# Multimodal social scenario perception model for initial human-robot interaction

Diego Cardoso Alves[1]
School of Electrical and
Computer Engineering
University of Campinas
Campinas, Brazil
Email: d189729@dac.unicamp.br

Paula Dornhofer Paro Costa
School of Electrical and
Computer Engineering
University of Campinas
Campinas, Brazil
Email: paulad@unicamp.br

***Abstract*—Human-robot interaction imposes many challenges and artificial intelligence researchers are demanded to improve scene perception, social navigation and engagement. Great attention is being dedicated to the development of computer vision and multimodal sensing approaches that are focused on the evolution of social robotic systems and the improvement of social model accuracy. Most recent works related to social robotics rely on the engagement process with a focus on maintaining a previously established conversation. This work brings up the study of initial human-robot interaction contexts, proposing a system that is able to analyze a social scenario through the detection and analysis of persons and surrounding features in a scene. RGB and depth frames, as well as audio data, were used in order to achieve better performance in indoor scene monitoring and human behavior analysis.**

## I. Introduction

Social robotics field has been discussed by several of the human-robot interaction (HRI) themes, demanding researchers to advance navigation, engagement and action decision systems. A central question in social robotics is how to promote a comfortable and engaging interaction between humans and intelligent robots, which are capable of performing tasks by sensing the environment, interacting with external sources and adapting their behaviour [1].

The aforementioned interaction should be able of supporting natural scenarios, without movement or conversation restrictions. Current social robotics applications attempt to achieve this state-of-the-art performance. However, in these applications, even common situations faced by humans characterize a difficult problem in robotics, since it requires the robot to detect persons and their location in the scene, to monitor their gaze, to infer their psychological state and to identify their interaction pattern. Moreover, there is an innate tendency of humans to anthropomorphize surrounding entities [2], especially those that seems to present emotional, sensitive and communicative abilities. As a consequence, robots that do not meet human expectations turn the interaction extremely frustrating.

The HRI research community has shown that recent machine learning and robot vision techniques can improve object detection, person facial and pose recognition, and surrounding analysis. Aspects of long-term conversation are frequently mentioned in the engagement process. Conversely, there is an under-explored field dedicated to the initial interaction analysis between robots and humans, which can be more resilient or resistant to be interrupted, specially by an unknown entity.

The ability to perform scene perception through the extraction of affective individual and group features during a first approximation, can help the robot to handle complex situations and to build a realistic relationship with humans. In this way, consecutive human-robot interactions can reduce false positives and avoid undesired initiatives. Furthermore, the social robotics field can benefit from the integration between initial and continuous engagement in order to complement consolidated studies and applications.

In this context, we need to have alternative strategies that do not wholly rely on longer-term trials but also yield insights that address the initial interaction study. This paper is one such effort since it presents an analysis model of typical social scenes with a different number of persons on it and derives affective labels, that could be used as a suitable interaction choice related to the humans in the scene.

More specifically, our work aims to present two main contributions: a social scene analysis and multi-target detection based on multimodal data; and a labeling approach for initial interaction contexts and human-aware robot behavior.

The methodology is based on macro features, that corresponds to the similar deductions observed by a human during initial social interactions. Some of these perceptions are the location of the group, their interaction level in the scene and receptiveness. Thus, the model can work over unknown (not trained) scenes, since it guarantees that there is no bias related to the number of persons or their intrinsic characteristics.

In order to improve the robustness of the model, we used RGB-D and audio data to increase the model effectiveness while dealing with complex situations in which only specific features were not sufficient or could not be analyzed due to occlusions or image quality limitations.

The figure 1 illustrates some features extracted by the robot-vision system and some significant attributes derived from them that were fundamental to process group interaction metrics. Our model combines these metrics, with distance

[1]This work is related to a Master's dissertation developed by the author.

