Towards Informing an Intuitive Mission Planning Interface for Autonomous Multi-Asset Teams via Image Descriptions

Lisa R. Le Vie[1§], Mary Carolyn Last[2¤], Bryan B. Barrows[3§], B. Danette Allen[4§]
[§]*NASA Langley Research Center, Hampton, VA, 23681, USA*
[¤]*Analytical Mechanics Associates, Inc, Hampton, VA, 23666, USA*

Establishing a basis for certification of autonomous systems using trust and trustworthiness is the focus of Autonomy Teaming and TRAjectories for Complex Trusted Operational Reliability (ATTRACTOR). The Human-Machine Interface (HMI) team is working to capture and utilize the multitude of ways in which humans are already comfortable communicating mission goals and translate that into an intuitive mission planning interface. Several input/output modalities (speech/audio, typing/text, touch, and gesture) are being considered and investigated in the context human-machine teaming for the ATTRACTOR design reference mission (DRM) of Search and Rescue or (more generally) intelligence, surveillance, and reconnaissance (ISR). The first of these investigations, the Human Informed Natural-language GANs Evaluation (HINGE) data collection effort, is aimed at building an image description database to train a Generative Adversarial Network (GAN). In addition to building an image description database, the HMI team was interested if, and how, modality (spoken vs. written) affects different aspects of the image description given. The results will be analyzed to better inform the designing of an interface for mission planning.

I. Nomenclature

ATTRACTOR	=	Autonomy Teaming and TRAjectories for Complex Trusted Operational Reliability
CAS	=	Convergent Aeronautics Solutions
CIDEr	=	Consensus-based Image Description Evaluation
DRM	=	design reference mission
GAN	=	Generative Adversarial Network
GUI	=	Graphical User Interface
HINGE	=	Human Informed Natural-language GANs Evaluation
HMI	=	Human-Machine Interface
LIWC	=	Linguistic Inquiry and Word
LSA	=	Latent Semantic Analysis
NUI	=	Natural User Interface
PASCAL	=	Pattern Analysis, Statistical Modelling and Computational Learning
SAR	=	Search and Rescue

[1] Aerospace Research Engineer, Crew Systems Aviation and Operations, MS 152, AIAA Member
[2] Human Factors Engineer, Crew Systems Aviation and Operations, MS 152, non-member
[3] Aerospace Research Engineer, Crew Systems Aviation and Operations, MS 152, non-member
[4] NASA Senior Technologist for Intelligent Flight Systems, MS 233, AIAA Senior Member

II. Introduction

As a new start in NASA's Convergent Aeronautics Solutions (CAS) project, this work aims to develop innovative human-machine interface (HMI) approaches in support of human-machine teaming for autonomous systems. More specifically, this research is part of the ATTRACTOR effort (Autonomy Teaming and TRAjectories for Complex Trusted Operational Reliability) where advances in human-machine teaming and HMI will enable multi-agent missions such as search and rescue (SAR) missions in challenging scenarios. The SAR missions envisioned will be led by a human commander that has multiple robotic teammates with autonomous capabilities. Robotics teammates will be equipped with the necessary payload to identify the target of the search and will work collaboratively with humans to succeed in their shared mission. As the ability to sense, perceive, and plan through complex scenarios improve, the demarcation between humans and autonomous robotic systems is expected to blur with humans and machines teaming in efficient and creative ways and playing increasing roles in our everyday lives.

ATTRACTOR's design reference mission is modeled after a SAR mission where a person or object is lost or missing in an outdoor setting. A team composed of a human mission commander and a number of autonomous assets (human or machine) is tasked to search an area of interest. This asset team may consist of ground or air vehicles, manned or unmanned. The mission commander requires a human-machine interface to communicate mission objectives and receive status information, all with the possibility of updating the mission as necessary with the team. The ultimate goal is humans working with machines as teammates, where the team's performance is better than either entity alone.

