Damaged Building Detection using Multiple Object Tracking and Decision Tree from Aerial Videos for Disaster Response

Shono Fujita
Graduate School of Informatics,
Kyoto University
fujita.shono.32x@st.kyoto-u.ac.jp

Michinori Hatayama
Disaster Prevention Research
Institute, Kyoto University
hatayama@dimsis.dpri.kyoto-u.ac.jp

Abstract

Information about damaged buildings is crucial in disaster response owing to the risk they pose, including property damage and loss of human lives. However, it is difficult to capture this information rapidly during disaster. This study developed an automatic model to detect buildings damaged by earthquakes from aerial videos. It is composed of multiple object tracking model of buildings, classification model of damage, and decision tree model to output final estimation by each track. This system considers; (1) detection of damaged and collapsed buildings such as pancake collapse of wooden buildings and significant roof damage, (2) input of time-series information to determine the extent of building damage, (3) less annotation labor to train datasets, and (4) effective usage of decision tree nodes for disaster response. The obtained results indicated that the average recall of three classes was 47.9%, average precision was 48.4%, and average F-measure was 45.7%.

Keywords: Aerial video, Damaged building, Deep learning, Earthquake, Disaster Response

1. Introduction

In past disasters, such as earthquakes and typhoons, several buildings were damaged, which caused excessive property damage and loss of human lives. In Japan, several wooden buildings have collapsed or been partially damaged by earthquakes. For example, the 2016 Kumamoto earthquake damaged approximately 200,000 buildings (Disaster response headquarters of Kumamoto prefecture, 2022). Accurate and timely information about damaged buildings is crucial for locating and rescuing individuals affected by disasters; estimate the volume of resources for disaster response in cities, prefectures, and nations; estimate the number of people who will live in refugee camps or temporary housing; and distribute support money corresponding to the damage level of each building. Therefore, understanding the extent and nature of damage to buildings is crucial for effective disaster response in affected areas. However, it is difficult to obtain this information. During the 2016 Kumamoto earthquake, because fire brigades could not obtain accurate information about the location of damaged buildings and victims, they investigated all buildings in the area. However, this operation required abundant manpower, support staff from other local governments, and various equipment (Kumamoto city fire department, 2018). In anticipation of future large-scale earthquake disasters that will overwhelm disaster response headquarters and fire brigades, we propose the use of advanced technology to augment these efforts. This study developed an automatic model to detect damaged and collapsed buildings during earthquakes using aerial videos, primarily drone videos, to provide information that is effective for disaster response.

Drone videos can collect extensive information from a bird's eye view at low cost in places that cannot be accessed by people. Currently, 52.9% (383/724) of fire departments in Japan have drones that can capture damage caused by earthquakes, landslides, fires, or toxic gas accidents (Fire and Disaster Management Agency in Japan, 2022). Moreover, disaster response organizations, such as local governments and fire brigades, use helicopters to capture disasters in aerial videos (Nazarov, 2011).

2. Related studies

Our previous study detected roof damage from aerial images using deep learning to ensure a more rapid and efficient investigation of building damage in Japan (Fujita and Hatayama, 2021, 2022). Miura et al. (2020) estimated damaged buildings by detecting damaged roofs and roofs covered with blue sheets from aerial images using deep learning. Calantropio et al. (2021) extracted building regions through deep-learning segmentation and detected damaged buildings. The obtained ortho-aerial images can detect roof damage, which can help assess the extent of building damage in large areas.

URI: https://hdl.handle.net/10125/106638
978-0-9981331-7-1

Table 1. Comparison table with related studies.

	Collapse detection	Considering Japanese feature	Distinction among "no damage," "damage," and "collapsed."	Using time-series information to estimate as a single track
Ours (2021, 2022) Miura et al. (2020) Calantropio et al. (2021)	×	✓ and ×	×	×
Pi et al. (2020) Qi et al. (2016) Zhu et al. (2020)	✓	×	✓ and ×	×
Ours (2023)	✓	✓	×	×
Naito et al. (2021)	✓	✓	✓	×
This study	✓	✓	✓	✓

However, these images cannot detect collapsed buildings without roof damage, where rescuers or injured people are likely located, because they cannot distinguish between the vertical displacements of collapsed buildings. This study used aerial videos, particularly drone videos, to automatically detect collapsed and damaged buildings through deep learning. Aerial videos can be used to detect individual collapsed buildings with vertical displacement and without roof damage from an oblique angle. Moreover, aerial videos are effective during the emergency phase because of their minimal calculation volume and filming labor.