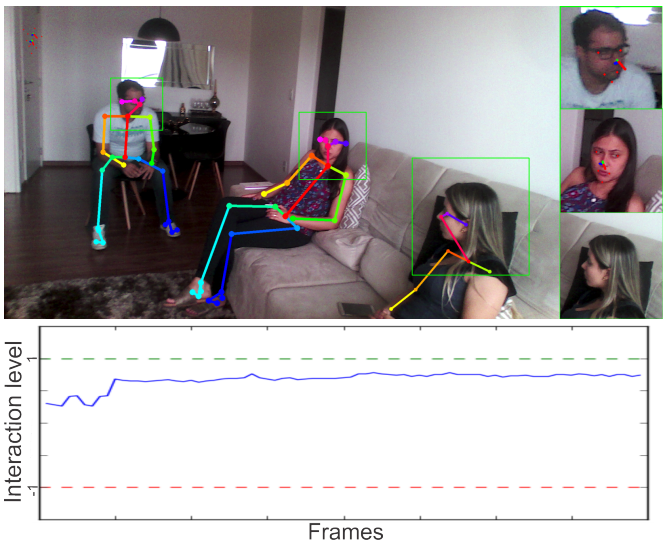


Fig. 1. Social robot scene perception: the interaction intensity graph shows the human-group mean value during the analysis of the social situation.

measurements, pose attributes and audio attributes to correctly detect the affective state of a social situation.

This work is structured as follows. Section II reviews relevant related works, highlighting the existing approaches of human-robot interaction and scene perception adaptation. Our approach for interaction context classification is described in Section III. Section IV brings an overview of the methodology chosen and the techniques used to improve the social robot initial engagement. Afterward, Section V describes the experiments performed, including the alignment between the current results and project goals. Finally, section VI wraps up with project conclusions, mentioning the contributions made and achieved goals. The expected improvements are also presented in this section.

## II. Related Work

Recognize other persons behavior prior to interacting with them is a natural human feeling [3], [4]. There is a great demand to create this capability in the context of social robotics since, in the near future, they will coexist with us in our environments and they will impact our daily routine. Considering this, interest has grown from diverse disciplines including: human-computer interaction [5], psychology [6], artificial intelligence [7]–[9], medicine [4] and safety [10].

### A. Affective Trust

Recent studies are usually focused on methods to keep human-robot engagement, ignoring the analysis of primary social context. Some prior research has identified the relevance of starting an interaction at the right time as one of the most important factors to increase the users affective trust [11]–[14].

In addition, social robots must provide confidence while interacting with humans as a crucial factor to keep an assertive engagement. In [15], the authors discussed that human trust in robots is essential because it directly impacts the results of human-robot interaction.

Regarding the interpersonal trust general concepts, social psychology has studied the main factors that contribute to its efficiency. According to some authors [16], there are two interpersonal aspects mentioned as cognitive and affective trust. Cognitive trust relies on the reliability and capability of a specific party, while affective trust is based on expected responses to the behavior of a party [17].

Existing studies relate the increasing of cognitive trust to the repetition of expected cycles during an interaction, while affective trust is based on the initial impressions [18], [19]. In this way, casual and spontaneous communication can improve user's affective confidence regarding a previous unknown entity, specially when it is a robot.

Applications based on surrounding and emotional state of humans in a scene should have focus on affective trust concepts when leading with initial interactions. Thus, to minimize disturbance and maximize action timing response rates, robot perception should be able to detect the most convenient social group situation.

### B. Human-aware interaction systems

A number of studies related to human-robot interaction are focused on improving long-term conversation, usually during 1-1 conversation [20]. This type of analysis is restricted to contacts already established between the person and the robot, without the human primary intention analysis.

Regarding the use of procedures to detect human behavior, some common features are usually retrieved from the engagement monitoring such as the facial expressions, the gaze behavior and the body position. However, the authors in [21] cite the importance of understanding and recognize unusual behaviors to prompt robot adaptation. The recognition of activities in daily living [22] and the unsupervised learning of temporal sequential actions [23] are some of them.

