

# A Cybernetic Approach to Characterization of Complex Sensory Environments: Implications for Human Robot Interaction

Kelly Dickerson<sup>(✉)</sup>, Jeremy Gaston, and Kelvin S. Oie

US Army Research Laboratory, Aberdeen Proving Ground, MD 21005, USA  
kelly.dickerson5.civ@mail.mil

**Abstract.** Humans are increasingly interacting and collaborating with robotic and intelligent agents. How to make these interactions as effective as possible remains, however, an open question. Here, we argue that consistent understandings of the environment on the part of the human and agent are critical for their interaction and basing these understandings on only the objective features of sensory inputs may be inadequate. To that end, the current paper presents a novel approach to more integrated characterizations of the sensory environment that encompass objective and subjective features of sensory inputs. We propose that an approach to signal and behavioral estimation consistent with the control and communication theoretic perspective of Cybernetics could inform human robot interaction (HRI) applications. Specifically, we offer a potential path forward for quantifying similarity in stimulus events that can lead to consistent understandings of the environment, which when applied to HRI can enhance human-agent communication in HRI applications.

**Keywords:** Cybernetics · Human robot interaction · Stimulus classification · Information theory

## 1 Introduction

In recent years, the US Army has seen an increased integration of highly skilled Soldiers with advanced technologies, with a significant emphasis in future interactions with robotic and other intelligent agents. To support these efforts, we have adopted the conceptual framework of Cybernetics. This transdisciplinary approach, popularized most notably by [1], is the “scientific study of control and communication in the animal and machine.” Cybernetics, therefore, encompasses the disciplines of control theory and communication (a.k.a., information) theory, as a general approach to understanding closed-loop, feedback systems [2, 3].

Feedback can be defined generally as the information that is generated by a system’s control actions and the resulting interactions with the external environment, which when sensed by the system can then be utilized in planning and executing future actions. In a complex system, such as the human brain, feedback operates at multiple levels; for example, feedback influences internal affective or physiological states, or the

interactions among single or small clusters of neurons. At the cognitive level, feedback influences intention and shapes action “closing” the loop that moves from intention to action, to sensing the outcome of action, to the comparison of an outcome to the original intention, and provides information to make appropriate adjustments in further action. This closed loop circularity is essential to “cybernetic” systems as it enables adaptation under complex and dynamic conditions.

However, understanding the mechanisms underlying the incorporation of feedback into higher-level human behaviors, such as decision-making, is more conceptually challenging than such an idealized loop might suggest. Human sensation and perception is often incomplete, inaccurate, and ambiguous, and it is not always possible to extract all of the information needed to support behavior directly from immediate sensing of the environment. This can create conditions of undesirable uncertainty about one’s relationship with the environment. In the face of this uncertainty, humans likely weigh the input from multiple modalities, and can use the combined or integrated perceptual result to guide action. The information that underlies this weighting in a highly complex real world interaction remains an open question. The answer to this question, of course, can depend strongly on the situational context, and humans are generally quite flexible in their putative information processing strategies. When a human must communicate their intentions to an agent, the information encoded in flexible combinations of modalities, stimulus features, and assumptions or expectations that typically serve human-to-human communications well are not available for decoding by the agent. This lack of effective communication can lead to significant misunderstandings between human users and the systems they rely on for successful task performance.

Here, we suggest, that in human-agent teaming situations, for effective communication between an agent and their human counterpart it is likely critical that their understandings of the environment be consistent with each other. For example, humans and robots, depending on their relative size, may understand the same physical objective differently: For a micro-autonomous system, a shoebox would be a significant obstacle, whereas for their human counterpart, it would not. However, for both the human and robot, the shoebox may pose a threat, because something could be contained within it that neither of their visual sensors would be able to detect. This is just a single example of the difficulties human-agent teams could face while navigating a complex and dynamic environment. However, creating consistent representations of the everyday environment is no small feat given the differences in sensing and perceptual architectures between humans and the myriad potential artificial agents with which they might team. The representations that typically underlie such understandings in robotic and other intelligent agent applications still mainly reflect low-level, quantitative aspects of their physical sensory inputs. By contrast, as argued above, understanding the environment for humans does not just comprise mappings of the immediate physical domain. Human representations, instead, typically reflect the integration of the more objective information based on current sensory inputs with more subjective information that strongly depends on assumptions or expectations derived from previous experiences in other contexts, which is difficult, if not impossible, to directly measure.

