# Estimation of Human Condition at Disaster Site Using Aerial Drone Images

Tomoki Arai[1,2] Kenji Iwata[1] Kensho Hara[1] Yutaka Satoh[1,2]
[1]Information Technology and Human Factors,
National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8565 Japan
[2]Graduate School of Science and Technology, University of Tsukuba
1-1-1 Tenoudai, Tsukuba, Ibaraki 305-8573 Japan
{t.arai, kenji.iwata, kensho.hara, yu.satou}@aist.go.jp

## Abstract

*Drones are being used to assess the situation in various disasters. In this study, we investigate a method to automatically estimate the damage status of people based on their actions in aerial drone images in order to understand disaster sites faster and save labor. We constructed a new dataset of aerial images of human actions in a hypothetical disaster that occurred in an urban area, and classified the human damage status using 3D ResNet. The results showed that the status with characteristic human actions could be classified with a recall rate of more than 80%, while other statuses with similar human actions could only be classified with a recall rate of about 50%. In addition, a cloud-based VR presentation application suggested the effectiveness of using drones to understand the disaster site and estimate the human condition.*

## 1. Introduction

When natural or man-made disasters occur, it is important to quickly assess the situation at the disaster site. Drones are useful for understanding disaster sites because they can quickly search large areas, gather information from a bird's eye view, and fly over areas too dangerous for humans to enter. Various studies have been conducted around the world to use drones to understand disaster sites, including a comparison of flood damage by Islam *et al*. [6], assessment of building collapse by Jiménez-Jiménez *et al*. [7], estimation of wildfire damage maps by Tran *et al*. [12], and investigating infrastructure damage from hurricanes by Schaefer *et al*. [11]. In Japan, aerial images using drones are being taken to understand actual disaster sites, and the number is increasing year by year [4].

In addition to understanding the damage to infrastructure and buildings, quickly understanding disaster sites is also important to saving lives. Assessing the condition of victims and rescuers helps determine the order of priority for rescue operations and select appropriate rescue methods. However, observers need to pay close attention to the people they identify in aerial images taken from an overhead view, because people may appear smaller than buildings and other objects, or their orientation, posture, and clothing may change the way they appear.

Therefore, this study investigates an approach to automatically estimate human disaster conditions to support rapid understanding of disaster sites using drones. To this end, a dataset of aerial videos of human actions was created based on a hypothetical disaster in an urban area.

## 2. Related work

### 2.1. Use of drone in search and rescue

Recently, the use of drones for search and rescue of disaster victims and people in distress has been explored. Qi *et al*. reported on a method that uses a multispectral camera and bio-radar to automatically identify the injured and camouflaged fakers [10]. Al-Naji *et al*. is working on detecting vital signs by detecting chest movements of lying survivors from images [1]. Ono *et al*. developed fallen persons detection and person shadow detection method using rotation-invariant features that take into account shadows and trees in the background region of persons [9].

These studies were conducted in situations where the subjects were unconscious, collapsed, and immobile. When focusing only on rescue, it is highly urgent to know the location of people who have collapsed and are immobilized, and to rescue them quickly. On the other hand, from the perspective of understanding disaster sites, it is important to understand not only those who have fallen, but also those who are evacuating to safer locations and those who are calling for help.

### 2.2. Drone aerial video dataset of human actions

The situation of the victims is related to their actions. If the injuries are minor, the victims can escape on their own. However, if the victims are seriously injured and have difficulty walking, they cannot escape on their own. Even if the injury is minor, if the victims are left at the scene, they can call for help by waving their hands or other means. Therefore, it is useful to recognize the actions of the victims in order to judging the human condition.