Traditional Japanese buildings often suffer from damaged roofs, particularly tile roofs, during earthquakes and typhoons. Prior to the amendment of the building standards in 1981, the tile roofs of buildings did not have to be fixed to the base of the roof. Currently, the number of tile roofs that sustain damage during disasters is reducing because of revised strict building standards, the use of light raw materials, and a decrease in the number of people using tile roofs. However, several old buildings, or buildings without sufficient countermeasures, may still be affected in the future. Various previous studies (Pi et al., 2020, Qi et al., 2016, Zhu et al., 2020) detected collapsed buildings using the debris and damaged parts as captured in videos. In Japan, even if buildings do not collapse, detection systems that focus on debris react overly to roof damage (Fujita and Hatayama, 2023). Additionally, Japanese buildings, particularly wooden ones, may cause pancake collapse, indicating that the collapsed building is crushed flat during earthquakes (Scawthorn and Yanev, 1995, Okada and Takai, 1999, Scawthorn, 2006). In this collapse pattern, the appearance of the building body, such as the walls and roof, may be intact, except for the crushed story. Because several of the crushed stories cannot be clearly seen, these buildings cannot be detected based solely on debris or damaged parts. Considering the above features of Japanese buildings, this study classified collapsed and damaged buildings using deep learning, regardless of the appearance of debris.

Our previous study (Fujita and Hatayama, 2023) detected collapsed buildings using a multiple-object tracking model of deep learning from aerial videos. The study indicated that the model required the enhancement of accuracy, and it was difficult to distinguish collapsed buildings from roof-damaged buildings. This study developed a model to detect three types of buildings: no damage, damage, and collapse, because mistaking between collapse and no damage classes may cause significant problems in making decisions during disaster response. The inclusion of a damage class can reduce this problem even if a mistake occurs. Moreover, some types of disaster responses require the estimation of damaged buildings. Finally, the model outputs several types of estimations based on these three classes.

Naito et al. (2021) detected damaged and collapsed buildings in oblique aerial images using object detection through deep learning. The study was based on images rather than videos. Although previous studies (Qi et al., 2016, Zhu et al., 2021) detected videos, they did not consider buildings as a single track, which refers to a series of objects in consecutive video frames. The consideration of buildings as one track is necessary to identify the number of damaged buildings in aerial videos. This study considers the buildings on one track in a video using time series information.

Table 1 shows the summarized information of the above. In comparison to previous research, this study takes into consideration the damage characteristics specific to Japan, classifies buildings into three categories: 'no damage,' 'damaged,' and 'collapsed' and outputs estimation results as a single track utilizing time-series information.

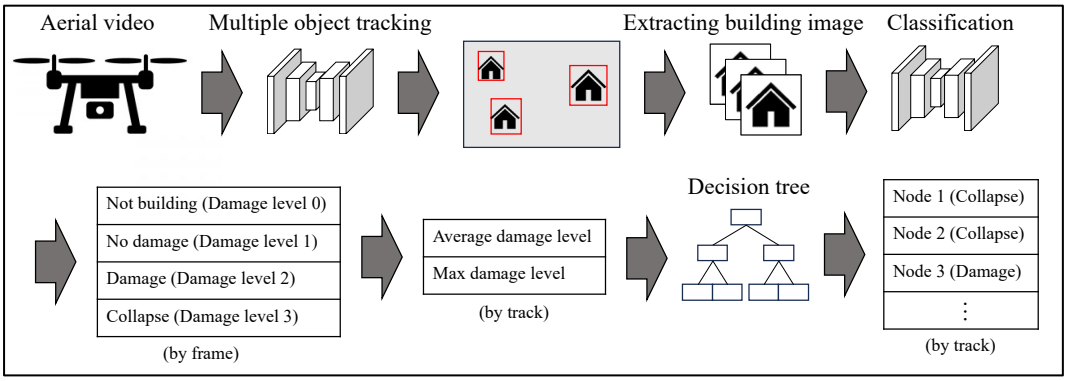

Figure 1. Flow of the proposed model.

3. Proposed system

Figure 1 shows the flow of the proposed system. This model does not operate in real-time; instead, a series of processes is executed after the video is filmed, and then the estimation results are outputted. A multiple object tracking model is used to track the same object using a bounding rectangle (bbox) in the video. Our previous study (Fujita and Hatayama, 2023) revealed that depicting the bounding rectangle of a building in annotating the training data of a video was time-consuming. Annotation means giving correct answers to each set of training data. It is important to decrease the labor and time required to create a dataset when creating machine learning systems during both normal times and disasters. If the machine learning model is trained after a disaster to adapt to new data of a disaster whose quality is different from that of the former disaster, the training dataset must be annotated rapidly. Therefore, the proposed system categorizes the data into three classes after extracting building images from videos because the labor required for annotation in the classification task is less than that in the tracking task. In this experiment, the average annotation time for the classification task was 0.798 s and that for the tracking task was 3.84 s. The proposed system uses fewer data for multiple object tracking and more data for classification tasks. After classifying the buildings in each frame, the system calculates the average and maximum damage levels in each track. We classified the damage level into four categories as follows: "Not building" (0), "No damage" (1), "Damage" (2), and "Collapse" (3). This study uses the average damage level to consider the time-series information for one track.

