Toward an Interactive Robot Docent: Estimating Museum Visitors' Comfort Level with Art

Ruikun Luo*†, Sabrina Benge‡, Natalie Vasher§, Grace VanderVliet§, John Turner§,
Maani Ghaffari*, and X. Jessie Yang†*
*Robotics Institute, University of Michigan, Ann Arbor, MI, US
†Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, US
‡Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, US
§University of Michigan Museum of Art, Ann Arbor, MI, US

***Abstract*—In museum education, a human docent inspires visitors by asking questions and encouraging active participation. However, museum visitors are sometimes shy to interact with and to ask questions to the human docent. We aim to build a robot docent that is able to interact with visitors to stimulate curiosity, imagination and individual expression. The present study focuses on the problem of estimating museum visitors' comfort level with art. Using the Amazon Mechanical Turk, we conducted a human-subject experiment with 215 participants. Each participant filled in a demographic survey and answered 6 questions about 3 art objects. Three museum education experts reviewed the answers and labeled each participant to be with a high or low level of comfort with art. With the three experts' classification, we found a Fleiss' Kappa score of 0.70 which indicated a substantial agreement. We used Logistic Regression to classify each participant's comfort level by analyzing their responses to the 6 questions. We tested two types of features - bag of words and word embeddings and two methods for extracting features - treating the 6 answers as one material and treating the 6 answers separately and concatenating feature vectors together. Our results show that both features can achieve an accuracy ranging from 67% to 76%.**

I. Introduction

Modern museums present exhibitions to stimulate visitor curiosity, imagination and individual expression, commonly aided by a docent [1]. The docent asks questions and encourages the active participation of the visitor. The interaction requires the docent to estimate the visitor's comfort level with the artwork in order to engage in a friendly and invigorating conversation. However, with human docents, visitors may feel uncomfortable to ask questions or to express themselves due to the preconceived notion of museums being exclusive to the wealthy and formally educated [4]. In collaboration with the University of Michigan Museum of Art (UMMA), we aim to design an interactive robot decent that is able to estimate a visitor's comfort level with art and in response to the comfort level, guide them adaptively. We define the comfort level with art as a visitor's ability to interpret, negotiate, and make meaning from an artwork.

Deploying robots in museums is not a new concept. However, prior research has focused on robot visual perception and trajectory planning [2, 11, 10], and human-like body movement and gesture design [5, 12]. For instance, Xia et al. [11], Chella and Macaluso [2] implemented visual cues to track and navigate the area inside a museum. Xia et al. [11] used a camera to detect marked objects in the world to plan and execute a trajectory based on a PID control algorithm. Chella and Macaluso [2] developed a movement algorithm with a camera and used a 2D simulation of the world to match with the environment to detect where the robot was and what items should be in the current location. Thrun et al. [10] deployed the robot Minerva within a museum that used probabilistic controls for navigation. Some researchers have also looked into robot gesture and movement design. The robot deployed at the Osaka Science Museum used various human-like movements such as shaking hands and leaning forward to engage the visitor and hold their attention [9]. Ghosh and Kuzuoka [5] dived deep into the body language element of maintaining engagement by means of eye contact and upper-body orientation throughout an interaction to engage and disengage with a visitor when appropriate.

There are limited amount of studies investigating verbal conversations between robots and museum visitors, and they were mostly limited to simple commands or scripted conversations [12]. Yamazaki et al. [12] implemented both verbal and non-verbal interactions with a robot in a museum including altering the robot language, reading visitors' body language, and analyzing visitors' speech answers to questions. The robot analyzed user expression and head direction to estimate visitors' interest in the conversation. In the experiment, the robot would gesture to a painting and ask questions with concrete answers such as what is the name of the painting. If the user correctly answered, the robot would show a positive body language such as nodding.

Our project focuses on verbal communication between visitors and the robot. Instead of relying on preset concrete answers like Yamazaki et al. [12]'s approach, we take a new approach to estimate visitors' comfort level with art by analyzing their answers to questions that are more natural and expressive. This paper presents our work in estimating visitors' comfort level with art based on their answers to the preset questions. Using the Amazon Mechanical Turk, we conducted a human-subject experiment with 215 participants. Each participant answered 6 questions about 3 art objects. The questions were screened and labeled by three experts in museum education. With the three experts' classification,

we found a Fleiss' Kappa score of 0.70 which indicated a substantial agreement. The final labels were determined by a voting mechanism. We tested the bag of words and word embedding features and two methods for extracting features from the 6 answers, i.e. treating the 6 answers separately or together. We used logistic regression as the classification algorithm. We also varied the feature dimensions to be 50, 100, 200 and 300. Experimental results show that both features can achieve accuracy ranging from 67% to 76%.