The studies related to crowd behavior monitoring also have their importance in social perception field, resulting in models that are able to re-plan movements and design collision-free movements. In [24], the authors brought the interaction force concept (attractive and repulsive), which is related to the tendency to keep a distance between individuals and avoid obstacles based on people velocity between frames and their direction in the scene. The work [25] also used different ways to maximize interaction and facilitate person localization in the scene, such as speaker location and gesture mapping.

The understanding of the affective scene state before initiating interaction with human groups is also mentioned in order to improve the social robot reliability. Recent works have used different techniques to understand humans and calculate the interaction forces between robot, human and obstacles.

In [26], some research was done to relate other works, trying to find a pattern to designate the best social situation and interruption labels. The paper brings a lot of ideas on how to model the problem and the most common features used in each sub-field. In [27], the authors used a multilayer perceptron,

having as input features related to speech, head pose and eye gaze, in order to classify seven levels of interruption. Despite of achieving good results, the experiment was restricted to few participants in controlled conditions and poses. Moreover, only RGB and audio data were used, without the distance measures proportioned by the depth sensors.

### C. RGB-D robot perception

RGB-D data has been used to improve human-robot interaction models, bringing advances in social perception and individual features detection. Applications that use this technology frequently performed better than traditional approaches based exclusively on RGB images.

Nowadays, RGB-D cameras have been used in robotic applications [28], expanding possibilities related to object texture information and people localization when compared to traditional optical images.

Current academic works had implemented RGB-D models able to contextualize the social scene situation according to people features with efficiency. In [29], a mobile robot improved navigation and behavior changes with the use of RGB-D data. The system contains manipulation skills and a vast set of tasks, tracking the person location with respect to the camera and making contact with the users in a home environment more assertively.

Applications using RGB-D camera images are constantly compared to approaches that use exclusively depth or color images. A study using default RGB-D settings [30] demonstrated that a robot could imitate human pose actions observed from a human teacher with significant improvement over other works with traditional cameras. In [31], they also explained the depth data importance and its use to develop a framework of real-life scenario for elderly subjects supported by an assistive bathing robot.

## III. Proposed Method

### A. Overview

This work presents a strategy to identify the affective level and interaction state during initial engagement with humans, asserting that the individuals are likely to be interrupted in order to start a conversation or receive a casual greeting. A RGB-D and audio dataset created from scratch was used to improve model performance since it contains recording specifications that guarantee that the attributes of interest were captured appropriately.

The affective scene perception module was designed to be:

- Robust: performs the analysis and classification of the social scene participants based on different situations trained by the model.
- Flexible: portable to run on untrained environments once the dataset used includes different types of rooms with a variable number of persons.
- Intuitive: abstracts low-level feature details and highlights the macro features to simulate a naturalistic way that a person would detect the social situation.

Our first approach to the problem focuses on the classification of the affective scene situation, in terms of how a group of persons is open to start a new interaction with a social robot. The labels represent the most suitable cases regarding human receptiveness:

- Active - An individual or a group of persons want to start an initial interaction or are actively demonstrating interest. In this case, the robot should perform an active interaction.
- Proactive - An individual or a group of persons may be open to starting an interaction but they are not demonstrating interest directly. In this case, the robot could perform a proactive interaction trying a future engagement.
- Passive - An individual or a group of persons do not want to start any initial interaction since they are not demonstrating interest. In this case, the robot must avoid interaction.

Based on these characteristics, our system pipeline is composed of three modules (Figure 2):

- Data collection module - Perform the tasks related to the acquisition and storage of frames and audio information. Intrinsic configurations are set to optimize the recording time and quality.
- Extract, Transform and Load (ETL) module - Responsible for the feature extraction based on raw video frames and audio segmentation; the attributes creation using filters and transformations; the loading of an unified feature vector.
- Classification module - Dataset labeling with the expected annotation for each sample; split into training, validation and testing data; final classification model design and implementation.

### B. Data collection module

The definition of camera and scenes configuration, as well as the recording details, were based on the characteristics in Section III-A. Hence, the scenario selection and the presence of the correct number of people in each sample was important to obtain a generic dataset. We executed various experiments in order to define the optimal parameters and the structure.