Finally, a decision tree was created to output several types of estimations. As the estimated information of damaged buildings is used in several cases, the required *accuracy*, *recall,* and *precision* vary for each case. These values were determined using Equations (1)–(3) and Table 2. For example,

- For rescue operations, it is desirable to increase the *recall* of collapsed buildings. Their aim is to locate and assist individuals in as many collapsed buildings as possible, even if many no collapsed buildings are also detected by mistake.
- It is desirable for cities, prefectures, and nations to estimate the volume of necessary resources to increase the *accuracy* of determining collapsed and damaged buildings and understand the scale and severity of the damage. They need an approximate figure of damaged and collapsed buildings without excessive overlook and mistaken detection.
- In Japan, a system determines the content of each type of support, such as a summary of the support money for victims corresponding to the building damage level (Disaster Management, Cabinet Office in Japan, 2020). Because investigating the damage level on-site is time-consuming, local governments currently investigate some buildings based on aerial images or photos taken by the victims at first and then other buildings on-site. Moreover, automatic judgement by artificial intelligence such as our system without human judgement would further accelerate the first photo judgement. In this case, they can grasp damage in later on-site investigations, even if judging from photos and artificial intelligence overlooks damage. Therefore, it is desirable to increase the *precision* of collapsed and damaged buildings because it is necessary to decrease mistaken detection (FP). If there are many mistaken detections, it will lead to excessive support for many no damage buildings.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Table 2. Confusion matrix.

		Estimation	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Considering the diverse applications in disaster response, this study uses a decision tree to output several types of results, including accuracy, recall, and precision, which are essential for addressing various requirements for effective disaster response. A decision tree is a model used to classify or predict values using conditional branches in multiple layers. Because this model outputs some nodes with different estimation probabilities, these outputs are expected to be effective for various types of disaster responses. Moreover, decision trees make it easier to interpret internal processes than other machine learning methods, such as neural networks. Therefore, this model can be effective in providing explanations for its answers and building trust among people during a disaster response. It is also this study's unique point to select a decision tree from some types of machine learning algorithm considering the utilization in disaster response aspect such as several outputs and interpretability.

Once damaged buildings have been identified from aerial videos, it is necessary for the effective disaster response to associate these buildings with accurate location information. Some previous studies (Qi et al., 2016) located a building in a video by determining parameters such as the height and location of the drone and camera angle. If these parameters cannot be obtained, location information can be obtained from three-dimensional (3D) point data and points with location information inputted by humans. 3D point data can be obtained from aerial images or aerial videos using structure from motion, as reported in previous studies (Yamazaki et al., 2017, Arko et al., 2014). The former can automatically estimate the position from various parameters, while the latter requires manual input of position information with four or more points by human. Therefore, this study did not include associated location information.

4. Experiment

4.1. Dataset

This study used aerial videos recorded by drones in Mashiki city, Kumamoto prefecture, following the 2016 Kumamoto earthquake. The 2016 Kumamoto earthquake occurred on April 14 and 16, 2016, with a maximum seismic intensity of 7 and moment magnitude scale of 7.0. The number of deaths reported was 273 and 198,258 buildings were damaged (Disaster response headquarters of Kumamoto prefecture in Japan, 2022), particularly wooden buildings. The frame rate of these videos is 30 fps, and the average altitude of the drone was about 60 ~ 75 m.

Figure 2 illustrates a breakdown of the dataset. We separated the frames of the aerial videos into three regions to create training, validation, and test datasets. The training data is used to update the parameters of the model, the validation data is used to determine the number of trainings, and the test data is used to evaluate the model. In the multiple object tracking model (4.2), the system uses part of the data of the three regions, considering the large labor required for tracking annotation. In creating the classification model (4.3), the system uses building images that have been cropped by the tracking model from all the data of the three regions. Then, average and maximum damage levels are calculated using output of the classification, and used to create the decision tree (4.4) in the proposed system. The decision tree uses all the data obtained from regions B and C.

