# A survey of technologies supporting design of a multimodal interactive robot for military communication

Sheuli Paul
*Autonomous and Radiology Technology Section,
Defence Research and Development Canada – Suffield Research Centre,
Ralston, Canada*

Received 2 November 2022
Revised 4 November 2022
15 August 2023
6 September 2023
10 September 2023
13 September 2023
Accepted 13 September 2023

## Abstract

**Purpose** – This paper presents a survey of research into interactive robotic systems for the purpose of identifying the state of the art capabilities as well as the extant gaps in this emerging field. Communication is multimodal. Multimodality is a representation of many modes chosen from rhetorical aspects for its communication potentials. The author seeks to define the available automation capabilities in communication using multimodalities that will support a proposed Interactive Robot System (IRS) as an AI mounted robotic platform to advance the speed and quality of military operational and tactical decision making.

**Design/methodology/approach** – This review will begin by presenting key developments in the robotic interaction field with the objective of identifying essential technological developments that set conditions for robotic platforms to function autonomously. After surveying the key aspects in Human Robot Interaction (HRI), Unmanned Autonomous System (UAS), visualization, Virtual Environment (VE) and prediction, the paper then proceeds to describe the gaps in the application areas that will require extension and integration to enable the prototyping of the IRS. A brief examination of other work in HRI-related fields concludes with a recapitulation of the IRS challenge that will set conditions for future success.

**Findings** – Using insights from a balanced cross section of sources from the government, academic, and commercial entities that contribute to HRI a multimodal IRS in military communication is introduced. Multimodal IRS (MIRS) in military communication has yet to be deployed.

**Research limitations/implications** – Multimodal robotic interface for the MIRS is an interdisciplinary endeavour. This is not realistic that one can comprehend all expert and related knowledge and skills to design and develop such multimodal interactive robotic interface. In this brief preliminary survey, the author has discussed extant AI, robotics, NLP, CV, VDM, and VE applications that is directly related to multimodal interaction. Each mode of this multimodal communication is an active research area. Multimodal human/military robot communication is the ultimate goal of this research.

**Practical implications** – A multimodal autonomous robot in military communication using speech, images, gestures, VST and VE has yet to be deployed. Autonomous multimodal communication is expected to open wider possibilities for all armed forces. Given the density of the land domain, the army is in a position to exploit the opportunities for human–machine teaming (HMT) exposure. Naval and air forces will adopt platform specific suites for specially selected operators to integrate with and leverage this emerging technology. The possession of a flexible communications means that readily adapts to virtual training will enhance planning and mission rehearsals tremendously.

**Social implications** – Interaction, perception, cognition and visualization based multimodal communication system is yet missing. Options to communicate, express and convey information in HMT setting with multiple options, suggestions and recommendations will certainly enhance military communication, strength, engagement, security, cognition, perception as well as the ability to act confidently for a successful mission.

**Originality/value** – The objective is to develop a multimodal autonomous interactive robot for military communications. This survey reports the state of the art, what exists and what is missing, what can be done and possibilities of extension that support the military in maintaining effective communication using

Journal of Defense Analytics and Logistics
Vol. 7 No. 2, 2023
pp. 156-193
Emerald Publishing Limited
2399-6439
DOI 10.1108/JDAL-11-2022-0010

multimodalities. There are some separate ongoing progresses, such as in machine-enabled speech, image recognition, tracking, visualizations for situational awareness, and virtual environments. At this time, there is no integrated approach for multimodal human robot interaction that proposes a flexible and agile communication. The report briefly introduces the research proposal about multimodal interactive robot in military communication.

**Keywords** Sensor fusion, Human–machine teaming (HMT), Multimodal communication, Multimodal interactive robotic system (MIRS), Spoken dialogue system (SDS), Visual story telling (VST), Virtual story telling (VST), Virtual personal assistant in VE

**Paper type** Literature review

## 1. Introduction

Communication is a fundamental human need. Human communication is interactive and multimodal. Effective communication in a time-critical setting is vital for all types of military operations (i.e. tactical engagements, Combat Search and Rescue (CSAR), Joint Intelligence Surveillance and Reconnaissance (JISR), Joint Terminal Attack Controller (JTAC) and Information Operations (IO)). Effective communication is dynamic, interactive and influential. There is a need to develop automation capabilities that will enable the military to successfully execute missions in complex environments while protecting the forces engaged in combat. This paper surveys the state of the art, identifying what exists and what are missing, what can be done and possibilities to extend existing technology. There is ongoing progress in machine-enabled speech, image recognition, visualizations for situational awareness and virtual environments (VEs). At this time, there is no integrated approach for multimodal human robot interaction (HRI) that achieves a flexible and agile communication.

Effective HRI requires an ability to understand, design and evaluate robotic systems with a goal to execute collaborative functions for or with humans. An essential element of the interaction process is a multimodal form of communication. Modalities include voice, gesture, image and graphics and data visualization. Challenges in these systems include maintaining communication, joint action and human-aware execution. They can be understood in terms of cognitive skills that they mandate (Bensalem *et al.*, 2008):

(1) A joint goal, which has been previously established and agreed upon (typically through dialogue);

(2) A physical environment, estimated through the robot’s exteroceptive sensing capabilities and augmented by inferences drawn from previous observations;

(3) A belief state that includes *a priori* common-sense knowledge and mental models of each of the agents involved (the robot and its human partners).

This paper presents developments in the robotic interaction field with the objective of identifying essential technological developments that set conditions for robotic platforms to function autonomously. This includes the key aspects in HRI, unmanned automatic system (UAS), visualization, VE and prediction. The paper then describes the aspects in the application areas that will require extension and integration to enable prototyping of the interactive robot system (IRS).

The rest of this paper is organized as follows: Section 2 defines key aspects of communication and military applications that benefit from its multimodal forms. Section 3 describes challenges and opportunities for this technology. Section 4 visits the critical threads in the literature to define the central parameters of multimodality and surveys the current state of the art. Section 5 adds some unifying thoughts to the survey and describes a multimodal interactive robotic systems (MIRS) concept. Finally, Section 6 presents some brief conclusions.

## 2. Multimodalities and communication

Multimodal communication comprises a combination of multiple heterogeneous sources for interactions. Effective communication often requires a meaningful representation of multiple data sources. Multimodal systems process information from different human communication channels at multiple levels of abstraction. These systems emphasize abstract levels of processing, explicit representations of the dialogue context, the user and investigations of the users' beliefs, intentions, attitudes, capabilities and preferences. These components are media, mode analysis and design, interaction and context management, user modelling and knowledge sources. Multi-media systems consist of various types of speech, graphical and direct manipulative interfaces with different modules (Jokinen and Raike, 2003). A conversational interactive interface leads to natural interactive systems. A multimodal signal stimulates the sensory system's response to the environment. A multimodal signal currently does not have a broadly established definition.

This section presents a brief discussion of multimodal human communication using various contextual metaphors and how this is used in human interactions. It clarifies how multimodal communication is a fundamental human capability and how this can be extended from human robot communication to military and robot communication.

Human interactions are multimodal. Each interaction uses multiple modes to listen, perceive, sense, gesture, visualize and taste. Such interactions are distinguished by modes, modalities and mediums. The modes are related to human sensory systems such as visual, auditory and tactile. The modality can be perceived as text, images and tactile sensation that are not easily represented internally by a machine. For a machine, the medium is an output device such as a screen, speaker or haptic technology, i.e. a feedback device. The medium can be an interactive modular system composed of independent elements. For instance, visualization is a type of media (Yau, 2013). Multimodal integration is a type of fusion engine. The meanings of input streams can vary with time, user and context (Rousseau *et al.*, 2006).

The images in Plates 1–12 show how humans communicate using different modes. Some humans are speaking, some are using hand gestures and others are using devices to portray visual descriptions. VR simulations and AR based heads-up displays assist in perceptions by portraying visual description and explanation through images, pictorial representations or clarifications of similar situations. Thus VR and AR devices are helpful for supporting

**Plate 1.**
Communication using texts, gestures, images and devices

**Plate 2.**
Face-to-face interactions using hands and digital media based pictorial representations

endurance and gaining confidence to embrace challenging situations. Furthermore, these devices are foreseen to improve performance in these situations via training or other forms of support for the human. This is accomplished with a user interacting with virtual media-based communications.

Interaction and contextual reflection 65% of the information in a human-human interaction is non-verbal e.g. mixed with visual cues and gestures that express human thoughts, reflect mood state, support a response, present complements, furnish accents and thus, adjust verbal information (Birdwhistell, 1970). The images in Plates 1–3 show face-to-face discussion, use of text, hand gestures, images on the phone, computer and tablet. The

**Plate 3.** Communications using gaze and activities

**Plate 4.** Communication via visualization or media, e.g. display

**Plate 5.** Interpretation, perceptions and interactions using multiple communicative modes and medias

**Plate 6.** Communication, perceptions, descriptions, clarifications, presentation using text, pictorial representations, media and gestures

participants use, speech, gesture, facial expression and different bodily position which influence interpretations. When people engage in face to face interaction, they may discuss information of very different kinds, which in modern contexts may be displayed on a variety of electronic devices such as phones and computers. Interaction and communication using different modalities and modes can vary substantially based on contextual situations.

### *2.1 Multimodal interaction, fusion and information processing*

An appreciation of how humans perceive and transmit signals through different modalities is central for designing a multimodal system. A key to this process is the interpretation of multimodal signals with an appropriate level of fusion and fission of information across the

U.S. Army veteran interacts with the Bravemind VR therapy to safely relive his deployment... [+]

PHOTO CREDIT: STEPHANIE DAVIS KLEINMAN FRENCH. PHOTO COPYRIGHT © ALBERT "SKIP" RIZZO

**Phobias & Anxiety Disorders**

**Plate 7.**
Military user with VR simulation and AR heads-up display

**Plate 8.**
Military communication: patrolling/surveillance or conducting training

**Plate 9.**
Soldier wearing a Visual Augmentation System and related hand held device while training

**Note(s):** Visualization based technology will be helpful in this case

**Plate 10.**
Military personnel on operation, e.g. advance to contact using vision and posture

modality channels. This is complicated by the differences in the character of the information in different input and output modality channels. These differences can be accentuated by how the messages are encoded into the modalities. Thus, for effective information presentation, there is a need for clarity in how different modalities complement and contradict each other. Furthermore, the means and methods for precise interpretation of information must be specified clearly.

Two pieces of information in different modalities complement each other if they contribute toward the same meaning but could not be easily interpreted in a meaningful way as standalone, unimodal information. Different types of modalities may complement each other. For instance, in a car collision, the common modalities (e.g. vision, audition and tactility) may convey the same message: we see one car approaching and hit a second car; we hear the sound of collision; we feel the vibration of impact. The perceptions of all these modalities convey the same information and complement each other. Alternatively, the information is redundant if the same complete information is conveyed in both modalities independently (e.g. the Tram 3B icon moving on the map at the point that denotes railway station along with simultaneous audio: "Tram 3B leaves railway station"). Information in different channels can also be contradictory. For example, a person could circle a location on a map and say simultaneously "this bus stop here", but the circle does not contain any references to bus stops. Although this is not reasonable as a system output, it may occur as a user input (Norris, 2004).

The fusion of information refers to the analysis and integration of input information from different modalities into a composite, meaningful form. The fission of information is the opposite process, which occurs during information generation. Fission refers to the division of information into appropriate modalities with a goal of generating an effective presentation. Fusion exists in three levels: lexical, syntactic and semantic. Lexical fusion happens on the hardware and software levels, for example, when selecting objects with the shift key down. Syntactic fusion involves combining data to form a syntactically complete command. Finally, semantic fusion concerns the detailed functionality of the interface and defines the meanings of the actions. Merging various inputs for information processing through fusion and combination of output media for the purpose of presentation through fission are fundamental

**Plate 11.**
Military deployed on field training, e.g. a temperate climate

**Plate 12.**
Military deployed on field training, e.g. an arid climate

concepts in multimodal communication. Concepts related to fusion and fission in multimodal and multimedia-based interaction are discussed in Maragos *et al.* (2008).

In multimodal language communication and interaction, the concepts and meanings are extracted from different modalities. For instance, speech, touch and vision can be combined to produce a single meaning or a representation for the user's action. Multimodal fusion can be applied to build a coherent and consistent interpretation of different modal sources. Fusion-based procedures may select the best candidate interpretation using models based on weights and combinations of information (Oviatt *et al.*, 2017). For example semantic fusion typically takes place in two levels: First, inputs are combined into events that are processed by the higher-level interpreter, which uses its knowledge about users' intention. Second, the context is analyzed to finalize and disambiguate the input (Ida, 2020).

HRI needs to take place in dynamic, partially unknown environments that are not currently designed for robots. The robot needs to understand and interpret a variety of situations with rich semantics, and conduct physical interactions with humans that require fine, low-latency yet socially acceptable control strategies. This requires natural and multimodal communication, which mandates common-sense knowledge and the representation of possibly divergent mental models. Such HRI is needed in situations, where human control is either infeasible or not cost-effective (Canal *et al.*, 2020; Rachid, 2022; Bonarini, 2020).

## 3. Challenges and potential impact

IRS that support effective military training in a VE has the potential to revolutionize military training. An IRS would enable the HMT to train together from the earliest phase of military training. The advantages that come from candidate selection, training intensity and mission rehearsal will be augmented through the use of the IRS, translating into a critical advantage over future adversaries. Different modes and medium are required to achieve this vision with confidence. Thus, the IRS can be seen as a valuable companion in many military activities. The multimodal autonomous communication capability of the IRS will enable the HMT to rapidly adapt to a dynamic combat environment and improve mission performance. The multimodal combination of audio, visual, graphics, imagery and digital modelling will help military forces maintain effective communications and face difficult situations. The additional advantage is the ability to conduct rapid mission rehearsals that will lead to improved outcomes and preservation of friendly forces.

Western military forces face a number of near term challenges arising from the imminent fielding of platforms enabled by AI capabilities. The autonomous mobile robot holds the prospect of assuming many of the difficult and dangerous tasks currently executed by humans. The integration of a protected platform with AI, which is then paired with selected human teams, will result in a potent and flexible capability.

Military commanders at all levels confront uncertainties during the execution of their missions. Potential adversaries are developing AI-infused systems and are now on the verge of fielding machines capable of conducting semi-independent offensive actions, selecting targets and taking lethal action without direct human input. The Canadian Armed Forces (CAF) has no experience in facing such capabilities and is now exposed to technical surprise on current operations. The CAF lacks an ability to explore the implications of these developments and consider near-term steps to mitigate the adversarial advantage.

Military commanders must comprehend and act across all domains of conflict with increasing amounts of data. Accordingly, the speed, complexity and breadth of operations are increasing at every level (strategic, operational and tactical). One of the major challenges is the ability to act quickly. This is directly related to the capability to process the flood of information. The data load of military operations requires that the commander's staff be able to process and analyze at a pace sufficient to provide situational awareness to a commander. However, perfection cannot be the objective: a commander's intuition remains vital to time-compressed decision making with uncertain battle field conditions. The most difficult challenge lies in integration of the aforementioned multimodal features to provide value to a military decision.