The remaining of this paper is organized as follows. Section II describes the method used in collecting human-subject data and in annotating the data by three museum education experts, and the Natural language processing(NLP) methods used to estimate visitors' comfort level with art. Section III presents the results and discusses the process of selection and additional methods attempted in NLP algorithms. Section IV summarizes our findings and Section V presents future work using NLP in museum robotics.

II. METHOD

In the present study, we classified visitors' comfort level with art into two levels, i.e experienced visitor and amateur, by analyzing their responses to 6 preset questions.

A. Data Collection, Annotation, and Adjudication

Together with two subject matter experts from UMMA, we selected three art objects for the present study: *Location Plan*, *Untitled Cube*, and *Nydia*. Also, we designed an online survey consisting of a demographic survey and 6 questions about the 3 art objects. The answers of the 6 questions were used to annotate the data by the experts, i.e, classify a person into high or low comfort level with art. Figure 1 shows the three selected art objects. The questions for the art objects were:

- *Location Plan* by Terry Winters:
  1) How would you describe these lines?
  2) Why might the rectangles be placed in these positions?
- *Untitled Cube* by Alvin D.Loving:
  1) What shapes do you see?
  2) How does the artist use color in this work?
- *Nydia* by Randolph Rogers:
  1) What do you think is happening in this sculpture?
  2) This sculpture is made of marble - why do you think the artist used this material?

We used the Amazon Mechanical Turk platform for data collection. We received 250 responses in total and 215 of them were considered valid answers.

Three experts in art museum education from UMMA first annotated the data set and classified the participants into high, medium and low comfort level. To measure agreement between the experts, we implemented Fleiss' Kappa and found a Kappa score of 0.5754 [3]. The Fleiss' Kappa measures agreement compared to agreement by chance. Therefore, by the commonly used scoring table from Landis and Koch [7], the experts were in moderate agreement (0.41-0.60). We used the voting mechanism to determine the final labels for every participant. 103 participants were labeled as low comfort level with art, 74 participants were labeled as medium comfort level and 35 high comfort level. Considering the small data set we had and the imbalanced data set i.e, we had only 35 participants labeled as high comfort level with art; we considered people with medium or high comfort level as experienced visitors and people with low comfort level as amateurs. Therefore our new Fleiss' Kappa score was 0.6957 which indicated a substantial agreement (0.61-0.80). As a result, the dataset consisted of 112 high (experienced visitor) comfort level with art and 103 low (amateur) comfort level.

Below shows an example of participants with high and low comfort level with art. For the question "How does the artist use color in this work?", a participant with high comfort level with art answered "The artist uses color to make each nook and cranny pop out some [some], adding 3D emphasis to the image. Color is also used to emphasize the square on the top, so it doesn't seem as plain." A participant with low comfort level with art answered "Very nice.".

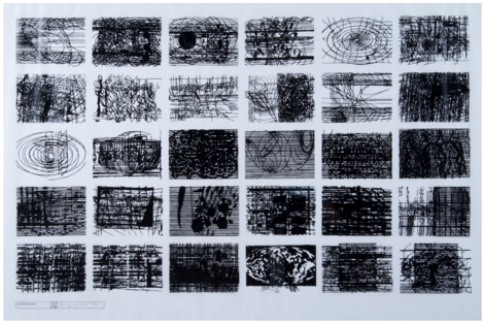
B. Algorithm

We used logistic regression for classification. We tested two types of features (i.e. bag of words and word embedding) and two methods for extracting features (i.e. treating all answers as one or differently). For word embedding, we used the mean of the embeddings of words in a sentence as the feature for a sentence [6] and used the pre-trained word embeddings proposed in [8].

