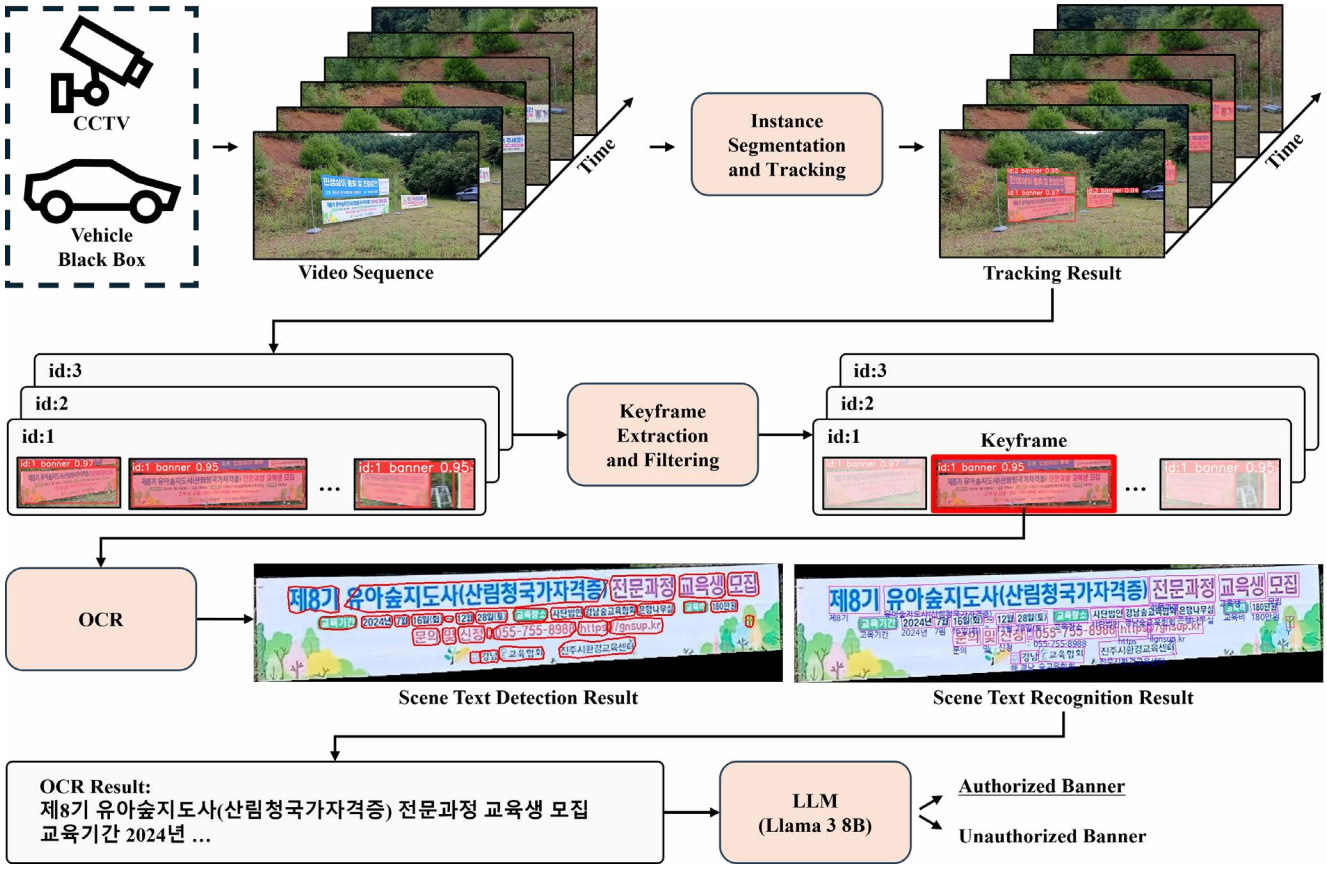

Fig. 2. Proposed banner analysis pipeline. It includes four stages: instance segmentation and (optional) tracking, keyframe extraction and filtering, OCR(text detection and recognition), and LLM-based classification, addressing challenges of unstructured layouts and noise.

To address these limitations, we propose an end-to-end framework that integrates robust Optical Character Recognition (OCR) with a powerful Large Language Model (LLM) to detect unauthorized banners more accurately by incorporating contextual understanding. Unlike traditional methods that rely solely on keyword matching, our approach analyzes the semantics and context of the text extracted from banners, enabling more robust detection even under occlusion or varying layouts.

