Künstl Intell (2017) 31:357–361
DOI 10.1007/s13218-017-0505-9

RESEARCH PROJECT

Red Hen Lab: Dataset and Tools for Multimodal Human Communication Research

Jungseock Joo[1] · **Francis F. Steen**[1] · **Mark Turner**[2]

Published online: 21 September 2017

Abstract Researchers in the fields of AI and Communication both study human communication, but despite the opportunities for collaboration, they rarely interact. Red Hen Lab is dedicated to bringing them together for research on multimodal communication, using multidisciplinary teams working on vast ecologically-valid datasets. This article introduces Red Hen Lab with some possibilities for collaboration, demonstrating the utility of a variety of machine learning and AI-based tools and methods to fundamental research questions in multimodal human communication. Supplemental materials are at http://babylon.library.ucla.edu/redhen/KI.

Keywords Multimodal communication · Non-verbal communication · Face and gesture

1 Introduction: Multimodal Communication

Human beings are evolved for elaborate multimodal communication. Cultures support this power. Communicating seems easy to human beings, just as seeing seems easy. But it is immensely complex, including not only vision but also movement, sound, interpersonal interaction, dynamic coordination across agents, conceiving of the intentions of other agents, and so on. Unlike vision, advanced multimodal communication is found in only human beings; there are no good animal models. This report summarizes our effort to gather and develop computational and statistical tools and datasets to help advance research into multimodal communication.

Red Hen Lab (http://redhenlab.org) is an international team of researchers collaborating on computational and statistical tools for locating, identifying, characterizing, and modeling multimodal communication in large datasets. Our core objectives are as follows:

1. Data: constructing a global network for live and timestamped capture of multimodal human communication in multiple languages to create a massive dataset of television news from around the world, including debate and talk shows, supplemented by social media data and digital collections of art works, film, recordings made in laboratories, etc.;
2. Tools: developing and applying a range of of computational and statistical methods, from advanced computer vision, audio signal analysis, and natural language processing to innovations in deep learning, adaptively assembled as modules in integrated high-performance computing pipelines;
3. Multimodal: detecting various universal and culturally variable components of multimodal communication—including speech, sound, gesture, facial movement, gaze direction, paralinguistic elements such as pauses and laughter, music, on-screen text, communicative images, cinematic constructions, and so on, and characterizing their complex communicative intent;
4. Indexed news events and metadata: identifying and recognizing named entities, characterizing events, and

✉ Jungseock Joo
jjoo@comm.ucla.edu

Francis F. Steen
steen@comm.ucla.edu

Mark Turner
turner@case.edu

[1] Communication, UCLA, Los Angeles, USA

[2] Cognitive Science, Case Western Reserve University, Cleveland, USA

clustering topics across sources and time zones to track the spatiotemporal development of events and stories, along with contrasts in narrative framing, selection and presentation bias, and strategies of causal reasoning;
5. Interdisciplinary collaboration: bringing together researchers from disparate disciplines that specialize on different aspects of the analysis of multimodal communication, fostering student involvement and voluntary contributions, and developing a global network with flexible strategies for funding, publications, and workshops.

2 Multimodal Communication in AI

Human communication is enormously diverse in topics and forms. Traditional research methods do not scale up and existing quantitative studies and datasets mainly focus only on verbal communication, ignoring vision, audio, or their interplays. Hence, we construct, collect, and tag large-scale datasets of multimodal human communication (e.g., news videos). Through these resources, we strive to support the fundamental research questions in communication: how humans rely on both verbal and non-verbal cues in order to communicate and interact with other humans; what *meanings* are carried by such multimodal messages; and how we decode and perceive these meanings from multimodal cues.

In artificial intelligence, multimodal human communication has been extensively studied in the context of human-robot interaction (HRI—[1, 2]) including collaborative robots relying on gesture-based interface [3, 4] or semantic learning of human gesture by reinforcement learning [5]. The fundamental objective of these works is to understand the meanings encoded in *human* behaviors such as gestures or speeches, which must be sensed and comprehended by *autonomous agents*. This is especially important as many practical applications and systems of AI, robotics, and machine learning are being deployed in workplace or at home. Such systems will be required to interact, communicate, and collaborate with humans and necessitate a natural interface between humans and agents. Therefore, research on multimodal human communication, e.g., detecting hand gestures, identifying their meanings, or understanding how they augment accompanying verbal speeches, will also provide a critical guideline to research towards various communicative AI systems.