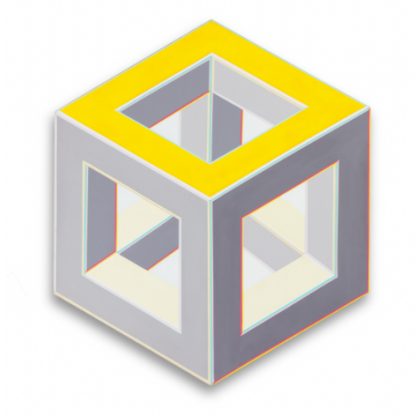
For each participant, let $X = \{x_1, x_2, ..., x_6\}$ represent the answers for the questions, where x_i represents the answer for the i th question. For bag of words features, we built up the vocabulary using the entire the data set and selected the key words based on a given dimension and the word frequency. $g(Y)$ is the feature extraction function for bag of words where Y represents a set of words. For the word embedding, we have $f(Y) = \frac{\sum_j^N \text{v2w}(w_j)}{N}$, where w_j is the word in Y, N is the number of words in Y and v2w() is the word2vec embedding.

We also tested two different ways for extracting features from the answers. The first method treated all the 6 answers as one answer, where we had $Y = X$ and the features for bag of words were $g(Y)$ and the features for the word embedding were $f(Y)$. For clarification purpose, we denoted this method as "one". The second method treated the 6 answers separately and we computed a vector for each answer and concatenated them together. Thus the features for the bag of words were $[g(x_1); g(x_2); ...g(x_6)]$ and the features for word embedding were $[f(x_1); f(x_2); ...; f(x_6)]$. We denoted this method as "concatenate".

We used pre-trained word embedding "Glove-wiki-gigaword" from [8] and tested different dimensions of the word features, i.e, 50, 100, 200 and 300.



(a)



(b)

(c)

Fig. 1: Three selected art objects. (a) Location Plan, Terry Winters (b) Untitled Cube, Alvin D.Loving (c) Nydia, Randolph Rogers.

TABLE I: Performance for different features and different numbers of word feature dimensions. "one": Treat all the 6 answers as one answer. "concatenate": Treat the 6 answers separately, compute a vector for each answer and concatenate them together.

Word Feature Dimension	Feature		Accuracy	Precision	Recall	F1
50	one	BoW	0.712 ± 0.004	0.716 ± 0.004	0.715 ± 0.004	0.711 ± 0.004
		Word2vec	0.679 ± 0.004	0.691 ± 0.005	0.678 ± 0.004	0.672 ± 0.004
	concatenate	BoW	0.720 ± 0.004	0.724 ± 0.004	0.723 ± 0.004	0.719 ± 0.004
		Word2vec	0.715 ± 0.004	0.726 ± 0.004	0.713 ± 0.004	0.708 ± 0.004
100	one	BoW	0.720 ± 0.004	0.725 ± 0.004	0.724 ± 0.004	0.719 ± 0.004
		Word2vec	0.709 ± 0.004	0.722 ± 0.004	0.708 ± 0.004	0.702 ± 0.004
	concatenate	BoW	0.726 ± 0.004	0.731 ± 0.004	0.730 ± 0.004	0.725 ± 0.004
		Word2vec	0.730 ± 0.004	0.745 ± 0.004	0.727 ± 0.004	0.722 ± 0.004
200	one	BoW	0.759 ± 0.004	0.764 ± 0.004	0.763 ± 0.004	0.758 ± 0.004
		Word2vec	0.711 ± 0.004	0.728 ± 0.004	0.709 ± 0.004	0.703 ± 0.004
	concatenate	BoW	0.751 ± 0.004	0.757 ± 0.004	0.755 ± 0.004	0.750 ± 0.004
		Word2vec	0.729 ± 0.004	0.745 ± 0.004	0.726 ± 0.004	0.721 ± 0.004
300	one	BoW	$\mathbf{0.765 \pm 0.004}$	$\mathbf{0.771 \pm 0.003}$	$\mathbf{0.769 \pm 0.003}$	$\mathbf{0.763 \pm 0.004}$
		Word2vec	0.709 ± 0.004	0.728 ± 0.004	0.706 ± 0.004	0.699 ± 0.004
	concatenate	BoW	0.759 ± 0.004	0.766 ± 0.003	0.764 ± 0.004	0.758 ± 0.004
		Word2vec	0.726 ± 0.004	0.744 ± 0.004	0.722 ± 0.004	0.716 ± 0.004

III. Results and Discussion

We run 100 holdouts for the two types of features and different numbers of word feature dimensions, i.e. 50, 100, 200 and 300. In each run of holdout, we randomly selected 130 participants as training data and the remaining 85 participants as testing data.