Though Graphical User Interfaces (GUI) have been researched and used for upwards of three decades [1], they are not necessarily the most efficient or intuitive option due to a bilateral communication barrier between human and machine. One step towards alleviating this barrier is the Natural User Interface (NUI) which enables a seamless interaction between humans and machines using speech, touch, and gestures/body movements that imitate real world interactions. [2] This multi-modal interaction is preferable as studies show that no one single modality can provide a complete natural interaction experience and users feel comfortable using a variety of modes when interacting with machines. [3] Additionally, a major strength of designing a multi-modal NUI is that the user can utilize any of the natural interactions interchangeably depending on their unique needs and those of the mission. [2]

Recently, speech-based natural language interfaces have become commonplace [1,2,3,4] as they allow users a natural, intuitive way to communicate with machines, whether at home, work, or in between. [4] Using a speech-based natural language interface is both convenient and efficient and has proven to increase situational awareness, usability, and satisfaction. They allow humans to interact with machine teammates as they would with human teammates. However, speech should not be relied on as the sole modality when interacting with machines if intuitive interaction is the goal. Instead, a multimodal approach using a combination of input/output modalities (speech/audio, typing/text, touch, and gesture) should be considered to facilitate easier and more intuitive communication between humans and machines. [4] Thus, creating a NUI that is multimodal, efficient, reliable and intuitive is the basis of the ATTRACTOR HMI team's research. This interface will be used to define the number of assets and their roles, the area of interest, the target of the mission, as well as enable interactions during the mission that will aid in the trust and trustworthiness that ATTRACTOR is researching in the hopes it can pave a path to certification.

III. Background

In a typical SAR mission, a description of the target is provided by either a photo and/or verbal narrative. When asked to describe a missing person (or a representative image), humans can give concise sentences that can identify the "who" (the most interesting object or person), "what" (the things the objects or people are doing), and "where" (the location of the action that is happening). Because people will omit what they judge to be less significant, the sentences are concise, but even then the target descriptions still tend to exhibit consensus. [5]

In order to facilitate multi-asset teaming between humans and autonomous systems, communication between the human and machine needs to be natural and intuitive for the human teammate. [6] Our first exploration, the Human Informed Natural-language GANs Evaluation (HINGE), is focused on defining the objective for the mission and thereby informing the interface being designed.

HINGE served a dual purpose effort, allowing the HMI and Target Description team to gather the human informed image description data needed to build a database to train a Generative Adversarial Network (GAN) for another aspect of ATTRACTOR. A GAN is a system, comprised of two neural networks, that learns to synthesize new data from a training set of existing real-world data. [8, 9] Human image descriptions will be used as features to train the GAN, with the original photo being the ground truth representation. Once trained, the GAN will take an image description from the user and the output will provide the SAR autonomous assets an internal representation of what is being sought during the mission.

In addition to providing image description data to the GANs being implemented for ATTRACTOR, the HMI team sought to determine if, and how, modality (spoken vs. written) affects different aspects of the image description given. To determine this, image description data was gathered from 53 participants, with 26 participants speaking their descriptions into an audio recorder (later transcribed by a researcher) and 27 participants typing their descriptions into an excel spreadsheet.

IV. Study Design

Five images were chosen from the PASCAL 50S dataset. This dataset is based on the UIUC Pascal Sentence Dataset [5], which pulled 1000 images from the larger 2008 PASCAL dataset. This larger image dataset was used as part of the PASCAL Visual Object Classes challenges that ran from 2006-2012 and contains at least one object from a set of twenty object classes (e.g., person, bird, dog, bicycle, bus, bottle, chair, potted plant). The UIUC Pascal Sentence Dataset had five sentence descriptions per image. This number was later increased to 50 description sentences per image by Vedantam et al. [10] This work created a new automated metric, the Consensus-based Image Description Evaluation (CIDEr), which improved reliability of image description evaluation due to the larger number of reference image descriptions. [5, 9-11] The images chosen all contained only one individual in an outdoor setting to limit the feature set when training the GAN. This decision was made to limit the feature set to have objects that ATTRACTOR would be detecting located in the scene, while minimizing non-target classes which introduces noise in the data with respect to the particular model being used. [7] During the course of the study, participants were asked to describe these images using either a speech or text interface.

