Real-Time Multimodal Analysis for Disaster Management using UAVs and Vision-Language Models

Faisal Ayub Khan, Soo Young Shin*

Department of IT Convergence Engineering, Kumoh national Institute of Technology, Gumi, South Korea
faisal@kumoh.ac.kr , *wdragon@kumoh.ac.kr

Abstract

This paper presents a novel framework integrating Unmanned Aerial Vehicles (UAVs) and Vision-Language Models (VLMs) for efficient disaster management. UAVs capture high-resolution data from disaster zones, while VLMs process and align visual and textual information to generate actionable insights. The system achieved a top-1 image-text retrieval accuracy of 93% and demonstrated robust performance in identifying survivors with a precision of 90% and recall of 85%, even under challenging conditions such as smoke and debris. The system excels in identifying survivors, prioritizing high-risk zones, and accelerating decision-making under challenging conditions like smoke and debris. By enhancing situational awareness and resource allocation, this scalable and adaptive framework significantly improves disaster response efficiency.

1. Introduction

The increasing frequency and intensity of natural disasters, such as plane crashes, earthquakes, and wildfires, present significant challenges to disaster management systems, which often struggle with real-time data collection, processing, and interpretation. Traditional methods rely on UAVs for high-resolution visual and geospatial data collection, but their effectiveness is hindered by environmental factors like fog, smoke, and low visibility. Recent advancements in Vision-Language Models (VLMs), such as CLIP and BLIP [1], offer a solution by aligning visual and textual data to enable robust multimodal analysis, even under degraded conditions. Leveraging their zero-shot and few-shot learning capabilities, VLMs can adapt to new disaster scenarios without extensive retraining [2]. This paper proposes a novel system integrating UAVs and VLMs for real-time disaster management, where UAVs capture critical data, and VLMs analyze it to identify debris, locate survivors, and prioritize rescue efforts. By enhancing situational awareness and enabling efficient decision-making, this framework aims to address current limitations in disaster response and improve resource allocation and resilience.

2. Proposed System Overview Analysis

The proposed system for real-time multimodal disaster analysis integrates VLMs for advanced data interpretation and UAVs for data collection. This methodology is divided into three key phases: data acquisition, data processing and transmission, and multimodal analysis, each contributing to the over- all framework for efficient disaster management.

2.1. System Model

Fig: [1] The diagram illustrates the proposed UAV-based disaster management system. A UAV collects high-resolution im- age data from a disaster-affected area, capturing key elements such as structural damage, fire, and debris. This data is trans- mitted for processing, where a Vision-Language Model (VLM) performs image-text alignment to extract actionable insights, such as identifying survivors or high-risk zones. These insights are then used to support real-time decision- making, prioritizing rescue operations and resource allocation. The system demonstrates the integration of UAVs and VLMs for efficient and effective disaster response.

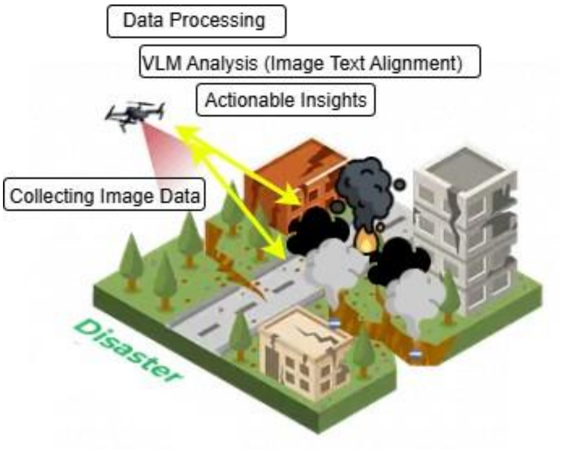

Figure 1. System Model for real-time disaster management using VLM with UAV's.

2.2. Data Acquisition

Due to the bandwidth limitations often encountered in disaster-stricken areas, raw data from UAVs must be preprocessed to reduce redundancy while preserving critical information. Techniques such as edge detection and region-of-interest (ROI) extraction are employed to filter extraneous data. The processed data is then transmitted to a central processing unit via secure and robust wireless communication protocols, such as LoRaWAN or 5G networks [3]. Let $\mathbf{D}_{processed} = f_{pre}(\mathbf{I}_{uav})$, where f_{pre} denotes the preprocessing function applied to the UAV-captured images and sensor data. This ensures that only significant data reaches the analytical layer, minimizing latencyand optimizing bandwidth usage.