Previous works on drone aerial video datasets of human actions include Okutama-Action by Barekatain *et al*. [2] and UAV-Human by Li*et al*. [8]. These datasets primarily record basic human behaviors such as walking and running, as well as human interactions with each other and with objects, and do not consider the background scenario in which the actions occur. However, the actions of disaster victims can be considered in the context of the disaster site. For example, the action of ”running” may have different purposes, such as ”running to save victims” or ”running away from a fire. Therefore, it is difficult to accurately grasp the human condition simply by systematically recognizing and classifying only human actions. Therefore, in order to understand the human condition at a disaster site, it is necessary to consider actions in the background scenario of the disaster.

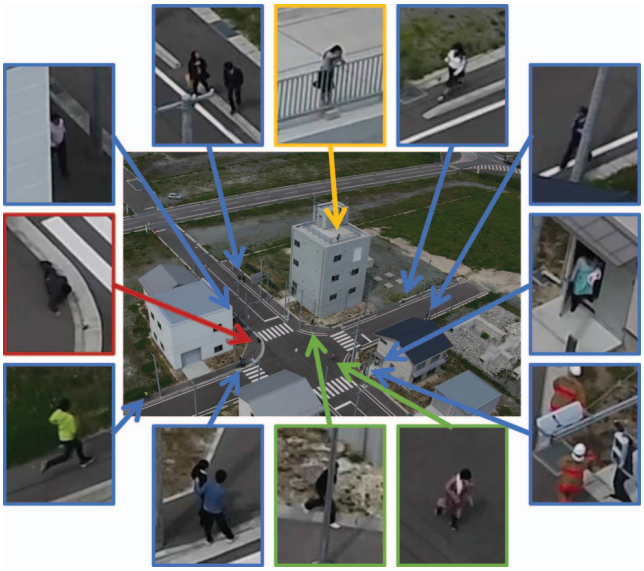

Figure 1. One scene image of the video included in the data set. The drone is turning around the scene at an altitude of 50m and taking video. The blue frame shows people in a safe situation, the green frame shows people evacuating, the yellow frame shows people calling for help, and the red frame shows people in a dangerous situation.

## 3. Proposed Dataset

In this study, a new dataset was constructed by capturing a series of human actions in a simulated disaster environment with a drone. The dataset consists of videos of the actions of disaster victims, rescuers, passersby, and onlookers in various disaster scenarios. We captured 62 videos under different disaster conditions, drone altitudes, and drone paths. In 45 of those videos, 20 extras played the actions of people expected in a disaster. All of the extras were dressed appropriately for their roles.

The disaster scenario for each video is shown in Table 1, and the shooting conditions in Table 2. An example dataset is shown in Figure 1.

StrongSORT with OSNet for Yolov5 [3] was used for each video, and bounding box coordinates and person tracking IDs were automatically generated for persons in each frame. Errors in the bounding box coordinates and person IDs were then manually corrected using a hand-crafted annotation tool, and a damage status label was assigned to each person.

## 4. Experimental setup

In this study, four damage statuses, ”safe,” ”evacuation,” ”call for help,” and ”emergency,” were defined as follow in order to clarify the priority order of rescue targets, with reference to triage in emergency medical care.

Safe: The person is safe and is not in any of the following three statuses. Evacuation: The person is evacuated from the dangerous area. Call for help: The person is calling for help. Emergency: The person cannot move on his/her own.

These damage statuses strongly influence human actions, but they do not always coincide. Therefore, this study attempts to directly estimate human condition labels using RGB images containing time-series changes caused by human actions.

In human action recognition, 3D CNNs are frequently used to convolve 2D spatial and 1D temporal information. Considering the relationship between human condition and human action, this study also uses a 3D CNN for damage status classification. In this experiment, 3D ResNet-50 by Hara *et al*. [5] was used as the 3D CNN.

The input clips were generated as follows. Each clip consists of 16 frames of one person. Each clip is cropped at square coordinates so that the person is in the center of the 9th frame. The other 15 frames are cropped at the same position and the annotation for the ninth frame is applied in the same way to the other frames in the whole clip.

