A practical mSVG interaction method for patrol, search, and rescue aerobots

Ayodeji O. Abioye[1], Stephen D. Prior[1], Glyn T. Thomas[1], Peter Saddington[2], & Sarvapali D. Ramchurn[1]
[1]Fac. of Eng. & the Env., University of Southampton, UK. [2]Tekever Ltd, Southampton, UK

Abstract

This paper briefly presents the multimodal speech and visual gesture (mSVG) control for aerobots at higher nCA autonomy levels, using a patrol, search, and rescue application example. The developed mSVG control architecture was presented and briefly discussed. This was successfully tested using both MATLAB simulation and python based ROS Gazebo UAV simulations. Some limitations were identified, which formed the basis for the further works presented.

KEYWORDS: mSVG, aerobot, HCI, visual gesture, speech

Introduction

This paper is interested in how the increasing leagues of human operators interact with small multi-rotor UAVs. According to Green *et al.* (2007), "It is clear that people use speech, gesture, gaze and non-verbal cues to communicate in the clearest possible fashion." Abioye *et al.* (2016, 2017) identified the need for smart and intuitive control interaction methods for aerobots (aerial robots) on higher nCA autonomy levels. Such aerobots could include a patrol, search, and rescue robot in the Alps, Southcentral Europe. If a UAV could be developed to patrol dangerous regions of the Alps, providing signposting to climbers, alerting search and rescue teams of any incident, and supporting search and rescue team operations; and if the patrol UAV is meant to interact with climbers when needed, perhaps an intangible HHI-like multimodal speech and visual gesture (mSVG) interaction method could proof very useful in such climber aerobotic interaction.

Literature Review - Multimodal Interfaces

Cacace, Finzi, and Lippiello (2016) investigated multimodal speech and gesture communication with multiple UAVs in a search and rescue mission, using the Julius framework (Lee, Kawahara and Shikano, 2001) and Myo device for speech and gesture respectively. Fernandez *et al.* (2016) investigated the use of natural user interfaces (NUIs) in the control of small UAVs using the Aerostack software framework. Harris and Barber (2014) and Barber et al. (2016) investigated the performance of a speech and gesture multimodal interface for a soldier-robot team communication during an ISR mission, even considering complex semantic navigation commands such as "*perch over there* (speech + pointing gesture), *on the tank to the right of the stone monument* (speech)" (Borkowski and Siemiatkowska, 2010; Barber, Howard and Walter, 2016). In a related research by Hill, Barber, and Evans (2015), the researchers suggested that multimodal speech and gesture communication was a means to achieving an enhanced naturalistic communication, reducing workload, and improving the human-robot communication experience. Kattoju *et al.* (2016) also investigated the effectiveness of speech and gesture communication in soldier-robot interaction. Cauchard *et al.* (2015) and Obaid *et al.* (2016) conducted elicitation study to determine intuitive gestures for controlling UAVs.

Research - Multimodal Speech and Visual Gesture (mSVG)

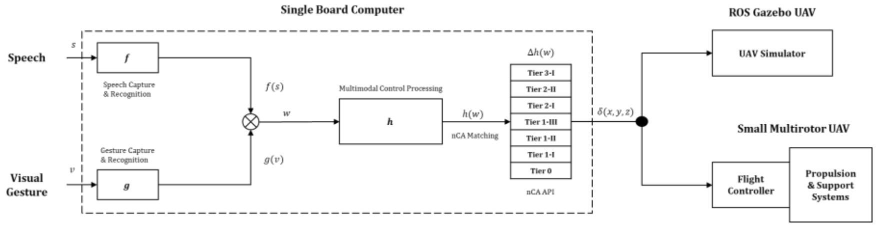

Figure 1: mSVG design architecture – control capture, processing, and execution.

The mSVG technique is basically the multimodal combination of speech and visual gesture, a method that leverages familiar human-human type interaction, in human aerobotic interaction. This combination could be sequential or complementary. The underlying architecture of how this technique is designed to work is as described in Figure 1. Speech is captured via a microphone, processed and recognised using the CMU Sphinx ASR with custom-defined phonetic dictionary containing only the set of command vocabulary, in order to increase recognition speed and accuracy. The speech input, $\mathcal{U}_{speech} = [s_1, s_2, s_3, \dots, s_n]$, is processed into the control symbol $f(s)$. Visual gesture is captured via a camera connected to the aerobot SBC computer. In the preliminary work, a simple finger-coded visual gesture control commands set was developed to be recognised through a combination of two OpenCV algorithms – Haar cascade for hand tracking and convex hull for finger counting. The visual gesture control command, $\mathcal{U}_{gesture} = [g_1, g_2, g_3, \dots, g_m]$, is also being processed into control symbols $g(v)$. These are then combined into a standardized control symbol, $w = f(s) + g(v)$, which is then passed into the multimodal control processing (MCP) framework. $h(w)$ is the resultant control output generated after the multimodal combination of both the speech and the visual gesture input. $\delta(x, y, z) = \Delta h(w)$ Where Δ is a function generated by the nCA API to modify the MCP output, $h(w)$, to enable compatibility with multiple nCA navigational control autonomy levels. $\delta(x, y, z)$ is the increment/decrement change in 3-dimensional position of the UAV with respect to its previous position. A mathematical set model was developed and used to describe the computational algorithm mapping speech and visual gestures control symbols to UAV control operations to be executed.

Results - MATLAB and ROS Gazebo Simulation

Based on the mathematical set model, the mSVG control navigation was simulated in MATLAB, which was then implemented in python for easy integration of algorithm on a single board computer (in this case, Odroid XU4 SBC), and simulated on a rotors gazebo firefly UAV simulator in an open world environment. In each case, a series of command such as 'go forward', 'go up half metre', 'go right one metre', 'hover at three metre', 'and', 'hover', 'go forward backward two half metre', 'patrol', etc. were successfully tested.