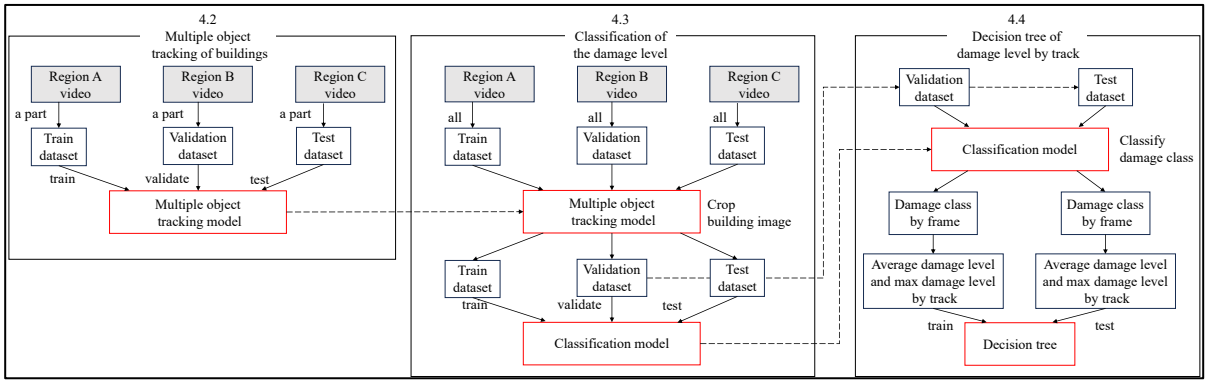

Figure 2. Breakdown of the dataset.

4.2. Multiple object tracking of buildings

Table 3. Number of dataset in tracking.

	Train	Validation	Test
Image	1,762	481	2,101
Bbox	118,948	93,268	131,886
Track	293	347	237

Table 4. Result under different conditions

		Condition A	Condition B
recall	0.620	0.806	0.950
precision	0.778	0.778	0.851

Figure 3. Captured image of the tracking annotation.

Figure 4. Example of tracking (blue: correct, red: estimation, green: estimation under condition A)

This study trained Bytetrack (Zhang et al., 2021) as a multiple object tracking model using the aforementioned dataset. Table 3 summarizes the number of images, bbox (bounding rectangle in each frame), and tracks of the dataset, and Figure 3 shows a captured image illustrating the annotation tool in use alongside the accompanying video. We annotated one image out of every ten frames and compensated for the remaining nine frames using linear interpolation. This annotation took approximately 3.84 s per bbox. The backbone for object detection was YOLOX-m (Ge et al., 2021), and fine-tuning was executed using a pre-trained backbone model with the COCO dataset, which is large scale dataset for image recognition. The input image size was 800 × 1440. In the test, we used a trained model of 43 epochs, because the accuracy for the validation data was highest.

Based on the test data, the recall by track was 62.0%, whereas its precision was 77.8% as shown in table 4. If one or more bboxes were detected correctly in each track, we considered the track as a correct estimation (TP). The threshold of Intersection over Union (IoU) between the correct and estimated bboxes, which measure the overlapping ratio, was 0.5. From the visualized estimation results, we observed that some of the estimated bboxes surrounded several correct buildings, as shown in Figure 4 left. We classified the data surrounding several correct buildings as correct data if the IoU exceeded 0.5 as shown in Figure 4 right, which is defined as condition A. Consequently, the recall by track increased to 80.6%, while precision by track remained at 77.8%. Under this condition A, the recall of no damage by track was 78.8%, the recall of damage by track was 86.5%, and the recall of collapse by track was 80.0%. This paper uses these values as recall of building tracking from now on. In addition to condition A, when restricted to data with a bbox size of 10,000 pixels or more and a bbox aspect ratio (height/width) of 2 or more, recall by track increased to 95.0% and precision increased to 85.1%, which is defined as condition B. The obtained results indicated that employing an effective filming method can significantly increase the accuracy of building detection. However, the number of ID switch, which changes the target object to other object in the middle of the track, was 414. This ID switch may have occurred especially in small buildings when the estimation bbox became large, included buildings next to the target building, and moved to the buildings. This number of ID switch was large comparing with 237, the number of tracks. Because this ID switch may cause an incorrect estimation of each track, the number must be reduced in the future.

4.3. Classification of the damage level

After tracking the buildings from the aerial video, we extracted the building part using a bbox in the frame out of every ten frames and classified them into four classes, as shown in Figure 5. In this study, the term “collapse” refers to buildings that have experienced severe structural damage resulting in the collapse of one or more stories. These buildings are of particular interest because they are potential locations where rescuers or injured people may be located. If several buildings appeared in a single image, a more serious damage state was assigned for the data. This annotation took 0.798 s per image using the Finder application of the Mac PC. Table 5 summarizes the number of images in the dataset used for classification. This study used ResNet50 (He et al., 2015) as the classification model, with an input size of 512 × 512. During training, we set the inverse of the composition of each class of data as the class weight of the loss function to address the imbalanced dataset. In the test, we used a trained model of 14 epochs, because the accuracy for the validation data was highest. The test

data exhibited an accuracy of 64.4%, average recall of 53.0%, and average F-measure of 50.4%. Table 6 summarizes the confusion matrix used in this experiment. The F-measure is the harmonic mean of accuracy and recall. These low values and training result of Figure 6 indicated that richer and more various training dataset is necessary to increase accuracy.