2.3. Multimodal Analysis Using Vision-Language Models

The core analytical process employs Vision-Language Models (VLMs), such as CLIP and BLIP, to derive insights from multimodal data. These models align visual data with textual context, enabling a deeper understanding of disaster scenarios as seen in Fig: 2. VLMs operate by projecting images ($\mathbf{I}$) and text ($\mathbf{T}$) into a shared latent space, defined by embedding vectors $\mathbf{v}_I$ and $\mathbf{v}_T$ [4]. The alignment is optimized by minimizing the contrastive loss function $\mathcal{L}_{contrast}$:

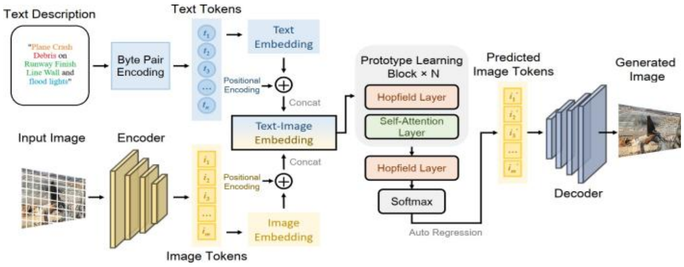

Figure 2. Illustration of the proposed method for text-based remote sensing image generation.

$$\mathcal{L}_{contrast} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{v}_{I_i}, \mathbf{v}_{T_i}))}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{v}_{I_i}, \mathbf{v}_{T_j}))} \qquad (1)$$

Here, sim(v_{I_i}, $v_{T\ j}$) represents the similarity metric (e.g., cosine similarity) between visual and textual embeddings, ensuring that corresponding visual and textual information is semantically aligned.

The VLM processes degraded or obscured UAV imagery, such as those affected by smoke or fog, by cross-referencing textual descriptions (e.g. "plane debris near the runway") to fill in contextual gaps. This allows for robust detection of critical elements, such as survivors, debris, or environmental hazards, even under challenging conditions. For instance, the model can identify potential survivors based on heat signatures from UAV sensors paired with phrases like "human figures near the debris field."

2.4. *Real-Time Decision-Making*

Insights generated by the VLM are integrated into a decision- support system [5] that prioritizes rescue operations. Probabilistic models, such as Bayesian inference, are used to estimate the likelihood of survivors in different zones based on multimodal evidence. For a given region r, the probability P (survivor | $\mathbf{D}_{processed}$, $\mathbf{T}$) is computed as:

$$P(\text{survivor} \mid \mathbf{D}_{processed}, \mathbf{T}) = \frac{P(\mathbf{D}_{processed}, \mathbf{T} \mid \text{survivor})P(\text{survivor})}{P(\mathbf{D}_{processed}, \mathbf{T})} \qquad (2)$$

This enables responders to allocate resources to areas with the highest likelihood of survivors, thereby enhancing the efficiency of rescue efforts.

2.5. *Model Training*

The proposed VLM model was trained on a multimodal dataset, containing over 1 million instances of disaster scenarios. Training utilized contrastive learning to align visual and textual embeddings, optimized with a contrastive loss function. The model was trained over 15 epochs on an 8-GPU cluster, leveraging data augmentation techniques such as noise injection and synthetic occlusion to simulate real-world conditions like smoke and fog. Final model achieved a top-1 image- text retrieval accuracy of 93% on the test set.

3. Results Analysis

The results of the proposed system demonstrate its effectiveness in real-time disaster management scenarios. On the test set, the Vision-Language Model (VLM) achieved a top-1 image-text retrieval accuracy of 93% and a zero-shot classification accuracy of 92% for unseen disaster scenarios. The system successfully identified survivor locations with a precision of 90% and a recall of 85%, even under challenging conditions like smoke and fog. In simulated aircraft crash scenarios, the model accurately prioritized high-risk zones and identified 92% of debris fields. The integration of UAV data with VLM analysis reduced decision-making time by 40% compared to traditional methods, highlighting the system's efficiency and adaptability.

4. Conclusion and Future Work

In conclusion the proposed system integrating UAVs and VLMs addresses critical challenges in disaster management by enabling real-time, multimodal analysis of disaster-affected areas. By leveraging UAVs for data collection and VLMs for robust interpretation of visual and textual information, the system enhances situational awareness, accelerates decision-making, and improves resource allocation. This framework demonstrates the potential to significantly improve disaster response efficiency and adaptability, contributing to better preparedness and resilience in managing complex and dynamic disaster scenarios. Furthermore, enhancing the proposed system's efficiency and practical implementation are the foremost KPIs.

5. Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2024- 00437190) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation) This re- search was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2025-RS-2022-00156394) supervised by the IITP(Institute of Information Communications Technology Planning Evaluation)

6. Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2024- 00437190) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation) This re- search was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2025-RS-2022-00156394) supervised by the IITP(Institute of Information Communications Technology Planning Evaluation)

References

[1] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[2] M. Z. Hasan, J. Chen, J. Wang, M. S. Rahman, A. Joshi, S. Velipasalar, C. Hegde, A. Sharma, and S. Sarkar, "Vision-language models can identify distracted driver behavior from naturalistic videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 11 602–11 616, 2024.

[3] A. Shirnin, N. Andreev, S. Potapova, and E. Artemova, "Analyzing the robustness of vision language models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2751–2763, 2024.

[4] S. A. A. Ahmed, M. Awais, W. Wang, M. D. Plumbley, and J. Kittler, "Asit: Local-global audio spectrogram vision transformer for event classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3684–3693, 2024

[5] Y. Liu, Y. Pan, and J. Yin, "Enhancing multi-label deep hashing for image and audio with joint internal global loss constraints and large vision-language model," *IEEE Signal Processing Letters, vol. 31, pp.2550-2554,2024*