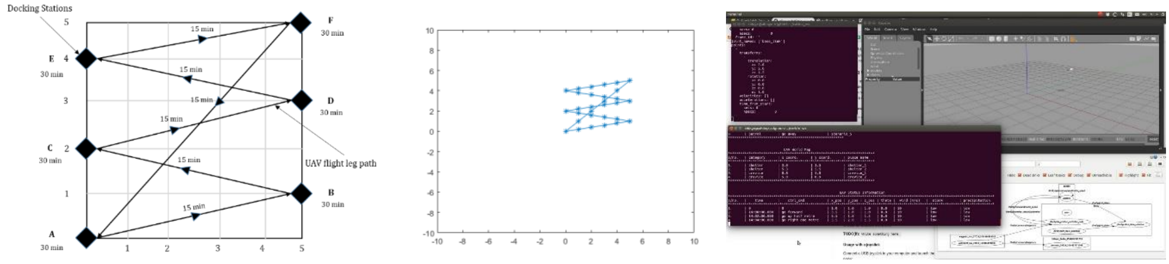

Figure 2: Specified aerobot patrol grid, MATLAB simulation, and ROS Gazebo Simulation

Discussion, Conclusion, and Further Works

The main limitations of the proposed system is 1) its susceptibility to speech corruption during capture, due to the noise generated by the multirotor propulsion systems and other loud ambient noise such as in stormy weathers, 2) the effect of poor visibility level on visual gesture capture, as could be the case at night, or in cloudy or misty weather. The next phase of this research is already underway to determine the range of effectiveness of the mSVG method under varying noise and visibility levels. This could inform the possibility of working around this or developing techniques that may extend this range, thereby extending the usefulness of the propose mSVG technique over a much wider application area. Also, a comparison of the mSVG and RC joystick in terms of training time, same nCA Tier task completion rate, and cognitive workload requirement, is currently being conducted.

References

Abioye, A. O., Prior, S. D., Thomas, G. T. and Saddington, P. (2016) ‘The Multimodal Edge of Human Aerobotic Interaction’, in Blashki, K. and Xiao, Y. (eds) *International Conferences Interfaces and Human Computer Interaction*. Madeira, Portugal: IADIS Press, pp. 243–248.

Abioye, A. O., Prior, S. D., Thomas, G. T., Saddington, P. and Ramchurn, S. D. (2017) ‘Multimodal Human Aerobotic Interaction’, in Isaías, P. (ed.) *Smart Technology Applications in Business Environments*. IGI Global, pp. 39–62.

Barber, D. J., Howard, T. M. and Walter, M. R. (2016) ‘A multimodal interface for real-time soldier-robot teaming’, 9837, p. 98370M. doi: 10.1117/12.2224401.

Borkowski, A. and Siemiatkowska, B. (2010) ‘Towards semantic navigation in mobile robotics’, *Graph Transformations and Model-Driven Engineering*, pp. 719–748.

Cacace, J., Finzi, A. and Lippiello, V. (2016) ‘Multimodal Interaction with Multiple Co-located Drones in Search and Rescue Missions’, *CoRR*, abs/1605.0, pp. 1–6.

Cauchard, J. R., Jane, L. E., Zhai, K. Y. and Landay, J. A. (2015) ‘Drone & me: an exploration into natural human-drone interaction’, *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 361–365. doi: 10.1145/2750858.2805823.

Fernandez, R. A. S., Sanchez-lopez, J. L., Sampedro, C., Bavle, H., Molina, M. and Campoy, P. (2016) ‘Natural User Interfaces for Human-Drone Multi-Modal Interaction’, in *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*. Arlington, VA USA: IEEE, pp. 1013–1022.

Green, S., Chen, X., Billinnghurst, M. and Chase, J. G. (2007) ‘Human Robot Collaboration: an Augmented Reality Approach a Literature Review and Analysis’, *Mechatronics*, 5(1), pp. 1–10. doi: 10.1115/DETC2007-34227.

Harris, J. and Barber, D. (2014) ‘Speech and Gesture Interfaces for Squad Level Human Robot Teaming’, in Karlsen, R. E., Gage, D. W., Shoemaker, C. M., and Gerhart, G. R. (eds) *Unmanned Systems Technology Xvi*. SPIE. doi: 10.1117/12.2052961.

Hill, S. G., Barber, D. and Evans, A. W. (2015) ‘Achieving the Vision of Effective Soldier-Robot Teaming : Recent Work in Multimodal Communication’, *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pp. 177–178. doi: 10.1145/2701973.2702026.

Kattoju, R. K., Barber, D. J., Abich, J. and Harris, J. (2016) ‘Technological evaluation of gesture and speech interfaces for enabling dismounted soldier-robot dialogue’, 9837, p. 98370N. doi: 10.1117/12.2223894.

Lee, a., Kawahara, T. and Shikano, K. (2001) ‘Julius — an Open Source Real-Time Large Vocabulary Recognition Engine’, *Eurospeech*, pp. 1691–1694.

Obaid, M., Kistler, F., Kasparaviciute, G., Yantaç, A. E. and Fjeld, M. (2016) ‘HowWould You Gesture Navigate a Drone? A User-Centered Approach to Control a Drone’, in *Proceedings of the 20th International Academic Mindtrek Conference*. Tampere, Finland: ACM New York, NY, USA, pp. 113–121. doi: 10.1145/2994310.2994348.