There has also been a great deal of research effort in the fields of machine learning, computer vision, natural language processing and multimedia to study multimodal human communication using automated computational methods, mostly based on data-driven approaches and supervised learning. For example, automated recognition of emotional or affective states from multiple cues such as facial expressions or speeches is a well-studied problem [6–10]. These studies typically extract features from multiple channels and combine them in a joint space to eventually output a discrete label to each example. While many studies in this area have mainly focused on classification tasks with a small number of pre-defined categories, the expressive power of multimodal communication is infinitely large and the interactions between multiple cues are too complex to be well understood by simple predictive approaches. Researchers drawing on Cognitive Film Studies usefully recruit audience comprehension and responses, reasoning with deep semantics about viewers' visuo-spatial narrative primitives [11, 12]. We provide an example analysis in Sec. 4.1 to illustrate the level of sophistication required in a typical gesture analysis.

Another line of research has therefore attempted to address the fluency and richness of multimodal communication by exploring the possibility of automatically generating multimodal cues (e.g., speech and gesture) [13, 14], transferring one to another [15], or learning the model from human performance [16]. Not surprisingly, their target applications are in the domains of HRI or virtual reality where the synthesized multimodal behaviors are employed by a robot or an agent to communicate with human users.

To sum up, multimodal communication is an active research area in communication and AI, and one of its core products is a multimodal interface between a human and an AI-based system ranging from robots to virtual characters. Such a system should be able to understand communicative intents of human speakers and their meanings and to communicate their messages back to humans with multimodal cues. The common challenge that both communities currently face is, however, the lack of sufficient data for multimodal human communication that can be shared by researchers. Red Hen collaboratively develops just such a massive global multimodal human communication dataset.

The remaining part of the paper will review our large-scale datasets and automated tools, which will foster interdisciplinary research on multimodal communication and dialogues.

3 NewsScape: Global TV News Archive

Red Hen's largest holding is NewsScape, an international TV news archive with computational tools for automatically detecting and annotating multimodal communicative events and topics in real time news streams. It now holds over 350,000 h of recordings in a growing number of languages: Russian, Arabic, Brazilian, and Portuguese, for example, have been added recently. The system automatically ingests and processes roughly 150 h of television news each day from miniature capture stations. To make the data

accessible, we extract closed captions, tags them for various linguistic features (grammar, parts of speech, lemmas, paralinguistic elements, conceptual frames, etc.), and then parse the news stream data into different topics by clustering news stories using multimodal cues [17]. We automatically detect and recognize faces, objects, and other entities in images and texts, semantically represented (e.g., who, when, where) and organized into hierarchical topic structures. These operations generate a great variety of visual and textual metadata for use in subsequent studies.

There are advantages to using TV news as Red Hen's main data. First, it provides multimodal data at massive scale so researchers can discover diverse patterns of communicative activities including conversations between people, public speeches of politicians, narrations of news anchors or reporters. Second, the news covers important events around the world, pertaining to contemporary or historical social issues, creating tight connections to research in broader social sciences. Third, researchers are permitted to use the news data freely for research, without concerns about copyright or privacy.

The recent breakthrough in machine learning and artificial intelligence has been largely attributed to the availability of large scale datasets [18–20]. Similarly, our news archive can also facilitate collaborative research in multimodal communication, AI and machine learning.

4 Our Projects

Several ongoing projects by linguists, political communication scholars, and computer scientists either rely on the Red Hen data or utilize Red Hen computational tools to analyze data. Due to the space limitation, we briefly summarize the main objectives and scopes of a few selected projects. More details can be found in the references.

- *Multimodal topic detection and tracking (text, image)* We parse the incoming news stream data by topic using multimodal cues [17]. We use deep learning-based image features and text features, semantically represented (e.g., who, when, where). We combine them in hierarchical topic structures. The news topic clusters are constantly discovered, modified, and merged into other topics in real time.
- *Syntactic parsing of news visuals (image, audio)* Red Hen uses computer vision modules to recognize on-screen text, faces, human attributes, and object and scene types. For audio, Red Hen uses forced alignment, speaker diarization, gender detection, speaker recognition, and acoustic fingerprinting. Red Hen uses traditional computer vision techniques and state-of-the-art deep learning models.
- *Analysis of Faces (image, text)* Human faces exhibit a wide variety of expressions and emotions. To recognize them, [21] developed a hierarchical model to judge the perceived personalities of politicians automatically from their facial photographs and detected traits. The results are predictive of election outcomes. This example illustrates how computational modeling can perform massive analysis of facial data and infer evaluation. [22] analyzed media bias in the selection the facial images of politicians. These studies investigate how advanced cognitive activities (inferring personality, evaluating performance) depend on nonverbal cues. They explore how mass multimodal media can influence public opinion through visual persuasion [23].
- *Multimodal Constructions (video, text)* [24] present an analysis of the way in which deictics in English versus Russian (e.g. "here" and "now") are used in news broadcasts to establish joint attention, in patterns of usage that differ from those in face-to-face communication. [25] further explore such extensions of grammatical constructions in multimodal broadcast settings. [26] explores patterns of deictic words and pointing gestures in such broadcasts, as when, e.g., the on-screen speaker says "If you have any questions ... we made a video and we will link to it, right there" and points to and looks up to her right. The word "HERE" then appears on-screen where she is looking and pointing (see the example at http://babylon.library.ucla.edu/redhen/KI).