The subjective attributes of human perceptual experiences conceived of here as cognitive features that vary across individuals, but with values bounded by prior

information from experience within relevant behavioral contexts. For example, the subjective experience of a cognitive feature, such as the perceived pleasantness of the sound of a car’s engine, may be low for someone unfamiliar with the loud and intermittently impulsive mechanical noise. However, when that engine sound has become associated with the returning home of a spouse or parent at the end of the day, those previous experiences might positively influence the perceived pleasantness of the input. In turn, this influence may also change subjective experiences across a variety of contexts. The example provided here is meant to illustrate that the influences of subjective experience on the representations that are fundamental to understanding in complex environments are likely pervasive. Still, the challenging task of quantifying subjective stimulus attributes is an area of research that has been somewhat lacking, with even less information available on how one would apply or translate this research in the context of enhancing human-robot interactions (HRI).

### 1.1 Objectives of This Paper

This paper will briefly discuss the relative strengths and weaknesses of human and systems approaches to stimulus classification from visual, auditory, and multimodal inputs. We will highlight the relatively limited research that compares human and agent performance on common tasks and discuss how these comparisons can support the development of better models of interaction between humans and agents. We then describe our work on characterizing the sensory environment, which has thus far focused largely on auditory environment quantification. This research on human performance in the context of subjective stimulus attributes is discussed with an eye towards using such an approach to improve models of multisensory integration for HRI applications.

## 2 Comparing Human and Machine Classification Approaches

One critical skill required for successful navigation of the environment is the ability to detect, localize, and recognize objects and boundaries. Using vision, humans perform this task seemingly effortlessly, while object localization and recognition in machine vision is resource-intensive and cannot yet match human performance in all conditions. For example, in a direct comparison of a robotic machine vision algorithm and human classification, [4] found that humans were 1.7% more accurate. This difference in accuracy may seem small, however, the source of accuracy differences was revealing. Specifically, when the algorithm made classification errors, it was found to be due to image features that humans typically have no difficulty processing, such as view point and color invariance [4, 5]. The algorithm also had difficulty classifying images that were graphic or symbolic representations of real objects (i.e. drawing of a coffee cup, or an image of a stuffed bear). Despite these limitations, machine vision is improving rapidly and there are emerging examples where Deep Learning approaches have exceeded the best human performance for specific image data sets. For example, [6] showed machine image classification for the ImageNet 2012 classification dataset that exceeded the best-reported human performance by more than 5%.

In audition, there are also examples of machine algorithms classifying environmental sound events; however, in almost every case the stimulus set is restricted to a homogeneous class of events, such as the detection and localization of gunfire events, [7, 8]. Under these conditions human and algorithm classification performance was comparable. However, in the majority of real world environments, the input event distribution is much more varied than that used in these studies. This is a pervasive issue for current machine language solutions: in application domains where the behavioral space intrinsically involves greater variety, machine language approaches cannot yet match human performance. For example, in speech recognition, the domain where the majority of auditory machine learning applications have been developed [9], humans typically do better than implemented machine approaches in terms of accuracy of recognized speech [10, 11]. However, as with machine vision, machine speech recognition systems are rapidly improving; a very recent application of machine learning approaches by Microsoft has yielded parity with humans in transcribing speech from the NIST 2000 speech test set [12].

Multimodal machine learning classifiers have also had demonstrated success, but with limited comparison to human performance. Importantly, in multimodal classification, the addition of redundant, as well as potentially unique, information from another modality is one obvious way to improve machine classification. For example, adding audio to visual information can improve classification by increasing bandwidth, accuracy, and decreasing processing time by using converging evidence to support classification of environmental objects under difficult or ambiguous conditions. Examples include the audio-visual classification of speech [13] and audio-visual and textual sentiment analysis [14], where in both cases the additional sensory cues led to better classification performance.

These direct comparisons between human and machine performance are useful in assessing advances in machine performance and potentially identifying where further advancements may occur. However, another approach to understanding the differences between human and machine performance is to examine their capabilities in the context of collaboration. One robust example of human-agent collaboration is the Human-Autonomous Image Labeler (HAIL) developed by the US Army Research Laboratory. The performance of the HAIL system depends critically on both human and computer vision systems [15]. Specifically, it takes advantage of the capability for rapid, but sometimes inaccurate classification of tens of thousands of images by computer vision agents, and couples that with the capability for very accurate, but much slower classification by human agents. The outcome is a very accurate and fast classification of a large set of images [15, 32] that, instead of highlighting the limitations of human and machine performance, takes advantage of the respective strengths of human and autonomous agents to increase the performance of the “system” as a whole.