B. Related Work

In this work, we propose a multi-stage pipeline for the automatic detection and classification of banners in complex urban environments. We utilize YOLOv8-seg [1] for accurate instance segmentation of banners, and ByteTrack [2] for robust multi-object tracking across frames. Text regions within the detected banners are identified using TextBPN++ [3], a scene text detector effective under diverse environmental conditions. For text recognition, we employ CDistNet [4], which is robust to distortions and font variations. Finally, Llama 3-8B [5], a large language model, interprets the recognized text to perform classification beyond simple keyword matching. This pipeline addresses challenges inherent in real-world banner analysis, achieving improved accuracy and robustness.

II. Proposed Method

In this section, we present our method for the detection of unauthorized banners. An overview of the entire pipeline is shown in Fig. 2.

A. Instance Segmentation and Tracking

Banners typically exhibit a rectangular and structured layout. However, distortions frequently occur when banners are captured from oblique or varying viewpoints. Most general-purpose OCR models are optimized for document-style inputs that assume clean, front-facing text [6]–[8]. Therefore, accurately localizing and cropping the banner region is critical for reliable recognition.

To this end, we adopt an instance segmentation approach rather than conventional (box-based) object detection. We use the YOLOv8-seg [1] model, which demonstrates strong performance in both object detection and instance segmentation tasks, to accurately segment banner regions.

When processing video input, banners are detected on a frame-by-frame basis. To associate the same banner across consecutive frames, we apply ByteTrack [2], a high-performance multi-object tracking algorithm. By combining instance segmentation and tracking, our system assigns consistent IDs to banners across frames, enabling robust temporal tracking in video sequences.

(a) Filtering edge (1)

(b) Filtering edge (2)

Fig. 3. Examples of banners located at the image boundary. Green box banners are filtered out due to partial occlusion or incomplete visibility, both of which can negatively impact OCR and classification performance.

(a) Filtering small (1)

(b) Filtering small (2)

Fig. 4. Examples of banners filtered out due to small size. Such instances(Yellow box) are often distant or irrelevant and may yield unreliable recognition or classification results.

B. Keyframe Extraction and Filtering

For image inputs, all detected banners are processed directly. In contrast, for video inputs, it is necessary to select a representative frame for each tracked banner instance. Although various keyframe extraction techniques have been proposed in previous studies [9], [10], we adopt a method based on two specific criteria.

First, we consider the confidence score provided by the instance segmentation model [1], as higher scores typically correspond to sharper and better-aligned frames. Second, we evaluate the aspect ratio of the segmented banner as an indicator of geometric distortion, since banners captured at oblique angles tend to exhibit abnormal aspect ratios. By combining these two cues, we select the frame that is both visually clear and front-facing as the keyframe.

After keyframe selection, we apply a filtering process to exclude banners unsuitable for downstream analysis. Specifically, we discard (i) banners that are partially outside the image boundary, as they are likely occluded or incomplete (see Fig. 3 green box), and (ii) banners that are significantly smaller than the largest detected instance in the same scene, which typically indicates distant or irrelevant instances (see Fig. 4 yellow box). This filtering step improves the quality of inputs for subsequent OCR and classification tasks.

C. OCR and LLM

Rather than developing a custom OCR system for banners, we adopt a modular pipeline based on robust, off-the-shelf models. The OCR pipeline consists of two stages: scene text detection and text recognition.

For the detection phase, we employ TextBPN++ [3], a state-of-the-art model designed to detect arbitrarily shaped text regions. Unlike conventional bounding-box-based methods, it generates polygonal boundaries through iterative refinement. This is advantageous for banners, where text often follows irregular layouts or appears distorted due to perspective.

In text recognition, we utilize CDistNet [4], a robust model that captures character-wise distance representations across multiple domains. CDistNet [4] is particularly effective in handling various fonts, sizes, and distortions common to real-world banner text.

Once the scene text is recognized, we incorporate a Large Language Model (LLM), specifically Llama 3-8B [5], to understand the context of the extracted text. Instead of relying solely on keyword matching or rule-based logic, the LLM offers context-aware interpretation of the banner's message. This enables high-level classification, such as political slogans, public service announcements, and general advertisements.

(a) Authorized banners (political and public banners)

(b) Unauthorized banners (general banners)

Fig. 5. Comparison between unauthorized and authorized banners. (a) Examples of authorized banners that comply with placement and content regulations. (b) Examples of banners labeled as unauthorized according to our heuristic labeling strategy, typically due to unregistered promotional content or improper placement. (Personal information has been redacted.) Since visual features alone are often insufficient to determine authorization status, labels are assigned based on banner category: political and public banners are considered authorized, while general banners are considered unauthorized.