For example, someone who cannot speak the local language will often find it difficult to be deployed alone. This barrier, however, can be alleviated by speech-to-speech (s2s) translation systems. Translation systems can also be integrated into communication tools to allow people who speak different languages to freely communicate with each other remotely. Speech technology can also help in unified messaging systems: a speech transcription subsystem can be used to convert voice messages left by a caller into text. The transcribed text can then be easily sent to the recipient through emails, instant messaging or short message, and conveniently consumed by the recipient. ASR technology can be used to dictate short messages to reduce the effort needed for the users to send short messages to others. Speech recognition technology can also be used to recognize and index speeches and lectures so that users can easily find the information that is critical to them (Yu *et al.*, 2015).

A robot that is capable of navigating in unknown urban environments without the use of GPS data or prior map knowledge will be very useful in military activities. Such a robot would retrieve directional information by interacting with humans or the military. It would store acquired information into a topological route graph, which can used to give feedback to the human and to navigate in unknown environments as discussed in the Autonomous City Explorer (ACE) project (Bauer *et al.*, 2009).

Logistics operations, recording and playback fidelity, record and book keeping and other CAF operations would benefit from multimodal robots as described in Pandya *et al.* (2019) and Mikušová *et al.* (2017).

The Future Operating Environment (FOE) will be dominated by automated systems. These systems will support decision making informed by multimodal military communications. The capabilities of an IRS are a vital step towards ensuring that a future military entity remains a relevant force.

A multimodal interface for an IRS has the advantage of increased usability and accessibility for both the military and civilians. For example, visually impaired users can rely on the voice modality and hearing-impaired users can rely on the visual modality. Haptic input/output with visualization mechanisms would expand usability even further. The communication between a human and a robot is currently an area with many new developments, theories and technologies. Emerging technologies such as AR, VR and XR can support the communication between human and the machine, providing immersive experiences for visualization and analysis. Other human machine interaction (HMI) issues related to skills such as dexterity and flexibility may enhance total productivity and decrease the human ergonomic stress. Existing solutions on VR and AR technologies in HRI and HRC are reviewed in Dianatfar *et al.* (2021).

## 4. Multimodal interactions and communications

Multimodality enables choosing any input or output modality or combination of input and output modalities for optimizing interaction. Modality is a particular way of representing input or output information in some physical medium, such as something touchable, olfaction or gustation. Multimodality is inherent in the normal interaction of living organisms, often expressed in poses or other actions.

Multimodal features may determine the extent to which an interface supports its users in completing their tasks efficiently, effectively and satisfactorily. The weaknesses of one modality can be offset by the strengths of another in a multimodal interface. Accessibility can determine how easy it is for people to interact with the robot. Multimodal interfaces can provide increased accessibility, demonstrated by capabilities such as: interpreting postures, gestures and facial expressions; filtering ambiguous and incomprehensible spoken utterances, which may include simultaneous pointing gestures or other shared visual input (Jokinen and Raike, 2003).

Common sensory modalities involved in multimodal systems are shown in Table 1. A human multimodal interaction involves a coordination of different anatomy and physiology of sensory organs (Schomaker, 1995).

The following subsections summarize literature related to multimodality in these critical technological areas: HMI, autonomous unmanned systems, visualization, emergent technology, spatial prediction, natural language processing (NLP) and computer vision (CV).

### *4.1 Multimodal human machine interaction*

The media are the material objects available for presenting and saving information. The physical components can be computer input and output devices such as a screen, microphone, speaker or pointer. The modes are the human mechanisms of perception engaged in processing the incoming information that uses modalities such as vision, audition, olfactory and touch. Modes, media and multimodal communication can be visualized in Figure 1. Multimodality is a means to express one's purpose in an effective way. Thus, more than one mode such as written and spoken text, moving or still images, sound or music are coordinated to communicate in a sensible and meaningful manner. In addition, modes could include gestures such as movement, facial expressions, body language, body position, physical arrangement and proximity. Contextual aspects of multimodal communication is described in Lutkewitte (2014), Fillmore (n.d) (unknown).

Bolt (1980) focused on spatial tasks and map-based applications to integrate speech and gestures in the MIT media lab. This is one of the early endeavours to design multimodal interfaces.

Different control modalities, such as gestures, CV and teleoperation, are included in some multimodal systems. The increasing use of robotic assets in the military, industry and even household settings has led to a growing need for developing more intuitive and flexible human–robot interfaces. An interactive simulation to facilitate the objective evaluation of multimodal HRIs was presented in Whitney (2019).

HMI can be based on combined modes of sight, hearing, touch, learning, and perception of the world. The machine may be a desktop computer or mobile phone or any emergent virtual or augmented reality related device or component of Internet of Things (IoT). The interaction can be defined as a successful exchange of data or instruction between the user and the technology. An extensive survey of verbal and non-verbal aspects of HRI, its historical introduction and human–robot system desiderata is provided in Mavridis (2015). This successful interaction can be based on the balance of the needs of a user and the technical capabilities of the machine. A multimodal interaction may be supported by a

**Table 1.**
Sensory modalities that are central to human systems

| Sensory perception | Sense organ | Modality |
|---|---|---|
| Sense of sight | Eyes | Visual |
| Sense of hearing | Ears | Auditory |
| Sense of touch | Skin | Tactile |
| Sense of balance | Organ of equilibrium | Vestibular |

**Note(s):** Modified version of Schomaker (1995)

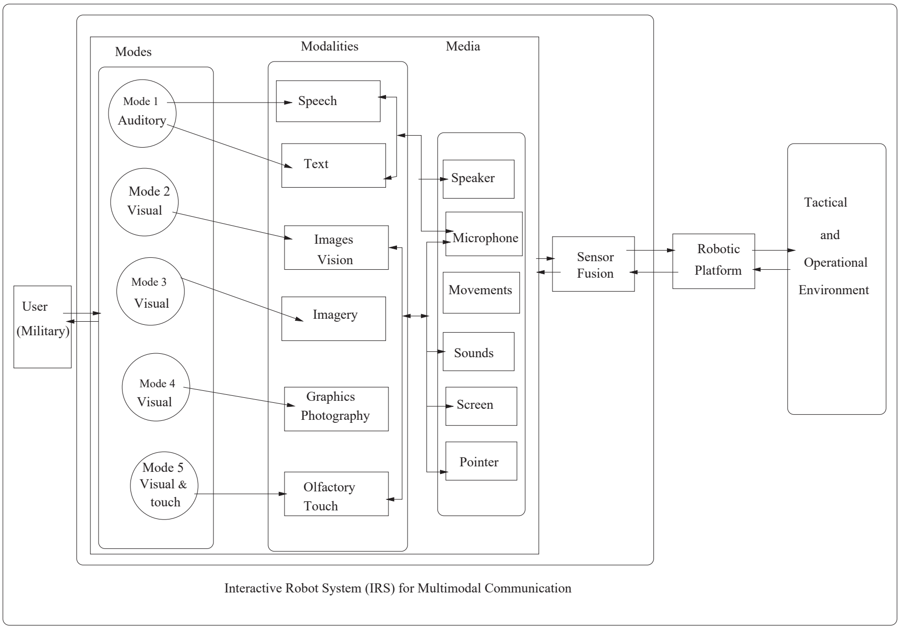

**Figure 1.**
Modes, modalities and media for multimodal communication

combination of images, audio, video, text or streaming audio-video in live or off-line form. A meaningful representation of such sources is essential to form a multimodal interactive system (Turk, 2014). Baltrusaitis *et al.* (2019) develops taxonomy of multimodal machine learning for building models, processing and relating information from multiple modalities. It includes a detailed analysis of early and late fusion, as well as a categorization of broader challenges in multimodal machine learning, namely: representation, translation, alignment, fusion, and co-learning.

Human machine communication (HMC) attempts to incorporate the nature of human perception and environmental interaction, as well as a basic reasoning capacity to draw inferences between acquired knowledge and its application (Brouwer and Harrington, 1994). This paper proposes a list of core challenges in communication and coordination, modelling approaches and infrastructure for optimized service. It recommends four main topics for the design of educational systems – a. Fundamental human perception and reasoning; b. New media: enabling technologies; c. Artificial Intelligence; d. Advanced applications.

Bauer *et al.* (2009) presented a HMC system in an ACE project. This system enabled the human to ask the robot for directions and stored the retrieved route information as internal knowledge. The system incorporated theories from linguistics in a mixed-modalities communication interface.

The idea of teaming between human and machine provides another perspective (O'Malley, 2007). Important features include the architecture of the team and task allocation. Supporting issues include automated systems perspectives such as theoretical analyses, laboratory experiments, modelling, simulation, field studies, ability, fatigue, neurological analysis, accident, mental workload, situation awareness, complacency and skill degradation.

A multimodal Scale Invariant Feature Transform (SIFT) algorithm was used to recognize hand-written characters in Guruprasad and Majumdar (2016). The shapes of the hands and characteristics such as formats, orientation, rotation, image formation, translation, scale and illumination were used to support multimodal recognition of handwriting.

The modes of interactions can be seen as an input, perception, output or control in the communication loop. Touch-based multimodal systems, interactions and applications using mainly a single mode are considered in Bezold and Minker (2011), Minker *et al.* (2006), Maragos *et al.* (2008) and Toselli *et al.* (2011). Communication based on automated capabilities was not considered in these studies.

A study on the human as a model operator and the behaviours in tactical environments such as in airborne early warning and control (AEW and C) was provided in Qureshi (1998), Bourdon and Kaluzny (2013), Caron and Kaluzny (2015) and Davy and Demczuk (2003). The main objective was to explore the cognitive aspects of human behaviour and possibilities of autonomous decision making capabilities applying control theory, task networks, human factors, and knowledge-based approaches. Yet all these studies did not intend to assist the fundamental multimodal HRI communication.

Huang *et al.* (2019) examined multimodal perception interfaces through a combination of human visual and haptic perception in guiding a coarse motor task. This was in the context of coordinating with the fine local motion performed by a robotic module running high-speed actuators and high-speed sensory feedback. The small sample population demonstrated reduced errors when human visual and haptic perceptions were combined with HRI. The results also suggested haptic perception as a major contributor to effective human response in fast motion situations. The focus was the integration of multimodal perception in human subjects for simple motor tasks. This type of multimodal integration may have a narrow application in specialized military robotic applications, such as ordnance disposal, but would not be as useful for dynamic tactical tasks.

Unhelkar *et al.* (2020) developed models for Bidirectional Communication Decision-Making in Sequential Human-Robot Collaborative (HRC) tasks. The objective of the models and algorithms was to leverage effective communication to attain high quality HMT. This research focused on sequential collaborative tasks that had well-known task dynamics and specified objectives. The results of the experiment showed that the CommPlan framework enabled robots to decide if, when, and what to communicate while performing sequential tasks in coordination with humans. The human robot team was able to communicate efficiently while collaboratively performing a well-defined task in a limited physical space (e.g. a meal preparation task). This verbal communication mode shows some promise to enhance routine interactions for repeated tasks; however, it would be of very limited value in the uncertain, complex environment of military operations.

Scimeca *et al.* (2020) proposed a future workshop to explore the use of AI, robotic simulation, haptic sensing and soft robotics technologies to improve the quality and efficiency of medical practitioner training as well as the creation of new tools for diagnosis and healthcare through the interaction of humans and robots in medical settings. It is likely that this workshop will push the frontiers of the use of soft robotics for training and diagnosis as well as improving the understanding of robotic simulation and visualization in the medical field. It remains to be seen if any insights concerning the robotic doctor/patient interaction from this workshop could be applicable for scenarios based on military capabilities, missions and activities.

Schoenherr *et al.* (2020) demonstrates an algorithm that produces adversarial examples against hybrid Automatic Speech Recognition (ASR) systems. In these examples, the systems remain robust in an over-the-air attack that is not adapted to the specific environment. The practical experiment revealed ASR systems are vulnerable in varying room setups. This included situations with no direct line-of-sight between speaker and microphone and utilized psycho-acoustic methods to hide changes of the original audio signal below the human audio threshold. An inconspicuous adversarial example can be used by an attacker for any target transcription to disrupt or misdirect an ASR function. A single modality robot communication capability would therefore be readily vulnerable in a contested environment.

Saad *et al.* (2020) employed an iterative Interaction-Design (ID) method in multimodal robot communication for a reception robot. A small sample population of industrial design students was engaged to use the ID tool to design person- and task-oriented communications. This method produced distinctive task- and person-oriented dialogue styles, to predict multimodal communicative behaviours. The resulting task-oriented style was a more formal, shorter and less chatty communication format. The study reinforces the value of extant verbal communication discipline well-established in demanding work conditions, such as Air Traffic Control and Fire Control nets.

Ermacora *et al.* (2015) explored the spectrum of autonomous operations by unmanned aerial vehicle (UAV) within a smart city scenario, where a cloud robotics service employed small UAVs for monitoring and surveillance in support of emergency management. The UAV operation was mediated by a cloud robotics platform according to the selected Level of Autonomy (LOA). Three levels of autonomy (tele-operation, mixed-initiative and full autonomy) for unmanned systems were employed to analyze the sliding autonomy approach in multiple scenarios. The rapid LOA selection required to adapt to dynamic combat conditions was totally dependent on a protected and pervasive cloud network and was therefore, unsuited to many military operations. It would support various routine base operations and allow a reduction in dedicated personnel for some functions.

Cockburn *et al.* (2013) presented an interactive simulation environment to facilitate the evaluation of multimodal HRI. Gesture controls and audio commands were used to direct robots in a simulation environment, challenging the robotic agents to interpret and respond correctly. The exploration of HRI control modalities of gestures, CV and teleoperation with the simulated environment is a valuable foundation for HMT experiments, and is adaptable to other interactive military training applications. This simulated research environment may support exploration of robot behaviour selection once it is validated in real world applications.

A prototype of a cognitive robotic assistant is designed to act proactively, adaptively and interactively with respect to humans with slight walking and cognitive difficulties in Fotinea *et al.* (2016). The key future capability is a multimodal action recognition system required to monitor, analyze and accurately predict user actions. This will require refined modelling of HRC in order to achieve an effective HRI. This research is not applicable to military capabilities such as injury recovery and rehabilitation due to the highly specialized nature of the assistive robotic application.

Lucignano *et al.* (2013) presented a Partially Observable Markov Decision Process (POMDP)-based dialogue system for bi-modal HRI to generate a robust interaction. An integrated architecture mounted on a mobile robot platform was tested in an HRI scenario to assess the overall performance with respect to baseline controllers. Voice and gesture based bi-modal interaction was more successful than gesture only, where the speech and gesture are naturally complementary. The simplicity of the case study was useful for exploring the probabilistic model. An extension of the concept is required to yield any relevance for military application.