A. Participants

Fifty-three volunteers, aged 21 to 74, were recruited outside the NASA Langley cafeteria during lunch over a four day period in May 2018. Fourteen females and 39 males participated. The participants' education levels ranged from "some college" to doctorate level education.

B. Procedures

Participants were randomly assigned into either the typed or verbal description group and then briefed on the task. The verbal group was handed five pictures and asked to state the image number and then give their descriptions based on the instructions below. The participants in the typed group were given the set of pictures and then asked to type their responses into an Excel spreadsheet where their descriptions were saved. Each participant saw each of the 5 images twice. Upon the first viewing of each image, the participant was told, "Using one sentence, please describe what is going on in this image." Upon the second viewing each person was told, "The person in this image went mission an hour after this picture was taken. Please describe them to help us find them." Participants were reminded to use one sentence in these descriptions as well. The verbal transcriptions were transcribed by a researcher and added to the corpus of typed image description responses for analysis. Preliminary analyses were conducted on Image 2 (See Fig.1.).

Preliminary descriptive analyses were conducted on Image 2 (see **Error! Reference source not found.**). A total of 2,200 words were collected from 105 image descriptions. (One person did not complete the second portion of the data collection) In total, there were 999 (45%) typed words when describing the second image, and researchers transcribed 1201 (55%) words that were given via audio recorded data. The following results were analyzed using the raw data, before misspellings and typos were corrected. Further analysis will be done on the clean data to include parts of speech and other linguistic analyses.

Fig. 1. HINGE Image 2 [11]

V. Results

A. Descriptive Analysis

The data were also broken down into the language used in the initial image description task, referred to as Context 1, as well as the language used after the second viewing of the image, referred to as Context 2. The dataset contains 105 image descriptions given by 53 participants, one participant did not complete the second image description task. A breakdown of the 353 unique words that made up the descriptions are shown in Figure 2. 1201 words were spoken from a dictionary of 238 unique words, 999 words were typed from a dictionary of 228 words. Of the total words from both spoken and typed there were 113 in common. Figure 3 shows the overlapping common word counts across context and modality.

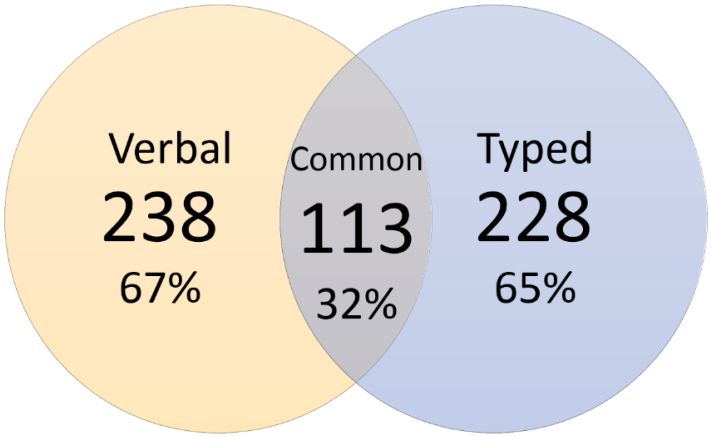

Fig. 2. Breakdown of Dictionary Word Count.

