2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

Learning to Assess Danger from Movies for Cooperative Escape Planning in Hazardous Environments

Vikram Shree, Sarah Allen, Beatriz Asfora, Jacopo Banfi, and Mark Campbell

***Abstract*— There has been a plethora of work towards improving robot perception and navigation, yet their application in hazardous environments, like during a fire or an earthquake, is still at a nascent stage. We hypothesize two key challenges here: first, it is difficult to replicate such scenarios in the real world, which is necessary for training and testing purposes. Second, current systems are not fully able to take advantage of the rich multi-modal data available in such hazardous environments. To address the first challenge, we propose to harness the enormous amount of visual content available in the form of movies and TV shows, and develop a dataset that can represent hazardous environments encountered in the real world. The data is annotated with high-level danger ratings for realistic disaster images, and corresponding keywords are provided that summarize the content of the scene. In response to the second challenge, we propose a multi-modal danger estimation pipeline for collaborative human-robot escape scenarios. Our Bayesian framework improves danger estimation by fusing information from robot's camera sensor and language inputs from the human. Furthermore, we augment the estimation module with a risk-aware planner that helps in identifying safer paths out of the dangerous environment. Through extensive simulations, we exhibit the advantages of our multi-modal perception framework that gets translated into tangible benefits such as higher success rate in a collaborative human-robot mission.**

I. Introduction

In the past decade, there has been a surge in the application of robotics in different avenues of our day-to-day life. Think, for example, of self-driving cars, cleaning robots, and robots as personal assistants. A few of these robots have achieved remarkable success while operating in organized environments with limited uncertainty like factories and homes. Yet, the deployment of robots during the World Trade Center disaster [1] and a more recent application during the Surfside condominium collapse [2], revealed significant untapped potential of current robotic systems, especially with regard to human-robot collaboration in search and rescue (SaR) missions. Often, people visiting public spaces like shopping malls, libraries, parks, etc. are unaware of their local map and exit points. Thus, in an emergency situation, a survivor can entrust a robot which has local map information to navigate out of the area. In turn, the robot can benefit from human's keen perception capability to ensure safety while navigating. To this end, we consider a modified version of the guide robot problem [3], adapted in the SaR situation where a robot helps a human in evacuating a hazardous environment.

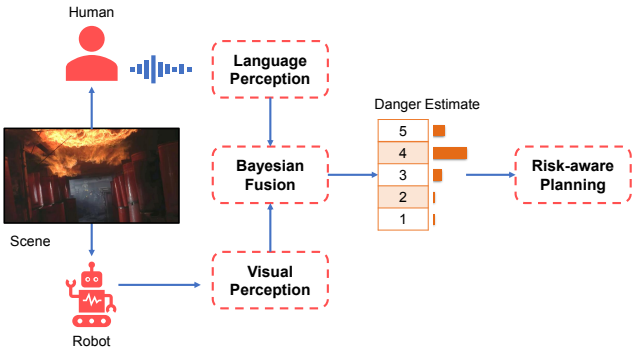


Fig. 1: System Architecture. The human and the robot operating in the same environment perceive the world through their respective senses. Both the modalities are processed separately to extract meaningful information about potential danger, which gets then fused into a final estimate which is passed to the risk-aware planner.

The 2018 DARPA Subterranean challenge [4] brought into light some of the key technical challenges that robots face when deployed in extreme environments [5]. Such environments possess severe obstruction for current perception systems due to low light conditions, sparse features, and presence of smoke, fire, fog, water, etc. Prior work has focused on extracting low-level features like occupancy grids [6] or feature maps [7] from such complex environments. However, planning an evacuation requires the agent to make decisions based on high-level scene attributes like "danger", as pointed out in [8]. We thus propose a principled way to assess danger in a scene and leverage the danger information for planning a safe evacuation mission out of the hazardous environment. Furthermore, while there are merits to having a fully autonomous system, we believe that there is abundant opportunity to benefit from collaboration between a robot and the human agent, e.g. building trust. Thus, we propose a collaborative estimation strategy that takes advantage of human perception, contributing towards more successful evacuation from the environment.

There has been substantial improvement in visual perception performance in the past years, achieved by the application of deep convolutional neural networks [9]–[12]. The improvement comes at the cost of collecting large training data that can aptly resemble the test environment. Ironically, replicating hazards is tough due to local policies regarding safety and the high cost investment needed. These limitations

V. Shree, S. Allen, B. Asfora, and M. Campbell are with the Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY USA. Email: {vs476, sea97, ba386, mc288}@cornell.edu. J. Banfi is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA USA. Email: jbanfi@mit.edu

Research supported by the NRI program of the National Science Foundation, award #1830497.

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9982279

Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.

have pushed researchers to test perception systems on simulated datasets [13]–[15] or model environments [16]. However, these alternatives are yet to attain the richness of the real world. To bridge this gap, we propose a scalable approach by leveraging images from movies which are significantly more photo-realistic than existing simulated datasets used in the community and can depict wider variety of scenarios as compared to the state-of-the-art model environments. Moreover, this route avoids the privacy infringement concerns that using actual images from disaster sites can have.

Our approach, sketched in Fig. 1, initially consists of independent vision and language perception modules, which estimate the danger based on images from robot's sensor and verbal input from the user, respectively. The estimated danger levels are then fused together to get an improved danger estimate. Finally, taking advantage of the danger estimate, we propose a risk-aware planner that maximizes the chances of survival for the human-robot team. In summary, this paper makes the following novel contributions:

1) Development of a visuo-lingual dataset for perception in hazardous environments, with images taken from mass media that closely depict real world scenarios. The dataset entails distribution of danger as well as associated word descriptions of each scene.
2) Performance assessment of representative machine learning models for danger estimation from images and language-input from humans.
3) A Bayesian fusion framework that capitalizes on the likelihood models for both sensing modalities: vision and language, resulting in superior danger estimation.
4) Extensive testing with the collected dataset to evaluate our risk-aware mission planner, showing that the proposed approach enhances the success rate on average by 19% points (compared to a baseline shortest path planner).

II. Related Work

A. Cooperative rescue missions

In [17], the authors provide a comprehensive outline of the technical challenges encountered in the domain of multi-robot SaR. They remark that in order to fully benefit from the multi-robot team, there should be mechanisms to fuse information from different agents, enabling superior scene awareness. Our present work contributes along this direction with a collaborative scene perception pipeline in a human-robot team setup, where information is fused across the visual and the language domain.

The idea of collaborative decision making is inspired by humans. In an unforeseen disaster event, if a group of people are trapped together, they naturally tend to join hands to escape that situation. Consequently, researchers have tried to incorporate collaboration in multi-robot teams [18] and in human-robot teams [19]. Still, prior studies suggest that human-robot teaming is relatively new in rescue missions because of interaction being a major bottleneck [20]. Our fusion framework allows the robot to account for human feedback, enabling superior danger awareness of the surrounding. This knowledge about the environment can be further used by the robot for planning purposes, for instance to identify safer routes to an exit.

B. Disaster scene understanding

Safe navigation in SaR missions relies heavily upon accurate scene understanding. There are several tasks aiming at scene understanding like object recognition, semantic segmentation, physics-based reasoning, 3D reconstruction etc., as mentioned in [21]. While some of these tasks involve low-level reasoning like 3D reconstruction, others need high-level scene awareness like physics-based reasoning. In this work, we intend to leverage a high-level attribute of the environment, the notion of scene danger.

Often, danger is associated with the presence of fire or smoke in the scene, thus, prompting researchers to identify them in a scene [22], [23]. In contrast to these methods, we propose a more holistic danger perception of the environment which is not just limited to fire and smoke. Given a camera image from the environment, our perception module leverages state-of-the-art classification networks [24]–[26] to predict a danger hypothesis. Furthermore, we embark on the opportunity to do collaborative perception for the task at hand and add a significant human component to get an updated danger representation of the environment. Previously, Ahmed et al. [27] introduced hybrid continuous-to-discrete likelihoods for fusing language data from humans by assuming a codebook consisting of a small set of words for the user to choose from. In contrast, our model provides the freedom to the user to choose any word from the vast English vocabulary. Similar work has been pursued in [8] where the authors propose an adaptable danger estimation pipeline that relies on an a priori list of danger descriptions from an expert. There are two key aspects that differentiates our current work from [8]. First, our model does not necessitate language input from the human for danger estimation and is capable of assessing danger solely from camera data. This is advantageous in a scenario where the human is unable to provide feedback regarding their surrounding, e.g. due to cognitive impairment. Second, our multi-modal Bayesian framework allows multiple online updates based on incoming language data from the human, which is in contrast to [8] where the authors assume an a priori set of danger descriptions from an expert, specific to the environment.

III. System Pipeline and Notation

In our modified guide robot problem, we assume that the robot is present in the vicinity of a human survivor, who can follow the robot's path. The robot is capable of perceiving the environment through its camera sensor, and receives language input from the human about their surrounding. Furthermore, we assume that the robot has knowledge of the metric map of the environment and its objective is to find the *best* path to an exit. The overall system can be divided into four major components as shown in Fig. 1: visual perception, language perception, Bayesian fusion, and

Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.

risk-aware planning. Following [8], we assume a 5-point danger scale: 1-low, 2-moderate, 3-high, 4-very high, and 5-extreme. Let us denote an image from the environment by I and its ground truth danger level by D. The visual perception module distills the key features of image I and predicts an estimate for danger $y_V \in \{1, \cdots, 5\}$.

Assuming language inputs consisting of a single word from the human, let us denote the word input by W. The language module predicts a danger estimate $y_L \in \{1, \cdots, 5\}$ based on the severity of word W. As a last step of the perception segment, the fusion module estimates a probability mass function (PMF) over the danger space, given the image and language input i.e. $\hat{D} = p(D = d|y_V, y_L)$, where $d \in \{1, \cdots, 5\}$.

We assume that the robot knows the start and goal locations. The planner capitalizes on the danger estimate from the perception segment and plans an escape path that maximizes the survival probability. The following sections will elaborate each segment of our system in greater detail, starting first with our hazardous environment dataset.

IV. Hazardous Environment Dataset

To get authentic perceptual data that can replicate hazardous environments in the real world, we pool images from the vast collection of video clips that are easily accessible on various online platforms like Netflix, Xfinity, and Amazon Prime. The images are then annotated with the help of Amazon Mechanical Turk (AMT) workers and post-processed to generate ground truth labels. The dataset is provisioned under the fair use clause of copyrighted material and is opensourced for free usage by only the research community[1].

A. Image Selection

We collect images from a wide range of movies and TV shows as candidates for our dataset. We initially select a set of 15 movies and 5 TV shows that embody scenes from variety of disaster scenarios, such as fire, flooding, and earthquakes. This is followed by capturing 75 small clips of the relevant sections in the movie/show with a screen capture software. Each clip (scene) is comprised of its unique set of visual and geographic attributes, distinguishing it from the other scenes. Images are then extracted automatically from these scenes at 2 frames per second. At last, we perform a manual check to get rid of images that are either redundant or blurred, resulting in a total of 1002 images.

B. Data annotation

For annotation, we used Amazon Mechanical Turk (AMT). For each image, an AMT user must provide a danger rating from 1 to 5 and at most three keywords describing what led them to choose that specific danger rating. Thus, given an image I, we can represent the input from a AMT user i as $(\eta_i, \mathbf{\Pi}_i)$, where η_i denotes the danger rating and $\mathbf{\Pi}_i$ is the set of keywords. To ensure data annotation quality, only AMT users with at least 1000 prior completed assignments and 98% approval rating were considered. Each image was examined by 15 unique AMT users, amounting to 355 unique AMT users.

[1]Dataset available at https://github.com/vikshree/hazard_dataset.git

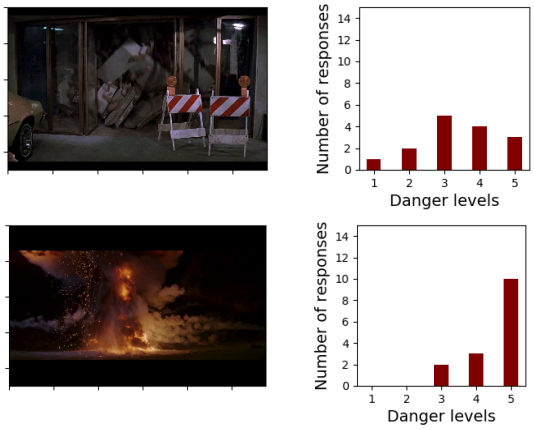
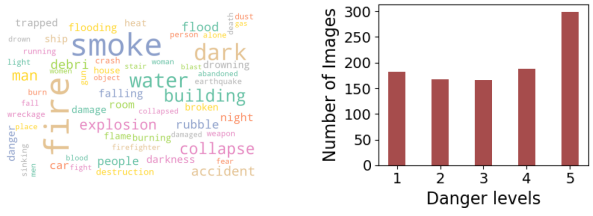


Fig. 2: Images from our hazardous environment dataset and corresponding ratings by AMT users. Best viewed in color.



(a) (b)

Fig. 3: (a) Frequent words appearing in the responses of AMT users. (b) Distribution of mode danger rating of images across for the whole dataset.

C. Data Statistics

Our data collection results in 15K danger ratings and 45K associated keywords from the AMT users, with 3K unique words in it. A few sample images are shown in Fig. 2. We observe higher consensus in the danger ratings for images with extremely low or extremely high danger, as indicated in Fig. 2. To gain further insight into the language data, we show the wordcloud in Fig. 3a depicting frequently used words. We find that factors like *fire, smoke, water, dark,* and *collapse*, play a key role in determining the danger rating.

Treating the *mode* of the 15 danger responses for an image as representative of its "true" danger level, we show the count of images in our dataset belonging to each "true" danger category in Fig. 3b. We observe that the number of images corresponding to different danger levels are comparable, thus, making the dataset well-balanced which is a crucial factor for training machine learning models.

V. Visual Perception

A. Task and Metrics

Given a local image I obtained from the robot's camera sensor, the goal of the vision module is to distill its key aspects, and predict a danger rating for the image $y_V \in$

Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.

$\{1, 2, \cdots, 5\}$. The model's ability to perceive entities that add to a person's notion of danger like fire, smoke, darkness, leakage, etc., is key to success in the task.

We use three metrics to capture the nuances in danger prediction. First, it is typical to use top-1 accuracy for classification tasks and is defined as the proportion of time danger prediction y_V matches the "true" danger of the image (assumed to be the *mode* $\tilde{D}$ of the AMT user-based danger ratings). Second, we use the root mean squared error of the predicted danger from the "true" danger of the image, i.e.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_{V,k} - \tilde{D}_k)^2}, \tag{1}$$

where, n is the total number of images. In addition to these standard metrics, we define a third one: the "off-by-1" accuracy. Off-by-1 accuracy is the proportion of time the danger estimate differs from $\tilde{D}$ at-most by 1 danger unit.

B. Performance Evaluation

1) Baselines: We evaluate the ability of state-of-the-art (SOTA) models to estimate danger in hazardous environments. We select four candidate models: VGGNet [24], ResNet [25], DenseNet [28], and EfficientNet [26]. Each of these candidates have certain fundamental traits that differentiate them from one another. VGGNet is one of the oldest deep neural networks for image processing. ResNet addresses the vanishing gradients problem in deep networks by adding identity connections between layers. DenseNet uses dense blocks that receive features from its preceding layers and also pass the processed features to all its subsequent layers, leading to stronger feature propagation. EfficientNet emphasizes scaling-up the network in a structured manner, leading to smaller yet effective models. The goal here is to assess performance and identify the most suitable model for our multi-modal danger assessment pipeline.

2) Dataset: We split our images and danger ratings data into train, validation, and test sets such that there are no overlapping scenes in two different sets. These sets were created manually, ensuring that the danger distribution for all three sets are similar. Ultimately, the train set consists of 795 images from 56 scenes, the validation set consists of 106 images from 10 scenes, and the test set consists of 101 images from 8 scenes.

3) Training: During training, the parameters for the last classification layer are tuned, while keeping the rest of the network frozen. This takes advantage from the rich knowledge of the pre-trained model, aligning well with our experiments where we observed higher performance across all metrics as compared to training the whole network. Given an image I, its ground truth danger PMF, denoted by $\mathbf{p} = [p_1, p_2, \cdots, p_5]$, is obtained by normalizing the corresponding ratings by the 15 users and is used for training the models. Each baseline network outputs a danger confidence $\mathbf{c} = [c_1, c_2, \cdots, c_5]$. The vision-based danger estimate is defined as $y_V = \arg\max c_i \ \forall \ i \in \{1, 2, \cdots, 5\}$. While training we minimize KL divergence of the danger PMF $\mathbf{p}$ from the model confidence $\mathbf{c}$, i.e.

$$D_{KL}(\mathbf{p}||\mathbf{c}) \equiv \sum_{i=1}^{5} p_i\big(\log p_i - \log c_i\big). \tag{2}$$

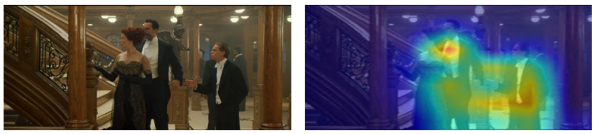
TABLE I: Visual danger assessment performance of SOTA networks. Best performance is shown in **bold**. The number next to the architecture denotes a particular version.

Model	Top-1	RMSE	Off-by-1
VGGNet-11	43.6	1.46	67.3
VGGNet-13	**47.5**	**1.24**	**79.2**
VGGNet-16	46.5	1.52	71.3
ResNet-50	39.6	1.55	68.3
ResNet-101	45.5	1.40	75.2
ResNet-152	42.6	1.59	70.3
DenseNet-121	30.7	1.78	63.4
DenseNet-169	38.6	1.73	65.3
DenseNet-201	32.7	1.76	63.4
EfficientNet-b0	44.6	1.45	72.3
EfficientNet-b1	31.7	1.85	61.4
EfficientNet-b2	34.7	1.56	68.3
EfficientNet-b3	36.6	1.58	72.3
EfficientNet-b4	**47.5**	1.49	72.3
EfficientNet-b5	34.7	1.70	63.4
Randomized	20.0	2.0	52.0

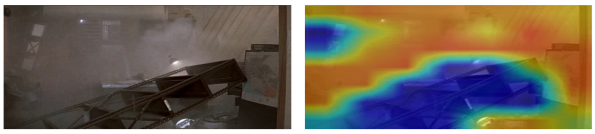
4) Results: We train each baseline for 50 epochs and the best model is identified as the one with highest top-1 accuracy on the validation set. The performance of the best model for each baseline on the test set is reported in Table I. All the models attain much higher top-1 accuracy compared to a randomized baseline, that would yield a top-1 accuracy of 20%. For example, VGGNet-13 is correct about 48% of the times in predicting the right danger level for an image and about 80% of the time its estimate is at-most off-by-1 from the correct answer. Although the models are competitive, we chose VGGNet-13 for successive sections of the paper because of its best performance across all three metrics.

For intuitive understanding of the model predictions, we leverage the Gradient-weighted Class Activation Mapping [29] (Grad-CAM) visualization tool. Grad-CAM produces a color map highlighting the important regions of the image that contributed towards the predicted result. This not only provides the much needed insight into the deep learning model, but could also help gain the trust of human when used in real-world missions. For example, we can observe that the image in Fig. 4 showing some people chatting has a low predicted danger. The smoke in Fig. 4b is a significant contributor for its danger prediction of 3. Finally, the flooded area in Fig. 4c prompts the model to predict extreme danger. The quantitative as well as the qualitative results demonstrate the ability of our baseline models to learn and predict danger from images. These baseline results are a solid precursor to specialized models for vision-based danger assessment, which is part of our future work.

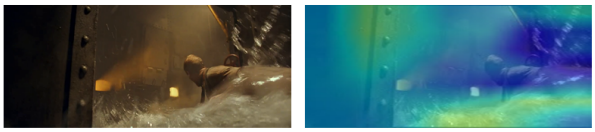
Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.



(a) Output Danger = 1



(b) Output Danger = 3



(c) Output Danger = 5

Fig. 4: Grad-CAM visualizations for test set images with VGGNet-13 model predictions. Red regions provide the highest contribution while blue regions the lowest. Best viewed in color.

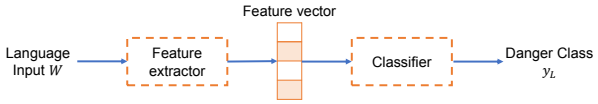


Fig. 5: Language perception pipeline. First, words are converted into features, followed by a classifier that predicts danger y_L.

VI. Language Perception

A. Task and Metrics

The language perception module takes into account the word input W from the human and accordingly predicts a danger rating $y_L \in \{1, 2, \cdots, 5\}$. The ability to understand what words people tend to use in hazardous environmental conditions is key to success in this task. It is intuitive to use the same metrics defined in section V-A for assessing performance of language perception models i.e. top-1 accuracy, root mean squared error (RMSE), and off-by-1 accuracy.

B. Performance Evaluation

1) Baselines: Following the mainstream approach in language classification, we first convert the words into a low-dimensional feature vector, which is then passed to a classifier [30]. This is shown in Fig. 5. In this work, we use GloVe features [31] for their superior performance reported in literature as compared to other representations like Word2Vec [32]. We test danger prediction with three well-known candidate classifiers: K nearest neighbor (KNN), logistic regression, and Support Vector Machine (SVM). The KNN algorithm predicts the danger class based on the k nearest matches of the word W in the training data; logistic regression aims at maximizing the conditional likelihood of the training data; SVM maximizes the margin between class variables, making it less prone to outliers compared to logistic regression.

TABLE II: Language perception performance. Best performance is shown in **bold**.

Model	Top-1	RMSE	Off-by-1
1-NN	27.2	1.72	64.5
3-NN	28.8	1.79	61.0
5-NN	30.8	1.70	66.3
11-NN	32.4	**1.58**	68.4
Logistic Regression	36.9	1.67	67.1
SVM (Linear)	37.1	1.68	66.8
SVM (Poly. kernel)	37.1	1.64	**68.5**
SVM (RBF kernel)	**37.6**	1.63	**68.5**
Randomized	20.0	2.0	52.0

TABLE III: Common words and their danger predictions generated from the SVM model.

Words	Danger output
people, gathering, dirty, night, tunnel	1
darkness, sewer, dust, cave, broken	2
debris, suffocation, wreckage, damage, violence	3
flood, accident, weapon, crash, freezing	4
fire, explosion, collapse, flooding, earthquake	5

2) Dataset: For evaluating performance of danger prediction based on language input, we use the keywords associated with the images and their corresponding danger ratings from the hazardous environment dataset. Since about 90% of the keywords consist of a single word, we assume one-word input from the human and ignore the sentences in our dataset. A typical training data point is of the form (W, η), where W is the word describing the scene and η is its danger level. We use the same splits for training, validation, and test set as chosen in the visual perception case, yielding a total of 31K, 4K, and 4K keywords in these sets, respectively.

3) Results: The results based on the test set are shown in Table II. SVM's compatibility with high-dimensional data renders it superior top-1 accuracy compared to other models. All three class of models achieve competitive performance in terms of RMSE and off-by-1 accuracy. It is interesting to note that the values of all the metrics for language-based danger predictor are significantly lower than the best visual perception model. This can be attributed to the richness of visual data when compared to single word inputs from the human. Since top-1 accuracy is widely accepted in the literature, we chose to use SVM (RBF kernel) for the subsequent sections.

Table III shows a few words frequently appearing in the test set along with their corresponding danger prediction by the model. Words with danger prediction of 1, such as people, gathering, dirty etc., align well with our intuitive sense of danger. Similarly, words that have a danger prediction of 5, such as fire, explosion, flooding etc, are also straightforward. Note that words with intermediate danger, such as debris, broken, crash etc. are in fact contentious for humans because of the lack of sufficient information that these words bear.

VII. Fused Danger Assessment

Vision and language perception modalities both have their own benefits. On one hand, image data is richer compared

Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.

to words. On the other hand, language modality can take advantage of human's keen perception to focus on key entities that contribute towards danger in the scene. Thus, it is natural to combine both modalities with the goal to achieve superior danger perception. Furthermore, as pointed out in the literature [33], accounting for human feedback enables a human-in-the-loop mission approach, thus promoting trust between the human and the robot. Hence, we now introduce our Bayesian fusion framework.

A. Bayesian fusion

The goal of the Bayesian fusion module is estimate $\hat{D} \equiv p(D = d|y_V, y_L)$, where $d \in \{1, \cdots, 5\}$. From Bayes rule, assuming conditional independence of predictions from the visual and language modules, given the scene danger d:

$$\begin{aligned} p(D = d|y_V, y_L) &\propto p(y_V, y_L|D = d)p(D = d) \\ &\propto \ p(y_V|D = d)p(y_L|D = d)p(D = d), \end{aligned} \tag{3}$$

where $p(y_V|D = d)$ and $p(y_L|D = d)$ denote the vision-based and language-based sensing likelihoods, respectively. Eq. (3) is key for our fusion module.

Note that until now, we have assumed a single word input from the human. However, this assumption can now be relaxed and Eq. (3) can be extended to m number of human inputs:

$$\begin{aligned} &\hat{D} \equiv p(D = d|y_V, y_L^1, \cdots, y_L^m) \\ \Rightarrow &\hat{D} \propto p(y_V|D = d)\Big(\prod_{k=1}^{m} p(y_L^k|D = d)\Big)p(D = d). \end{aligned} \tag{4}$$

As evident from Eq. 4, we need the likelihood functions for visual and language perception models. First, consider the vision-based danger likelihood function and denote it by $l_V^{i,j} \equiv p(y_V = i|D = j)$, where $i, j \in \{1, \cdots, 5\}$. The likelihood function can be calculated from the validation set:

$$\begin{aligned} l_V^{i,j} &\equiv p(y_V = i|D = j) \\ &\approx \frac{\text{\# images with ``true'' danger } j \text{ and prediction } i}{\text{\# images with ``true'' danger level } j}. \end{aligned} \tag{5}$$

In the lack of a large validation set, as in our case, Eq. (5) may overfit to the set and cause poor performance on the test set. To avoid this problem, we apply K-fold cross-validation strategy where depending on the selection of the validation set, we get K different estimates for the likelihood function $l_V^{i,j}$. Thereafter, the *mean* likelihood function is obtained, i.e. $\hat{l}_V^{i,j} = \frac{1}{K}\sum_{k=1}^{K} l_{V,k}^{i,j}$. We follow the same strategy to obtain the *mean* language-based danger likelihood function $\hat{l}_L^{i,j}$.

B. Performance evaluation

We apply 9-fold cross-validation to calculate the *mean* likelihood functions $\hat{l}_V^{i,j}$ and $\hat{l}_L^{i,j}$. See Fig. 6. Two crucial observations can be made here. First, we found that the maximum of the *mean* likelihood function occurs at the "true" danger level, i.e. $\arg\max_j \hat{l}_V^{i,j} = \arg\max_j \hat{l}_L^{i,j} = i$. Second, the values of likelihood functions at the "true" danger level, i.e. $\hat{l}_V^{i,i}$ and $\hat{l}_L^{i,i}$, are higher at the two extremes i.e. $d = 1$ and 5. This is because both images and words with danger values at the extremes are easier to classify by our models as compared to the ones that belong to moderate danger.

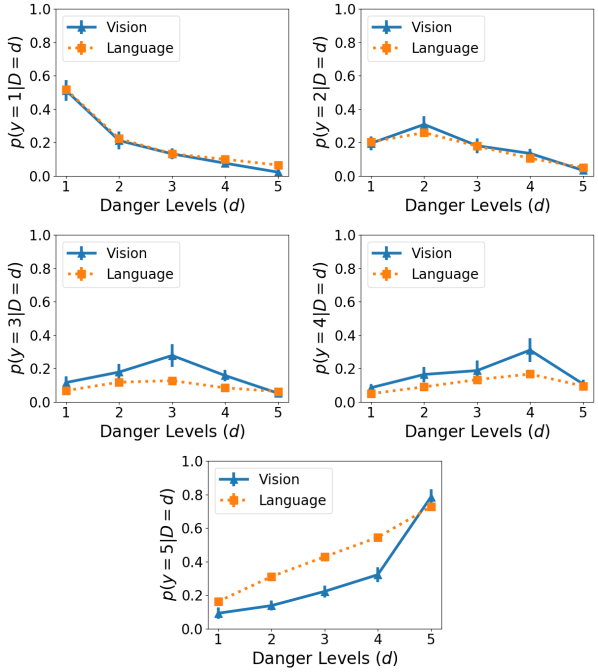


Fig. 6: Vision-based and Language-based likelihood function estimated from 9-fold validation.

To evaluate the performance of the fused danger estimate, we use the hazardous environment dataset with the same test set as chosen in section V-B.2. This helps us to gauge the relative change in performance by fusing the two modalities. Given an image I and corresponding set of keywords $\mathbf{\Pi}$, we simulate input from the human by randomly sampling keywords from $\mathbf{\Pi}$ without replacement. To evaluate the estimation performance, we use the maximum a posteriori (MAP) estimate $\hat{D}_{\text{MAP}} = \arg\max_d p(D = d|y_V, y_L^1, \cdots, y_L^m)$. Comparing the MAP estimate $\hat{D}_{\text{MAP}}$ with the "true" danger $\tilde{D}$, we were able to evaluate the top-1 accuracy, off-by-1 accuracy, and RMSE with different number of word inputs from the human. For visual perception, we use the VGG-13 model, while for language perception we use the SVM (with RBF kernel) model. The results, presented in Table IV, reveal that all the three metrics are maximized by leveraging the multi-modal pipeline. Specifically, combining visual data with 5 words from the human improves top-1 accuracy by 1% point and RMSE by 17%, compared to using any of the single modalities. Note that although the Bayesian model is tested with 5 words, it can incorporate even higher number of word inputs from human. However, special care should be taken to avoid redundant information because it can violate the independence assumption used in Eq. (4), leading the Bayesian model to overtrust the data coming from the human.

Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.

TABLE IV: Multi-modal danger assessment performance. Best performance is shown in **bold**.

Modality	Top-1	RMSE	Off-by-1
Vision only	47.5	1.24	79.2
Language-only	37.6	1.63	68.5
VL 1-word	46.5	1.30	78.2
VL 2-words	47.5	1.13	**84.2**
VL 3-words	44.6	1.27	80.2
VL 4-words	47.5	1.11	83.2
VL 5-words	**48.5**	**1.03**	83.2

VIII. Risk-aware Planning

With the capability to estimate danger, we now introduce the escape route planning problem.