Anjomshoae *et al.* (2019) conducted a Systematic Literature Review (SLR) of goal-driven eXplainable Artificial Intelligence (XAI) to support transparency and trustworthiness for AI applications. It focused on the explanation of robot behaviours to human users, and suggests that some focus on inter-robot explanation is required. It underscores several known shortfalls: AI-enabled military applications will require full transparency at all stages, so explainable and due diligence methods are built into all decision processes. The main findings include (1) a considerable portion of existing research only focuses on conceptual studies or lack evaluations or tackle relatively simple scenarios; (2) almost all deal with robots/agents explaining their behaviours to the human users and very few address inter-robot (inter-agent) explainability. Finally, (3) while providing explanations to non-expert users has been outlined as a necessity, only few research studies address the issues of personalization and context-awareness. The issue of full transparency, it seems, is an open topic with varying perspectives.

Aubert *et al.* (2018) evaluated motion-based and display-based communication modalities within a HMT that was required to complete a collaborative manufacturing task. This approach to evaluate communication of intent used various penalties for failing to safely coordinate movement within a well-defined space. A moderate sample of teams found that while multimodal communication of intent did not significantly improve upon unimodal approaches, the successful communication of intent can facilitate fluent coordination in concurrent team operations. However, there is no space for ambiguous communication of intent within a military operation.

Heard *et al.* (2019) presented a diagnostic workload assessment algorithm that estimates workload using results from peer-based and supervisory-based observation. This study used wearable sensors that capture physiological workload indicators that are sensitive to environmental effects. The algorithm correctly classified workload 90% of the time when trained on data from the same human robot teaming paradigm. It is feasible that military operations in permissive environments could employ such algorithms to adapt and allocate tasks, but in contested military environments, soldiers are expected to function at the maximum physical and cognitive workload. The multiple layers of real-time remote data feedback are also susceptible to adversarial intercept, geo-location and jamming.

Meng *et al.* (2020) proposed a fused bi-modal architecture of speech and gesture. A Convolutional Neural Network (CNN) and Baidu API are employed to recognize and forward match speech and gesture. This information is then processed by an algorithm to fuse the output into an intention category. The operator's intention is determined by judging the fill integrity of intents in the intention slot. The experiment showed that the bi-modal fusion architecture was an improvement on unimodal means of communication in recognition accuracy and efficiency. Such algorithms may have application in low critical circumstances; however, even the most mundane military tasks cannot be susceptible to algorithmic adjudication.

Shu *et al.* (2019) proposed an architecture for safe human robot collaboration on a coordinated and repetitive physical task. The HRI was initiated within a Virtual Reality (VR) simulation and followed by implementation of the tasks with real world robotic applications under VR supervision. This simulated world allowed for modification and rapid prototyping of multimodal communication options where tasks can be first executed in the VR simulation with different input and feedback channels to identify and validate the most efficient communication means for the HMT. This architecture appears to have ready application for the rehearsal of coordinated defined tasks with high degree of repetition which would lend itself well to a manufacturing workplace, but not a dynamic tactical military environment.

Efthimiou *et al.* (2019) presented the principles and technologies that form the basis of the i-Walk platform HRI environment. Multimodal communication patterns from live human interaction in a rehabilitation context enriched human robot communication. This was achieved by increasing the naturalness in interaction from the robot side with respect to its comprehension and reaction capabilities. This approach will require continued refinement of human robot communication modelling in order to achieve an effective HRI. This research is less applicable to tactical military capabilities due to the highly specialized nature of assisted robotic applications.

Kaindl *et al.* (2008) introduced a semi-autonomous mobile robot in a 2D space focusing on multimodal human and robotic communication. A visual dialogue based robot and human communication is discussed in Zhou *et al.* (2019). Multimodal interactive interfaces using speech and gestures are presented in Lunghi *et al.* (2019) and Fuhrman *et al.* (2019); however, these are not military context oriented. For incremental language generation, vision and audio are used in HRC in Yu *et al.* (2015) and Yu and Tapus (2020).

A multimodal interaction model that fused gesture, speech and pressure information was presented in Zeng and Feng (2020). This included simple interaction information set to identify users' intention. This multimodal intelligent interactive virtual experimental platform (MIIVEP) was tested with middle school teachers and students.

Vision and hearing modes are used in information transmission and communication in Hou *et al.* (2020). A neural network supported this single mode information identification.

A multimodal human–robot interface for a cooperative team of robots operating in a hazardous environment was demonstrated in Lunghi *et al.* (2019). This HRI facilitates the homogeneous control of a heterogeneous set of robots. The operator was capable of entering in the loop between the HRI and the CERNbot to customize the control commands during operations specific to the CERN facility. This can be a useful reference for design of a military robot multimodal communication in a contaminated or challenging environment.

In Robotic and Strategy (2017), the US Army Robotic and Autonomous Systems Strategy has stated the potential impact and necessity of human–computer interaction (HCI) in future intelligence analysis capabilities (FIAC). This study highlighted the operation of smart-room, collaborative working environments and various hardware display technologies (large group display, flexible display, wearable display, mixed/augmented reality, mobile and ubiquitous computing capabilities). It also examined the potential use of multimodal interactive technologies such as touch tables, speech and gesture recognition, bio-metrics, intelligent and adaptive user interfaces and advanced information visualizations.

Distributed mobile sensor networks, integrated communications and visualization technology are discussed as HMT capabilities in Mortimer and Elliott (2017). This research hypothesizes that tactical situational awareness (SA) can be improved if HRCs are prioritized for single or multi-sensory display according to importance and appropriateness. Some attention management issues were identified during task re-engagement and some guidelines relevant to tactile cues within multi-sensory bidirectional HRCs were offered. A novel neural network architecture called Global Workspace Network (GWN) is introduced in Bao *et al.* (2020). This may be capable of handling dynamic and unspecified uncertainties by applying multimodal data fusion. Behaviour based processes of an autonomous system model claiming trustworthiness is presented in Wang *et al.* (2020). Yet details of acquiring trustworthiness are not presented. Trust is closely related to communication which is often not addressed in the existing research.

Human perception of robots using a smart speaker embodiment, and how this affects the frequency of user interaction is discussed in Kontogiorgos *et al.* (2020). This study used a human-like robot embodiment teleoperation task using the Augmented Reality Interface for Teleoperation via the internet (ARITI) software framework. This architecture was tested in its ability to achieve safe human robot collaboration by using a simple repetitive task use-case which is jointly executed by the human and the robot. A nut was held by the human and a bolt was turned into the nut by the robot. The task was executed in the VR simulation with different input and feedback channels (multimodal) in order to identify the optimal communication between the human and the robot (Kontogiorgos *et al.*, 2020). A semi-immersive VR/AR platform was designed to assist human piloting of robotic platforms in Boudoin *et al.* (2008). Smart tracking and smart action, and a 'Follow-Me' technique were integrated with a large scale (ARITI framework) multimodal HRI environment. Inconsistencies in the use of hand and eye in the virtual scene has been studied in Li *et al.*

(2020), which applied gesture and voice based multi-channel fusion to efficiently complete a man-machine collaborative task via an interaction algorithm.

A multimodal interactive robotic framework based on a three-step approach: multimodal recognition, intention interpretation and prioritized task execution was presented in Iba *et al.* (2005). The multimodal recognition module translated hand gestures and spontaneous speech into a structured symbolic data stream without abstracting user intent. The interpretation module on the intents chose user input as primitives. This considered a task based user input, the current state of the system, and concurrent system state to prioritize robot sensing. The framework used a vacuum-cleaning robot to exemplify the interactive programming and the plan recognition.

The concept of HRI is applicable to various types of work such as search and rescue missions, industrial arms management, medical tasks, education or entertainment. For these purposes, the robot control system enables the human and robot to work together. The links for the system and operator must be identified through a control architecture. Another critical part is the system control algorithms which assess machine decisions for a single task. A sensor fusion based robotic system architecture for motion control is presented in Ruiz and Chandrasekaran (2020). A sensor fusion can integrate the information from sensors by using a data association approach (Xue *et al.*, 2020).

Robots in logistic services such as intelligent hoisting, assembling car bodies and welding parts of a car are investigated in Mikušová *et al.* (2017). The findings estimated that the robot and robotized workplace is four to six times more efficient than purely human service.

A recording and playback system was developed and tested in a surgical operation environment in Pandya *et al.* (2019). This was applied in several areas such as training of surgeons; collection of learning data for the development of advanced control algorithms and intelligent autonomous behaviours; and use as a “black box” for retrospective error analysis.

HMT is a core theme of the US Department of Defense vision of future warfare. A successful collaboration between humans and intelligent machines depends on trust. Trust is a complex and multilayered concept. Autonomy and AI technologies are consistently considered valuable in military systems and missions (Konaev and Chahal, 2021).

Explainability and trust, description and adaptation of technologies in a social and organizational context are essential in fostering trust in technology as discussed in Canal *et al.* (2020). Conventional and standard interfaces from skeleton tracking, to face recognition, to NLP for HRI scenarios have been considered for Robot Operating Systems (ROS). Construction of complex multimodal pipelines for HRI and social signal processing has been developed to bridge the HRI gap in the ROS ecosystem in Mohamed and Lemaignan (2021). The HRC requires robots with explicit internal models of a human and capacities to achieve tasks effectively with a human partner.

A cognitive robotic decision framework for situation assessment and context management, goals and plans management, refinement, execution and monitoring to share space and tasks with a human partner is developed in Alami *et al.* (2011).

### *4.2 Autonomous unmanned systems*

Unmanned Ground Vehicles (UGV) have advantages over people in a number of different applications, ranging from sentry duty, scouting hazardous areas, convoying goods and supplies over long distances and exploring caves and tunnels. Despite the growing technical integration (e.g. artificial intelligence (AI), machine learning (ML), CV related technology) to unmanned vehicles, the state of technology is still far from achieving some basic communication needs. As CV algorithms are implemented in hardware, the UGV is becoming partially autonomous, yet it is difficult to successfully issue higher level commands to a UGV, such as “patrol this corridor” or “move to this position while avoiding obstacles” (Anderson *et al.*, 2006).

Computer vision algorithms for analyzing Electro-Optical (EO) and Infrared (IR) Full Motion Video (FMV), 3D Light Detection and Ranging (LiDAR) and Wide Area Motion Imagery (WAMI) sensor data are used in Lavigne *et al.* (2019). Broadly speaking, this technology is related to tracking, insight seeking and surveillance in challenging environments.

A streamlined unmanned underwater vehicle (UUV) based deck with a submerged submarine in high sea littoral water was proposed in Watt *et al.* (2015). The solution used an automated active dock to correct for transverse relative motion between the vehicles. This concept was evaluated and tested by building and testing individual components to characterize their performance, errors, limitations and cost. A dynamic simulation model of automated docking of a UAV to a slowly moving submarine in littoral conditions was presented in Roy *et al.* (2017).

Some separate applications such as dynamic reactive behaviours and capabilities in Platforms for Ambulating Wheel (PAW) robots were studied in Broten *et al.* (2004, 2005, 2007), Monckton *et al.* (2005), and Shaker *et al.* (1992). Similarly, developments in reinforcement learning in autonomous climbing and obstacle avoidance robotic vehicles were explored in Vincent (2008). A Robot Operating System (ROS) autonomous mobile ground rover platform for cultivation and fertilization was reported in Post *et al.* (2017). The challenges of mobile agricultural robots are navigation and localization, object recognition and mapping as well as path planning algorithms. However, in this limited case the rover successfully navigated in the environment and performed its tasks. A preliminary human activity recognition for HRI, including human fall detection, ambient intelligence, video indexing, content-based video analytics, robotics and visual surveillance was conducted in Harriott and Adams (2010). All of these investigated particular capabilities separately. None of these investigated how the military communication would be improved by applying automation capabilities.

A robot is teamed with bystanders to accomplish victim triage, search, a hazard reading and a hazard sampling task in Harriott *et al.* (2011). This study used Human Performance Moderator Functions (HPMFs) to predict human performance in various roles for a peer-based human robot team. Factors such as fatigue, stress, injury, dehydration, weather and cognitive workload were considered for the HPMFs. Ingrand and Ghallab (2017) discussed the interaction between a person and a robot without any training and the influence of HPMFs in cognition and performance management. A similar concept was attempted with trained response personnel in Arrabito *et al.* (2010) and Bray-Miners *et al.* (2012). Four different types of gesture based 3D Lidar interfaces for a UGV was investigated in Kealey and Collier (2020).

Human factors relating to the operation of unmanned aircraft systems are discussed in Kaluzny (2012), Arrabito *et al.* (2010), Cahrbonneau and Legault (2017), Giang *et al.* (2010), Kim and Hmam (2009) and Bourdon and Kaluzny (2013). A literature review on human activity recognition (HAR) by feature extraction, object segmentation, bag of words and Hidden Markov Model (HMM) is reported in Ruitang *et al.* (2007). This survey of literature describes longer-term research on speech recognition techniques and the systematic progress towards features and requirements of intelligent robots. These include speech interaction, dialogue management and embedded system implementation of intelligent-robot-oriented speech interactive techniques.

An autonomous orbital robot with a panoramic camera to improve position accuracy using a distributed control system hierarchy was proposed in Chngchun *et al.* (2015). Tanaka (2016) provided a survey of human–machine systems and discussed the growing involvement of intelligent robots in space, ground, marine, social, industrial, legal service and domestic service. This work provides an outline of a HMS, state-of-the-art research and open questions with the aim of identifying what should/could be done to achieve further innovation in the HMS field.

An autonomy strategic challenge (ASC) explored the need for sensor and image analysis and related technology, gaps between sensing and control, missing data analytics and AI technologies in Monckton (2019). It was loosely organized into live and synthetic experiments geared towards true multiple unmanned air, land, and maritime vehicles coordinated by mission-oriented control. Decision support-based socio-technical systems have been developed for battle management systems in Allouche and Turgeon (2020).

To support unmanned aircraft systems crew operations, an authority pathway based weapon engagement system is designed and developed in Fang and Hou (2021). This assists personnel in following the laws of armed conflict and standard operating procedures. Such systems can be properly used only when there is clear communication involved. This fundamental component is yet missing.

A UGV and unmanned aerial vehicle (UAV) have potential impact in reconnaissance, surveillance and target acquisition platforms for the battlefield of tomorrow in military communication (Moreau, 2011). A testing approach in unmanned UAV systems for local-area surveillance augmented by automated data association is discussed in Anderson (2009). A statistical hypothesis-based method is used to achieve track-to-track automated data association. Another approach to mobile robot navigation based on Q-learning and W-learning is investigated in complex terrains in Vincent (2006). The study narrowly investigated robot navigation in terrains using the selected models.

An impressive variety of approaches is evident in the research supporting autonomy in robotic systems. As we discovered in this part of the survey, none of these achieve a primary goal of assisting the military create multimodal interactive robot communication.