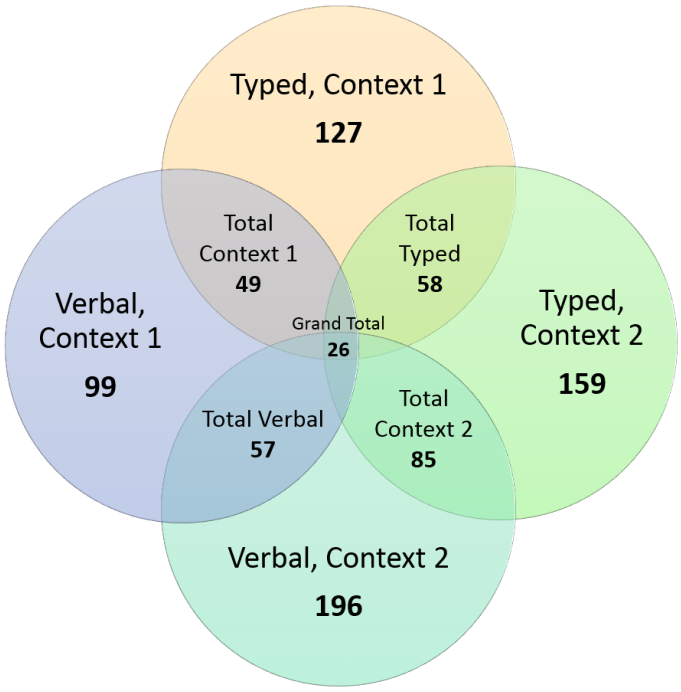

Fig. 3. Breakdown of Dictionary Word Count by Context and Modality.

Though the differences are slight, participants used more words total and more unique dictionary words when inputting descriptions with a speech interface than with a text interface. This is partially explained by the presence of common hesitation markers (uh, um) and hedgewords (just) that are used in spoken English but omitted in text. However, the difference in amount of input may also speak to a willingness in users to provide more description depending on format. Slow typers or those more generally familiar with speaking than typing may prefer to provide information using this modality. It may be the case that the more information provided to a machine teammate, the better its analysis will be.

For this initial analysis of the data, focus was given to dictionary words that were used in both verbal and text modalities with a frequency greater than ten. This value was chosen as an initial point of comparison because initial analysis of Context 1 demonstrated a gap between a frequency of 5 and a frequency of 10, indicating that this value was a noteworthy cutoff point. Context 1 had ten words in common between verbal and text modalities with a frequency greater than ten: “and,” “man,” “is,” “in,” “on,” “beach,” “walking,” “dog,” “the,” and “a.” Context 2 had 11 words: “brown,” “dark,” “hat,” “the,” “pants,” “wearing,” “jacket,” “with,” “black,” “and,” and “a.”

The ten most frequent words used in both verbal and typed contexts are striking to the human language producer in that they form a coherent description: Man is on the beach and walking a dog (in). Despite modality of input, users seem to generate a consensus in the description of the image. If descriptive content can be easily agreed upon it may be similarly easy to allow machine teammates to generate comparable output in order to participate in a natural manner with human teammates. Moreover, if input content can be predicted in such a manner if may be easy to allow machine teammates to process image description input in a meaningful way.

The most frequent word use in both verbal and typed inputs modalities for Context 1, however, seem to form less of a coherent sentence but rather descriptive phrases. Noticeable here is the presence of a higher number of descriptive words (brown, dark, black). This betrays a shift in focus from a full descriptive sentence, present in Context 1, to maximizing descriptive information, a switch perhaps expected given the emergent situation of Context 2.

B. Attribute Analysis

In addition to analyzing the data based on frequency of occurrence and commonality between modalities and contexts, the content of the descriptions was also analyzed. Image descriptions given as sentences can be difficult to measure since sentences are fluid and can be quite different when describing the same image. Also, the wide variety of human vocabulary and the tendency to substitute “puppy” or “animal” for “dog,” and “car,” “truck,” or “bicycle” for “vehicle” can all lead to many different words being used to describe the same image. This was a defining feature the team noticed when categorizing the words given to describe Image 2.

If research uncovers the structure and common elements that are needed to impart the information, human operators can use fragmented commands to pass on the search target's description to the asset team. In part of the preliminary analysis of the image descriptions given for Image 2, we broke down the descriptions into six attributes. These attributes could be further broken down in the future to determine other relationships and interdependencies. Those attributes were: gender, age, clothing type, clothing color, map, and related objects or animals. Map and related objects or animals were further broken down into two sub categories that we hope will aid SAR team members in determining a probable location of the target. For instance, if the description contains information about an activity or landmark, then that may narrow the area to be search. Likewise, if a related object is found, the target object may be closer than if no related objects had been found. Future work is being considered to explore these ideas.