## 2.1 The Problem of Similarity

Classification performance by intelligent agents, humans, or human agent teams is, in many cases, negatively impacted when the to-be-classified content is highly similar to background or distractor information [16]. Although some image algorithms excel at

classifying highly similar images, such as those that could be part of the same fine-grained (local level) category, many developed algorithms tend not to be robust to suboptimal viewing conditions and could still produce significant classification errors. Multimodal cuing could aid classification by using auditory information to disambiguate highly similar visual inputs and support discrimination; for example, two dogs of different breeds likely have distinctive barks.

Indeed, for humans, it is well-known that visual task performance is often augmented by the presence of auditory information. This multisensory enhancement effect [see 19] is possible due, to the fact that many objects in the environment are only fully described by the combination of distinct auditory and visual features. This suggests that processes for audiovisual integration can capitalize on informational redundancy to reduce uncertainty in perceptual estimates, enhancing the resultant representation of the world and making it both more coherent and more robust [17, 18]. More generally, human perceptual systems combine and integrate information from their multiple different sensory modalities, which reduces the variance and, generally, increases the reliability of perceptual estimates that support the higher cognitive functions, including decision making.

There is clearly value in adapting these strategies for HRI applications, yet it remains unclear whether multisensory perception based solely on current sensory inputs provide adequate information for complex decision making. Indeed, [4, 33] (see also [18] for review) have shown that some of the efficient and robust sensory combinations and integrations underlying humans' higher-level perceptual capabilities rely on representations of prior experiences, and that these subjective attributes may not be easily translated from human to non-human systems. For example, in social interactions using text-based communication (i.e., IM), the presence of punctuation can influence perceived sincerity of a comment in younger users who are used to crafting text in the absence of traditional punctuation [20]. Similarly, in reading, the same words can convey different senses of urgency depending on the contextual framing provided by story narration. A reader can perceive the activity of a character as urgent if the narrator uses language that suggests fast movements, but perceived urgency is limited when the narrator uses language to suggest slower movements or does not describe the rate of activity [21].

### 3 Characterizing the Sensory Environment

As discussed above, humans use information from multiple modalities to reduce uncertainty in perceptual estimates, which supports efficient decision-making. Multisensory integration is often biased based on previous experiences with a given object in a particular context [19]. This bias can manifest in two opposing ways, as a performance decrement or as a performance enhancement, depending on how the information available is combined or integrated with prior information from previous experience in the representation of the current situation. However, as alluded to above, sometimes complex and high dimensional, experience-based factors can be difficult to define and are resistant to direct measurement.

In the human multisensory perception literature, Bayesian approaches have emerged as an important tool in understanding how multimodal sensory cues can be integrated.

There are a number of examples where Bayesian maximum likelihood estimation (MLE) models predict multisensory integration [9, 20, 22, 23]. Maximum likelihood estimation models of multisensory integration [22] (also known as Bayesian inference models) maximize the *maximum a posteriori* (MAP) estimate associated with a particular response by dynamically updating the maximum likelihood functions associated with the sensory cues. Across trials, the resultant MAP predictions are weighted sums of the unimodal sensory inputs, where the weight reflects the relative cue uncertainty gathered from previous trials. However, the Bayesian priors are theoretical values that estimate the magnitude of the impact of previous experiences with a given set of multimodal cues and these estimates are not necessarily based on the real distribution of experiences. It is not clear how well this approach would extend to dynamically changing multimodal cues, or even if this approach could scale up to real world audio visual events. A prerequisite to evaluate the possibility of quantifying experience-based factors using a Bayesian approach would be a better understanding of the physical and perceived signal qualities of stimuli in the environment. [24] found that informational and contextual factors affect listeners' ability to identify environmental sounds. For example, they found that the presence of a competing background, particularly a background with overlapping information reduced sound identification accuracy. Similar effects have emerged in our own work, [25] found that identification accuracy was better when sounds had a clear originating event ("concrete") than when the link between the sound generating object and the sound produced was less obvious ("abstract").