### *4.3 Visualization and visual analytics*

Humans are responsive to visualizations, which can have a dramatic impact. Visual thinking and the visual representation of data, interpretation, and hypothesis building can be thought of as VST. Graphics, pictures and imagery all help in perception and cognition through powerful human visualization perception processes. Visual analytics (VA) attempts to leverage data visualization and support analytical reasoning. A typical approach to this is to collect, process and present data using visual forms. Then visual perception is analyzed and knowledge is discovered to visually reveal the story hidden in the data.

The fundamental VA steps such as data collection and processing are essential for the preliminary visual representation. This analysis strives to assign a useful data model such as through clustering or classification.

A large proportion of human brain power is dedicated to processing the signals from vision. Visualization of information helps in perception and reduces cognitive stress. Visualization helps in exploring large data sets and gives insight into the data. The human vision is the most powerful perception tool among the five human senses: Sight – 70%, hearing – 20%, smell –5%, touch – 4% and taste – 1% (Stangor and Walinga, 2010).

An immersive interactive VA interface was used in Cybulski *et al.* (2014). An interactive visual analytics metaphor was explored for designing and communicating using visual representations. Three nodes, two machine tools equipped with sensors and measurement systems, a collaborative robot with 6 degrees of freedom, a human machine interface for condition monitoring and visualization are some features and items in this framework. This framework was able to manage events and alarms in machine tools and driver robot manipulators using a Raspberry Pi 2 and cloud-based communication protocol system.

A deep survey of visual object tracking in a surveillance environment was conducted in Kaur *et al.* (2018). This study used divergent techniques to investigate human activity recognition and interaction, ambient intelligence, video indexing, content-based video analytics, robotics and visual surveillance conducted for the purpose of object tracking.

The visualization of a geospatial active vessel trace was discussed in Gouin *et al.* (2011). The VA was applied in support of maritime situation awareness, perception and comprehension (Stewart, 2012).

A sensemaking support system (S3) prototype was developed based on visual encoding, visual summary cards, a record browser, magnetic grid, multiple timelines, a graph analyzer and intelligence studies in Lavigne and Gouin (2011a, b) and Lavigne *et al.* (2011a). A sense-making toolbox with animated map and timeline, visual summary and magnet grid used as visual widgets for situational awareness was developed in Varga and Lavigne (2016). Social network analysis (SNA) in a non-maritime context such as an extended graph analyzer widget and a generic graph analyzer widget in counter-insurgency are applications of VA (Lavigne, 2014a).

A number of studies consider anomaly detectors, historical models of data streams, prediction of future trends, identifying anomalies in data streams, anomaly correlation, filter potential for false positives and network alert messages. For instance, VA was applied in visual anomaly detection, maritime domain situational analysis, maritime threats, coastal safety and Maritime Security Operation Centers (MSOC) including a JavaScript based interactive prototype in Hall *et al.* (2014a) and Venour and Roodnick (2011). Maritime Visual Analytics Prototype (MVAP) considered the trajectory, a specialized widget and factual information exploration (Lavigne, 2014b). The MSOC was later extended to social network analysis in a counter-insurgency context as MVAP in Lavigne (2014c) and Lavigne *et al.* (2011b). This was further extended to maritime domain applications. The focus was Visualizing Normal Maritime Behaviour (VNMB), Surveillance and Anomaly Detection (SAD),and Collaborative VA of a Vessel of Interest (CVAV) (Davenport *et al.*, 2013). A sense making support system for joint intelligence collection and capability was discussed in Lavigne *et al.* (2011, 2012, 2019), Riveiro (2011) and Lavigne (2015). The maritime domain awareness of adjacent and bordering seas, ocean and navigated waterways, including some maritime-related activities, infrastructure, people, cargo, vessels and other conveyances were part of this development. Dynamics of Counter Improvised Explosive Devices (E-IED) applying story telling exploratory visual analytics was reported in Lavigne *et al.* (2020). This enables a storytelling feature to provide insights to the user. The VA exploratory tool can provide a geospatial view of IED incidents, as well as incident type in text form and visual forms. These approaches do not address prediction which is an essential element for emergency events or military activities.

A vision-based framework for automatically recovering an AUV by another AUV in shallow water is proposed in Liu *et al.* (2019). This framework contains a detection phase for the robust detection of underwater landmarks mounted on the docking station in shallow water and a pose-estimation phase for estimating the pose between AUVs and underwater landmarks. This approach has been outperformed in terms of remote landmark detection. Liu's vision-based framework with acoustic sensors in field experiments demonstrated its effectiveness in the automated recovery of AUVs.

Trajectories, interactive Kohonen maps for trajectory aggregation, clustering trajectories for ship density regular traffic, vessel movement patterns using hybrid velocity signatures to detect an anomaly and spline trajectory based clustering for coastal surveillance are discussed in Anderson (2009) and Davenport *et al.* (2013). An improved prediction inspired algorithm and pattern based learning model for real-time maritime domain information tracking is presented in Rhodes *et al.* (2007). This approach demonstrated improved prediction accuracy. A rule based expert system and vessel motion is applied to maritime domain anomaly detection and situation awareness in Roy (2009). A literature and product review on VA for maritime awareness is presented in Davenport (2009). This review considered approximately 70 papers from global VA researchers, research groups, journals and conferences and fifteen VA products. It focuses on VA patterns, topics and needed tools for the Recognized Maritime Picture (RMP). A prototype of expert knowledge acquisition via social networking tools and technology for the Canadian Forces (CF) is investigated in Crebolder *et al.* (2014). Social networking is seen as a useful and supportive tool for the CF in these studies.

VA was applied in cyber-security to recognize risks, and protect against cyber-threats, enable key aspects of digital forensic processes and information discovery in Goodall and Sowul (2009). ManyNets for forensic activities, history trees and a Starlight visual information system analysis tool were used in Lavigne and Gouin (2015). A visual form of successful and missed terrorism attacks across the world since the 1970s was described in Hall *et al.* (2014b).

Situational awareness (SA) is the perception of the environmental elements and events with respect to time and space. In an effort to improve SA, the Army might deliver swarms of multiple small robots to an area of operations in advance of the maneuver of forces during operations (Lapinski, 2009). The authors applied VA to exploit interactive visualization and human cognition abilities to sustain SA in this contested operational environment.

Visual analytics have been explored extensively for use in different military applications; but visualization and VST for communication purposes paired with a multimodal communicative interface has not been achieved.

### *4.4 Emergent technology*

Multimodal communication involves multimodal perception, multimedia and the virtual environments of virtual reality (VR) and augmented reality (AR).

Emergence of new Information and Communication Technologies (ICT), such as VR and AR and Mixed Reality (XR), have created the current era of HRI and HRC.

AR is intended to provide the user an interactive experience within a real-world environment, where computer generated images are overlaid. This can be accomplished using multiple sensory modalities, including visual, auditory, haptic, somato-sensory and olfactory. The potential HMT application of AR and VR based communication, AR-based behaviour explanation, AR/VR for robot testing and diagnostics, VR for HRI human-subject experimentation, efficient representations for AR/VR, AR-enabled robot control as well as architecture for AR/VR have been explored in Williams *et al.* (2018) and Stedmon *et al.* (2013).

VR provides an immersive environment where objects are modelled by computer graphics. It creates a virtual environment for users to experience, observe and interact with virtual objects as a means to better perceive a real environment. In contrast, AR overlays a VE over the real world. AR features such as spatial mapping, audio and visual feedback, enable the simulation of realistic applications. In robotics, VR and AR technologies are used to study and prototype costly concepts and evaluate their validity with much lower expenses in a safe user environment.

An efficient, coordinated emergent solution using a multi-agent shepherding task was demonstrated in Nalepka *et al.* (2018). Shepherding refers to the guidance required for a multi-robot cooperative task. This study also deals with a dynamic and deformable environment (referring to terrain deformation or destruction). For this shepherding task evaluation, participants were engaged in a virtual world and the emergence of oscillatory-like behaviour in the virtual avatar inhibited efficient, coordinated behaviours.

Emergent technologies have been investigated in various contexts: A virtual environment for investigating human trust and human robot interaction in the context of a disaster is laid out in Ablavsky *et al.* (2002) and Crebolder *et al.* (2014). An obstacle avoidance strategy using emergent technology to discover an alternative path to a selected destination was studied in Yang *et al.* (2014). A stationary integrated video and active surveillance with GPS to GPS denied tracking was introduced in Maciejewsk *et al.* (2008). To enhance usability and interactions, emergent technology was deployed by the Royal Canadian Navy (RCN) (Karle *et al.*, 2018).

Virtual assistance in multimodal autonomous interactions to enhance military robot communication and provide support as needed has not yet been achieved.

### *4.5 Spatial prediction*

Spatial analysis, which focuses on location-based data, often processes geographic data (Esri, 2018). Geographic data can include geographic location, raster data (which consists of vector layers such as rectangular or triangular or circular or as a mixture of these) and natural environmental types based on elevation, temperature, precipitation and location. These might include roads, terrain, canals or census area. Spatial data exploration is a process of interaction with a collection of data and maps in order to visualize and explore geographic information. It is often an iterative interaction with information drawn from tables, charts, graphs and multimedia.

Logistic regression (LR), support vector machine (SVM) and random forest (FR) were used in predictions of traffic violations in Hana *et al.* (2008). An association rule, namely the Forward Prediction (FP)–Growth algorithm for describing the set of classes in a relevant association, is applied in geo-spatial semantics. This used Web Ontology Language (OWL) for human-computer interaction to feature the forest, semi-natural areas, wetlands and water-bodies for spatial prediction (Mc Cutchan and Giannopoulos, 2018).

A systematic review of different spatial methods such as spatial statistics, spatial econometrics, data-mining, ML, CV, remote sensing, geographic information science and spatial database by different research communities is provided in Jiang (2018). A spatial modelling, scene understanding, navigation and action-generation based autonomous robot is discussed in Stachniss (2009). A robot with a Hidden Markov Model (HMM) based high level interactions such as dyadic dialog between the presenter and a listener and vice versa is investigated in Mead *et al.* (2011). The objective of this study was to provide a social robot with the ability to recognize and produce appropriate social behaviour (Huang *et al.*, 2017). An interactive human–machine user interface basing predictions on eye movement gaze features is proposed in Hu *et al.* (2019). This applied SVM clustering analysis for spatial prediction, and the accuracy rate was 64%. A high correlation between haptic information and perception of the real and virtual environment was observed, while a visual-anchored prediction based on haptic information on the perceived presence of the VE was investigated in Gall and Latoschik (2018).

Deep learning networks (DNN) have been applied to estimate the indoor partial space for an air conditioner to blow air selectively into the main living area of residents. The DNN learns the human body detection saliency map to estimate the living or non-living area of the residents by accumulating sequential predictions (Cho and Lee, 2018).

For location prediction, a temporal spatial Bayesian model was applied to locate and predict locations for selected friends in social networks in Jia *et al.* (2016). Spatial autocorrelation and analysis for highly correlated large dimensional spatial data predictions has been successfully applied for many different system settings in Diaz (2007).

### *4.6 Natural language processing (NLP)*

ASR, automatic speech synthesis, machine language translation (MLT), machine language understanding (MLU), text-to-speech (TTS) and natural language generation (NLG) are sub-areas in NLP. NLP, ASR, and TTS were applied in short message service (SMS) applications: reading a weather map aloud, reading emails and news, and updating business news. NLP is applied to convert syntax to semantics to achieve effective communication (Wiriyathammabhum *et al.*, 2017). A spoken utterance is typically closely related to a context. Utterances are influenced by the expression, an event, the intention and related situation, e.g. how a particular speaker acts and speaks in a specific event, at a certain time in that particular position at that location. A simulated AI-based chatbot and military dialogue for a robotic system is introduced by the author. This is being extended and integrated with real robots for military communication.

VibLive is a text dependent voice user interface which detects the liveliness of voice. It has been used as a voice user interface within the Internet of Things (IOT) environment obtaining 97% accuracy. It verifies live users and detects spoofing attacks without requiring users to enroll specific passphrases (Zhang *et al.*, 2020).

NLP is also applied for interactive HMC and interaction in Brouwer and Harrington (1994). A multi-tasking ASR speech recognizer called Elektrobit Virtual Assistant was implemented applying delayed neural network (NN) and integrated into a Kaldi speech recognition development tool (Ranzenberger *et al.*, 2018). The objective was to assist with maps, search engines, web browser applications, phone applications, radio stations and speech dialogue. Speech recognition for controlling a robot was developed in Kumar *et al.* (2012).

Intelligent and interactive robots using dialog and speech for robotic research are not new. With speech as one of the most common communication means, research in speech technology has a very long history. A speech recognition system applying HMM, neural net, speech signal processing, noise reduction and filtering for noisy speech enhancement was applied in an educational setting and showed 85% recognition accuracy (Ruitang *et al.*, 2007). Industrial noisy speech recognition applying adaptive signal processing, pattern recognition and HMM is described in Richter *et al.* (2022). CNN and recurrent neural network (RNN), as an end-to-end deep NN, were used for ASR in Song and Cai (2015). In this work, the framework classification was done for the CNN, and connectionist temporal classification (CTC) was used for decoding. In the CNN and RNN based speech recognition, the TIMIT database is used for experiments and was implemented by using a python library called SAIL. Based on the results in that research, the recognition accuracy is good, but the complexities and difficulties in classification are high. GMM and HMM-based classifiers are predominantly used in ASR research. The latest progress of deep learning NLP with respect to modelling, learning and reasoning perspectives is discussed in Young *et al.* (2018) and Zhou *et al.* (2020). For rich-source tasks with enough training data, supervised-learning has been shown to be effective. According to these studies, for low-resource tasks with little training data, semi supervised and unsupervised learning, multitask learning, transfer learning and active learning can be used. These methods can either generate more pseudo training data for model training or leverage the knowledge learned from existing models and tasks. Logic and reasoning can extend NLP applications to advanced multitasking NLP stages combining multiple modes and explaining these from different communication perspectives. Reasoning can either be performed explicitly by designing specific inference models, or performed implicitly by end-to-end training. The state-of-the-art results on these tasks are achieved by end-to-end neural models. Poor performance in these systems may be due to existing knowledge bases suffering from low coverage of open-domain natural language texts for tasks requiring processing of both text and visual content. Typical (non-neural and neural) inference methods including integer linear programming (ILP) and Markov logic networks (MLNs) have been successfully used in various NLP tasks, such as Question and Answering (QA), dialog systems and information extraction.

Human–machine speech can be used in the classroom for education. This requires a multimodal enabled interactive intelligence which has a time-aware, stream-based programming model for parallel coordinated computation as well as a set of tools for data visualization, processing and learning. AI tools support the estimation of alignment probabilities between an observable and a hidden sequence in Bohus *et al.* (2017). An intelligent robot oriented speech interaction tool for speech synthesis to convert normal text consisting of words into speech was introduced in Ruitang *et al.* (2007). Investigation of multimodal features of speech, supporting spoken dialogue in assembly and emergency situations for HRI applying deep learning technology is ongoing in Paul *et al.* (2022a) and Paul *et al.* (2022b).