We used the Intel Realsense R200 to capture the video frames, which is focused on medium distances and can record RGB-D with a frame rate up to 60 frames per second (fps) [32]. Since this work is not worried about small transitions in terms of milliseconds, the default value considered was 30 fps. Moreover, we only stored significant information, so RGB and depth frames were collected with the maximum resolution of 1920x1080 for RGB frame and 640x480 for depth frames.

The Intel RealSense SDK documentation describes the configuration options to use the R200 camera on different situations. We analyzed each possibility through the open-source tool called cpp-config-ui [33], which is provided by the Intel development kit. Thus, we could study the main camera characteristics and select the best parameters during a real-time frame comparison.

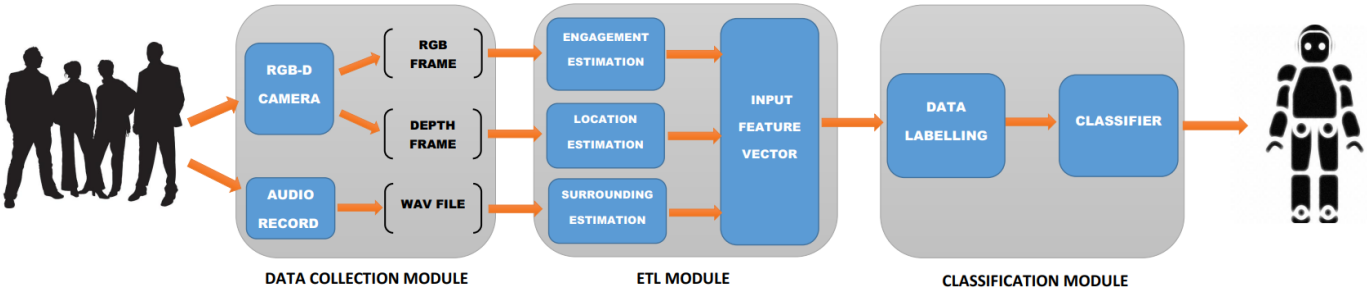


Fig. 2. Social robot scene perception system

The training dataset contains images representing the robot vision perspective of approximately one meter high, avoiding occlusions and maximizing the field of view in common indoor environments. In this way, both standing and sitting persons could be detected properly to improve body and facial tracking quality. Moreover, each sample corresponds to a 7-second RGB-D recording of a scene. This time period was used based on exploratory tests endorsing it as a reasonable time frame to recognize individual and surrounding features, as well as analyze the social cues in the scene.

The samples were collected in dierent indoor environments, with variations in natural luminosity, trying to ensure that the system becomes generic and robust.The recording sessions were carried out in the following simulated environments:

- Social: Public indoor places such as gym and game room.
- Home: Indoor rooms such as living room and TV room.
- Workplace: Indoor rooms such as office and a meeting room.

Given that the dataset was created from scratch, the data labelling of the expected social robot reaction to each scene was performed by three annotators avoiding biased results. The labelling process was based on the raw collected samples and the most frequent value annotation represented the chosen class for the respective scene.

### C. ETL module

Regarding the data complexity, a proof of concept system was developed to define the best features to work with. As a consequence, many attributes based on face detection and body tracking were discarded, since they contained a high percentage of missing values and a low importance to the model.

The audio files collected correspond to the entire samples and include moments with diverse sound background and conversations. The audio features extraction is based on the determination of overall metrics for surrounding, measuring the conversation-level and noise-level present in the scene. The intrinsic characteristics of the audio waveform were analyzed such as the energy, the RMS (Root-Mean-Square), the zero crossing rate and the spectral centroid. We chose these features because they can provide a general context of how the audio signal behaves during each frame.