The results of [24, 25] suggest that subjective and contextual factors convey important task relevant information. To better understand the content of environmental sounds, and the interaction between these subjective factors and object stimulus parameters researchers have applied stimulus classification techniques to environmental sound perception. These techniques offer a method for addressing possible limitations in the way Bayesian priors are estimated for real stimulus events. [26] used listener defined similarity scores, as well as objective measures of spectral and temporal features of sounds, to create a classification space for a large set of common environmental sounds. Dickerson et al. [27] extended this approach by using prior subjective ratings of stimulus similarity to characterize human listener performance on a variety of different behavioral tasks. It is possible that data of this nature could be used to improve Bayesian estimation techniques and could provide consistent and meaningful feedback about the environment to both human and non-human intelligent agents. In several related experiments, we have further extended our understandings of and approaches to quantifying the relationship between informational and contextual effects using the construct of similarity. [28] found that, for change discrimination, similarity in perceived loudness influenced the likelihood of noticing that an element within a scene had changed. This effect of similarity also manifested in the identity relationships among the 25 signals in their stimulus set, with a linear relationship found between the likelihood of change discrimination and the overall similarity (defined via user rating and a multidimensional scaling (MDS) analysis) of the sounds in the stimulus set. This relationship between identifiability, similarity, and perceptual performance is not particularly compelling on its own; the well-established informational masking literature would likely predict some of these effects [29]. However, this becomes more compelling when we examine the robustness

of these trends across changes in methods and paradigms. [27] found that similarity among sounds in a scene affected both change discrimination and change localization performance, where increasing similarity decreased accuracy. [30] found that this effect extended to performance on cued-recall tasks, as well. It was further revealed that a complex interaction exists between user ratings of identifiability and similarity and later memory performance: Sound sources group together based on identifiability and category membership, but sounds within a tightly clustered group were more poorly recalled.

The work from our group, along with Bayesian approaches to multisensory integration suggests that subjective information quantified in the manner discussed in this paper provides a tractable method for including subjective information in Bayesian prior estimation. By using this type of information in the development of machine sensing approaches, it becomes possible for man and machine to have a deeper and more consistent understanding of their operational environment, potentially reducing the workload associated with communicating information between agent and human teammate.

## 4 Conclusions and Future Directions

The research highlighted here suggests that in order to accurately and meaningfully represent the environment to both man and machine, more information than the direct sensory stream may be required. By quantifying subjective attributes, such as similarity, that relate complex features across objective and subjective perceptual estimates, researchers can develop a better understanding of the feedback that guides behavior under the complex and dynamic conditions of the real world. Additionally, the research summarized here converges on emerging perspectives in multisensory integration, that, rather than separating out each sensory stream for modular and potentially parallel processing, the auditory and visual information are processed together as a holistic object [31]. This perspective suggests that there may be value in the further uncertainty reductions and saliency gains in including subjective factors in the characterization of stimulus events. Future research will focus on continuing to evaluate how humans and intelligent agents complete tasks in isolation and in cooperation in order to uncover the stimulus-related objective and subjective factors producing efficient and accurate behavior.

## References

1. Wiener, N.: *Cybernetics: Control and Communication in the Animal and the Machine*. Wiley, New York (1948)
2. Seising, R.: Cybernetics, system (s) theory, information theory and fuzzy sets and systems in the 1950s and 1960s. *Inf. Sci.* **180**, 4459–4476 (2010)
3. Dubberly, H., Pangaro, P.: Cybernetics and service-craft: language for behavior-focused design. *Kybernetes* **36**, 1301–1317 (2007)
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015)
5. Biederman, I., Bar, M.: One-shot viewpoint invariance in matching novel objects. *Vis. Res.* **39**, 2885–2899 (1999)