Statistical classification for task-oriented dialogue between soldiers and robots is introduced in Quach (2021) and Army (2020). This robot enabling conversational AI can communicate and perform actions based on underlying intents.

### *4.7 Computer vision (CV)*

In order to detect human motion and enable control, a CV-based robotic pet interaction system was developed in Mihara *et al.* (2000). The Motion Processor (MP) and Region of Interest (ROI) were used in lego mindstorm robotics to detect human actions and control a robotic pet. A survey of robotics applications clarifies the relationship between CV and NLP as well as related applications in Wiriyathammabhum *et al.* (2017). A survey of robots and workplace, industries, logistics and healthcare was conducted revealing that North America lags other parts of the world in having robots in the workplace (IFR, 2018).

How statistical methods, deep learning and large scale neural networks have been applied in machine intelligence, speech recognition, CV, language understanding and translation, robotics and HRI are briefly covered in Dean (2016).

The perception of robotic interaction and human social impact was examined in Fraune and Sabanovic (2014). The study concluded that there is a requirement for the development of more robust HMC. The study included participants from different age groups and used robots for several different types of communications.

An AI based autonomous chess playing robot on a Robot Operating System (ROS), hand detection and chess movement tracking was developed in Rath *et al.* (2019). This autonomous chess player accomplished 50 independent successful moves in an indoor environment. CV was also employed in the real-time autonomous video enhancement (RAVE) system developed in Carloni *et al.* (2013).

CV-related capabilities for the recognition of automated facial mimicry with laughing facial expressions were developed in Paetzel *et al.* (2017). A vision sensor for detecting the distance between surrounding buildings and a vehicle was developed in Kang *et al.* (2014).

Real-time tracking for three dimensional computer graphics (3D-CG) vision was introduced to locate the operator's object in Mitsuishi *et al.* (1997).

HRI in the context of robotic musicianship was explored in Cicconet *et al.* (2013). This work investigated computer vision, mainly with visual cues, to anticipate a robotic response to human gestures. The use case measured the time difference between a human player and a robot with the robot performing better.

A low-complexity lidar gesture recognition system for mobile robot control in a real world teleoperation setting has been demonstrated in Chamorro *et al.* (2021). A gesture-based HRI applying a NN classifier for gesture recognition was implemented in Trigueiros *et al.* (2015). A gesture recognition system applying the HMM was developed in Acharya and Pant (2015). The processing tasks included background subtraction, segmentation, feature extraction and classification. The result was compared with multilayer neural networks (MLP). The HMM-based recognition system demonstrated 90% accuracy and the MLP based recognition system obtained 60% accuracy in Tanguay (1995) and Marcel *et al.* (2000). Yet none of these gesture recognition investigations are adequate to meet the multimodal communication need. Substantial modifications and additional poses can be some options to extend this to gesture-based communication.

### *4.8 Overview*

This section presents an overview of the research of multimodal interaction using various modalities and multimodal fusion of speech and gesture for various applications and HRI-based on sensing and perception. The application domain of existing research in general is very restricted and mainly limited to indoor environments. Significant progress has been made in some areas, yet these technologies are not yet available to be used in complex emergency situations, outdoor complex military missions or natural catastrophes such as wild fires, floods and hurricanes.

A summary of technologies, modalities, technical aspects and application domains is shown in Table 2.

## 5. Analysis, gaps and a design concept

A mobile robot requires a mechanism of locomotion that enables it to move unbounded throughout its environment. Even though significant progress has been made, navigation remains one of the most challenging tasks for the mobile robot. Navigation is closely related to perception, which requires sensors, extracting useful and meaningful information, determining localization with respect to a robot's position in the environment, cognition for path planning to decide how to achieve its goal and finally, motion control that allows the robot to manipulate its motor for desired trajectories. Proper coordination of different sensing and fusion are critical to managing navigation better.

Visual analytics have been applied in airborne surveillance and the maritime domain. However, visual analytics and VST are yet to be extensively applied in decision making to capture dynamic changes in military situation management as well as to predict and assist in managing emergency situations. Multimodal autonomous interactive communication can be a practical companion in emergency situation management.

Multimodal fusion for multimodal learning based multimodal autonomous military robots in communication for complex situations are yet to be deployed. Up to this date, based on the available knowledge and publicly available documentation, a vigorous active research and development in UAS and HRI systems to support better decision making, are yet to be implemented. Autonomous multimodal communication and general autonomous communication do not currently exist in the civilian and military domain. A dialogue-based autonomous mobile robot has yet to go beyond simple message transactions.

### *5.1 Objectives and proposed approach*

The primary objective of the IRS is to establish a test-bed capability to enhance the tactical execution of missions in complex environments while protecting the employed force. Furthermore multimodal autonomous communications deployed onto tactical platforms should improve the speed and quality of decision making in the FOE. This is a demanding and complex problem space that will require a significant ongoing investment by the military, emergency management team, first responder and soldier. Knowledge base locomotion, multimodal fusion, multimodal machine learning, sensing as well as perception, virtual assistance and multimodal HRI are central components of such a system. The major components of the proposed system are:

**Table 2.**
Overview of major topics

| Technology | Modalities and technical aspect | Application domains |
|---|---|---|
| Multimodal interaction | Speech, gesture, vision | Mainly indoors |
| Sensing and perception | Vision, visual and textural | Mainly indoors |
| HRI | Speech, text, gestures, vision and images | Mainly indoors |
| SA, ET | Visualization, CV, AR, VR | Mainly Indoors |
| UAS, HRI | VA, SSDS, gesture, speech | Mainly Indoors |

(1) Multimodal interactive robot that communicates using multiple modes, senses and perceives and is mobile

(2) Virtual environment for personal assistance

(3) Multimodal user interface

*5.1.1 Multimodal interactive robot.* Figure 2 shows the proposed IRS framework: A user (e.g. military team leader) communicates with the robot using spoken commands, gestures, images and interacting with the VST and virtual personal assistance as needed. The communication modes are consistently coordinated by the sensor fusion. The goal is to make the robot communicative with the military using the following modalities:

(1) Verbal, non-verbal, text, graphics, images and sense based modalities:

- ***Spoken Dialogue System:*** This assists two-way communication between the military user and the robotic system using verbal and non-verbal commands. Figure 3 is an architectural view of a spoken dialogue system. Here user speech is captured, processed and transformed into a set of features correlated with uttered speech via a language model in an ASR system. NLG and natural language understanding (NLU) provide semantics and logical representations of text based on the dialogue manager. The Text-to-Speech Synthesis (TTS) system transforms the textual output of the NLG into acoustic signals. The Dialogue Manager (DM) is responsible for bidirectional connections of the acoustic front-end namely ASR, TTS and semantics and logical NLG and NLU components. Domain-based applications, such as HRC, are linked and connected via the DM (Paul *et al.*, 2022a).

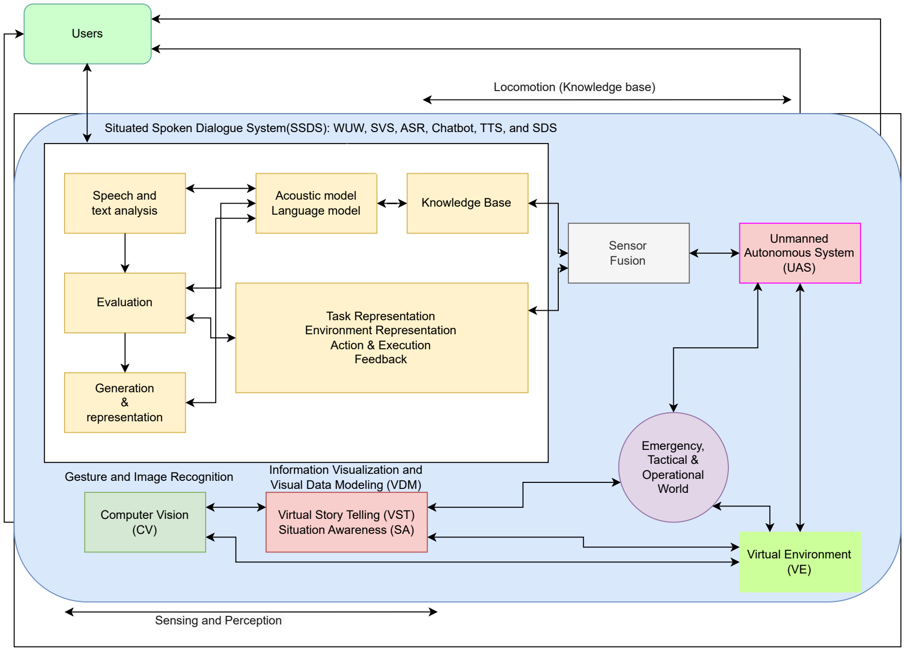
Users
Locomotion (Knowledge base)
Situated Spoken Dialogue System(SSDS): WUW, SVS, ASR, Chatbot, TTS, and SDS
Speech and text analysis
Acoustic model Language model
Knowledge Base
Evaluation
Task Representation Environment Representation Action & Execution Feedback
Generation & representation
Sensor Fusion
Unmanned Autonomous System (UAS)
Emergency, Tactical & Operational World
Gesture and Image Recognition
Information Visualization and Visual Data Modeling (VDM)
Computer Vision (CV)
Virtual Story Telling (VST) Situation Awareness (SA)
Virtual Environment (VE)
Sensing and Perception

**Figure 2.** An overview of multimodal interactive robot in military communication

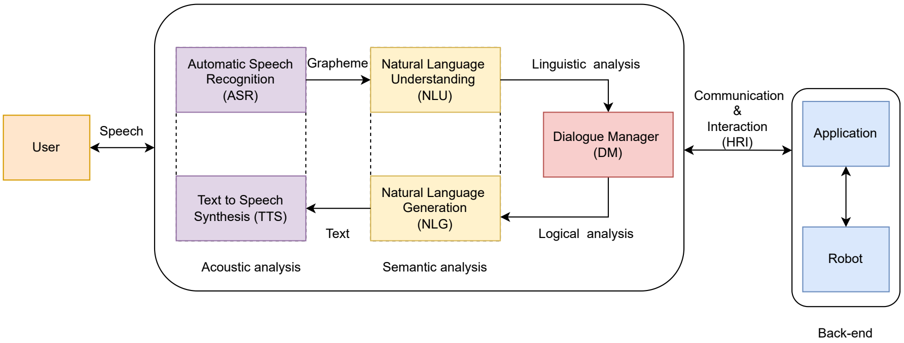
Automatic Speech Recognition (ASR)
Grapheme
Natural Language Understanding (NLU)
Linguistic analysis
Dialogue Manager (DM)
Communication & Interaction (HRI)
Application
User
Speech
Text to Speech Synthesis (TTS)
Text
Natural Language Generation (NLG)
Logical analysis
Robot
Acoustic analysis
Semantic analysis
Back-end

**Figure 3.**
Abstract architecture of a spoken dialogue system

- ***Gestural:*** The military user communicates with the IRS using a sequence of natural gestures which the robot can act upon via this component. Gesture-based interactions will be beneficial for the operation of robots conducting tasks in non-standard, uncontrolled situations and loud environments. This will have its largest impact in situations where gestures combined with speech emphasize the meaning of the message, particularly when speech is not viable. Hand movements and pointing to something are typical gestures in human interactions. Such interactions between the military user and the robot will establish a natural and intuitive experience in practice.

Established military hand gestures include: “STOP”, “ENEMY”, “FREEZE”, “HALT”, “MOVE LEFT”, “MOVE RIGHT”, “ASSEMBLE”, “RALLY”, “JOIN ME”, “INCREASE SPEED”, “LINE FORMATION”, “VEE FORMATION”, “ECHELON LEFT or RIGHT”, “CONTACT LEFT or RIGHT”, “ENEMY in SIGHT”, “QUICK TIME”, “MAP CHECK”, “DANGER AREA”, “STOP, LOOK, LISTEN, SMELL (SLLS)”, “RIFLE”. Gestures can be static (i.e. motionless) or dynamic (i.e. including movement). The IRS must recognize these established gesture-based commands and transform them into intended actions.

- ***Object Recognition and Detection:*** This component describes the environment, allowing focused actions. The object recognition problem can be defined as a labeling problem based on models of known objects. Object recognition enables computers to recognize different categories of objects in images e.g. appearance, shape, location, any parts or any particular characteristics. A major goal is to minimize risk by recognizing danger. This component helps the military user geo-locate weapons, dangerous emissions and contaminated zones.

- ***Visual Story Telling for Visualization and Prediction:*** This component allows visualization of scenarios, supporting decision-making through predictions and descriptions of the environment. This has two main components:

  (1) Visualization

  (2) Visual data mining

The VST fills a role as visual media in multimodal communication, assisting with visualization, cognition and perception. Some basic VST building blocks are shown in Figure 4. These include information processing, predictions and descriptions. The VST captures abstract data, encodes data for visualization, transforms them for cognition and perception to discover knowledge and create hypotheses as shown in Figure 5.

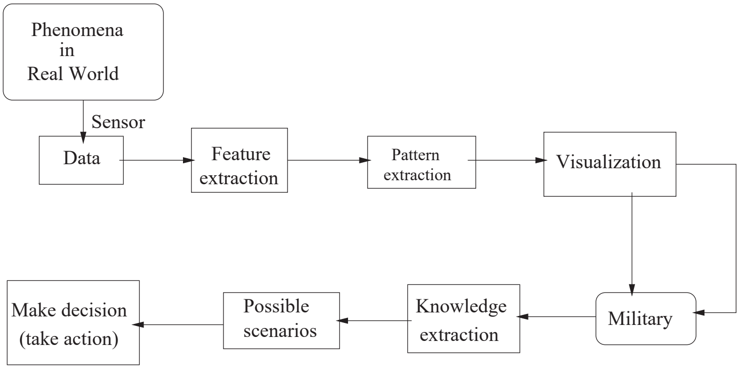

**Figure 4.**
Visualization in the VST building block

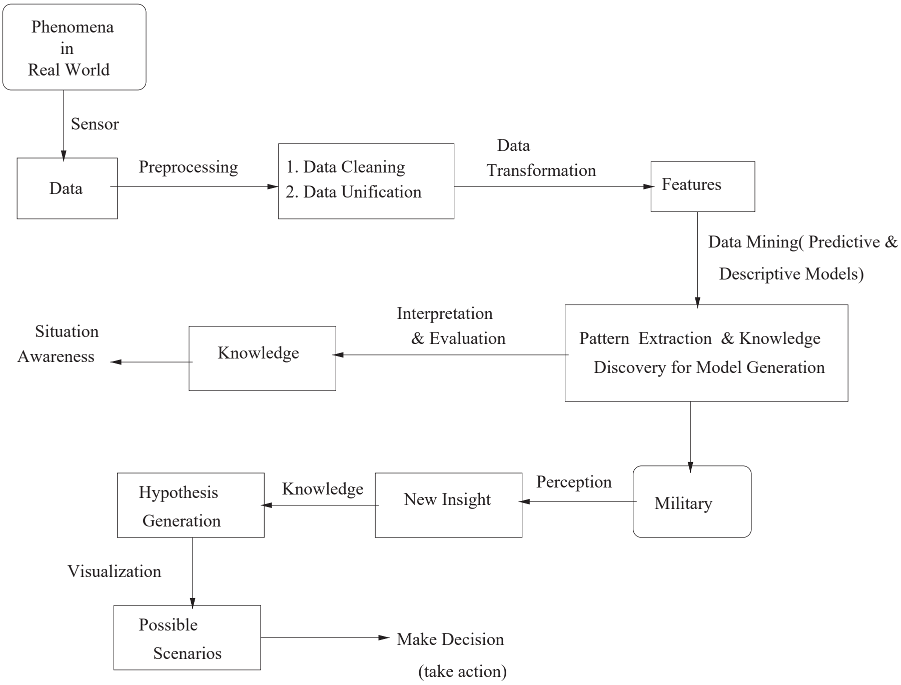

**Figure 5.**
Visual story telling: visualization and prediction

- ***Virtual Environment based Personal Assistance:*** The capability of the VE to provide personal assistance is central to its value to the military user. It supports high-level skills training and retention. The VE often aims to provide a synthetic experience for its users in this training mode. The experience may be illusory or virtual: the sensory stimulation to the user is simulated and generated by the system.