Regarding the image feature extraction, the OpenFace application [34] was used as an improvement attempt, using techniques that could verify facial reference points, head position, facial action units and eye-gaze direction. In this way, some characteristics were collected and analyzed as:

- Head location: location of each head in the scene with respect to the camera.
- Head pose: head rotation in radians around X,Y,Z axes.
- Facial 2D landmarks: location of 68 landmark points in pixel, corresponding to pairs of (x,y) values.
- Facial Action Units: intensity (from 0 to 5) of 17 action units and the presence (present or absent) of 18 action units.
- Eye region 2D landmarks: location of 2D eye region landmarks in pixels.
- Eye gaze direction: eye gaze direction in radians averaged for both eyes.

Some of these features had to be disregarded since they require a high-quality image with good illumination condition, correct focus and with persons close to the camera. Therefore, some samples with several people located at a medium distance from the camera or hidden by some object in the scene, caused unsatisfactory results during the feature extraction, especially when estimating the region of the eyes and the action units. Even with the use of large images, the detection of each face region considering different head poses is a common difficulty in the field of computer vision. With the resolution of 1920x1080, humans in a profile position prevent the accuracy of the face detection model from being efficient.

Pursuing a better solution for face detection, we used body tracking through the detection of joints, which has been a good approach to verify person location. Hence, our method has used the well-known OpenPose [35] solution to estimate 25 body keypoints with good evaluation when having multiple people on the scene. Consequently, given the results of the body position, the face image could be cropped to the next phase of the process.

In possession of the appropriate cropped face of each subject presented in the social scene and its distance to the camera (provided by the depth sensor), the head pose detection and the facial landmarks were implemented. Thus, we end up with the following intermediate features required for the affective scene detection model:

- RGB-D features: The number of persons inside the scene region, their face bounding boxes, body tracking and distance to the robot. The union of these characteristics contributes as additional context information regarding each individual.
- Facial Analysis: Location of 2D facial landmarks and pose features (head location with respect to the camera and radian rotation), for each subject recognized.
- Audio features: The audio signal energy and RMS, the zero crossing rate and the spectral centroid.

The affective scene detection model must be generic to classify situations, not depending on the number of people and their individual characteristics. Based on this, the attributes were created considering the main goal of mapping the different types and levels of interaction between individuals and their relationship with the robot presence.

The intermediate features previously described were used as input and the data modeling methodology applied some transformations to derive some representative attributes. We also reused some features in the final dataset that were already suitable for the entire environment such as the number of persons.

Firstly, the head pose was estimated using the facial landmarks and by fitting them to a generic 3D face model with use of Levenberg-Marquardt Optimization [36] of Direct Linear Transform [37] approach. This method uses the idea of getting 3D points in world coordinates and transforming to 3D points in camera coordinates, with the premise that some variables are already known. Some parameters were considered in order to perform this estimation:

- Intrinsic parameters of the camera: The camera was considered calibrated, the focal length approximated by the width of the image in pixels, the optical center by the center of the image and assumed that radial distortion does not exist.
- 3D landmarks locations of the same points: The 3D points location correspondent to the respective 2D feature points. These values were predefined based on the arbitrary reference frame and represented in world coordinates.

In possession of the head pose angles named pitch (transverse axis), yaw (vertical axis) and roll (longitudinal axis), the interaction intensity between scene participants could be calculated based on the following descriptive macro attributes:

- Human-group interaction level: Measures the intensity of the engagement between a group of humans presented in the scene, ranging from -1 to +1.
- Human-robot interaction level: Measures the intensity of the interaction interest between a group of humans and the social robot presented in the scene, ranging from 0 to +1.

The notion of parallelism or similarity between vectors was based on the cosines law by using their sum and difference calculation. The intensities derived from human pose make use of these concepts to obtain a relevant metric.