One hundred and five descriptions were categorized by their attributes. The team noted words that describe each of the attribute groups, for example the gender ("man," "woman," "his") or age ("20s," "young") of the missing person. The results are shown in in Table 1 below. Preliminary analysis found that people tended to use the same or very similar words for these attributes. Noteworthy findings include the similar number of mentions of each attribute category between verbal and typed responses. As expected, Context 1 descriptions provided more information about the general location, activity, and instance of another subject in the image (a pet dog). Context 2 descriptions veered more toward describing clothing type and color to aid finding the fictitious missing subject. The team was surprised to see such a small amount of mentions of the subject's age and last seen location in Context 2, the SAR context. Though age was only mentioned 15 out of 52 times, approximate age is an important description metric that can be used to predict a missing person's behavior. For instance, a missing child or senior adult will likely shelter in place and wait for assistance whereas teenagers and adults have been found up to 17 miles from their last known location. [12] After quantifying attributes for one image, the team will examine the remaining images as these tallies could point to how a brief fill-in-the-blank form, drop-down list, outline, or prompt of some kind would be beneficial for the user interface to gather as much data as quickly as possible. This may also aid in training the GAN as Sun, et al., used color, shape and material attributes to train a classifier in their work on attribute based object identification. [13]

Table 1. Attribute Analysis

					Map Information		Relevant Objects or Animals	
	Gender	Age	Clothing Type	Clothing Color	Location/ Landmarks	Activity	Animate	Inanimate
Typed Total	36	8	31	29	32	32	40	4
Verbal Total	43	9	32	27	29	34	38	3
Context 1 Total	36	3	12	9	51	53	53	4
Context 2 Total	42	15	50	46	9	11	23	4
GRAND TOTAL	78	18	62	55	60	64	76	8

C. Parts of Speech

To complement analysis of language frequency, commonality, and context, investigation of the syntax of image descriptions may provide additional information. Though the results are only preliminary, initial analysis suggests critical differences in the number and percentage of parts of speech used in text and verbal descriptions as well as between the two contexts. More adjectives were used in Context 2 descriptions, which may be indicative of a different style of language use for this different context. Future work may look at instances of pronouns and words used across multiple parts of speech in order to identify other differences in language styles that may contribute to better communication with machine team members in various situations. [14, 15]

D. Future Work

The initial examination of these data provides a promising analysis of the impact of different input modalities based on intended goal (image description or person identification). Such interesting preliminary results suggest that further research is merited. Additional areas of analysis may include the CIDEr metric, Latent Semantic Analysis (LSA) and Linguistic Inquiry and Word Count (LIWC), and semantic context analysis. These analyses perform better when the data is error free.

CIDEr is a metric that measures the similarity between sentences using word and phrase frequency to determine consensus. In addition to use of the CIDEr metric, previously established linguistic analysis tools may help to further

identify patterns in image descriptions. [6] LSA and LIWC may provide tools for incorporating semantic information or further diving into word count tools for determining the success of image description or even just predicting patterns.

Additionally, analysis of the dictionary words used in image description removes the language from its semantic context: When the word "one" appears, does it refer to the number 1 or is it instead used as a third person pronoun? Future work could work to take into account the semantic context of the language to determine whether that has any impact on the overall analysis.

Initial analysis was performed over the raw data from the user study. While such analysis highlighted interesting results, further work on sanitizing the data may make any inferences clearer. Work on accounting for typos and spelling differences within the text data as well as hedge words and hesitations in the speech data may provide for a better comparison. Moreover, while all due care was afforded the transcription of the speech data, any transcription process may provide a point for introducing errors into the system. Establishing a system for checking and correcting transcription errors, or even automating the transcription process to ensure standardization, may further reduce errors.