(2) ***Sensor Fusion:*** Sensor fusion is employed to support and augment decision making capability and achieve effective human-human and HMI and

communication. Sensor fusion has two main functions: 1. Coordination of sensor data and 2. Capture of human motion to coordinate this to robotic motion and action.

The component must activate the correct sensor(s) in order to achieve effective collaboration. The preliminary components of this are Control unit, multi-sensor data fusion and motion control. It selects the mode and purpose of the sensing signal to be communicated via a sensor in order to direct the information properly. It decides when to initiate and when to end the action of the particular sensor. It also fuses information to support learning. It leverages human motion for human robot collaboration and to support learning.

Interactions between human and robot can be initiated by the robot or by the human user. When approached by a human user, the robot must understand the users' intent followed by a decision on how to proceed, for example, deciding whether an interaction with a specific person should be launched.

In sensor fusion, available information is organized so that it is in an accessible form and to support decision making.

The decision making capability is captured from information by analyzing the features and classifying the behavioral sense. This capability to capture human behavioral intentions allows the robot to perform naturally and interact intelligently. Development and implementation of these features are not simple, requiring substantial analysis and investigation. Figure 6 depicts how the modified version of (Wu *et al.*, 2022) is accomplished in multiple iterative steps.

(1) Associate and coordinate multi-sensor data for fusion

(2) Integrate and update robot status based on multi-sensor fusion

(3) Human and robot communicate via motion capture and learning

Based on the real world situation and scenario, these features will be modified, updated and extended as necessary.

*5.1.2 Multimodal user interface.* Figure 7 shows the multimodal user interface for use in a tactical military environment. The multimodal interface is designated to increase usability and accessibility. The benefits of a multimodal interactive interface could be most valuable to physically impaired users. For example, visually impaired users can rely on the voice modality while hearing-impaired users can rely on the visual modality. These users will benefit from both speech and some haptic input/output along with visualization mechanisms.

### *5.2 Discussion*

There are diverse research streams and development efforts addressing C4ISR (Command, Control, Communication, Coordination, Intelligence, Surveillance and Target acquisition). These involve the coordination of ISR assets, synthesis of intelligence and implementation of machine aids supporting decision making for commanders. These streams have primarily

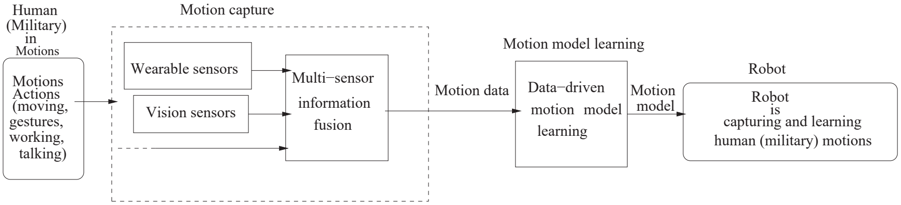

**Figure 6.**
Fusion of human motion capture and learning

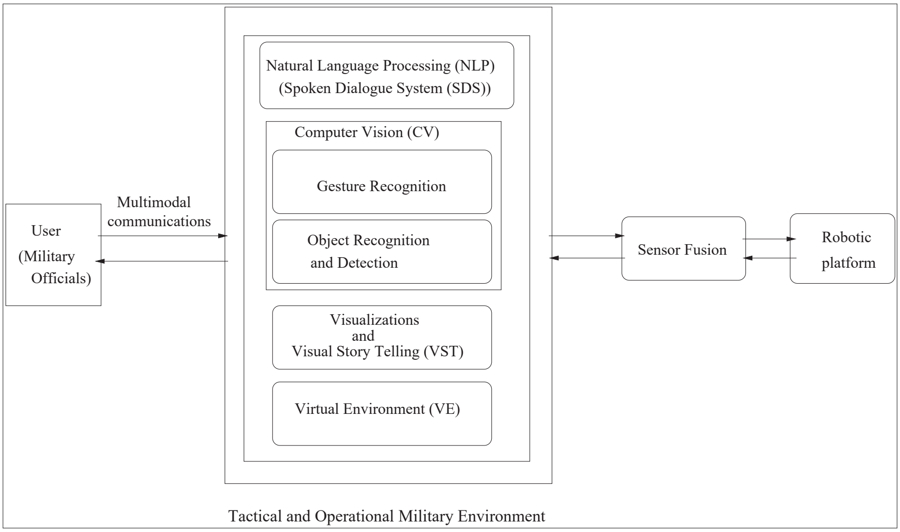

**Figure 7.**
Multimodal interactive robotic interface (MIRI) for military communications

been focused activities, such as improved surveillance and terrain mapping, improved collation of intelligence materials and computer analysis and modelling programs available to higher commanders as aids to decision making. A unified "Interactive Multimodal Robot in Military Communication" provides a synthesis of these types of activities, applicable at a much more tactical level. Rather than being used at the Battle Group or Brigade Group levels (e.g. for mission planning), the paradigm discussed here is most applicable to Section, Platoon or Company level operations. At this level the focus will be for planning and execution of operations.

Based on a review of the available applications of robotic systems with visual analytics in airborne surveillance and the maritime domain, these have yet to be applied in decision making to capture dynamic changes in military situation management and first responder assistance. Multimodal interactive communication has the potential to yield a significant improvement to the central features of effective communication in the planning and execution of military missions and manned-unmanned teaming. Based on the survey presented here, we believe there is an opportunity to combine multiple modes for HRI, apply artificial intelligence and leverage natural language processing for the design of military and civilian communication. We assess that an interactive robot for autonomous multimodal communication will be beneficial in a variety of civilian and military roles.

## 6. Conclusion

A multimodal autonomous robot in military communication using speech, images, gestures, VST and VE has yet to be deployed. Autonomous multimodal communication is expected to open wider possibilities for all armed forces. All branches of military organizations can potentially benefit from these developments, with platform-specific suites for specially selected operators. The flexible communication capability supports virtual training, which will enhance planning and mission rehearsals tremendously.

Each mode of multimodal communication is an active research area. Blending these developments into a functional and useful system is a substantial and daunting task. The introduction of the interactive robot concept for autonomous multimodal communications will require a series of prototyping and development iterations.

## References

Ablavsky, V., Snorrason, M. and Taylor, C.J. (2002), "Real-time autonomous video enhancement system (RAVE)", *II – II, Proceedings International Conference on Image Processing, IEEE* **2**.

Acharya, M. and Pant, D.R. (2015), "Computer vision based hand gesture recognition for speech disabled persons", *Journal of the Institute of Engineering, IOE*, Vol. 11 No. 1, pp. 30-35.

Alami, R., Warnier, M., Guitton, J., Lemaignan, S. and Sisbot, E.A. (2011), "When the robot considers the human", *International Symposium on Robotics Research (ISRR)*.

Allouche, M.K. and Turgeon, S. (2020), *Towards Self-Explaining Naval Battle Management Systems, Technical Report DRDC-RDDC-2020-R007, DRDC - Valcartier Research Centre*, DRDC - Valcartier Research Centre, Valcartier, Quebec.

Anderson, M. (2009), *Exploring the Contribution that Small Uav Can Make towards Maritime Force Protection, Ttcp Group Mar Action Group Ag-10 (Maritime Security, Area and Port Force Protection*, Defence Technology Agency, Auckland (New Zealand), Technical report.

Anderson, J., Lee, D.-J., Schoenberger, R., Wei, Z. and Archibald, J. (2006), "Semi-autonomous unmanned ground vehicle control system", *Unmanned Systems Technology VIII*, Vol. 6230, pp. 551-561, SPIE.

Anjomshoae, S., Najjar, A., Calvaresi, D. and Främling, K. (2019), "Explainable agents and robots: results from a systematic literature review", *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, Montreal, Canada, May 13-17, 2019, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1078-1088.

Army, U. (2020), "Army research enables conversational ai between soldiers, robots", available at: https://www.army.mil/article/237580/army_research_enables_conversational_ai_between_soldiers_robots.

Arrabito, G.R., Ho, G., Lambert, A., Rutley, M., Keillor, J., Chiu, A., Au, H. and Hou, M. (2010), *Human Factors Issues for Controlling Uninhabited Aerial Vehicles: Preliminary Findings in Support of the Canadian Forces Joint Unmanned Aerial Vehicle Surveillance Target Acquisition System Project*, DRDC - Toronto Research Centre.

Aubert, M.C., Bader, H. and Hauser, K. (2018), "Designing multimodal intent communication strategies for conflict avoidance in industrial human-robot teams", *IEEE International Symposium WeCT2.6 on Robot and Human Interactive Communication*, Vol. 27.

Baltrusaitis, T., Ahuja, C. and Morency, L.P. (2019), "Multimodal machine learning: a survey and taxonomy", *IEEE Computer Society*, Vol. 41, pp. 423-443.

Bao, C., Fountas, Z., Olugbade, T. and Bianchi-Berthouze, N. (2020), "Multimodal data fusion based on the global workspace theory", *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 414-422.

Bauer, A., Wollherr, D. and Buss, M. (2009), "Information retrieval system for human-robot communication - asking for directions", *IEEE International Conference on Robotics and Automation*, ICRA, Kobe.

Bensalem, S., Ingrand, F. and Sifakis, J. (2008), "Autonomous robot software design challenge", *Proceedings of Sixth IARP-IEEE/RAS-EURON Joint Workshop on Technical Challenge for Dependable Robots in Human Environments*.

Bezold, M. and Minker, W. (2011), *Adaptive Multimodal Interactive Systems*, Springer, New York.

Birdwhistell, R.L. (1970), *Kinesics and Context Essays on Body Motion Communication*, Penn Press, Pennsylvania.

Bohus, D., Andrist, S. and Jalobeanu, M. (2017), "Rapid development of multimodal interactive systems: a demonstration of platform for situated intelligence", *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Bolt, R.A. (1980), "'put-that-there' voice and gesture at the graphics interface", *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 262-270.

Bonarini, A. (2020), "Communication in human-robot interaction", *Current Robotics Reports*, Vol. 1, pp. 279-285.

Boudoin, P., Domingues, C., Otmane, S., Ouramdane, N. and Mallem, M. (2008), "Towards multimodal human-robot interaction in large scale virtual environment", *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 359-366.

Bourdon, S. and Kaluzny, B. (2013), *Development of a Capability-based Bidder Evaluation Tool for the Fixed-wing Search and Rescue Replacement Project*.

Bray-Miners, J., Ste-Croix, C. and Morton, A. (2012), *The Human-Robot Interaction Laboratory Pilot Study Experimental Plan*, Human Systems, Guelph, Ontario.

Broten, G., Monckton, S., Giesbrecht, J., Verret, S., Collier, J. and Digney, B. (2004), *Towards Distributed Intelligence - a High Level Definition, Technical Report*, DRDC-Suffield Research Centre, Suffield.

Broten, G., Giesbrecht, J. and Monckton, S. (2005), *World Representation Using Terrain Maps: Enabling High-Speed Navigation, Technical Report DRDC-SUFFIELD-TM-2005-248*, DRDC-Suffield Research Centre, Suffield.

Broten, G., Monckton, S.P., Giesbrecht, J. and Collier, J. (2007), *Towards Framework-Based UxV Software Systems: an Applied Research Prospective, Technical Report DRDC-SUFFIELD-TR-2004-287*, DRDC-Suffield Research Centre, Suffield.

Brouwer, M.D. and Harrington, T. (1994), *Human-machine Communication for Educational Systems Design*, NATO Advanced Study Institute (ASI), Eindhoven, Netherlands, Technical report.

Cahrbonneau, R.J. and Legault, S.R. (2017), *Test Plan for Fall 2017 Radar Cross Section Calibration Experiments with an Unmanned Aircraft System, Technical Report*, DRDC-Ottawa Research Centre, Ottawa.

Canal, G., Borgo, R., Coles, A., Drakeb, A., Huynha, D., Keller, P., Krivic, S., Luff, P., ain Mahesara, Q., Moreaua, L., Parsons, S., Patel, M. and Sklar, E. (2020), "Building trust in human-machine partnerships", *Computer Law & Security Review*, Vol. 39 Nos 5-7, 105489, doi: 10.1016/j.clsr.2020.105489.

Carloni, R., Lippiello, V.V., Massimo, D., Fumagalli, F., Mersha, A.Y., Stramigioli, S., Bruno, S. and Siciliano, B. (2013), *Realtime Autonomous Video Enhancement System (RAVE)*, Vol. 2, IEEE Robotics and Automation Magazine, Rochester, NY, pp. 22-31.

Caron, J.D. and Kaluzny, B.L. (2015), *Simulation Tool for Optimizing Non-combatant Evacuation (STONE): Optimization of Evacuation Time and Transportation Resource Utilization, Technical Report*, Defence Research and Development Canada, Centre for Operational Research and Analysis, DRDC-Suffield Research Centre, Suffield.

Chamorro, S., Collier, J. and Grondin, F. (2021), "Neural network based lidar gesture recognition for realtime robot teleoperation", *IEEE International Conference on Safety, Security, and Rescue Robotics*.

Chngchun, L., Xiaodong, Z., Dong, P. and Liu, X. (2015), "Research of visual servo control system for space intelligent robot", *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Cho, J. and Lee, M. (2018), "Estimation of user-indoor spatial information using deep neural networks selective ventilation for living area estimated by deep neural network", *2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, IEEE, pp. 1-4.