Each of the four classes had 1000 clips. The number of data for training, validation and testing was 8:1:1. The number of data for training and validation was randomly selected from Patterns A, C, D, and E, and the number of data for testing was randomly selected from Pattern B in order to avoid using clips that closely resemble the training and test data.

| Pattern Name | Disaster scenario |
|---|---|
| A | A fire has broken out in a building and people are stranded on the third floor.<br>On the ground, there are fallen people and onlookers. |
| B | A person is stranded on the roof of a building.<br>On the ground, there are fallen people and onlookers. |
| C | A fire has broken out in a building and many people are evacuated from the building.<br>There are people left on the second floor, third floor, and roof of the building, respectively. |
| D | Many people are on the ground or evacuated after the fire. |
| E | There are people calling for help from different parts of the building. |

Table 1. Major assumed disaster scenarios in each video.

| Terms | Value |
|---|---|
| Location | Urban field<br>at Fukushima Robot Test Field |
| Drone altitude | 10m, 20m, 30m, 50m |
| Drone path | Straight ahead<br>Turning (small, large radius) |
| Video Resolution | 4K (3840x2160px) |
| Frame rate | 30fps |
| Average video duration | 1min 55sec |

Table 2. Shooting conditions of the videos in our dataset.

For training, the images were cropped at random positions and resized to 112x112px. The image was then flipped horizontally at a ratio of 50%. The parameters were set as follows: the loss function was set to cross-entropy loss, the optimisation method was set to Momentum SGD (momentum = 0.9, dampening = 0, weight decay = 0.001), the learning rate was set to 0.1, the batch size was set to 128, and the number of epochs was set to 200. The learning rate was reduced to 1/10 for every 50 epochs. These hyper-parameters were the default values of the 3D ResNet implementation by Hara *et al*. [5].

During verification and testing, only resizing to 112×112px was performed.

For comparison with the case of no time series data, experiments were also carried out using 2D ResNet. The input image was the 9th frame of each clip input to the 3D ResNet and the same parameters were used in the experiment.

## 5. Experimental results and discussion

When assisting in the understanding of a disaster scene, an incorrect assessment may result in people at risk going unnoticed. Therefore, in this study, the recall indicator was used to assess classification performance.

A set of 4000 randomly selected clips was used as one set and the results of the three sets were averaged. The average recall of the test data is shown in Table 3 and the confusion matrix is shown in Figure 2.

The 3D ResNet classification results for each class showed that ”safe” and ”call for help” had recall values of 80% or higher, while ”evacuation” and ”emergency” had recall values of about 50%.

| Class label | 3D [%] | 2D [%] |
|---|---|---|
| Safe | 84.33±3.56 | 84.67±3.77 |
| Evacuation | 50.67±0.41 | 50.00±0.82 |
| Call for help | 93.33±1.47 | 85.33±2.62 |
| Emergency | 46.67±4.32 | 40.00±9.93 |

Table 3. Recall of human damage status classification in test data.

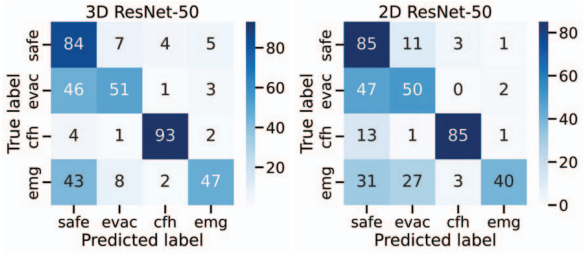


Figure 2. Confusion matrix in 3D/2D ResNet. ”evac” stands for ”evacuation”, ”cfh” for ”call for help”, and ”emg” for ”emergency”.

The reason for this is that ”evacuation” and ”emergency” are often misclassified as ” safe,” as shown in the confusion matrix. The human actions included in ”safe” are diverse. Although the purposes of the actions are different, the class includes common actions with ”evacuation” and ”emergency”, so it is considered difficult to distinguish them. For example, there is the action ”running to rescue” in ”safe” and the action ”running to escape” in ”evacuation”. Also ”safe” includes action such as being onlookers and not moving, while ”emergency” includes action such as being injured and unable to move. On the other hand, ”call for help” has unique actions such as waving and is easy to distinguish.