III. Experimental Results

A. Dataset

To develop and evaluate our system, we constructed two types of banner datasets: one for segmentation model training and the other for end-to-end evaluation.

For segmentation model training, we collected 250 banner images from the Web manually annotating 308 banner instances. This dataset was split into 160 images for train, 40 images for validation, and 50 images for test.

To evaluate the full pipeline, we collected 1,031 banner images captured in real-world urban environments. These images contain a total of 1,438 annotated banner instances, each labeled as one of three categories: *political*, *general*, or *public*. Unlike curated Web data, this dataset captures real-world complexities, such as occlusions, low resolution, and varying lighting conditions.

While our ultimate goal is **to detect unauthorized banners**, their authorization status is not always visually discernible. authorization status often depends on contextual factors such as installation permits, location compliance, or administrative approval—all of which are unavailable in image data alone. Therefore, we adopt a *heuristic label assignment* strategy: banners labeled as *political* or *public* are treated as authorized, whereas those labeled as *general* are treated as unauthorized for the purpose of model training and evaluation. This approach serves as a practical surrogate for banner authorization within our vision-based framework.

B. Quantitative Evaluation

We first evaluated the instance segmentation performance of the proposed pipeline. The model achieved high localization accuracy, with a Mask Average Precision (AP) of 0.945 and a Box AP of 0.925 at an IoU threshold of 0.5. These results confirm that the segmentation module reliably detects and accurately delineates banner regions, providing a strong foundation for subsequent recognition and classification tasks.

Next, we quantitatively evaluated the multi-class classification performance across three categories *political*, *general*, and *public*. Fig. 6 shows the confusion matrix, which visually summarizes classification results and misclassifications among classes. Notably, most classification errors occur between the *general* and *public* classes, suggesting some overlap in visual features or textual content that introduces ambiguity.

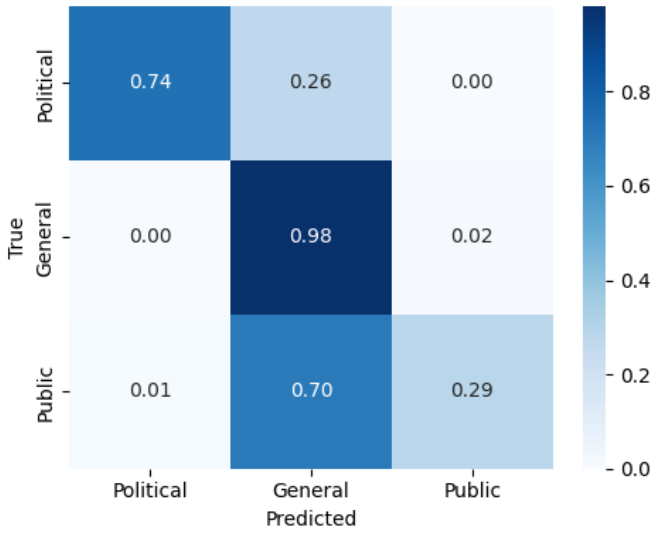

Fig. 6. Confusion matrix for multi-class classification.

TABLE I
Evaluation Metrics for Multi-class Classification

Class	Precision	Recall	F1-score	Sample Num
Political	0.918	0.735	0.816	155
General	0.685	0.980	0.804	797
Public	0.907	0.286	0.429	412
Accuracy		**0.743**		1364

Table I provides detailed metrics such as precision, recall, and F1-score for each class. The overall classification accuracy reaches 0.743, demonstrating the model's solid performance in distinguishing between banner categories. The *general* class achieves the highest recall of 0.980, indicating that most of the *general* banners are correctly identified. In contrast, the *public* class has a lower recall of 0.286, likely due to its visual

similarity to the *general* class, which causes frequent misclassifications. The *political* class exhibits a balanced performance with high precision (0.918) and reasonable recall (0.735), highlighting effective identification of *political* banners.

Taken together, these results demonstrate that the proposed pipeline robustly localizes and classifies banners under real-world conditions. The combination of the confusion matrix and quantitative metrics offers a comprehensive view of the model's strengths and areas for improvement, particularly in recognizing visually similar categories.

C. Binary Classification Performance

As described in the dataset section, the ultimate goal of our system is to detect unauthorized banners. However, since the authorization status of a banner cannot be determined solely from visual information—requiring contextual factors such as installation permits or administrative approval—we employ a heuristic label assignment for binary classification. Specifically, banners annotated as *political* or *public* are grouped into a single *authorized* class, while those labeled as *general* are considered *unauthorized*.