6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: *Proceedings of the IEEE International Conference on Computer Vision* (2015)
7. Lopez-Morillas, J., Canadas-Quesada, F.J., Vera-Candeas, P., Ruiz-Reyes, N., Mata-Campos, R., Montiel-Zafra, V.: Gunshot detection and localization based on non-negative matrix factorization and SRP-hat. In: *Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 1–5. IEEE (2016)
8. Khalid, M.A., Babar, M.I.K., Zafar, M.H., Zuhairi, M.F.: Gunshot detection and localization using sensor networks. In: *Smart Instrumentation, Measurement and Applications (ICSIMA)*, pp. 1–6. IEEE (2013)
9. Deng, L., Li, X.: Machine learning paradigms for speech recognition: an overview. *IEEE Trans. Audio Speech Lang. Process.* **21**, 1060–1089 (2013)
10. Lippmann, R.P.: Speech recognition by machines and humans. *Speech Commun.* **22**, 1–15 (1997)
11. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvett, D., Rose, R.: Automatic speech recognition and speech variability: a review. *Speech Commun.* **49**, 763–786 (2007)
12. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G.: Achieving Human Parity in Conversational Speech Recognition. Microsoft Research Technical Report MSR-TR-2016-71, February 2017
13. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y. Multimodal deep learning. In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (2011)
14. Poria, S., Cambria, E., Howard, N., Huang, G.B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* **174**, 50–59 (2016)
15. Saproo, S., Faller, J., Shih, V., Sajda, P., Waytowich, N.R., Bohannon, A., Jangraw, D.: Cortically coupled computing: a new paradigm for synergistic human-machine interaction. *Computer* **49**, 60–68 (2016)
16. Brooks, J., Slayback, D., Shih, B., Marathe, A., Lawhern, V., Lance, B.J.: Target class induction through image feedback manipulation in rapid serial visual presentation experiments. In: *2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1047–1052, October 2015
17. Burr, D., Alais, D.: Combining visual and auditory information. *Progr. Brain Res.* **155**, 243–258 (2006)
18. Ernst, M.O., Bühlhoff, H.H.: Merging the senses into a robust percept. *Trends Cogn. Sci.* **8**, 162–169 (2004)
19. Shams, L., Seitz, A.R.: Benefits of multisensory learning. *Trends Cogn. Sci.* **12**, 411–417 (2008)
20. Gunraj, D.N., Drumm-Hewitt, A.M., Dashow, E.M., Upadhyay, S.S.N., Klin, C.M.: Texting insincerely: the role of the period in text messaging. *Comput. Hum. Behav.* **55**, 1067–1075 (2016)
21. Gunraj, D.N., Drumm-Hewitt, A.M., Klin, C.M.: Embodiment during reading: Simulating a story character’s linguistic actions. *J. Exp. Psychol.: Learn. Mem. Cogn.* **40**, 364–375 (2014)
22. Angelaki, D.E., Gu, Y., DeAngelis, G.C.: Multisensory integration: psychophysics, neurophysiology, and computation. *Curr. Opin. Neurobiol.* **19**, 452–458 (2009)
23. Roach, N.W., Heron, J., McGraw, P.V.: Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proc. R. Soc. Lond. B: Biol. Sci.* **273**(1598), 2159–2168 (2006)
24. Leech, R., Gygi, B., Aydelott, J., Dick, F.: Informational factors in identifying environmental sounds in natural auditory scenes. *J. Acoust. Soc. Am.* **126**, 3147–3155 (2009)



25. Dickerson, K., Foots, A., Gaston, J.: The influence of concreteness on identification and response confidence for common environmental sounds. *PLoS ONE* (under review)
26. Gygi, B., Kidd, G.R., Watson, C.S.: Similarity and categorization of environmental sounds. *Atten. Percept. Psychophys.* **69**, 839–855 (2007)
27. Dickerson, K., Gaston, J., Foots, A., Mermagen, T.: Sound source similarity influences change perception during complex scene perception. *J. Acoust. Soc. Am.* **137**, 2226 (2015)
28. Gaston, J., Dickerson, K., Hipp D., Gerhardstein, P.: Change deafness for real spatialized environmental scenes. *Cogn. Res.: Princ. Implic.* (in press)
29. Dickerson, K., Gaston, J.R.: Did you hear that? The role of stimulus similarity and uncertainty in auditory change deafness. *Front. Psychol.* **5**, 1–5 (2014)
30. Dickerson, K., Sherry, L., Gaston, J.: The relationship between perceived pleasantness and memory for environmental sounds. *J. Acoust. Soc. Am.* **140**(4), 3390 (2016)
31. Ramenahalli, S., Mendat, D.R., Dura-Bernal, S., Culurciello, E., Niebur, E., Anderou, A.: Audio-visual saliency map: overview, basic models and hardware implementation. In: 2013 47th Annual Conference on Information Sciences and Systems (CISS), pp. 1–6. IEEE (2013)
32. Slayback, D., Files, B., Lance, B., Brooks, J.: Effects of image presentation highlighting and accuracy on target class induction (in preparation)
33. Ernst, M.O., Banks, M.S.: What determines dominance of vision over haptics? In: *Proceedings of the Annual Psychonomics Meeting* (2000)

Advances in Human Factors in Robots and Unmanned  
Systems

Proceedings of the AHFE 2017 International  
Conference on Human Factors in Robots and  
Unmanned Systems, July 17–21, 2017, The Westin  
Bonaventure Hotel, Los Angeles, California, USA

Jessie Chen, Y.-S. (Ed.)

2018, XII, 358 p. 142 illus., Softcover

ISBN: 978-3-319-60383-4