4.1 An Example Analysis: Viewpoint and Epistemic Stance in TV News

This section presents a concrete example of our analysis on multimodal communication using video footage from our dataset and demonstrates how our automated tagging tool is used for the analysis.

Speakers express their viewpoint, attitude, or perspective on the meaning of what they are saying, often framing its source. Linguists use the terms *viewpoint, epistemic stance* and *evidentials* for such patterns of expression. The stance a speaker or a listener adopts towards some content can be expressed wordlessly, with a shrug, a stare, a gasp, a wave of the hand, a smack, a tearful eye, a hollow laugh or delicate modulations of the speed, pitch, and quality of the voice. A speaker in the act of presenting claims may for instance indicate epistemic distance—that is, a viewpoint of doubt or distrust—from these claims. This epistemic distance is crucial to the communication, but is often irretrievably lost in a mere verbal transcript.

A broad spectrum of auditory and visual forms is available to the speaker to convey the speaker's viewpoint on the content they are presenting. There are common facial

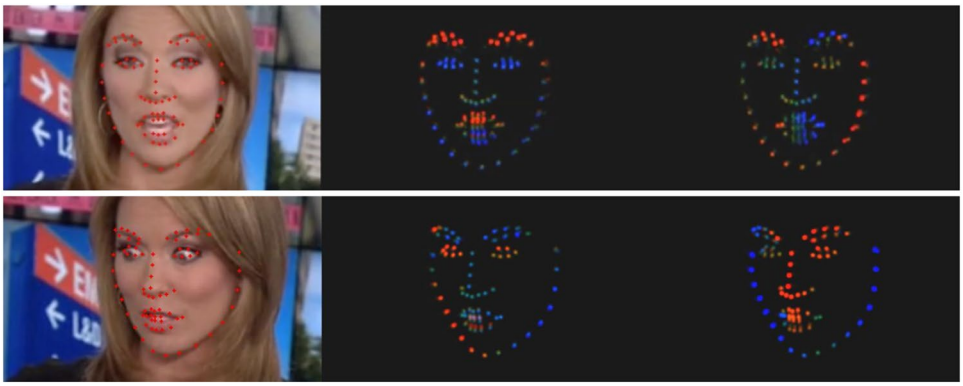

Fig. 1 Facial expressions of Brooke Baldwin captured by automated facial keypoint detection. (Left) original images (middle) vertical movements (red: upward, blue: downward) (right) lateral movements (red: left, blue: right). **a** Baldwin raises her eyelid. **b** She executes a side-eye and shakes her head to her right. The video is at http://babylon.library.ucla.edu/redhen/KI

expressions for indicating epistemic distance, for example, such as side-eye, in which the speaker looks to the side at the moment of expressing the apex of the content. Other such facial expressions include fluttering the eyelids, or directing the gaze to the side and down, or lifting the head slightly to the side. Prosody for such viewpoint includes emphasizing a word with a brief pause before and after the word and an elongation of the word with very brief pauses between syllables.

In the supplemental video, consider CNN newscaster Brooke Baldwin's use of multimodal expressions around "inconsistencies" at 48 s: the word itself is said with a pause before and after and with an elongation including slight separations between the syllables-in–con–SIS–tencies with the pitch highest on "sis". The human ear is highly sensitive to such modulations. Now add to this a pitch of rise to "sis" and then a fall. In parallel, she flutters her eyelids, makes a sweeping left–right motion with her hand, and quickly shakes her head back and forth. The whole performance enhances and underlines the verbal message, mapping the discrepancy between the hospital's claim and the CDC's findings onto an abstract space where they are compared and found to differ. Every native speaker knows these multimodal forms and uses them for semantic interpretation.

In order to detect such subtle facial motions of speakers, we developed a tool for automated facial landmark detection from videos. Figure 1 shows two example frames with facial expressions detected by automated facial keypoint detection. By tracking facial keypoints, we recognize subtle facial motions such as raised eyelid or head shaking. We used the implementation of [21]. The model first locates a mean shape of keypoints in the facial region and iteratively refines their locations by computing the first order gradients of the objective function with respect to the current locations, approximated on the basis of the image features evaluated at those locations.