Crowdsourcing to gather additional image description data is also being discussed. [16, 17]

VI. Conclusion

How humans respond to different input modalities is impacted by many disparate factors. Familiarity with an interface is a prime reason for preferring one to another, which in turn is often affected by experience with a particular system as well as various demographic data (age, years in field, location). Identifying the ways in which language changes with different input modalities and with different objectives allows for more intuitive and accurate communication with machine team members, improving interface design and furthering ATTRACTOR's goal of establishing a basis for certification of human-machine (autonomous system) teams.

Acknowledgments

The authors would like to thank Erica Meszaros, the Autonomy Incubator, the ATTRACTOR team and Crew Systems and Aviation Operations Branch at NASA Langley for their support.

References

[1] Medioni, G., Kang, S. B., *Emerging Topics in Computer Vision*, Upper Saddle River, NJ, USA Prentice Hall PTR, 2004.

[2] Fernández, R. A. S., Sanchez-Lopez, J. L., Sampedro, C., Bavle, H., Molina M., and Campoy P., "Natural User Interfaces for Human-Drone Multi-modal Interaction," *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, Arlington, VA, 2016, pp. 1013-1022.

doi: 10.1109/ICUAS.2016.7502665

[3] Cauchard, J. R., E, J. L., Zhai, K. Y., and Landay, J. A., "Drone & Me: An Exploration Into Natural Human-Drone Interaction," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 2015, pp. 361-365.

doi: 10.1145/2750858.2805823

[4] Meszaros, E. L., Chandarana, M., Trujillo A., and Allen, B. D., "Speech-Based Natural Language Interface for UAV Trajectory Generation," *2017 International Conference on Unmanned Aircraft Systems (ICUAS)*, Miami, FL, USA, 2017, pp. 46-55.

doi: 10.1109/ICUAS.2017.7991401

[5] Farhadi, A., Hejrati, M., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D., "Every Picture Tells a Story: Generating Sentences for Images". *European conference on computer vision*, pp. 15-29. Springer, Berlin, Heidelberg, 2010.

[6] Meszaros, E. L., Le Vie, L. R., Allen, B. D., "Trusted Communication: Utilizing Speech Communication to

Enhance Human-Machine Teaming Success," *AIAA Aviation and Aeronautics Forum and Exposition (AIAA AVIATION 2018)* (to be published).

[7] Ecker, J., Allen, B. D., "Goal Detection via Mental Representation," *AIAA Aviation and Aeronautics Forum and Exposition (AIAA AVIATION 2018)* (to be published).

[8] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." arXiv preprint arXiv:1711.10485. 2017

[9] Wang, J. K., and Robert Gaizauskas. "Cross-Validating Image Description Datasets and Evaluation Metrics." *Proceedings of the 10th Language Resources and Evaluation Conference*. European Language Resources Association, 2016.

[10] Vedantam, R., Zitnick, C.L., Parikh, D. "CIDEr: Consensus-Based Image Description Evaluation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566-4575.

[11] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., and Zisserman, A., "The PASCAL Visual Object Classes Challenge: A Retrospective." *International journal of computer vision*, *111*(1), 2015, pp.98-136.

[12] *National Search and Rescue Manual*, February 2018. URL:https://natsar.amsa.gov.au/documents/Land-Operations/LSOM_Appendix_I.pdf [retrieved 12 March, 2018].

[13] Sun, Y., Bo, L. and Fox, D., "Attribute Based Object Identification," *International Conference on Robotics and Automation (ICRA), IEEE*, 2013, pp. 2096-2103.

[14] Toutanova, D. K., Manning, C., and Singer, Y., Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259

[15] Fellbaum, C., *WordNet*, John Wiley & Sons, Inc., 1998.

[16] Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J., "Collecting Image Annotations Using Amazon's Mechanical Turk." *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010.

[17] Vedantam, R., Zitnick, C.L. and Parikh, D., "Collecting Image Description Datasets using Crowdsourcing," *arXiv preprint arXiv:1411.3041*. 2014.