Semantic analysis of manual communication using monocular vision and the structure of American Sign Language

Human communication is fundamentally multi-modal. Leveraging as many modes as possible can have a significant impact on human-computer and human-robot interaction. I propose as my research topic the study of non-verbal, manual communication. To demonstrate the efficacy of my work, I will initially focus on the translation of American Sign Language (ASL), a well-studied structured language with readily available and well-annotated corpora. Additionally, solving this problem on accessible monocular hardware has the potential for immediate humanitarian impact. But beyond the initial goal, my hope in the long term is to generalize this process for any context, allowing for more productive symbiosis between humans and robots.

The first important problem is the computer vision challenge: how can one convert a single video stream without depth information into a form that is both rich in information and easily understandable by a machine? Analysis of sign languages in the latter half of the 20th century attempted to identify *cheremes*, or perceptually distinct units of hand motion analogous to the phonemes of spoken language [1]. Just as phonemes are composed of different elements (tongue shape, lip position, tongue position, etc.), cheremes are also composed of individually measurable elements. Currently, the literature identifies five such elements: hand shape, palm orientation, hand location (relative to each other and the rest of the body), hand motion, and facial expression, a proxy for emotion. Despite being non-manual, emotion is as important to sign language as tone and emphasis are to spoken language. Recent advances in the design of convolutional neural networks (CNNs) in both industry and academia alike have made progress quantifying some of these elements individually. Mask regional CNNs have become efficient at identifying multiple objects' shapes and locations in still images [2], and 3D CNNs have been used to analyze the emotional content of movements [3]. To solve the computer vision challenge, I plan to combine and expand upon these techniques, taking advantage of the specificity afforded by focusing solely on human hands and faces. The second important task is the generation of a semantically rich representation from the information-rich metrics output by the computer vision algorithm. A possible starting point is an algorithm like Word2Vec, which converts written words into a vector representation with a correlation between Euclidean distance and semantic similarity [4]. However, such an algorithm benefits from an enormous amount of written language, dwarfing any existing ASL corpus, that can be used for training. Another possible approach is to formulate the problem as a sort of translation game, where models compete to translate each other's outputs. This allows for the competitive reinforcement learning that was so successful in game-playing machines like AlphaGo [5], though the formulation would be a significant undertaking unto itself. Regardless of the final approach, I am confident that the recent advances in machine learning have put a solution within reach.

This particular research topic benefits from the immediately apparent impact of improving translation between sign languages and spoken languages. Current industry solutions require dedicated, immobile hardware [6, 7], require hardware with minimal penetration due to cost and specificity [7, 8, 9], or must neglect some chereme elements [8, 9]. Solving the monocular vision challenge alone can expand the availability of translation to the billions of individuals with a webcam or a smart phone. Even a crude solution can greatly improve sign language education tools, allowing for feedback without the need for a human instructor. Additionally, the semantic analysis of the signs themselves could enable the translation of a given sign language to another country's spoken language and vice versa without the need for intermediate translation. This halves the number of translation steps in which meaning can be lost and increases international accessibility for the deaf and hard-of-hearing.

Successful research on this topic can also have an effect on the field of human-robot interaction. Current research explores the two problems of conveyance of intent [10] and ascertainment of intent from non-verbal cues [11]. Even without the structure of a sign language, the latter still relies on the conversion of cheremes to a form robots can understand, and the former requires the inverse conversion. Again, solving the vision challenge and inverting the model can provide a powerful tool for tackling both of these problems. As for semantic analysis, while the sign language vector representations may be of limited use here (outside of the obvious application to interaction between robots and the deaf), the generalized process can be used to analyze a given gesture "language" in any context, given a well-constructed corpus of examples. And again, an inversion of any model can be used by robots to help convey their own intent in the same context.

In summary, I propose as my research topic the derivation of intent from manual cues using monocular vision. It can be broken down into the two complimentary problems of visual identification of chereme elements and the semantic analysis of those cheremes. While this topic has two challenges, tackling either by itself can still have a significant impact on both accessibility for the hard-of-hearing and the field of human-robot interaction. Human communication is fundamentally multi-modal. It is my hope that my contributions here will, directly or indirectly, advance the human condition by taking advantage of as many modes as possible.

[1] Liddell, S., Johnson, R. American Sign Language: The phonological base. 1989. Sign Language Studies 64. 197–277.

[2] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. arXiv:1703.06870, 2017.

[3] Chen, W., Picard, R. Predicting perceived emotions in animated GIFs with 3D convolutional neural networks. Proc. IEEE International Symposium on Multimedia, December 2016.

[4] Mikolov et al. Efficient estimation of word representation in vector space. arXiv:1301.3781, September 2013.

[5] Silver, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.

[6] www.kintrans.com

[7] www.signall.us

[8] SignAloud, www.youtube.com/watch?v=l01sdzJHCCM

[9] www.motionsavvy.com

[10] Unhelkar et al, Human-Robot Co-Navigation using Anticipatory Indicators of Human Walking Motion. IEEE International Conference on Robotics and Automation (ICRA), 2015.

[11] Huang et al, Enabling Robots to Communicate Their Objectives. Robotics: Science and Systems (R:SS), 2017.