Table 5. Number of images in the dataset used for classification.

	Train	Validation	Test
Not building	724	289	1,837
No damage	40,815	17,070	12,459
Damage	17,329	13,938	6,208
Collapse	4,242	1,817	1,126

Table 6. Confusion matrix of classification.

		Estimated			
		Not building	No damage	damage	collapse
Actual	Not building	587	1,014	188	48
	No damage	1,189	8,901	1815	554
	Damage	171	1,255	3,943	839
	Collapse	20	296	303	507

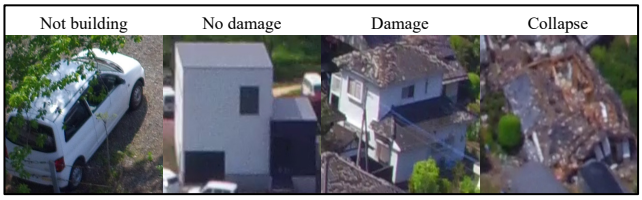

Figure 5. Example of four classes.

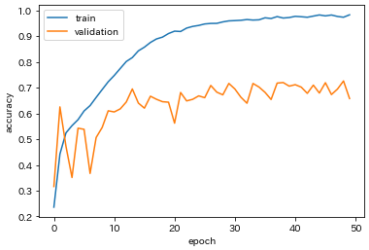

Figure 6. Accuracy of train and validation data

4.4. Decision tree of damage level by track

After classifying the building damage in each frame, the damage levels were categorized into four: "Not building" (0), "No damage" (1), "Damage" (2), and "Collapse" (3). In this study, the maximum damage level for each track was considered as the correct damage level for the track. This is because more severe damage can sometimes appear from different angles, even if these buildings are the same. In this section, we describe the development of a decision tree to estimate the damage level of each track. The input of the decision tree uses the maximum damage level for the same reason as the above correct damage level for each track and the average damage level to use time-series information between different frames like majority rule. Figure 7 illustrates an example of this calculation method being applied, while table 7 lists the number of data points. In this training, we set the inverse of the composition of each class of data as the class weight of the loss function to address the imbalanced dataset. Moreover, we set 4 as the maximum depth of the decision tree and 16 as the maximum number of leaf nodes.

Table 8 lists the confusion matrix, while Figure 8 illustrates the decision tree constructed using the training data. Considering the recall of multiple object tracking of buildings reported in Section 4.2, the final average recall was 47.9% and the final average precision was 48.4%. The final average F-measure was 45.7%. Figure 9 illustrates an example of visualizing this estimation in a video.

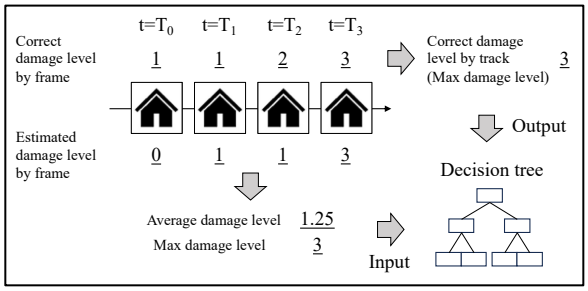

Figure 7. Rule of input and output from the damage level.

Table 7. Number of dataset tracks in the decision tree.

	train	test
Not building	49	168
No damage	999	791
Damage	1,153	482
Collapse	125	80

4.5. Discussion

4.5.1. Comparison between models with two and three and classes. In our previous study (Fujita and Hatayama, 2023), which detected only collapsed buildings, the recall of collapse by track was 36.5%, precision was 35.5%, and the F-measure was 36.0%. Considering the 80.0% recall of the collapse of tracking buildings obtained in Section 4.2, the final recall of collapse by track of the proposed model was 46.0%, precision was 24.2%, and the F-measure was 31.7%. This reduction of F-measure indicates that adding "damage" class to "no damage" and "collapse" does not necessarily increase accuracy.

Table 8. Confusion matrix of the decision tree.

		Estimated						
		Not building	No damage	damage	collapse	Total / Average	Recall	Recall + Recall of building Tracking
Actual	Not building	79	45	41	3	168	0.470	0.470
	No damage	131	426	191	43	791	0.539	0.425
	Damage	10	61	313	98	482	0.649	0.561
	Collapse	3	6	25	46	80	0.575	0.460
	Total / Average	223	538	570	190	1521	0.558	0.479
	Precision	0.354	0.792	0.549	0.242	0.484		
	F-measure	0.404	0.641	0.595	0.341	0.495		
	F-measure + Recall of building Tracking	0.404	0.553	0.555	0.317	0.457		