Under this scheme, the model achieves an overall accuracy of **0.743**, demonstrating effective discrimination between authorized and unauthorized banners in most cases. Notably, the classifier attains a high recall of **0.980** for the *unauthorized* class, which is critical in enforcement scenarios to minimize missed detections (i.e., false negatives).

The precision for unauthorized banners is **0.703**, indicating that while most unauthorized banners are correctly identified, some authorized banners are occasionally misclassified as unauthorized (i.e., false positives). The resulting F1-score of **0.820** reflects a balanced trade-off between precision and recall.

This high recall ensures reliable flagging of potential unauthorized banners, reducing the chance of overlooking violations. To further reduce false alarms and unnecessary interventions, future work may focus on enhancing precision through improved feature representations or integration of additional contextual cues.

To further validate the effectiveness of our method, we compared it with Llama 3.2-11B [11], [12], which is one of the most advanced open-weight Vision-Language Models (VLMs). The model was evaluated in a zero-shot setting, where each banner image was accompanied with a prompt such as:

> "Given an image of a banner, classify it into one of the following categories based on its content and visual elements: Political Party, Public, or General. Let's think step by step."

TABLE II
PERFORMANCE FOR UNAUTHORIZED BANNER CLASSIFICATION

Metric	Accuracy	Precision	Recall	F1-score
Our Method	**0.743**	**0.703**	**0.980**	**0.820**
Llama 3.2-11B [12]	0.496	0.696	0.273	0.393

Table II summarizes the performance comparison between our method and the VLM. Our method consistently outperformed the VLM across all metrics.

IV. CONCLUSION

We proposed an automated system for detecting unauthorized banners by integrating state-of-the-art instance segmentation, tracking, OCR, and large language models. The system demonstrated robust performance across both image and video inputs, achieving a segmentation Mask AP of 0.945 and 74.3% classification accuracy across three classes. In binary classification, the system achieved a high recall of 98.0% for unauthorized banners, ensuring minimal false negatives. These results highlight the viability of combining vision-language models in a modular framework for real-world enforcement tasks. Future work will focus on improving precision, supporting multilingual input, and enabling real-time deployment in urban environments.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00255968) grant funded by the Korea government(MSIT). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-2020-0-01461) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

REFERENCES

[1] G. Jocher, A. Chaurasia, Laughing, J. Fang, Y. Patel, and V. Dibia, "Ultralytics YOLOv8," https://github.com/ultralytics/ultralytics, 2023, accessed: 2025-07-01.

[2] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," 2022.

[3] S.-X. Zhang, C. Yang, X. Zhu, and X.-C. Yin, "Arbitrary shape text detection via boundary transformer," *IEEE Transactions on Multimedia*, vol. 26, pp. 1747–1760, 2024.

[4] T. Zheng, Z. Chen, S. Fang, H. Xie, and Y.-G. Jiang, "Cdistnet: Perceiving multi-domain character distance for robust text recognition," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 300–318, 2024.

[5] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[6] G. Kim, T. Hong, M. Yim, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "Donut: Document understanding transformer without ocr," *arXiv preprint arXiv:2111.15664*, vol. 7, no. 15, p. 2, 2021.

[7] H. Wei, L. Kong, J. Chen, L. Zhao, Z. Ge, J. Yang, J. Sun, C. Han, and X. Zhang, "Vary: Scaling up the vision vocabulary for large vision-language models," *arXiv preprint arXiv:2312.06109*, 2023.

[8] H. Wei, C. Liu, J. Chen, J. Wang, L. Kong, Y. Xu, Z. Ge, L. Zhao, J. Sun, Y. Peng *et al.*, "General ocr theory: Towards ocr-2.0 via a unified end-to-end model," *arXiv preprint arXiv:2409.01704*, 2024.

[9] K. Tan, Y. Zhou, Q. Xia, R. Liu, and Y. Chen, "Large model based sequential keyframe extraction for video summarization," in *Proceedings of the International Conference on Computing, Machine Learning and Data Science*, 2024, pp. 1–5.

[10] H. Tang, L. Ding, S. Wu, B. Ren, N. Sebe, and P. Rota, "Deep unsupervised key frame extraction for efficient video classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3, pp. 1–17, 2023.

[11] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, and A. Fan, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[12] Meta Platforms, Inc., "Meta Llama-3.2-11B-Vision-Instruct," Hugging Face model card, Sep. 2024, released September 25, 2024; instruction-tuned multimodal model (text + image) with 11 billion parameters under the Llama 3.2 Community License. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct