2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Social Context Perception for Mobile Robots

Aastha Nigam and Laurel D. Riek[1]

***Abstract*— As robots enter human spaces, unique perception challenges are emerging. Sensing human activity, adapting to highly dynamic environments, and acting coherently and contingently is challenging when robots transition from structured environments to human-centric ones. We approach this problem by employing *context-based perception*, a biologically-inspired, low-cost approach to sensing that leverages noisy, global features. Across several months, our mobile robot collected real-world, multimodal data from multi-use locations; where the same space might be used for many different activities. We then ran a series of unimodal and multimodal classification experiments. We successfully classified several aspects of situational context from noisy data, and, to our knowledge are the first group to do so. This work represents an important step toward enabling robots that can readily leverage context to solve perceptual tasks.**

I. INTRODUCTION

When robots solve localization problems, they typically rely on traditional techniques. However, localization is more than coordinates and semantic mapping; truly knowing where you are requires an understanding of spatio-temporal context. Most *human social environments* (HSEs), where robots are likely to operate proximately with humans, are multi-purpose, where the same physical space is used for many different activities [1]. Humans are able to resolve this problem by taking a context-based approach - they watch, learn, and adapt to change dynamically. However, robots do not yet have this ability.

One approach toward solving these problems is to design situated sensing algorithms that employ *global feature processing*, based on naturalistic data. This is a biologically-inspired approach to sensing, and is how many animals and insects solve tasks ranging from planning and sensing to pick-and-place. In robotics and computer vision, several researchers have utilized these techniques for low-cost sensing and localization tasks (c.f., [2]–[5]). We build upon this previous work, by looking to solve the problem of *context-based perception*.

When robots solve sensing challenges in HSEs, they typically employ content-based algorithms, which assume rigid, static contexts. While this approach may work for structured, predictable environments, it is not suitable for unstructured, dynamic, HSEs. Other fields, such as multimedia, have recognized this shortcoming of a content-based approach, and have seen great strides in solving intractable problems by taking context into account [6].

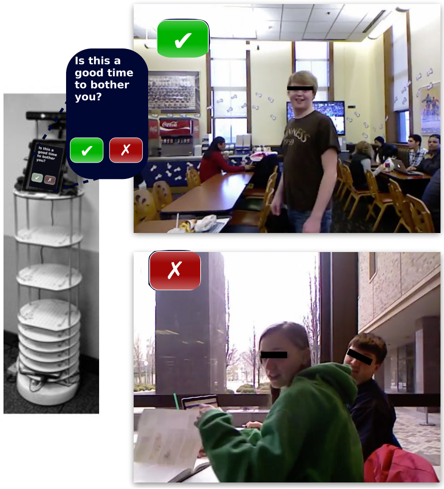

Fig. 1: A robot approached hundreds of participants in real-world settings, asking for permission to interrupt them. An on-board Kinect captured RGB-D and audio data. Two images are shown from the robot's perspective: in the dining hall where a participant agreed to be interrupted, and in the library where participants declined.

We argue similar gains can be achieved in tackling challenging problems in robot perception by taking advantage of context.

This paper presents a foray into context-based perception methods for mobile robots operating in HSEs. Section II-D describes our initial work in this domain, where we applied a top-down situated learning model to teach a robot both to learn situational contexts and behavior propriety [7]. In Section III, we describe significant improvements to this approach through new feature selection techniques and classification algorithms. Section III-D explores alternate multimodal fusion techniques, and shows how introducing audio features earlier in the fusion process can improve classification accuracy. Finally, in Section IV we discuss the importance of these findings for the broader robotics community.

II. BACKGROUND AND INITIAL WORK

A. Defining Context

While definitions of context vary across fields, in computing Dey [8] defines context as “any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.” Since Dey's seminal work, context has surged in popularity in the technology sector, to the point where several have called 2014, “The Year of Context” [9], [10].

[1] The authors are with Department of Computer Science and Engineering, University of Notre Dame, USA {anigam, lriek}@nd.edu

978-1-4799-9993-4/15/$31.00 ©2015 IEEE

Daiss, CEO of one of the largest contextual computing companies, recently suggested the following components are required for realizing "contextual computing experiences": the correct input signals, information pertinent to a person's situation, and the ability to provide the right experience [9]. While Daiss was discussing mobile computing platforms, the same ideas resonate for mobile robotics platforms operating in HSEs. A robot requires: the ability to understand its inputs, correctly identifying the current situation it and its interactants are in, and the ability to perform contextually appropriate actions.

In robotics, substantial work has been performed in sensing and decision-making tied to sensing. However, the majority of this work has been content-driven, as opposed to context-driven, and often assumes rigid, static contexts [1], [6], [11], [12]. In social robotics these assumptions are particularly problematic, because HSEs are often in flux, highly fluid and variable in nature, making traditional control paradigms unsuitable [1]. Furthermore, the work is usually representationally-based. As Dourish [13] suggests, these approaches to context are problematic and likely to fail, because the context problem is not representational, but interactional.

We define the social context for an agent (robot), P, in a given environment, E, as the disjoint union of several subsets: the situational context as a function of E, the social role of P in E, P's cultural norms (irrespective of E); E's cultural norms (irrespective of agents in P); and the social norms for P in E [11], [12].

In this paper, we focus on situational context, which deals with the perception of the environment. We consider situational context to include contextual elements which comprise both the physical and social environment. This includes the physical location, the time of day, observable behaviors in the scene, the social event taking place, etc.

Situational context may be useful for robot perception for two reasons. First, while context is a broad topic to consider in computing, situational context is a sufficiently narrow sub-concept to make headway in studying. Second, there are multiple studies in neuroscience and psychology that support the idea that situational context, or environmental cues, have a large role in the formation and recall of memories, essentially modulating how we perceive and remember the world. If humans can efficiently utilize situational contextual cues to perceive the world, it seems reasonable to also consider exploring them in robotics.

B. Behavioral Propriety

Lieberman writes "People look to the social environment and external context to guide their behavior, particularly when the appropriate course of action is ambiguous or undefined." [14] While this complex set of interactions is difficult to explain neurologically in humans, neuroscientists have also shown in rats that a range of similar background contextual signals representing familiar places activates neural circuitry associated with learned behavioral outputs for that situation [15].

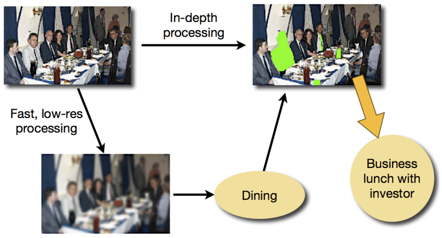

Fig. 2: An example of Bar's model of visual context [16]. In-depth processing such as person and object recognition takes time, while top-down low-resolution processing occurs simultaneously, providing high-level information which is incorporated before the final perception is formed.

For example, rats conditioned to associate a tone with a shock who were then rehabilitated in a different context still showed a fear response in the original context. This work demonstrated that there is a neurological basis for the association between context and behavior.

Propriety, or the state of observing socially accepted behavioral norms, follows from, and is amplified by, this ability we have to link contexts to behaviors. Knowing that we link these concepts, the question which naturally arises is "How?" Bar proposes one answer, which is a neurological model of visual context [16]. In Bar's model, gist-based, thin-sliced, contextual associations are made based on the scene as a whole. *Thin-slicing* is the ability to analyze patterns after only a short period of exposure. Later, this initial sense of context is combined with more sensitive, but computationally intensive, cues such as the appearance of particular objects in the scene. A pictorial representation of this process can be seen in Figure 2.

Much computer vision and robotics research deals with this second aspect of context: the ability to detect and recognize salient objects. However, less work has been done in taking the quicker, top-down approach.

In this paper, we use thin-slicing methods and non-expert features to learn a particular contextual association, an *appropriateness function*, based on real-world data collected from a mobile robot. An appropriateness function defines a mapping from a sensed context to an appropriateness value for some behavior. In our case, the appropriateness value is binary, indicating that either the act of interruption is appropriate or inappropriate.

C. Situated Learning

In this paper, we focus on top-down situated learning with an updatable learning model to learn an appropriateness function. Situated learning, also known as active learning, is a form of supervised learning where labels are acquired by asking a user for them. By learning directly from the environment, we can be certain to acquire valid labels and can more easily move toward a fully autonomous learning robot by pairing this with an updatable learning model.

The ability to learn online is particularly important for helping robots adapt to fluid social environments. Because of the wide range of situations possible in even a simple

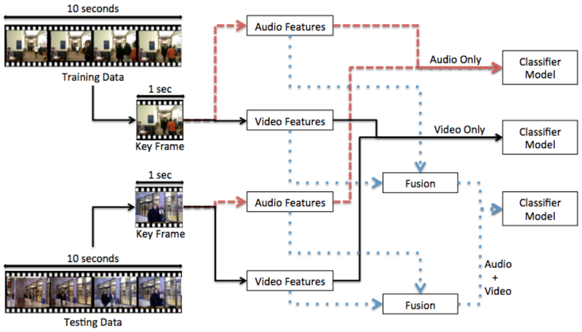

Fig. 3: An illustration of our approach. From our training and testing datasets, we first select a 10*s* frame right before the interaction began, and then extract a keyframe. For each frame, we compute both the audio and video features. For audio-only classification, we trained our model on the audio features of the training dataset and tested them using audio features from the testing set. We also followed this model for video-only classification. For audio-video multimodal fusion, we concatenate the audio and features, then build a classifier model from the training dataset. We then evaluate the model against the testing dataset.

environment, it is impossible to categorically learn an appropriateness function for each one. Furthermore, as the number of contextual elements in a scene increases, the number of relationships between them grows exponentially. Therefore, we need to employ situated learning.

There are a few challenges for this form of learning. When dealing with real-world data, the chance of a large noise-to-signal ratio is high. There are also few existing corpora, so data collection becomes arduous for the initial training phase where algorithm parameters are being tweaked. However, because propriety can only be defined by those within a given context, the positive aspects of situated learning far outweigh the negative.

D. Initial Work

Our initial work aimed to establish the importance of situational context to determine behavioral propriety for robots [7], [12]. We define an *appropriateness function* to determine whether a robot should interrupt a person in a public place, and evaluated it based on contextual cues. The function was validated using data collected from real world, multi-context settings, and used a top-down, situated learning model.

1) Data Collection: We collected naturalistic data from geographically distributed indoor settings around our university campus. Since the ultimate aim for our work is for a robot to be able to conduct itself in real-world settings, it was necessary for our data to reflect the world outside the laboratory.

All data came from GPS-ambiguous (and, in some cases, GPS-denied) locations. We collected data from two multi-use locations - the on-campus student center, and informal locations throughout the campus library, including a cafe and study area called "the fishbowl". Both the student center and library have multiple situational contexts within them: study contexts, dining contexts, and waiting (lobby) contexts.

These three contexts included varied levels of ambient noise and different types of interactions between people. In the study contexts, people were observed reading, using computers, and generally not making much noise. In the dining contexts, people ate meals, talked with friends, and were overall louder compared to the study contexts. In the lobby contexts, people waited for various reasons, walked through the area, and chatted with one another. This context was the loudest of the three.

The data were collected using a modified Turtlebot robot. This is an open-source software and hardware platform consisting of an iRobot Create, a Microsoft Kinect v.1, and an ASUS laptop running the Robot Operating System (ROS) on Ubuntu Linux [7]. The robot was modified in order to stand and sense at human height and interact with participants audio-visually via a tablet interface and portable speakers. These adaptations enabled us to perform multimodal data collection in an ecologically valid way with low-cost hardware.

Fig. 1 shows an example of the setup for the data collection, and a view of the data from the robot's perspective. We collected data from three sources: the Kinect (RGB, depth), the tablet (labels from participants), and a voice recorded mounted to the top of the robot (audio). All data was carefully time-synchronized [17].

Solely the robot's motion was controlled by a remote teleoperator (Wizard). In terms of Wizard production variables [18], the teleoperator had complete control over the robot's physical movements (forward, backward, left, right), and controlled the robot's motion via line-of-sight. The teleoperator also controlled when the robot began its interactive activities, i.e., asking a user if it was a good time to bother them. This decision was accomplished by line-of-sight, and pressing a button to start the application. Once the robot began its Android application, it was fully autonomous in terms of its speech and visual display.

To create a robust data set, we collected data across several days, several months apart. Our data collection occurred at different hours of the day (morning, evening, and night), and during different busy and non-busy periods (e.g., lunchtime, during scheduled class times, etc.). In total, we collected 169 interactions between the robot and participants, split into two sets. *Dataset 1* contained 60 interactions, *Dataset 2* contained 109 interactions.

2) Analysis and Results: Full details of our analysis are described in Hayes et al. [7], but we briefly describe the techniques here to set the stage for our new work. Our approach consisted of two phases. The first phase (learning) involved gathering data from interruptions to train our appropriateness function. The second phase (validation) tested the accuracy of our appropriateness function when the robot was once again placed in these locations. We also attempted to classify the situational context the robot was in (study, dining, or lobby).

Each interruption was represented by 10 seconds of data from both the recorded video and audio prior to the interruption. We used one keyframe per second to represent the visual component of our audio-visual feature vector, and used each second of audio to represent the other. We then used principal component analysis (PCA) to reduce the dimensionality of each vector. (PCA reduces dimensionality in a lossy way and is analogous to the fast, low-dimension processing in Bar's model [19].) We performed classification using 10-fold cross-validation and a linear SVM.

Our initial cross-validation results yielded an accuracy of 87.16% for detecting which context the robot was in (library, study, or dining, chance = 33%), and 80.5% for detecting the appropriateness of interrupting someone (chance = 50%). However, when we tried training on *Dataset 1* and learning on *Dataset 2*, our accuracy dropped to 52.11% and 58.07% respectively. Since the ultimate goal of our work is to build robust, context-based perception and planing algorithms that work online in real-time, we wanted to improve our results. The next section describes how we accomplished this.

III. METHODOLOGY

We ran a series of four experiments to improve our prior results [7]. We were particularly interested in exploring feature selection, classifier suitability, and multimodal fusion. Across all four experiments, we used the aforementioned datasets described in Section II-D. We performed a pre-processing step on the training data *Dataset 1* and testing data *Dataset 2* prior to running our experiments, similar to our previous work. For each of the interactions, we selected a 10*s* time window from both the audio and video data immediately before the robot interacted with the person, and again sampled 1 keyframe per second. This yielded 600 training samples and 1090 testing samples. See Tables I and II.

We engaged in a symmetric validation and comparison method through the use of the three classifiers: Naive Bayes, Support Vector Machines (SVM) and Decision Trees (J48). These are three commonly used classifiers, and are well suited to our data [20]. As before, we classified the context the robot was in (study, dining, or lobby), and the appropriateness of interrupting a person in the space (appropriate or inappropriate).

We ran a total of four experiments. In Experiment 1, we replicated our prior approach to multimodal feature selection, but employed different classifiers. In Experiment 2, we employed more descriptive audio features, unimodaly, to explore their classification success. In Experiment 3, we employed a new video feature descriptor, GIST, unimodally, to explore its classifiation success. Finally, in Experiment 4, we employed multimodal fusion to combine the results of Experiments 2 and 3 to see if overall classification performance improved.

A. Experiment 1: Replication of prior work with new classifiers

In our previous work, described in Sect. II-D, our approach was evaluated using just one classifier: SVM. One of our first steps was to explore classification performance using different SVM kernels and other classifiers. This step would be helpful to understand if our approach is robust and how sensitive it is to other classifiers.

TABLE I: Distribution of training and testing samples across the three different contexts: Study, Dining and Lobby.

	Train	Test
Study	80	100
Dining	150	180
Lobby	370	810
Total	600	1090

We followed the same methodology for feature extraction and feature selection. In the validation stage, we tested our approach using Naive Bayes, Support Vector Machines, and Decision Trees.

1) Feature Extraction: We extracted the same features as in our initial work. For each 1*s* keyframe, we had one video image and one audio sample. From each video image, we extracted the greyscale intensity values. All of these values were concatenated into the feature vector. Similarly, the amplitude values from the audio sample were read and concatenated into the same vector. Thus, early fusion was employed (see Sect. III-D).

2) Feature Selection: We performed feature selection by employing PCA to reduce the dimensionality of the fused feature vector. To decide on the best feature dimension size, we experimented on sizes varying from 50-500 features using cross-validation accuracy as our measurement. We found that 150 features yielded the best accuracy, and thus reduced our dataset to the top 150 features with the highest eigenvalues. After performing the dimensionality reduction, the size of our training dataset was 600 x 150 and testing dataset was 1090 x 150.

3) Results: Three classifiers were employed: Naive Bayes, SVM and Decision Trees (J48). The training data obtained in the previous step was used to train the classification models and tested with the testing dataset.

The cross validation results and the classification results from the testing data can be seen in Tables III and IV respectively. One can see that by changing the SVM kernel, there was an overall improvement in cross-validation accuracies. However, when testing our models on unseen data, we found similar results as in the prior work.

4) Discussion: From this experiment, we found that we can detect both context and appropriateness using generic global features. Features as simple as audio amplitude and grey-value intensity can help in classification better than chance, particularly for classifiers such as Naive Bayes.

However, we still sought to improve these results. In particular, we wanted to explore how audio and video features individually contribute to classification accuracy. We also wanted to explore more descriptive (but still basic) audio and visual features.

B. Experiment 2: Alternate Audio Features

Previously, we found that the fusion of audio and video is helpful in determining situational context and

TABLE II: Appropriateness distribution of training and testing samples.

	Train	Test
Inappropriate	370	770
Appropriate	230	320
Total	600	1090

TABLE III: The cross validation results from Experiment 1, using multimodal fusion on the training dataset.

	Naive Bayes	SVM	Decision Trees
Context	87.83%	98.33%	99 %
Appropriateness	65.33 %	98%	92.33%

appropriateness. However, the individual contribution of the audio alone was not clear. Furthermore, we wanted to explore if more descriptive (though still fast to compute) audio features would yield higher classification accuracies.

1) Feature extraction: Initially, we only used amplitude values, which can be strongly affected by noise and thus give misleading information. Therefore, we chose standard audio features that have been used in the scene understanding literature [21]. These features can be broadly divided into 2 categories: Volume and Frequency.

The volume distribution captures the temporal variation in an audio sample. For each keyframe in our data, we used the Matlab Audio Analysis Library[22] to compute the following volume features:

- **Volume mean**: Average of the amplitude across the audio.
- **Volume standard deviation**: Standard deviation of amplitude for the sample.
- **Volume Dynamic Range (VDR)**: VDR is defined as: *(max(audio) - min(audio)) / max(audio)*
- **Silence ratio**: Time when the volume was below 0.3 times the average volume.

We also extracted the following frequency features:

- **Frequency centroid**: Computes the spectral centroid from the Fourier transform of the audio signal.
- **Frequency bandwidth**: Difference between the maximum and minimum frequency.
- **Feature energy**: The energy of the audio frame.

2) Feature Selection: Next, we sought to establish which of the aforementioned audio features were the most distinctive. Building on prior work [23], we used information gain as a criteria for feature selection. It is defined as follows, where H is the information entropy:

$$InformationGain(Class, Attribute) = \\ H(Class) - H(Class|Attribute) \quad (1)$$

We noted the information gain score for each feature and chose those features which had a positive score. There were a few features that did not contribute significantly to the class, which were discarded.

3) Results: The same three classifiers were used again: Naive Bayes, KNN, and SVM. Using the training data obtained in the previous step, we trained our classification models and tested them with the testing data. The results obtained by using only the audio features without performing feature selection and after applying information gain as a selection metric, can be seen in Tables V and VI (see columns labeled "Audio Only").

TABLE IV: Results from Experiment 1, applying the model learned from the training dataset to the testing dataset.

	Naive Bayes	SVM	Decision Trees
Context	51.65%	54.95%	60.36%
Appropriateness	46.88%	52.29%	54.95%

Audio features alone yielded an 74.31% accuracy for predicting the context using the SVM classifier. The best accuracy for appropriateness using the audio features was 70.27% using decision trees. Even after performing feature selection using information gain, there is only a marginal difference for appropriateness and context prediction. Since initially we had seven features, selecting from those does not improve our performance significantly.

4) Discussion: Our results show that audio features alone are able to accurately predict context when using SVM as a classifier. The seven features we selected (volume and frequency) appear to capture the essence of context well. Interestingly, there was little difference overall with and without feature selection. This may be due to the fact that we selected three features (from seven), which is unlikely to make a big difference.

Overall, audio-alone outperformed our initial results using multimodal fusion. This suggests audio plays an important role in identifying situational context. We next look to see if we can improve our results further with new visual descriptors.

C. Experiment 3: Alternate Video Features

Thus far, we have seen that audio is important for detecting context and appropriateness. We also wanted to explore the importance of video alone. Furthermore, in our previous work we used intensity as a feature, which, like amplitude, can be misleading. Intensity values can vary due to factors such as artificial lighting, camera positions and amount of natural light, none of which are necessarily important in detecting context.

We sought to use features that would make our approach illumination invariant, that would have good discriminative power yet be simple to calculate. Thus, we instead employed the GIST descriptor instead of intensity values. GIST perceives the scene as a representation of dimensions, such as naturalness, openness, roughness, expansion, and ruggedness to capture the spatial structure [24]. GIST has been used extensively in the literature for both scene and object recognition. It is a global descriptor shown to be less affected by illumination and small transitions. It captures the background, and neglects changes in the foreground [24], [25]. Given our biologically-inspired approach (c.f. Bar [16]), GIST is well-suited to our problem.

1) Feature extraction: We applied the GIST descriptor to the video data, and computed the GIST descriptor using code provided by Olivia and Torrabla [24]. The GIST descriptor contained 512 features.

2) Feature Selection: After computing the GIST descriptor, we sought to explore the contribution of these features toward predicting context and appopriateness. As explained in Sect. III-B.2, we wanted to measure the goodness of each feature and so also used information gain to determine which of the 512 features contribute the most to the class. We again only included features who had a positive information gain score.

3) Results: The results can be seen in tables V and VI (column labeled "Video Only"). Table V shows results without feature selection, and Table VI shows results after applying information gain as a feature selection metric. Without feature selection, we found that context is more sensitive to video features compared to appropriateness. We obtained an accuracy of 77.61% for context prediction using SVM. However, the accuracy for appropriateness classification was 54.86% using SVM.

After performing feature selection on the GIST features using information gain, we found that the performance of video features with respect to context and appropriateness changed only marginally. The SVM accuracy on video features increased to 78.07%, only a marginal increase.

4) Discussion: For context, we found that video features performed well compared to audio features, particularly for SVM (77.61% and 78.07% after feature selection respectively). This is understandable, as GIST has been shown to be an effective scene descriptor. However, for appropriateness classification, video-only features did not perform as well as audio-only features. This finding also is unsurprising, as GIST is generally more robust for scene understanding (as opposed to robot social behavior).

D. Experiment 4: Multimodal Fusion

Previously, we explored understanding situational context using unimodal classification. Here, we seek to explore the effects of multimodality. This approach is intuitive, because it is how biological creatures (like humans) quickly solve contextual understanding problems [19].

There are two types of multimodal fusion techniques: early and late [26]. In our work, we employ early fusion, where we fuse the unimodal features into a multimodal feature vector, and then use the multimodal feature vector to train the classifier. In contrast, with late fusion, classifiers are trained independently using the unimodal features, and then unimodal scores are fused together [26]. In our former work we found our data to be better suited for early fusion, so elected to employ it again here (albeit with improved features).

We computed the audio and video feature vectors individually. Using early fusion, we concatenated the two vectors before building our classifier model on the training data. Here, we have seven audio features (volume and frequency related), and 512 video features (GIST descriptor). After concatenating them, we ended up with 600 training samples and 1090 testing samples, each with 519 features.

We performed early fusion after completing feature extraction. We computed the seven audio features and one video feature (GIST descriptor) individually. We concatenated feature vectors and then we applied information gain to select the most informative features. We again only included those which had a positive score.

Then, the 600 instances of training data were used to train the three classifier models. The classifier models were evaluated using the 1090 test instances.

1) Results: Tables V and VI show the results for audio and video multimodal fusion (see "Audio+Video"). For context prediction, SVM yielded 74.95% without feature selection, and 75.13% with feature selection. For appropriateness, SVM yielded the best accuracy at 56.69%, which declined slightly with feature selection (55.13%).

2) Discussion: Fusion appears to perform best for context prediction using SVM after performing feature selection. However, video-alone still appears to outperform fusion for context. This may be attributed to the fact that the video features are more descriptive. For audio and visual fusion, feature selection does not appear to affect the performance of predicting appropriateness.

IV. GENERAL DISCUSSION

Our work shows that it is possible for a robot to understand situational context solely from global audio and video features. To our knowledge, we are the first group to do this. This is an exciting finding for robotics, as robots operating in HSEs need to be able to fluidly assess their situation and make appropriate decisions, particularly in unstructured environments.

We found significant improvement over our prior results [7]. Previously, using multimodal fusion and PCA-based dimensionality reduction, we found an accuracy of 52.11% for appropriateness classification using an SVM. We were able to improve this to 70.64% using only audio features (Naive Bayes, without feature selection). We also obtained an accuracy of 78.07% using video-only features for context prediction, which was encouraging (SVM, with feature selection).

We have explored the use of both unimodal and multimodal features, particularly with using more descriptive features (volume and frequency for audio, and the GIST descriptor for video). We found these features performed better than simply using amplitude and image intensity values in our previous work. We also found that using the new audio features alone without any feature selection was able to correctly predict appropriateness at 70.64% (Naive Bayes, without feature selection). Feature selection itself was not very effective for audio only features because of the low dimensionality. However, using only video features yielded at best a 78.07% accuracy for context (SVM,

TABLE V: Results obtained on the test dataset without feature selection.

	Naive Bayes			SVM			Decision Trees		
	Audio Only	Video Only	Audio + Video	Audio Only	Video Only	Audio + Video	Audio Only	Video Only	Audio + Video
Context	11.28%	69.90%	69.35%	74.31%	77.61%	74.95%	47.70%	65.59%	65.50%
Appro.	70.64%	40.45%	40.91%	70.18%	54.86%	56.69%	70.27%	48.34%	48.62%

TABLE VI: Results obtained on the test dataset using information gain for feature selection.

	Naive Bayes			SVM			Decision Trees		
	Audio Only	Video Only	Audio + Video	Audio Only	Video Only	Audio + Video	Audio Only	Video Only	Audio + Video
Context	11.46%	69.44%	68.89%	74.31%	78.07%	75.13%	45.41%	65.59%	65.59%
Appro.	69.54%	39.26%	39.81%	70.09%	54.12%	55.13%	70.27%	47.06%	47.06%

with feature selection), and 70.64% for appropriateness (Naive Bayes, without feature selection).

While these results are encouraging, there is scope for further improvements. One way to improve accuracy may be to use more descriptive (but still inexpensive to calculate) features. For example, Scale Invariant Feature Transform (SIFT) and its variants are commonly used for recognizing scenes and objects [27]. There are also more comprehensive audio features that may be well-suited to this problem domain (c.f. [21]). Another avenue to explore would be weighted fusion, since there is a skew in the data toward more video than audio features. Since video-only features performed quite well, it might make sense to give it a higher weight during fusion. Finally, it would be worth exploring varying the kernels and optimizations in our classifiers. This may lead to improved results.