A. Planning Problem and Solution Approach

Let us represent the environment as a directed graph $G = (V, A)$, where vertices V represent a set of locations in the environment, and arcs A represent the possibility to travel between two locations. The starting vertex of the human-robot team is denoted by $v_s \in V$ and the goal location (for example, one of the building exits) is denoted by $v_g \in V$. Let us define a parameter $\tau \in \{1, ..., 5\}$, denoting the level of danger that the human-robot team can tolerate. Accordingly, we define the survival probability of traveling along arc $(i, j) \in A$ as $s_{ij} = p(D_j \leq \tau)$, where D_j is the ground truth danger of the destination node j. Assuming these events to be independent, the survival probability s_π along a graph path $\pi = [(v_s = v_1, v_2), \ldots, (v_{k-1}, v_k = v_g)]$ connecting start and goal vertices can be expressed as

$$s_\pi = \prod_{i=1}^{k} s_{v_i v_{i+1}}. \tag{6}$$

Note that the independence assumption might not always hold in practice, and more sophisticated models could be built to account for spatial dependencies in the danger map. If all the survival probabilities were known exactly and in advance, the path π^* maximizing the the overall probability of survival, i.e.

$$\pi^* \in \arg\max_{\pi} \; s_\pi, \tag{7}$$

could be easily obtained by computing the shortest path between v_s and v_g on a weighted version of the graph G, with weights w_{ij} computed as $w_{ij} = -\log s_{ij}$. However, the robot does not have access to the true survival probabilities s_{ij}, and it must rely on the danger estimate $\hat{D}$ to establish an approximation for the survival probability $\hat{s}_{ij} = p(\hat{D} \leq \tau)$. We assume that the robot-human team can only access data of the neighboring vertices of G. Furthermore, the unexplored vertices are assigned a uniform prior danger distribution.

Given the problem inputs described above, the goal of the planning module is to compute a policy that maximizes the mission success rate, i.e. the probability of reaching v_g without traversing an arc having a ground truth danger level higher than τ. Since the survival probability in Eq. (7) depends on robot's belief of danger, we hypothesize that superior danger perception with multi-modal sensing can enable the team to avoid hazardous exposures, leading to higher mission success.

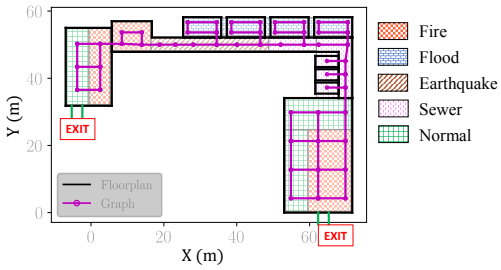


Fig. 7: Graphical representation of the environment used in our simulations, with danger map.

To tackle the planning problem described above, we use the following receding-horizon planning heuristic: at each planning iteration, the team moves along the first arc of the safest path π^* computed as described above, but replacing s_{ij} with $\hat{s}_{ij}$. When the corresponding destination vertex is reached, the team updates the danger estimate of the neighboring vertices. This process is repeated until the team ends up in a vertex with an intolerable danger level (in which case the mission counts as a failure), or the the goal node v_g is reached.

B. Simulation Enironment

We use the School environment from [8] and abstract it into the final graph shown in Fig. 7, consisting of a total of $n = 54$ nodes and two exits. We manually assign scene characteristics e.g. fire, flood, etc., for different segments of the environment and accordingly associate each node with an image from the test set of the hazardous environment dataset.

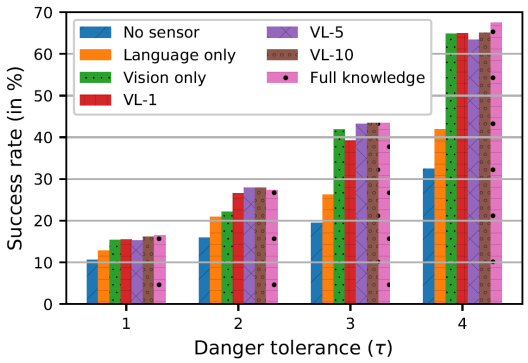


Fig. 8: Planning results: Success rate with different sensing modalities. Results are obtained from 1000 simulation runs.

C. Results and Discussion

We perform 1000 Monte Carlo simulations of the combined planning and danger estimation framework with different sensor modalities and for different tolerable danger levels τ. The "no-sensor" case is when the robot is unable to perceive danger. The "full-knowledge" case refers to the hypothetical situation when the robot is aware of the ground truth danger map of the whole environment. During simulations, when the team is moving from node i to j, the survival of the team is sampled from PMF of s_{ij}. A

Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.

successful mission is the one where the team survives the whole mission and reaches the exit. Note that we do not report results for $\tau = 5$ since it corresponds to the unrealistic case when the team can survive extreme danger.

In Fig. 8, we observe that the "no sensor" case has lowest success rate amongst all methods for every value of τ. This indicates that the ability to perceive and account for danger is critical for mission success in disaster scenarios. A further analysis of different modalities provides some intriguing insights. The "vision only" sensing consistently outperforms "language only" sensing. As mentioned earlier, this is because of the richness of visual data compared to single word inputs. Occasionally, the "vision-only" success rate is even competitive to the multi-modal method, implying that the perception advantage shown in Table IV does not always yield similar benefits in terms of mission success. Nonetheless, across all simulation scenarios, the highest success rate is achieved in VL-10 case, by fusing information between the vision and language domain. The results support our hypothesis that improved perception leads to more successful missions. In fact, sometimes VL-10 case even outperforms the model with full knowledge. This is probably because of higher danger estimate reported by the model, compared to ground truth danger, thus, forcing the robot to choose a conservative route.