Also, when compared to the 2D ResNet results, the 3D results show that the recall of the ”safe” and ”evacuation” classes remains the same, while the recall of the ”call for help” and ”emergency” classes has improved. Comparison of the confusion matrices in 3D and 2D showed that there was no significant difference in the classification results for ”safe” and ”evacuation”, indicating that recall had not changed. On the other hand, ”call for help” and ”emergency” with improved recall both showed a reduction in

misinterpretation of the 3D results, especially ”emergency”, with a significant reduction in misinterpretation to ”evacuation”. The reason for this is thought to be that 2D ResNet, which does not learn time- series data, could not distinguish between the states of moving for evacuation and injured and not moving, as both are captured as still images. In 3D ResNet, which learns time-series data, ”emergency” that cannot move on their own are misidentified as unmoving onlookers and classified as ”safe”.

These results suggest that time-series image learning with 3D ResNet strongly captures human actions and does not adequately capture features such as the purpose of the action, which is sufficient for estimating the human condition. Therefore, it is important to add features such as location relative to the disaster area to achieve higher accuracy in disaster situation classification. We consider that this would make it possible to distinguish, for example, whether the purpose of the action of ”running” is to help the victim or to escape from the disaster.

## 6. Real-time verification

A demonstration experiment using the damage status classification method developed in this study was conducted at the Fukushima Robot Test Field. We developed a system that transmits real-time images from an actual drone, performs image recognition in the cloud, and presents the results on the VR space of a client in a remote location. The configuration of the equipment and software is shown in Figure 3. The drone is equipped with a camera (KODAK PIXPR0 4KVR360) capable of 360-degree omnidirectional 4K video, which is transmitted from the Jetson Xavier NX to a video relay server in the cloud via Wifi or 5G communication in 4K H.265 RTSP. The image recognition program that performs damage status classification and smoke/flame detection runs on the cloud. Images are acquired in real time from a video relay server in the cloud and recognition processing is performed. The VR presentation program on the client side is implemented using Unity. The bounding box coordinates and category ID are acquired from the image recognition program via ROS2 communication, superimposed on the omnidirectional image obtained from the image relay server, and presented to the remote operator as a VR video.

An example of the operation is shown in Figure 4. In this example, smoke is emitted from a window on the third floor of a building by a smoke generator, and two people on the roof are waving for help. Bounding boxes make it easy for the operator to see the results of the waving person and smoke detection.

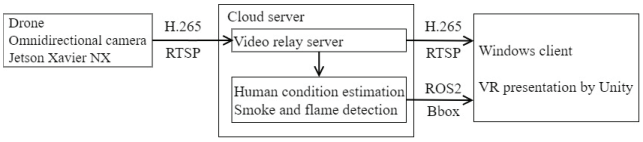


Figure 3. Configuration of real-time damage status classification using the cloud.


Figure 4. Situation presentation by VR.The results detected as ”call for help” are shown as ”SOS” in red BBOXes. The detected smoke is highlighted with a purple BBOX.

## 7. Conclusion and outlook

In this study, a dataset of human actions in a hypothetical disaster taken aerially by a drone was constructed. By learning time-series images using 3D ResNet, we found that the damage status of a person can be classified into ”safe” and ”call for help” with high recall. The VR presentation application also suggested the effectiveness of using drones to assess the situation at disaster sites.

In the future, we plan to improve the classification by using changes in a person’s position as an additional feature. A change in a person’s position is, for example, ”evacuate” means moving away from a fire and ”safe” means moving closer to a fallen person. We believe that by recognizing these changes, the system will be able to learn that the same action has different states. We will also explore changes in performance by searching for hyper-parameters or changing the training video pattern and 3D CNN model.