The human-group interaction level estimation was based on the mean of derived principal axis intensities $I_{yg}$, $I_{pg}$ and $I_{rg}$ which represents respectively the yaw, pitch and roll group engagement intensities. These values were calculated using the following equations, assuming $N$ the number of persons in the scene and $Y_i$, $P_i$, $R_i$ representing respectively the yaw, pitch and roll relative angles to the human in analysis:

$$I_{yg} = -\frac{\sum_{i=1}^{N-1} cos(Y_i - Y_{i+1})}{N-1}, \tag{1}$$

$$I_{pg} = \frac{\sum_{i=1}^{N-1} cos(P_i - P_{i+1})}{N-1}, \tag{2}$$

$$I_{rg} = \frac{\sum_{i=1}^{N-1} cos(R_i - R_{i+1})}{N-1}. \tag{3}$$

The human-robot interaction level was calculated using the mean of derived principal axis intensities $I_{yh}$, $I_{ph}$ and $I_{rh}$ which represents respectively the yaw, pitch and roll group interest in interact with the social robot. The following equations describes the estimation basis:

$$I_{yh} = \frac{\sum_{i=1}^{N} cos(Y_i)}{N}, \tag{4}$$

$$I_{ph} = \frac{\sum_{i=1}^{N} cos(P_i)}{N}, \tag{5}$$

$$I_{rh} = \frac{\sum_{i=1}^{N} cos(R_i)}{N}. \tag{6}$$

Based on these attributes and some information derived from intermediate features, the final unified feature vector loaded into the classifier contains:

- Number of persons in the scene;
- Relative distance from the closest person;
- Rotation angles yaw, pitch and roll from the closest person;
- Human-group interaction level based on yaw, pitch and roll angles;
- Human-robot interaction level based on yaw, pitch and roll angles;
- Median, Amplitude and Standard deviation of the audio energy.
- Median, Amplitude and Standard deviation of the audio RMS.
- Median, Amplitude and Standard deviation of the audio zero crossing rate.
- Median, Amplitude and Standard deviation of the audio spectral centroid.

### D. Classification module

Since the main model goal was to achieve, with satisfactory results, an affective classification of the scene being able to provide a robust social situation detection, we used different machine learning algorithms with tuning steps such as grid-search and cross-validation. Due to the existence of four-dimensional information, the spatiotemporal domain was considered and consecutive frames had to be analyzed as a sample sequence while designing the neural network structure. However, we also implemented the traditional classification algorithms such as SVM (Support Vector Machines) and LDA (Linear Discriminant Analysis), in order to compare the frame-by-frame baseline result and to evaluate the developed methodology.

Our methodology consisted in using the combination of three neural networks of the same architecture to designate the final output, focused on the development of a bagging ensemble, in which each learner corresponded to a classifier based on a specific time frame. The MLP (Multi-Layer Perceptron) and the ELM (Extreme Learning Machine) neural networks were used as classifiers during the experiments.

Since our training dataset contained short-time samples, we decided to split the frames into only three parts. First, we created partial datasets based on these sub-samples so that each new sample contains part of the entire information but with the same output label. Then, we fit a neural network classifier for each of these samples. Finally, we aggregated them such that we retrieved the average of their outputs probabilities, obtaining an ensemble model that analyzes the social scene in dierent moments in time and assigns an unique interaction strategy (soft-voting technique).

## IV. EVALUATION AND EXPERIMENTS

In this section, we present metrics obtained from different algorithms during cross-validation results. As mentioned in section III-A, a classifier was trained to predict the labels active, proactive and passive.

Our approach took 70% of the dataset as training data, 20% as validation data and the remaining 10% as test data, in order to potentiate the evaluation metric during unseen situations. In order to have a model less sensitive to the scale of features, the standardization method was applied.

In the beginning, we tested the Support Vector Machine (SVM) multi-class classifier, that designates the class with the greatest margin from other classes as being the correct result. The weak version achieved the precision of 72.4% while the optimized 73.3%. Thereat, we used the Linear Discriminant Analysis (LDA) without tuned hyper-parameters (the classifier has few adjustable options), obtaining 72% as a result.

Later, because the previous classifiers did not take into account the spatiotemporal domain, we used neural networks having as input partial vectors of stacked features. Thus, they could indirectly assimilate the entire sample as a video sequence using their internal characteristics after the application of an ensemble methodology.

Based on this, we first applied the ELM model, reaching 74.2% accuracy. As the number of macro attributes enabled a heavier training, we also used the MLP, which in this case resulted in the best solution compared to the others (Table I): an accuracy of 75.7% with the three-layer configuration and 500 as the number of maximum iterations. Since the MLP model took only some extra training time to increase 1.5 % of precision, we chose it as final classifier.