In the future, we plan to explore additional contexts, as well as models of spatio-temporal data [12]. Furthermore, we are currently working to adopt our approach to work online on the robot, so it can classifying contexts on the fly and use that to inform its motion.

REFERENCES

[1] L. Riek, "The Social Co-Robotics Problem Space: Six Key Challenges," *In Proceedings of Robotics: Science, and Systems (RSS), Robotics Challenges and Visions*, 2013.

[2] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, 2007.

[3] T. Stone, M. Mangan, P. Ardin, and B. Webb, "Sky segmentation with ultraviolet images can be used for navigation," *Robotics, Science, and Systems (RSS)*, 2014.

[4] R. Iser and F. M. Wahl, "AntSLAM: global map optimization using swarm intelligence," *IEEE ICRA*, 2010.

[5] M. J. Milford and G. F. Wyeth, "Mapping a suburb with a single camera using a biologically inspired SLAM system," *Robotics, IEEE Transactions on*, vol. 24, no. 5, pp. 1038–1053, 2008.

[6] R. Jain and P. Sinha, "Content without context is meaningless," in *Proceedings of the international conference on Multimedia - MM '10*. New York, New York, USA: ACM Press, 2010, p. 1259.

[7] C. J. Hayes, M. F. O'Connor, and L. D. Riek, "Avoiding robot faux pas: using social context to teach robots behavioral propriety," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot interaction (HRI)*. ACM, 2014, pp. 168–169.

[8] A. K. Dey, "Understanding and using context," *Personal and Ubiquitous Computing*, 2001.

[9] M. Panzarino, "Yahoo Girds Its Loins For The Battle Over Your Home Screen," *Tech Crunch*, Jan. 2014.

[10] S. Shank, "Big Idea 2014: The Year of Context," *LinkedIn: Big Ideas 2014*, Dec. 2013.

[11] L. Riek and P. Robinson, "Challenges and Opportunities in Building Socially Intelligent Machines," *IEEE Signal Processing*, 2011.

[12] M. F. O'Connor and L. D. Riek, "Detecting social context: A method for social event classification using naturalistic multimodal data," *In Proc. of the 11th IEEE Int'l Conference and Workshops on Automatic Face and Gesture Recognition (F&G)*, 2015.

[13] P. Dourish, "What we talk about when we talk about context," *Personal and Ubiquitous Computing*, vol. 8, no. 1, pp. 19–30, 2004.

[14] M. D. Lieberman, "Neural bases of situational context effects on social perception," *Soc Cogn Affect Neurosci*, 2006.

[15] R. G. Phillips and J. E. LeDoux, "Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning." *Behavioral neuroscience*, vol. 106, no. 2, pp. 274–285, 1992.

[16] M. Bar, "Visual objects in context." *Nature reviews. Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[17] T. Iqbal and L. D. Riek, "A Method for Automatic Detection of Psychomotor Entrainment," *IEEE T on Affective Computing*, 2015.

[18] L. Riek, "Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines," *Journal of Human Robot Interaction*, vol. 1, no. 1, pp. 119–136, 2012.

[19] M. Bar, "Visual objects in context." *Nature reviews. Neuroscience*, vol. 5, no. 8, pp. 617–29, Aug. 2004. [Online]. Available:

[20] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757 – 1771, March 2004. [Online]. Available:

[21] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," in *Journal of VLSI Signal Processing System*, 1998, pp. 61–79.

[22] T. Giannakopoulos and A. Pikrakis, Eds., *Introduction to Audio Analysis*. Oxford: Academic Press, 2014.

[23] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, 2003.

[24] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.

[25] P. S. Sowjanya and R. Mishra, "Video show boundary detection–comparison of color histogram and gist method," *Journal of Research Engineering and Applied Sciences*, vol. 1, 2012.

[26] C. G. M. Snoek, "Early versus late fusion in semantic video analysis," in *In ACM Multimedia*, 2005, pp. 399–402.

[27] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.