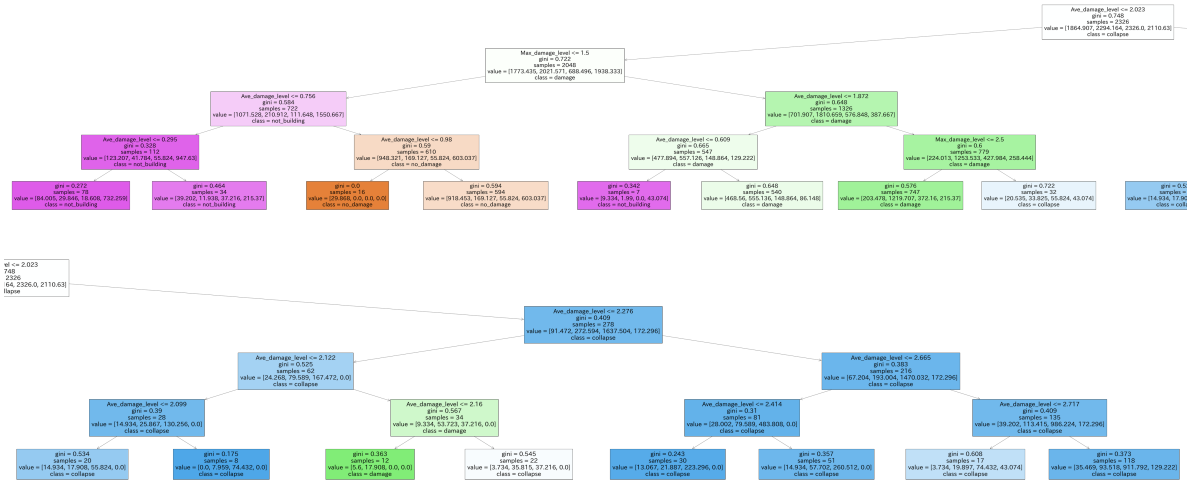

Figure 8. Constructed decision tree.

Figure 9. Example of visualizing estimation (red bbox: collapse, yellow bbox: damage).

Table 9. Amount of data and precision of each decision tree node.

Node No.	Estimated class	Train		Test		Difference of precision
		The amount of data	Precision	The amount of data	Precision	
1	Not building	78	0.218	123	0.407	0.189
2	Not building	34	0.147	92	0.304	0.157
3	No damage	16	1.00	56	0.786	-0.214
4	No damage	594	0.828	482	0.793	-0.036
5	Not building	7	0.143	8	0.125	-0.018
6	Damage	540	0.517	316	0.421	-0.096
7	Damage	747	0.821	247	0.713	-0.108
8	Collapse	32	0.094	28	0.250	0.156
9	Collapse	20	0.150	15	0.067	-0.083
10	Collapse	8	0.500	6	0.00	-0.500
11	Damage	12	0.750	7	0.571	-0.179
12	Collapse	22	0.091	13	0.077	-0.100
13	Collapse	30	0.400	20	0.300	-0.075
14	Collapse	51	0.275	25	0.200	-0.074
15	Collapse	17	0.235	5	0.600	0.365
16	Collapse	118	0.415	78	0.295	-0.120
Average absolute value			0.411		0.369	0.151

However, 25 of the 34 overlooked data points listed in Table 8 were estimated as the damage class. This indicates that the proposed model is effective in reducing false negatives, particularly where collapse data is incorrectly classified as no damage.

4.5.2. Evaluation of time-series information. If this system uses only the maximum damage level without the average damage level and decision tree, the recall of collapse by track was 56.0% (0.800 × 0.700) and precision was 18.4%. In this calculation method, all mistaken estimations of collapse cause an overestimation of the damage level, an increase in the recall of collapse, and a decrease in the precision of collapse. The F-measure of the four classes in the proposed system using the average damage level, maximum damage level, and decision tree was 45.7%, and the F-measure of the four classes in the calculation method above, using only the maximum damage level, was 43.8%. These results indicate that the time-series information of different frames using the average damage level is slightly more effective for estimating the building damage.