Regarding the advantages of using multimodal data in our application, the Table II shows the results obtained from the comparison methodology that consisted of the training of three models. These models considered different feature sets, following the same definition of cross-validation defined previously.

TABLE I
CLASSIFIERS PERFORMANCE

| Classifier | Precision | Recall | F1-Score |
|---|---|---|---|
| SVM default version | 72.40% | 71.20% | 71.20% |
| SVM grid-search version | 73.30% | 73.30% | 73.30% |
| LDA | 72.00% | 71.50% | 71.70% |
| ELM | 74.20% | 74.10% | 74.10% |
| MLP | 75.70% | 75.20% | 75.45% |

TABLE II
COMPARISON RESULT AMONG FEATURE SETS.

| Feature Set | Precision | Recall | F1-Score |
|---|---|---|---|
| Audio | 52.80% | 52.30% | 52.55% |
| RGB-D | 71.20% | 70.80% | 71.00% |
| Multimodal | 75.70% | 75.20% | 75.45% |

## V. CONCLUSIONS

In this paper, we have described the social robotic system capable of classifying the affective scene based on multi-person interactions and contextual cues. Our current results show that the system accuracy was up to 75.70% at detecting the correct label on initial interaction situations. We also demonstrated that the analysis of each sample as a sequence enhances the accuracy, enabling the correlation between frames.

The multimodal data analysis of the scene and people features increased human-aware robot perception in cases of complex scenarios and short time of reaction. The raw features study and the modeling of novel attributes related to engagement intensities were feasible due to the use of a multimodal dataset created from scratch.

The future improvements of this work would include the creation of additional audio attributes to extract relevant information related to the detection of conversation and noise in the environment. In addition, a discrepancy value detection module will be created to filter out the erroneous intensities calculated across dierent frames, reducing the number of false positives.

## REFERENCES


[1] I. O. for Standardization, "Robots and robotic devices - vocabulary," *ISO/TC 299 Robotics*, vol. 2, Mar. 2012.

[2] M. Hutson, *The 7 Laws of Magical Thinking: How Irrational Beliefs Keep Us Happy, Healthy and Sane*. Plume, Jan. 2012.

[3] J. Fogarty, S. Hudson, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee, and J. Yang, "Predicting human interruptibility with sensors," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 1, pp. 119–146, 2005.

[4] J. Rivera, "A socio-technical systems approach to studying interruptions: Understanding the interrupter's perspective," *Applied ergonomics*, vol. 45, no. 3, pp. 747–756, 2014.

[5] D. McFarlane and K. Latorella, "The scope and importance of human interruption in human-computer interaction design," *Human-Computer Interaction*, vol. 17, no. 1, pp. 1–61, 2002.

[6] G. Miller, "The smartphone psychology manifesto," *Perspectives on psychological science*, vol. 7, no. 3, pp. 221–237, 2012.

[7] C. Roda and J. Thomas, "Attention aware systems: Theories, applications, and research agenda," *Computers in Human Behavior*, vol. 22, no. 4, pp. 557–587, 2006.

[8] A. Campbell and T. Choudhury, "From smart to cognitive phones," *IEEE Pervasive Computing*, vol. 11, no. 3, pp. 7–11, 2012.

[9] V. Pejovic and M. Musolesi, "Anticipatory mobile computing: A survey of the state of the art and research challenges," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 47, 2015.

[10] S. Kim, J. Chun, and A. Dey, "Sensors know when to interrupt you in the car: Detecting driver interruptibility through monitoring of peripheral interactions," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 487–496.

[11] J. Ho and S. Intille, "Using context-aware computing to reduce the perceived burden of interruptions from mobile devices," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 909–918.

[12] N. Lathia, K. Rachuri, C. Mascolo, and P. Rentfrow, "Contextual dissonance: Design bias in sensor-based experience sampling methods," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 183–192.