Table I shows the performance for the different features and different numbers of word feature dimensions, and the top performance is in bold. Overall, the bag of words features outperform the word embedding features for different word feature dimensions and different ways to extract features. Given the small dataset, we cannot train or fine-tune the word embeddings based on the collected data. The pre-trained word embeddings on Wikipedia data set may not represent the art related context well. Higher dimension leads to better performance and there is more improvement when feature dimension increases from 50 to 200 than from 200 to 300. Overall, "one" outperforms "concatenate" for different feature dimensions and different features. One possible reason is that some of the answers are extremely short and thus none of the

words in those answers are in the word embedding vocabulary which results in empty feature vectors for these answers. In such cases, we used zero vector for these answers. This problem is more severe for the "concatenate" method. There are 34 participants with at least one out of the 6 answers resulting in zero vectors. Further research is needed to deal with such a problem and to investigate how to utilize shorter responses from the visitors for estimating their comfort level with art and how to encourage visitors to be more expressive.

IV. Conclusion and Future Work

In this paper, we focused on the problem of estimating human visitors' comfort level with art by analyzing their responses to pre-set questions about certain art objects. We conducted a human-subject experiment and collected 215 valid responses from the Amazon Mechanical Turk. Three museum education experts labeled the participants to be with high or low comfort level with art. We used Logistic Regression to classify the participants into high and low comfort levels. We tested two different types of features and two ways to extract features and different word feature dimensions. The result showed that the bag of words features and treating the answers as one feature vector performs best with larger word feature dimensions, i.e 200 and 300, showing a maximum accuracy of 76%.

With the classification algorithm initially developed and tested, future work includes testing with a real robot, gathering more data from museum visitors, and expanding the annotation categories. In our next study, a NAO robot will be programmed as the robot docent to interact with a visitor through verbal conversation using google speech-to-text and DialogFlow. The visitor will be asked the same questions about the three art objects and we will be expanding our data set in a real environment. The new dataset will be annotated and the algorithm will be tested to further validate its usage in museum robot applications. Further design solutions will also be explored to deal with misclassification.

References

[1] Daniel H. Bowen, Jay P. Greene, and Brian Kisida. Learning to think critically: A visual art experiment. *Educational Researcher*, 43(1):37–44, 2014.

[2] Antonio Chella and Irene Macaluso. The perception loop in cicerobot, a museum guide robot. *Neurocomputing*, 72(4):760 – 766, 2009.

[3] Brian G. Dates and Jason E. King. Spss algorithms for bootstrapping and jackknifing, 2008. URL http://www.ccitonline.org/jking/homepage/interrater.html.

[4] Darby English. Don't be intimidated by museums. they belong to everyone. *The Guardian*, May 2015. URL https://www.theguardian.com/commentisfree/2015/may/31/museums-not-white-spaces-belong-everyone.

[5] Madhumita Ghosh and Hideaki Kuzuoka. An ethnomethodological study of a museum guide robot's attempt at engagement and disengagement. *J. Robotics*, 2014:876439:1–876439:20, 2014.

[6] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. Evorus: A crowd-powered conversational assistant built to automate itself over time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 295. ACM, 2018.

[7] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529310.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[9] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Interactive humanoid robots for a science museum. *IEEE Intelligent Systems*, 22(2):25–32, 2007.

[10] Sebastian Thrun, Michael Beetz, Maren Bennewitz, Wolfram Burgard, Armin Cremers, Frank Dellaert, Dieter Fox, Dirk Hhnel, Charles R. Rosenberg, Nicholas Roy, Jamieson Schulte, and Dirk Schulz. Probabilistic algorithms and the interactive museum tour-guide robot minerva. *The International Journal of Robotics Research*, 19:972–999, 11 2000.

[11] Wen T. Xia, Yan Y. Wang, Zhi G. Huang, Hao Guan, and Ping C. Li. Trajectory control of museum commentary robot based on machine vision. *Applied Mechanics and Materials*, 615:145–148, 08 2014.

[12] Keiichi Yamazaki, Akiko Yamazaki, Mai Okada, Yoshinori Kuno, Yoshinori Kobayashi, Yosuke Hoshi, Karola Pitsch, Paul Luff, Dirk vom Lehn, and Christian Heath. Revealing gauguin: Engaging visitors in robot guide's explanation in an art museum. pages 1437–1446, 04 2009.