4.5.3. Evaluation of annotation labor. Naito et al. (2021) detected no damage, damaged, or collapsed buildings using object detection by deep learning from aerial images. Because the definitions of collapse in their study and this study are different, we compared the average recall of two classes: no damage and damaged buildings, including collapsed buildings. Because their study did not describe the precision of each class, we used the recall for comparison. At first, with respect to building tracking, their recall of buildings was 82.0%, that of this study was 82.7%. In calculating this, we used 0.1 as threshold of IoU which their study adopted. Next, with respect to classification of buildings, their recall of no damaged buildings was 70.3%, their recall of damaged buildings, including collapsed buildings, was 62.4%, and the average recall was 66.4%. In this study, the recall of buildings no damage was 53.9%, the recall of damaged buildings including collapsed buildings was 77.8% ((313 + 98 + 25 + 46)/(482 + 80)), and the average recall was 66.5%. The annotation time was expected to be approximately 184 h, as determined from the number of training data points and the average time of tracking annotation (172,668 × 3.84s). This study's annotation time was expected to be approximately 49.0 h as determined from the number of training data and average time of tracking and classification annotation ((21,221 × 3.84 s) + (96,224 × 0.987 s)). These results indicate that the proposed model has almost the same accuracy as that of a previous study (Naito et al., 2021), although the annotation labor of the proposed model was approximately one-quarter that reported in the previous study. If damaged and collapsed buildings were directly detected from aerial videos without the two processes of building tracking and classification, as in our previous study (Fujita and Hatayama, 2023), the annotation time was expected to be approximately 103 h (96,224 × 3.84 s). This demonstrates that the proposed model, which combines the two processes of building tracking and classification, achieved high accuracy with less annotation labor. This training took 49.0 h; however, when developers train the model with historical data before a disaster and then retrain it with a small amount of new data after the disaster, they can complete the annotation in less time to adapt to new disaster data.

4.5.4. Evaluation of a decision tree node.

Table 9 lists the estimated class, amount of data, and precision by each node. The average absolute value of difference of precision between train and test data was 0.151, indicating that the precision values for both

the training and test data were comparatively close. When the amount of data in one node was less, the precision difference tended to be significant. For instance, to determine the approximate figure of damaged and collapsed buildings, it is effective to use all nodes. To capture as many collapsed buildings as possible in case of rescue operation with abundant manpower, it is effective to use nodes 6–16, whose recall was 71.0% and precision was 9.34%. To decrease mistaken detection of collapsed buildings in case of rescue operation with limited manpower, it is effective to use node 15–16, whose recall was 26.0% and precision was 31.3%. By providing information on recall and precision of training data, users can make informed decisions depending on their specific requirements. To decrease mistaken detection of collapsed and damaged buildings, appropriate nodes can be decided in the same way. To utilize the estimation results of this model without human judgement, it is necessary to devise selection of output nodes to increase accuracy, recall or precision of damaged or collapsed buildings.

Figure 10 shows damaged buildings estimated in node 16. We observed that the buildings classified under node 16 exhibit a higher damage level because this node has the highest damage average level in decision tree. This is effective for emergency operation such as rescue even if these buildings have not collapsed. This result can be attributed to: input of damage average level, which incorporates time-series information, and decision tree, which can output several types of nodes.

Figure 10. Damage data estimated as collapse in Node. 16.

5. Conclusion

In this study, an automatic model was developed to detect damaged buildings during earthquakes using aerial videos. The proposed model was composed of a multiple object tracking model of buildings, a classification of damage model, and a decision-tree model to output the final estimation for each track. To focus on collapsed and damaged buildings such as roof damage and pancake collapse in Japan, this study used a deep learning classification model instead of debris feature detection. Our previous study (Fujita and Hatayama, 2023) did not classify the "damage" class and did not use time-series information in each track. Therefore, this study categorized three classes as: "no damage", "damage", and "collapse" using time-series information without significant labor of annotation. Moreover, this study used a decision tree model to output various types of estimations. From the results of the multiple object tracking of buildings, the recall by each track of buildings was 62.0%, and the precision was 77.8%. Moreover, the recall was 80.6% if the estimation surrounding multiple buildings was also considered correct and 95.0% if the filming method was appropriate. Finally, the decision tree estimated the damage class for each track. The average recall was 47.9%, the average precision was 48.4%, and the average F-measure was 45.7%, considering the recall of multiple object tracking of buildings.

In the Discussion section, we drew the following conclusions;

- Adding "damage" class to "no damage" and "collapse" classes prevents misinterpretation of some collapse data estimated as no damage.
- The average damage level as time-series information in each track was slightly more effective in estimating the building damage.
- Separating multiple object tracking and classification tasks, as in the proposed model, decreases annotation labor while maintaining accuracy.
- Decision tree nodes are effective for various types of disaster response.

In the future, multiple object tracking models will need to develop recall and ID switches. An ID switch may cause an incorrect estimation of each track in the decision tree. Based on the results of ID switch, tracking buildings in urban areas with higher building densities may prove to be more challenging compared to Kumamoto Prefecture. Evaluating our systems under various conditions, including different locations, weather conditions, and time frames, will be necessary. To increase the recall, an analysis of drone flying and filming methods suitable for tracking with the appropriate size and aspect ratio of the detected bbox is necessary. Through this study, we have been able to analyze damage detection models and efficient annotation methods. Therefore, it is crucial for the future to collect additional available video data to enhance accuracy. Furthermore, setting acceptable accuracy levels based on feedback from disaster response stakeholders is necessary. To emphasize the effect of decreasing annotation labor, we need to evaluate a machine-learning model trained after a disaster using the disaster data to adapt to new data quality using fine-tuning, domain adaptation, or human-in-the-loop. Thereafter, the proposed model can be combined with geospatial information to produce maps that facilitate a disaster response. Finally, organizing the workflow of disaster response, using case of damaged building estimation, and outputting the proposed model,

which is appropriate for each case, is necessary to consider social implementation during a disaster. Then, we need to conduct an implementation test with a disaster response organization and gather feedback on our system.