[13] R. Harr and V. Kaptelinin, "Interrupting or not: exploring the effect of social context on interrupters decision making," in *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*. ACM, 2012, pp. 707–710.

[14] S. Iqbal and E. Horvitz, "Notifications and awareness: a field study of alert usage and preferences," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010, pp. 27–30.

[15] D. Billings, K. Schaefer, J. Chen, and P. Hancock, "Human-robot interaction: developing trust in robots," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 2012, pp. 109–110.

[16] J. Lewis and A. Weigert, "Trust as a social reality," *Social forces*, vol. 63, no. 4, pp. 967–985, 1985.

[17] J. Rempel, J. Holmes, and M. Zanna, "Trust in close relationships." *Journal of personality and social psychology*, vol. 49, no. 1, p. 95, 1985.

[18] P. Hancock, D. Billings, K. Schaefer, J. Chen, E. Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, pp. 517–527, 10 2011.

[19] K. Schaefer, T. Sanders, R. Yordon, D. Billings, and P. Hancock, "Classification of robot form: Factors predicting perceived trustworthiness," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1, pp. 1548–1552, 2012.

[20] M. Ahmad, O. Mubin, and J. Orlando, "A systematic review of adaptivity in human-robot interaction," *Multidisciplinary Digital Publishing Institute*, vol. 1, pp. 1–14, jul 2017.

[21] P. Caleb, S. Dogramadzi, A. Huijnen, and H. Heuvel, "Exploiting ability for human adaptation to facilitate improved human-robot interaction and acceptance," *The Information Society*, vol. 34, no. 3, pp. 153–165, 2018.

[22] H. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[23] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "Watch-n-patch: Unsupervised understanding of actions and relations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4362–4370.

[24] W. Samsudin and K. Ghazali, "Crowd behavior monitoring using self-adaptive social force model," *Mekatronika*, vol. 1, no. 1, pp. 64–72, 2019.

[25] A. Tsiami, P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6568–6572.

[26] L. Turner, S. Allen, and R. Whitaker, "Interruptibility prediction for ubiquitous systems: conventions and new directions from a growing field," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2015, pp. 801–812.

[27] O. Palinko, K. Ogawa, Y. Yoshikawa, and H. Ishiguro, "How should a robot interrupt a conversation between multiple humans," in *International Conference on Social Robotics*. Springer, 2018, pp. 149–159.

[28] E. Lachat, H. Macher, T. Landes, and P. Grussenmeyer, "Assessment and calibration of a rgb-d camera (kinect v2 sensor) towards a potential use for close-range 3d modeling," *Remote Sensing*, vol. 7, pp. 13 070–13 097, oct 2015.

[29] P. Puente, M. Bajones, C. Reuther, D. Wolf, D. Fischinger, and M. Vincze, "Robot navigation in domestic environments: Experiences using rgb-d sensors in real homes," *Journal of Intelligent and Robotic Systems*, jun 2018.

[30] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, mar 2018.

[31] N. Zlatintsi, I. Rodomagoulakis, P. Koutras, A. Dometios, V. Pitsikalis, C. Tzafestas, and P. Maragos, "Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3171–3175, apr 2018.

[32] *Intel Realsense camera R200 datasheet*, Intel RealSense Technology, 2017.

[33] I. Realsense. (2018, jul) Librealsense. Github. [Online]. Available: https://github.com/IntelRealSense/librealsense/blob/v1.12.1/examples/cpp-config-ui.cpp

[34] B. Tadas, Z. Amir, C. Yao, and M. Louis-Philippe, "Openface 2.0: Facial behavior analysis toolkit," *IEEE International Conference on Automatic Face and Gesture Recognition*, may 2018.

[35] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, apr 2017.

[36] S. Roweis, "Levenberg-marquardt optimization," *University Of Toronto*, 1996.

[37] Y. I. Abdel-Aziz and H. M. Karara, "Direct linear transformation from comparator coordinates into object space coordinates in closerange photogrammetry," *Symposium on CloseRange Photogrammetry*, 1971.