Cicconet, M., Bretan, M. and Weinberg, G. (2013), "Human-robot percussion ensemble: anticipation on the basis of visual cues", *IEEE Robotics & Automation Magazine*, Vol. 20 No. 4, pp. 105-110.

Cockburn, J., Solomon, Y., Kapadia, M. and Badler, N. (2013), *Multi-modal Human Robot Interaction in a Simulation Environment*, Dept. of CIS, University of Pennsylvania, Philadelphia, PA, Pennsylvania, Technical report.

Crebolder, J., Randall, T., Hunter, A. and Coates, C. (2014), *Investigating Virtual Social Networking in the Context of Military Interoperability, Technical Report*, DRDC-atlantic Research Centre, Atlanta, Toronto.

Cybulski, J., Keller, S. and Saundage, D. (2014), "Metaphors in interactive visual analytics", *Proceedings of the 7th International Symposium on Visual Information Communication and Interaction*, Vol. 14, pp. 212-215.

Davenport, M. (2009), *Literature and Product Review of Visual Analytics for Maritime Awareness*, DRDC-Valcartier Research Centre, Valcartier, Quebec, Technical Report ADA554351.

Davenport, M., Lavigne, V. and Gouin, D. (2013), *Design of a Maritime Domain Awareness Visual Analytics Prototype (Dmvap), Technical Report DRDC-VALCARTIER-CR-2011-221, DRDC-Valcartier Research Centre*, Salience Analytics, Vancouver, BC, (CAN).

Davy, J. and Demczuk, V. (2003), *Human Factors Aspects of Hard Target Weapons Systems*, Defence Science and Technology Organization, Australia, Technical Report DSTO-TN-0499, Australian Government Deptartment of Defence.

Dean, J. (2016), "Building machine learning systems that understand", *International Conference on Management of Data*, SIGMOD.

Dianatfar, M., Latokartano, J. and Lanz, M. (2021), "Review on existing vr/ar solutions in human–robot collaboration", *Procedia CIRP*, Vol. 97, pp. 407-411.

Diaz, F. (2007), "Performance prediction using spatial autocorrelation", *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Vol. 30, pp. 583-590.

Efthimiou, E., Fotinea, S.-E., Vacalopoulou, A., Papageorgiou, X.S., Karavasili, A. and Goulas, T. (2019), "User centered design in practice, adapting HRI to real user needs", *ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, NY, USA, Vol. 12, pp. 425-429.

Ermacora, G., Rosa, S. and Bona, B. (2015), "Sliding autonomy in cloud robotics services for smart city applications", *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pp. 155-156.

Esri (2018), "How to perform spatial analysis", available at: https://www.esri.com/arcgis-blog/products/product/analytics/how-to-perform-spatial-analysis/.

Fang, S. and Hou, M. (2021), *Authority Pathway Operating Manual for Unmanned Aircraft System Weapon Engagement by Allied Impact Command and Control Station*, DRDC- Toronto Research Centre, Toronto, DRDC-RDDC-2021-D061".

Fotinea, S.-E., Efthimiou, E., Goulas, T., Dimou, A.-L., Tzafestas, C. and Pitsikalis, V. (2016), "The mobot human-robot interaction: showcasing assistive hri", *Proceedings of the 20th Pan-Hellenic Conference on Informatics*, pp. 1-6.

Fraune, M.R. and Sabanovic, S. (2014), "Robot gossip: effects of mode of robot communication on human perceptions of robots", *ACM/IEEE International Conference on Human-Robot Interaction*, Bielefeld, Germany.

Fillmore, A. (n.d.), "Multi-modal communication: writing in five modes", unnknown OPEN ENGLISH SLCC, available at: https://pressbooks.pub/openenglishatslcc/chapter/multi-modal-communication-writing-in-five-modes/.

Fuhrman, T., Schneider, D., Altenberg, F., Nguyen, T., Blasen, S., Constantin, S. and Waibe, A. (2019), "An interactive indoor drone assistant", *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 6052-6057.

Gall, D. and Latoschik, M.E. (2018), "The effect of haptic prediction accuracy on presence", *2018 IEEE conference on virtual reality and 3D user interfaces (VR)*, IEEE, pp. 73-80.

Giang, W., Santhakumaran, S., Masnavi, E., Glussich, D., Kline, J., Chui, F., Burns, C., Histon, J. and Jelek, J. (2010), *Multimodal Interfaces: Literature Review of Ecological Interface Design, Multimodal Perception and Attention, and Intelligent Adaptive Multimodal Interfaces, Technical Report DRDC-TORONTO-CR-2010-051*, DRDC -Toronto Research Centre, Waterloo Univ, Waterloo Ont (CAN) Advanced Interface Design Laboratory, Toronto.

Goodall, J.R. and Sowul, M. (2009), "Viassist: visual analytics for cyber defense", *2009 IEEE conference on technologies for homeland security*, IEEE, pp. 143-150.

Gouin, D., Lavigne, V. and Davenport, M. (2011), *Towards Collaborative Visual Analytics of a Vessel of Interest*, Technical Report DRDC-RDDC-2020-N076, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Guruprasad, P. and Majumdar, J. (2016), *Multimodal Recognition Framework: an Accurate and Powerful Nandinagari Handwritten Character Recognition Model*, Vol. 12, Elsevier, ScienceDirect, pp. 836-844.

Hall, E., Bozowsky, N., Davenport, M. and Wright, W. (2014a), *Maritime Analytics Prototype Final Development Report, Technical Report*, DRDC-Valcartier Research Centre, Toronto, Ontario.

Hall, E., Davenport, M., Bozowsky, N. and Wright, W. (2014b), *Maritime Visual Analytics Prototype Phase 3 Validation Final Report*, DRDC-Valcartier Research Centre, Valcartier, Quebec, Technical Report AD1004162.

Hana, R.O.A., Freitas, C.O.A., Oliveira, L. and Bortolozzi, F. (2008), "Crime scene classification", *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC)*, pp. 16-20.

Harriott, C.E. and Adams, J.A. (2010), "Human performance moderator functions for human-robot peer-based teams", *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Vol. 5, IEEE Computer Soceitty, Osaka, Japan.

Harriott, C.E., Zhang, T. and Adams, J.A. (2011), "Evaluating the applicability of current models of workload to peer-based human-robot teams", *Proceedings of the 6th international conference on Human-robot interaction*, pp. 45-52.

Heard, J., Heald, R., Harriott, C.E. and Adams, J.A. (2019), "A diagnostic human workload assessment algorithm for collaborative and supervisory human–robot teams", *ACM Transactions on Human-Robot Interaction (THRI)*, Vol. 8 No. 2, pp. 1-30.

Hou, Y., Feng, Z. and Xu, T. (2020), "Decision making of mobile robot based on multimodal fusion", *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering*, pp. 243-246.

Hu, B., Liu, X., Wang, W., Cai, R., Li, F. and Yuan, S. (2019), "Prediction of interaction intention based on eye movement gaze feature", *IEEE Joint International Information Technology and Artificial Intelligence Conference*, Chongqing, China, Vol. 8.

Huang, X., Liang, J.L., Wang, Z. and Guo, J. (2017), "Spatial hidden Markov chain models for estimation of pretroleum reservoir categorial variables", *Journal of Petroleum Exploration and Production Technology*, Vol. 7, pp. 11-22.

Huang, S., Ishikawa, M. and Yamakawa, Y. (2019), "Human-robot interaction and collaborative manipulation with multimodal perception interface for human", *Proceedings of the 7th International Conference on Human-Agent Interaction*, pp. 289-291.

Iba, S., Paredis, C.J.J. and Khosla, P. (2005), *Interactive Multi-Modal Robot Programming*, SAGE Publications.

Ida, N. (2020), *Sensors, Actuators, and Their Interfaces*, 2nd ed., The Institution of Engineering and Technology, London.

IFR (2018), *Robots and the Workplace of the Future*, International Federation of Robotics, Frankfurt, available at: https://ifr.org/ifr-press-releases/news/robots-double-worldwide-by-2020.

Ingrand, F. and Ghallab, M. (2017), "Deliberation for autonomous robots: a survey", *Elsvier Artificial Intelligence*, Vol. 247, pp. 10-44.

Jia, Y., Wang, Y., Jin, X. and Cheng, X. (2016), "Location prediction: a temporal-spatial bayesian model", *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 7 No. 3, pp. 1-25.

Jiang, Z. (2018), "A survey on spatial prediction methods", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31 No. 9, pp. 1645-1664.

Jokinen, K. and Raike, A. (2003), "Multimodality–technology, visions and demands for the future", *Proceedings of the 1st Nordic Symposium on Multimodal Interfaces*, pp. 239-251.

Kaindl, H., Falb, J. and Bogdan, C. (2008), "Multimodal communication involving movements of a robot", *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pp. 3213-3218.

Kaluzny, B.L. (2012), A canadian f-35a joint strike fighter cost estimation model.

Kang, S.J., Lim, J., Kim, H.S., Lee, J.Y., Choi, K.H., Yoo, W.J. and Lee, H.K. (2014), "Estimation of relative distance from surrounding buildings by a vision sensor in urban area", *International Conference on Control, Automation and Systems (ICCAS 2014*, p. 14.

Karle, J., Randall, T., Frank, G. and Coady, D. (2018), *Emergent Display Technologies Developing Use-Case Prototypes for Military Command Teams in Virtual Environments, Technical Report*, Defence-Atlantic Research Centre, Dartmouth, Nova Scotia.

Kaur, K., Khurana, R. and Kushwaha, A.K.S. (2018), "Deep survey on visual object tracking in surveillance environment", *International Conference on Research in Intelligent and Computing in Engineering*, RICE).

Kealey, R. and Collier, J. (2020), *Using 3d Lidar for Gesture Recognition, Technical Report DRDC-RDDC-2020-R089*, DRDC-Suffield Research Centre, Suffield.

Kim, J. and Hmam, H. (2009), *Landmark-based Navigation of an Unmanned Ground Vehicle (UGV), DSTO Weapons Systems Division, Technical Report*, DSTO Weapons Systems Division, Edinburgh.

Konaev, M. and Chahal, H. (2021), *Building Trust in Human-Machine Teams*, Brookings Institution, Washington, D.C.

Kontogiorgos, D., Van Waveren, S., Wallberg, O., Pereira, A., Leite, I. and Gustafson, J. (2020), "Embodiment effects in interactions with failing robots", *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1-14.

Kumar, A., Rao, K.M. and Krishna, A.B. (2012), "Speech recognition system for controlling the robot", *International Journal of Engineering Research and Technology (IJERT)*, Vol. 1 No. 7.

Lapinski, A.L.S. (2009), *Situation Awareness Assistance for Pilots and Aircrew: Intelligent Software Agents for Net-Enabled Operations*, DRDC-Atlantic Research Centre, Dartmouth, Nova Scotia.

Lavigne, V. (2014a), "Graph analyzer widget : closer to agility through sense-making", *Proceedings of International Command and Control Research and Technology Symposium (ICCRT) (DRDC-RDDC-2014-P10)*.

Lavigne, V. (2014b), *Interactive Visualization Applications for Maritime Anomaly Detection and Analysis, Technical Report DRDC-RDDC-2014-P97*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V. (2014c), *Maritime Visual Analytics Prototype Final Report, Technical Report DRDC-RDDC-2014-P94*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V. (2015), *Visual Analytics Capability: Integration Strategy for the Sensemaking Support System*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V. and Gouin, D. (2015), *Visual Analytics for Cyber Security and Intelligence, Technical Report DRDC-RDDC-2015-P011*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V., Gouin, D. and Davenport, M. (2011), *Visual Analytics for Maritime Domain Awareness, Technical Report DRDC-VALCARTIER-SL-2011-568*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V., Lecocq, R. and Gouin, D. (2011a), *Counter-Insurgency Visualization Strategies, Technical Report*, DRDC-Valcartier Research Centre, Valcartier QUE (CAN).

Lavigne, V., Lecocq, R. and Gouin, D. (2011b), *Counter-insurgency Visualization Strategies, Technical Report DRDC-VALCARTIER-SL-2011-567*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V. and Gouin, D. (2011a), *Visual Analytics Applications for Cyber Security and Intelligence, Technical Report*, DRDC-Valcartier Research Centre, Valcartier QUE.

Lavigne, V. and Gouin, D. (2011b), *Visual Analytics for Defence and Security Applications, Technical Report DRDC-VALCARTIER-TM-2011-186*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V., Gouin, D. and Davenport, M. (2012), *Visual Analytics and Collaborative Technologies for the Maritime Domain, Technical Report DRDC-VALCARTIER-TR-2012-424*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Lavigne, V., Mokhtari, M., Breton, M., Martineau, E. and Fournier, J. (2019), *Exploitation of Sensor Data Using Artificial Intelligence in Battlefield Sensemaking*, DRDC-Valcartier Research Centre, Valcartier, Quebec, Technical Report STO-MP-IST-178.

Lavigne, V., Jayaram, S. and Panga, M. (2020), *Story Telling Exploratory Visual Analytics for Counter Improvised Explosive Device Insidents, Technical Report DRDC-RDDC-2020-N079*, DRDC-Valcartier Research Centre, Valcartier, Quebec.

Li, J., Feng, Z. and Yang, X. (2020), "Multi-channel human-computer cooperative interaction algorithm in virtual scene", *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering*, pp. 217-221.

Liu, S., Xu, H., Lin, Y. and Gao, L. (2019), "Visual navigation for recovering an auv by another auv in shallow water", *Sensors*, Vol. 19 No. 8, p. 1889.

Lucignano, L., Cutugno, F., Rossi, S. and Finzi, A. (2013), "A dialogue system for multimodal human-robot interaction", in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction,* pp. 197-204.

Lunghi, G., Martin, R., Castro, M.D., Masi, A. and Sanz, P.J. (2019), "Multimodal human-robot interface for accessible remote robotic interventions in hazardous environments", *IEEE Access*, Vol. 7, pp. 127290-127319.

Lutkewitte, C. (2014), *Multimodal Composition: A Critical Sourcebook*, Bedford, St. Martin's.

Maciejewsk, R., Kim, S.Y., King-Smith, K.O., Klosterman, N., Mikkilineni, A.K., Ebert, D.S., Delp, J.F.T. and Collins (2008), "Situational awareness and visual analytics for emergency response and training", *IEEE Conference on Technologies for Homeland Security*, Waltham, MA, USA.

Maragos, P., Potamianos, A. and Gros, P. (2008), *Multimodal Processing and Interaction: Audio, Video, Text (Multimedia Systems and Application*, Springer, New York, NY.

Marcel, S., Bernier, O., Viallet, J.-E. and Collobert, D. (2000), "Hand gesture recognition using input-output hidden Markov models", *proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, IEEE, pp. 456-461.