5 Conclusion

Red Hen deploys the contributions of researchers from complementary fields, from AI and statistics to linguistics and political communication, to create rich datasets of parsed and intelligible multimodal communication and to develop tools to process these data and any other data susceptible to such analysis. Red Hen's social organization and computational tools are designed for reliable and cumulative progress in a dynamic and extremely challenging field: the systematic understanding of the full complexity of human multimodal communication. The study of how human beings make meaning and interpret forms depends upon such collaboration.

References

1. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3):143–166
2. Jaimes A, Sebe N (2007) Multimodal human–computer interaction: a survey. Comput Vis Image Underst 108(1):116–134
3. Ende T, Haddadin S, Parusel S, Wüsthoff T, Hassenzahl M, Albu-Schäffer A (2011) A human-centered approach to robot gesture based communication within collaborative working processes. In: Intelligent robots and systems (IROS), 2011 IEEE/RSJ international conference on, pp 3367–3374. IEEE
4. Gleeson B, MacLean K, Haddadi A, Croft E, Alcazar J (2013) Gestures for industry: intuitive human–robot communication from human observation. In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, pp 349–356. IEEE Press
5. Yanik PM, Manganelli J, Merino J, Threatt AL, Brooks JO, Green KE, Walker ID (2014) A gesture learning interface for simulated robot path shaping with a human teacher. IEEE Trans Hum Mach Syst 44(1):41–54
6. Chen LS, Huang TS (2000) Emotional expressions in audiovisual human computer interaction. In: Multimedia and Expo, 2000. ICME 2000. 2000 IEEE international conference on, vol 1, pp 423–426. IEEE

7. Busso C, Deng Z, Yildirim S, Bulut M, Lee CM, Kazemzadeh A, Lee S, Neumann U, Narayanan S (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on multimodal interfaces, pp 205–211. ACM
8. Pantic M, Sebe N, Cohn JF, Huang T (2005) Affective multimodal human–computer interaction. In: Proceedings of the 13th annual ACM international conference on multimedia, pp 669–676. ACM
9. Caridakis G, Castellano G, Kessous L, Raouzaiou A, Malatesta L, Asteriadis S, Karpouzis K (2007) Multimodal emotion recognition from expressive faces, body gestures and speech. Artificial intelligence and innovations 2007: from theory to applications, pp 375–388
10. Soleymani M, Pantic M, Pun T (2012) Multimodal emotion recognition in response to videos. IEEE Trans Affect Comput 3(2):211–223
11. Suchan J, Bhatt M (2016) Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies. IJCAI, pp 2633–2639
12. Suchan J, Bhatt M (2016) The geometry of a scene: on deep semantics for visual perception driven cognitive film studies. WACV, pp 1–9
13. Cassell J, Kopp S, Tepper P, Ferriman K, Striegnitz K (2007) Trading spaces: how humans and humanoids use speech and gesture to give directions. Conversational informatics, pp 133–160
14. Kopp S, Bergmann K, Wachsmuth I (2008) Multimodal communication from multimodal thinking towards an integrated model of speech and gesture production. Int J Semant Comput 2(01):115–136
15. Marsella S, Xu Y, Lhommet M, Feng A, Scherer S, Shapiro A (2013) Virtual character performance from speech. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation, pp 25–35. ACM
16. Huang C-M, Mutlu B (2014) Learning-based modeling of multimodal behaviors for humanlike robots. In: Proceedings of the 2014 ACM/IEEE international conference on human–robot interaction, pp 57–64. ACM
17. Li W, Joo J, Qi H, Zhu S-C (2017) Joint image-text news topic detection and tracking by multimodal topic and-or graph. IEEE Trans Multimedia 19(2):367–381
18. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Computer vision and pattern recognition
19. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732
20. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision, pp 740–755
21. Joo J, Steen FF, Zhu S-C (2015) Automated facial trait judgment and election outcome prediction: social dimensions of face. In: Proceedings of the IEEE international conference on computer vision, pp 3712–3720
22. Groeling T, Li W, Joo J, Steen FF (2016) Visualizing presidential elections. In: APSA annual meeting
23. Joo J, Li W, Steen FF, Zhu S-C (2014) Visual persuasion: inferring communicative intents of images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 216–223
24. Tore N, Anna E, Janda LA, Makarova A, Steen F, Turner M (2013) How "here" and "now" in Russian and English establish joint attention in TV news broadcasts. Russ Linguist 37(3):229–251
25. Steen FF, Turner M (2013) Multimodal construction grammar. In: Borkent M, Barbara D, Jennifer H (eds) Language and the creative mind. CSLI Publications, University of Chicago Press. Stanford, CA, pp 255–274
26. Turner M (2017) Multimodal form-meaning pairs for blended classic joint attention. Linguist Vanguard. doi:10.1515/lingvan-2016-0043