## Acknowledgments

This paper is based on results obtained from a project, JPNP21004, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

## References

[1] Ali Al-Naji, Asanka G. Perera, Saleem Latteef Mohammed, and Javaan Chahl. Life signs detector using a drone in disaster zones. *Remote Sensing*, 11(20), 2019. 1

[2] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2153–2160, 2017. 2

[3] Mikel Broström. Real-time multi-camera multi-object tracker using yolov5 and strongsort with osnet. https://github.com/mikel-brostrom/Yolov5_StrongSORT_OSNet, 2022. 2

[4] Fire and Ambulance Service Division. [Survey on the use of unmanned aerial vehicles in disasters] Mujin kokuki no saigai ji ni okeru katsuyo jokyo tyosa ni tsuite (in Japanese). In *Syobo no ugoki*, volume 610, pages 14–15. Fire and Disaster Management Agency of the Ministry of Internal Affairs and Communications, 2022. 1

[5] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[6] Md Azharul Islam, Shawkh Ibne Rashid, Niamat Ullah Ibne Hossain, Robert Fleming, and Alexandr Sokolov. An integrated convolutional neural network and sorting algorithm for image classification for efficient flood disaster management. *Decision Analytics Journal*, 7:100225, 2023. 1

[7] Sergio Iván Jiménez-Jiménez, Waldo Ojeda-Bustamante, Ronald Ernesto Ontiveros-Capurata, and Mariana de Jesús Marcial-Pablo. Rapid urban flood damage assessment using high resolution remote sensing data and an object-based approach. *Geomatics, Natural Hazards and Risk*, 11(1):906–927, 2020. 1

[8] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16266–16275, June 2021. 2

[9] Taisei Ono, Haruka Egawa, Yuta Ono, Seiji Ishikawa, and Joo kooi Tan. Detection of fallen persons and person shadows from drone images. *Proceedings of International Conference on Artificial Life & Robotics (ICAROB2023)*, pages 890–894, 2023. 1

[10] Fugui Qi, Mingming Zhu, Zhao Li, Tao Lei, Juanjuan Xia, Linyuan Zhang, Yili Yan, Jianqi Wang, and Guohua Lu. Automatic air-to-ground recognition of outdoor injured human targets based on uav bimodal information: The explore study. *Applied Sciences*, 12(7), 2022. 1

[11] Martin Schaefer, Richard Teeuw, Simon Day, Dimitrios Zekkos, Paul Weber, Toby Meredith, and Cees J. van Westen. Low-cost uav surveys of hurricane damage in dominica: automated processing with co-registration of pre-hurricane imagery for change analysis. *Natural Hazards*, 11(20):755–784, 2020. 1

[12] Dai Quoc Tran, Minsoo Park, Daekyo Jung, and Seunghee Park. Damage-map estimation using uav images and deep learning algorithms for disaster management system. *Remote Sensing*, 12(24), 2020. 1

## Appendix

| Role | Attire | #people |
|---|---|---|
| Male student | Short-sleeved plain clothes | 1 |
| Male student | Long-sleeved plain clothes | 1 |
| Female student | Short-sleeved plain clothes | 1 |
| Female student | Long-sleeved plain clothes | 1 |
| Male office worker | Suit | 3 |
| Female office worker | Suit | 2 |
| Woman with child | Plain clothes | 1 |
| Ordinary man | Loungewear | 2 |
| Ordinary woman | Loungewear | 1 |
| Elderly woman | Long sleeve plain clothes | 1 |
| Male runner | Runner's style | 1 |
| Firefighter | Fireproof clothing | 2 |
| Police officer | Police uniform | 1 |
| Ordinary man | Long sleeve plain clothes | 1 |
| Elderly man | Plain clothes | 1 |

Table 4. List of extras in the videos of the proposed dataset. This table shows the assumed characters and their attire of the extras who played rescuers, victims, etc. in the videos of the proposed dataset.

Figure 5. Examples of "safe" class.

Figure 6. Examples of ”evacuation” class.