Mavridis, N. (2015), "A review of verbal and non-verbal human–robot interactive communication", *Robotics and Autonomous Systems*, Vol. 63, pp. 22-35.

Mc Cutchan, M. and Giannopoulos, I. (2018), "Geospatial semantics for spatial prediction (short paper)", *10th International Conference on Geographic Information Science (GIScience 2018)', Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.*

Mead, R., Atrach, A. and Mataric, M.J. (2011), "Recognition of spatial dynamics for predicting social interaction", *Proceedings of the 6th international conference on Human-robot interaction*, Vol. 6, pp. 201-202.

Meng, J., Feng, Z. and Xu, T. (2020), "A method of fusing gesture and speech for human-robot interaction", *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering*, pp. 265-269.

Mihara, I., Yamauchi, Y. and Doi, M. (2000), *A Vision-Based Pet Robot Interaction, Human Factors in Computing Systems*, CHI, ACM, CHI'00 Extended Abstracts on Human Factors in Computing Systems, The Hague, pp. 105-106.

Mikušová, N., Čujan, Z. and Tomková, E. (2017), "Robotization of logistics processes", *MATEC web of conferences*, EDP Sciences, Vol. 134, 00038.

Minker, W., Buehle, D. and Dybkaer, L. (2006), *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, Springer, Dordrecht.

Mitsuishi, M., Chotoku, Y. and Nagao, T. (1997), Tele-guidance system using 3D-CG and real-time tracking vision, Vol. 6, *IEEE International Workshop on Robot and Human Communication, RO-MAN*, Sendai, Japan.

Mohamed, Y. and Lemaignan, S. (2021), "Ros for human-robot interaction", *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE/RSJ.

Monckton, S.P. (2019), "TTCP Autonomy strategic challenge autonomous warrior 18", DRDC- Suffield Research Centre, DRDC-RDDC-2019-D104.

Monckton, S.P., Vincent, I. and Broten, G. (2005), *A Prototype Vehicle Geometry Server: Design and Development of the ModelServer CORBA Service*, Technical report, DRDC-Suffield Research Centre, Suffield.

Moreau, M.D.M. (2011), *Unmanned Ground Vehicles in Support of Irregular Wardare: A Non-lethal Approach, Master of Military Studies Research Paper, USMC Command and Staff College*, Marine Corps University.

Mortimer, B.J.P. and Elliott, L.R. (2017), "Information transfer within human robot teams: multimodal attention management in human-robot interaction", *IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA).*

Nalepka, P., Kallen, R.W., Lamb, M. and Richardson, M.J. (2018), "Emergence of efficient, coordinated solutions despite differences in agent ability during human-machine interaction: demonstration using a multiagent 'shepherding' task", *International Conference on Intelligent Virtual Agents*, Vol. 18, pp. 337-338.

Norris, S. (2004), *Analyzing Multimodal Interaction : A Methodological Framework*, Routledge, Taylor and Francis Group, NY, London.

Oviatt, S., Schuller, B., Cohen, P.R., Sonntag, D., Potamianos, G. and Krueger, A. (2017), *The Handbook of Multimodal-Multisensor Interfaces*, Vol. 1, Association of Computing Society (ACM).

O'Malley, M.K. (2007), *Principles of Human-Machine Interfaces and Interactions*, Artech House Publishers.

Paetzel, M., Varni, G., I, H., Chetouani, M., Peters, C. and Castellano, G. (2017), "Investigating the influence of embodiment on facial mimicry in HRI using computer vision-based measures", *IEEE International Symposium on Robot and Human Interactive Communication*, Vol. 26.

Pandya, A., Eslamian, S., Ying, H., Nokleby, M. and Reisner, L.A. (2019), "A Robotic Recording and Playback Platform for Training Surgeons and Learning Autonomous Behaviors Using the da Vinci Surgical System", *MDPI, Robotics*, Vol. 8, p. 9.

Paul, S., Sintek, M., Këpuska, V., Silaghi, M. and Robertson, L. (2022a), "Intent based multimodal speech and gesture fusion for human-robot communication in assembly situation", *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 760-763.

Paul, S., Sintek, M., Silaghi, M., Këpuska, V. and Robertson, L. (2022b), "A novel multimodal situated spoken dialog system for human robot communication in emergency evacuation", *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 1660-1665.

Post, M.A., Bianco, A. and Yan, X.T. (2017), "Autonomous navigation with ROS for a mobile robot in agricultural fields", *International Conference on Informatics in Control, Antomation and Robotics (ICINCO)*, Vol. 2, pp. 79-87.

Quach, K. (2021), "US Army develops natural-language voice-command AI for robots", tanks, etc. For search'n'rescue. For now, available at: https://www.theregister.com/2021/04/21/military_judi_ai/.

Qureshi, Z.H. (1998), "Human performance modelling in tactical environments", *Proceedings Australian Computer Human Interaction*, IEEE Explore, Adelaide, SA.

Rachid, A. (2022), *On human models for collaborative robots*, available at: https://homepages.laas.fr/rachid/drupal/node/22.

Ranzenberger, T., Christian, H. and Gallwitz, F. (2018), "Integration of a kaldi speech recognizer into a speech dialog system for automotive infotainment applications", Conference on Electronic Speech Signal Processing(ESSB).

Rath, P.K., Mahapatro, N., Nath, P. and Dash, R. (2019), "Autonomous chess playing robot", 28th IEEE International Conference on Robot and Human Interactive Communication, Vol. 28.

Rhodes, B.J., Bomberger, N.A. and Zandipour, M. (2007), "Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness", *2007 10th International Conference on Information Fusion*, pp. 1-8, IEEE.

Richter, M.M., Paul, S., Këpuska, V. and Silaghi, M. (2022), *Signal Processing and Machine Learning with Applications*, Springer Verlag, Heidelberg.

Riveiro, M. (2011), *Visual Analytics for Maritime Anomaly Detection, Master's Thesis, Oerebro Studies in Technology*, Öerebro University.

Robotic, T.U.A. and Strategy, A.S. (2017), Maneuver, aviation, and soldier division army capabilities integration center u.s. army training and doctrine command.

Rousseau, C., Bellik, Y., Vernier, F. and Bazalgette, D. (2006), "A framework for the intelligent multimodal presentation of information", *Signal Processing*, Vol. 86 No. 12, pp. 3696-3713, doi: 10.1016/j.sigpro.2006.02.041.

Roy, J. (2009), *Rule Based Expert System for Maritime Anomaly Detection*, Vol. 12, SPIE Defense, Security, and Sensing, Orlando, FL, United States.

Roy, A., Steinke, D. and Nicoll, R. (2017), "Dynamic simulation of the automated docking of a UAV to a slowly moving submarine in littoral conditions: part1: sensor and control system modelling : part 2; Modelling the Vehicles and the dock; Part 3: software Manual, Technical report", *DRDC-Atlantic Research Centre and and Dynamic System Analysis Limited, Victoria BC (Canada)*, DRDC-Suffield Research Centre.

Ruitang, M., Hongqian, Y. and Chengrong, G. (2007), "Research on intelligent-robot-oriented speech interactive technology and module", *IEEE International Conference on Control and Automation*.

Ruiz, A.Y.R. and Chandrasekaran, B. (2020), *Implementation of a Sensor Fusion Based Robotic System Architecture for Motion Control Using Human-Robot Interaction*, IEEE/RSJ, Honolulu, HI, USA.

Saad, E., Broekens, J. and Neerincx, M.A. (2020), "An iterative interaction-design method for multi-modal robot communication", *IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, Vol. 29.

Schoenherr, L., Eisenhofer, T., Zeiler, S., Holz, T. and Kolossa, D. (2020), "Imperio: robust over-the-air adversarial examples for automatic speech recognition systems", *Annual Computer Security Applications Conference (ACSAC)*.

Schomaker, L. (1995), *A Report of the ESPRIT Project 8579 MIAMI*, Met Lit. opg, [S.n.], Met lit. opg.

Scimeca, L., Iida, F., Maiolino, P. and Nanayakkara, T. (2020), "Human-robot medical interaction", *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 660-661.

Shaker, S., Bovey, R., Dwinnell, L., O"Neill, M., Martin, R. and *et al.* (1992), Unmanned ground vehicle users for the U.S. marine corps and U.S. Army, Marine corps combat development command, Technical report, available at: https://www.spiedigitallibrary.org/

Shu, B., Sziehig, G. and Pieters, R. (2019), "Architecture for safe human-robot collaboration: multi-modal communication in virtual reality for efficient task execution", *IEEE International Symposium on Industrial Electronics (ISIE)*, Vancouver, BC, Canada, Vol. 28, IEEE.

Song, W. and Cai, J. (2015), "End-to-end deep neural network for automatic speech recognition", *Standford CS224D Reports*, pp. 1-8, available at: https://api.semanticscholar.org/CorpusID:7743432

Stachniss, C. (2009), *Spatial Modeling and Robot Navigation*, Cumulative habilitation thesis of the Albert-Ludwigs-University Freiburg im Breisgau, Germany.

Stangor, C. and Walinga, J. (2010), "Introduction to psychology", 1st edn, available at: https://opentextbc.ca/introductiontopsychology"

Stedmon, A., Ryan, B., Fryer, P. and McMillan, A. (2013), "Human factors and the human domain: exploring aspects of human geography and human terrain in a military context", *HCI International, Proceedings, Part II*, USA, Vol. 8020, LAS Vegas, NV, pp. 302-311.

Stewart, S. (2012), "A practical guide to situational awareness", *Security Weekly*, Vol. 16, pp. 47-52.

Tanaka, Y. (2016), "Toward further innovation in human-machine systems", *International Workshop on Computational Intelligence and Applications (IWCIA)*, IEEE, Osaka, Vol. 9.

Tanguay, D.O. (1995), "Hidden markov models for gesture recognition", *Master Thesis, Computer Science, Massachusetts Institute of Technology*, Cambridge, Massachusetts.

Toselli, A.H., Vidal, E. and Casacuberta, F. (2011), *Multimodal Interactive Pattern Recognition and Applications*, Springer Science & Business Media, NY.

Trigueiros, P., Ribeiro, F. and Reis, L.P. (2015), "Hand gesture recognition system based in computer vision and machine learning", *Developments in Medical Image Processing and Computational Vision*, pp. 355-377.

Turk, M. (2014), "Multimodal interaction: a review", *Pattern Recognition Letters*, Vol. 36, pp. 189-195.

Unhelkar, V.V., Li, S. and Shah, J.A. (2020), *Decision-Making for Bidirectional Communication in Sequential Human-Robot Collaborative Tasks*, ACM, HRI.

Varga, M. and Lavigne, V. (2016), "Application of visual analytics to maritime domain analysis", Technical Report DRDC-RDDC-2016-P115.

Venour, C. and Roodnick, D. (2011), *Design of a Maritime Domain Awareness Visual Analytics Prototype (DMVAP), Technical Report DRDC-VALCARTIER-CR-2011-238*, DRDC-Valcartier Research Centre, Quebec.

Vincent, I. (2006), *Reinforcement Learning in Mobile Robot, Technical Report DRDC-SUFFIELD-TM-2007-190*, DRDC Suffield Research Centre, Suffield.

Vincent, I. (2008), *Control Algorithms for a Shape-Shifting Tracked Robotic Vehicle Climbing Obstacles*, DRDC-suffield Research Centre, DRDC, Suffield, Technical Report DRDC-SUFFIELD-TR-2008-123.

Wang, Y., Yanushkevich, S., Hou, M., Plataniotis, K., Coates, M., Gavrilova, M., Hu, Y., Karray, F., Leung, H., Mohammadi, A. and Kwong, S. (2020), "A tripartite theory of trustworthiness for autonomous systems, number DRDC-RDDC-2021-P021", IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3375-3380.

Watt, G.D., Roy, A.R., Currie, J. and Gillis, C.B. (2015), "A concept for docking a uuv with a slowly moving submarine under waves", *IEEE Journal of Oceanic Engineering*, Vol. 41 No. 2, pp. 471-498.

Whitney, D.F. (2019), *Multimodal Human-Robot Interaction with Decision Theory and Mixed Reality, PhD Thesis*, Brown University, Rhode Island.

Williams, T., Szafir, D., Chakraborti, T. and Ben Amor, H. (2018), "Virtual, augmented, and mixed reality for human-robot interaction", *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 403-404.

Wiriyathammabhum, P., Summers-Stay, D., Fermueller, C. and Aloimonos, Y. (2017), "Computer vision and Natural Language Processing: recent approaches in multimedia and robotics", *ACM Computing Surveys (CSUR)*, Vol. 49 No. 71, pp. 1-44.

Wu, M., Taetz, B., He, Y., Bleser, G. and Liu, S. (2022), "An adaptive learning and control framework based on dynamic movement primitives with application to human–robot handovers", *Robotics and Autonomous Systems*, Vol. 148, 103935.

Xue, T., Wang, W., Ma, J., Liu, W., Pan, Z. and Han, M. (2020), "Progress and prospects of multimodal fusion methods in physical human–robot interaction: a review", *IEEE Sensors Journal*, Vol. 20 No. 18, pp. 10355-10370.

Yang, D., Schafer, J., Wang, S. and Ganz, A. (2014), "Autonomous mobile platform for enhanced situational awareness in mass casualty incidents", *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 898-901, IEEE.

Yau, N. (2013), *Data Points: Visualization that Means Something*, John Wiley & Sons, Indianapolis.

Young, T., Hazarika, D., Poria, S. and Cambria, E. (2018), "Recent trends in deep learning based natural language processing", *IEEE Computational Intelligence Magazine*, Vol. 13, pp. 55-75.

Yu, C. and Tapus, A. (2020), "Multimodal emotion recognition with thermal and rgb-d cameras for human-robot interaction", *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 532-534.

Yu, Z., Bohus, D. and Horvitz, E. (2015), "Incremental coordination attention-centric speech production in a physically situated conversational agent", *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, Vol. 16, pp. 402-406.

Zeng, B. and Feng, Z. (2020), "Research on intelligent experimental equipment and key algorithms based on multimodal fusion perception", *IEEE Access*, Vol. 8, pp. 142507-142520.

Zhang, L., Tan, S., Wang, Z., Ren, Y., Wang, Z. and Yang, J. (2020), "Viblive: a continuous liveness detection for secure voice user interface in iot environment", *Annual Computer Security Applications Conference (ACSAC)*, Vol. 20, pp. 884-896.

Zhou, M., Arnold, J. and Yu, Z. (2019), "Building task-oriented visual dialog systems through alternative optimization between dialogue policy and language generation", *Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 9, arXiv preprint arXiv:1909.05365.

Zhou, M., Duan, N., Liu, S. and Shum, H.-Y. (2020), "Progress in neural nlp: modeling, learning, and reasoning", *Engineering*, Vol. 6 No. 3, pp. 275-290.

**Corresponding author**
Sheuli Paul can be contacted at: sheuli.paul@drdc-rddc.gc.ca