IX. Conclusion

Our work demonstrates that leveraging the vast collection of visual content from mass media can enable perception systems to function in disaster scenarios. The hazardous environment dataset paves the way for development and testing of future danger assessment pipelines. Further, we show through simulations that compared to an autonomous robot that only relies on a single sensing modality, a collaborative robot that takes into account the feedback from human user is better equipped to estimate danger. Finally, our risk-aware planning framework translates the improvements in danger assessment into tangible metrics such as higher mission success rate, which is critical in search and rescue operations.

References

[1] J. Casper and R. R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 33, no. 3, pp. 367–385, 2003.

[2] R. R. Murphy, "How robots helped out after the surfside condo collapse," *IEEE Spectrum*, 2021. [Online]. Available: https://spectrum.ieee.org/building-collapse-surfside-robots

[3] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "An affective guide robot in a shopping mall," in *HRI*, 2009, pp. 173–180.

[4] DARPA, "DARPA subterranean challenge," 2018. [Online]. Available: https://www.subtchallenge.com

[5] A. Agha *et al.*, "Nebula: Quest for robotic autonomy in challenging environments; team costar at the darpa subterranean challenge," *Journal of Field Robotics*, 2021.

[6] A.-A. Agha-Mohammadi, E. Heiden, K. Hausman, and G. Sukhatme, "Confidence-rich grid mapping," *IJRR*, vol. 38, no. 12-13, pp. 1352–1374, 2019.

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[8] V. Shree, B. Asfora, R. Zheng, S. Hong, J. Banfi, and M. Campbell, "Exploiting natural language for efficient risk-aware multi-robot sar planning," *RA-L*, vol. 6, no. 2, pp. 3152–3159, 2021.

[9] S. Milz, G. Arbeiter, C. Witt, B. Abdallah, and S. Yogamani, "Visual slam for automated driving: Exploring the applications of deep learning," in *CVPR Workshops*, 2018, pp. 247–257.

[10] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Ng, "Convolutional-recursive deep learning for 3d object classification," *Advances in Neural Information Processing Systems*, vol. 25, pp. 656–664, 2012.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[12] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," *ECCV*, pp. 628–644, 2016.

[13] H.-G. Jeon, S. Im, B.-U. Lee, D.-G. Choi, M. Hebert, and I. S. Kweon, "Disc: A large-scale virtual dataset for simulating disaster scenarios," *IROS*, pp. 187–194, 2019.

[14] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," *IROS*, pp. 4909–4916, 2020.

[15] P. Kirsanov *et al.*, "Discoman: Dataset of indoor scenes for odometry, mapping and navigation," *IROS*, pp. 2470–2477, 2019.

[16] J.-H. Kim and B. Y. Lattimer, "Real-time probabilistic classification of fire and smoke using thermal imagery for intelligent firefighting robot," *Fire Safety Journal*, vol. 72, pp. 40–49, 2015.

[17] J. P. Queralta *et al.*, "Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision," *IEEE Access*, vol. 8, pp. 191 617–191 643, 2020.

[18] A. Stroupe, T. Huntsberger, A. Okon, H. Aghazarian, and M. Robinson, "Behavior-based multi-robot collaboration for autonomous construction tasks," *IROS*, pp. 1495–1500, 2005.

[19] B. Chandrasekaran and J. M. Conrad, "Human-robot collaboration: A survey," *SoutheastCon 2015*, pp. 1–8, 2015.

[20] G.-J. M. Kruijff *et al.*, "Experience in system design for human-robot teaming in urban search and rescue," in *Field and Service Robotics*. Springer, 2014, pp. 111–125.

[21] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3d for autonomous agents: A survey," *IEEE access*, vol. 7, pp. 1859–1887, 2018.

[22] P. Li and W. Zhao, "Image fire detection algorithms based on convolutional neural networks," *Case Studies in Thermal Engineering*, vol. 19, no. 100625, 2020.

[23] A. Gaur, A. Singh, A. Kumar, A. Kumar, and K. Kapoor, "Video flame and smoke based fire detection algorithms: A literature review," *Fire Technology*, vol. 56, no. 5, pp. 1943–1980, 2020.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[26] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[27] N. R. Ahmed, E. M. Sample, and M. Campbell, "Bayesian multicategorical soft data fusion for human–robot collaboration," *IEEE Transactions on Robotics*, vol. 29, no. 1, pp. 189–206, 2012.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[30] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis lectures on human language technologies*, vol. 10, no. 1, p. 92, 2017.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the conference on empirical methods in natural language processing*, 2014, pp. 1532–1543.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ICLR*, 2013.

[33] K. Eder, C. Harper, and U. Leonards, "Towards the safety of human-in-the-loop robotics: Challenges and opportunities for safety assurance of robotic co-workers'," *International Symposium on Robot and Human Interactive Communication*, pp. 660–665, 2014.

Authorized licensed use limited to: Cornell University Library. Downloaded on December 29,2023 at 21:54:52 UTC from IEEE Xplore. Restrictions apply.