Figure 7. Examples of ”call for help” class.

Figure 8. Examples of ”emergency” class.

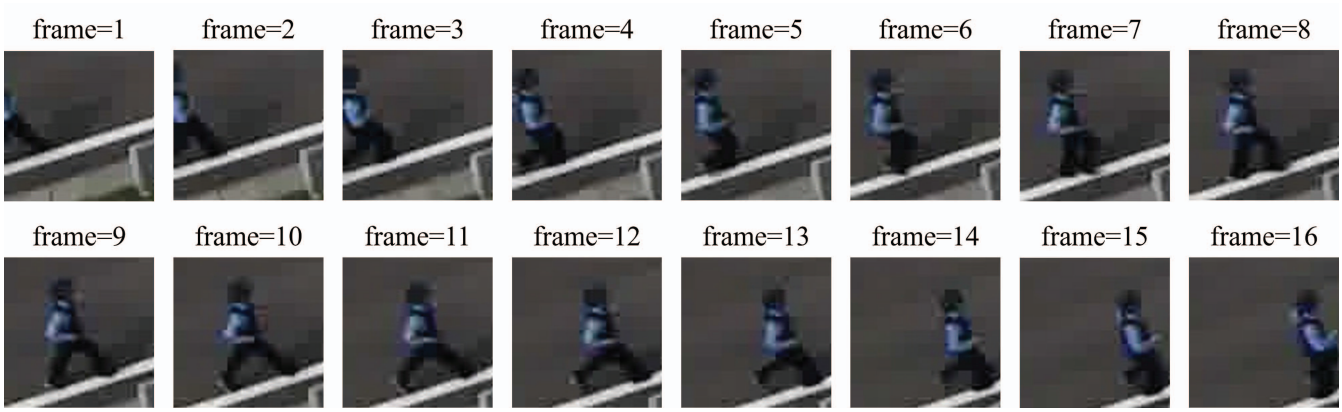


Figure 9. Example of a ”safe” class clip.


Figure 10. Example of a ”evacuation” class clip.


Figure 11. Example of a ”call for help” class clip.

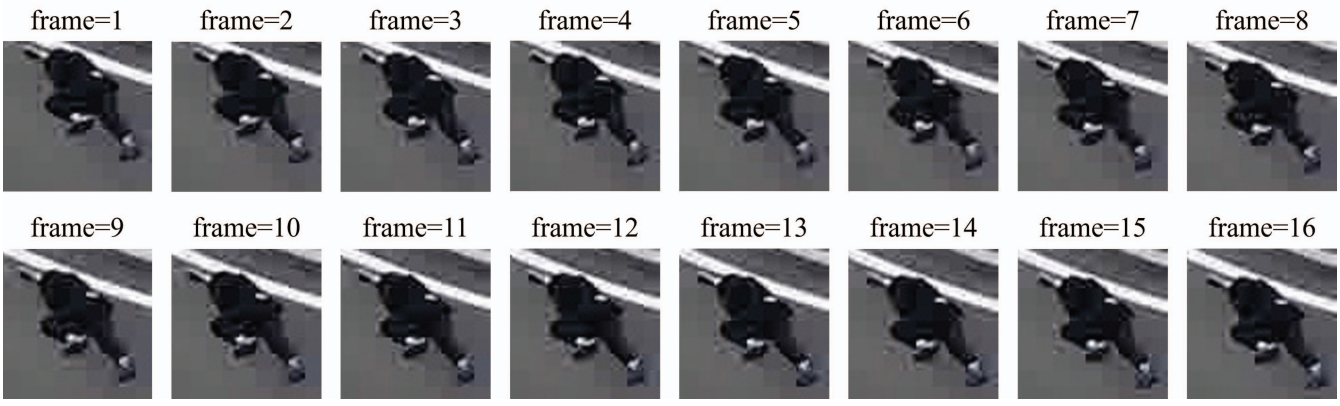


Figure 12. Example of a ”emergency” class clip.