6. Acknowledgements

This work is supported by JSPS KAKENHI Grant Number JP 22KJ1918.

7. References

Calantropio, A., Chiabrando, F., Codastefano, M., Bourke, E. (2021). Deep learning for automatic building damage assessment: application in post-disaster scenarios using UAV data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-1-2021, 113–120

Disaster Management, Cabinet Office in Japan (2020). Guidelines of the Operation of Criteria for Building Damage Investigation in Disasters (in Japanese)

Disaster response headquarters of Kumamoto prefecture in Japan (2022). No. 325 damage report of the 2016 Kumamoto Earthquake (in Japanese). https://www.pref.kumamoto.jp/uploaded/attachment/182677.pdf

Fire and Disaster Management Agency in Japan (2022). Notice about promotion of drone usage of the fire brigade in disaster response (in Japanese). https://www.fdma.go.jp/laws/tutatsu/items/040331_drone.pdf

Fujita, S., and Hatyama, M. (2021). Estimation method for roof‐damaged buildings from aero-photo images during earthquakes using deep learning. Information Systems Frontiers 25(1), 351-363

Fujita, S., and Hatyama, M. (2022). Automatic calculation of damage rate of roofs based on image segmentation. 6th IFIP WG 5.15 International Conference, ITDRR 2021, pp. 3-22.

Fujita, S., and Hatyama, M. (2022). Rapid and accurate detection of building damage investigation using an automatic method to calculate roof damage rate. IDRiM Journal 12(1), 89–111

Fujita, S., and Hatyama, M. (2023). Collapsed Building Detection Using Multiple Object Tracking from Aerial Videos and Analysis of Effective Filming Techniques of Drones. 6th IFIP WG 5.15 International Conference, ITDRR 2022, pp. 118-135

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. (2021) Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430

He, K., Zhang, X., Ren, S., Sun, J. (2015) Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385

Kumamoto city fire department (2018). Activity record journal of Kumamoto city fire department in the 2016 Kumamoto Earthquake (in Japanese). https://www.city.kumamoto.jp/common/UploadFileDsp.aspx?c_id=5&id=19060&sub_id=1&flid=134936

Lucieer, A., Jong, M., S., Turner, D. (2013). Mapping landslide displacements using Structure from Motion (SfM) and image correlation of multi-temporal UAV photography. Progress in Physical Geography, Vol. 38(1) 97–116

Miura, H., Aridome, T., Matsuoka, M. (2020). Deep learning-based identification of collapsed, non-collapsed and blue tarp-covered buildings from post-disaster aerial images. Remote Sensing 12(12)

Naito, S., Tomozawa, H., Mori, Y., Nakamura, H., Fujiwara, H. (2021). Development of the Building Damage Detection Model using Oblique Aerial Photograpgy and Deep Learning (in Japanese). Intelligence, Informatics and Infrastructure 2 (J2), 211-222

Nazarov, E. (2011). Emergency Response management in Japan, Final Research report. Disaster Reduction Center, FY2011A Program

Okada, S., Takai, N. (1999) Classifications of structural types and damage patterns of buildings for earthquake field investigation. 12th World Conference on Earthquake Engineering

Pi, Y., Nath, D., N., Behzadan, H., A. (2020). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. Advanced Engineering Informatics ,43

Qi, J., Song, D., Shang, H., Wang, N., Hua, C., Wu, C., Qi, X., Han, J. (2016). Search and rescue rotary‐wing UAV and its application to the Lushan Ms 7.0 earthquake. Journal of Field Robotics 33(3), 290-321

Scawthorn, C. (2006). Building aspects of the 2004 Niigata Ken Chuetsu, Japan, Earthquake. Earthquake Spectra 22(1), 75-88

Scawthorn, C. and Yanev, I., P. (1995). Preliminary report 17 January 1995, Hyogo-ken Nambu, Japanese earthquake. Engineering Structures 17(3), 146-157

Yamazaki, F., Kubo, K., Tanabe, R., Liu, W. (2017) Damage assessment and 3d modeling by UAV flights after the 2016 Kumamoto, Japan earthquake. The IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3182–3185

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X. (2021). Bytetrack: Multi-object tracking by associating every detection box. arXiv preprint arXiv:2110.06864

Zhu, X., Liang, J., Hauptmann, A. (2021) Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos. In: the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2023-2032