International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Collaborative Human-Robot Interaction Using Multi-Model AI Systems

Princi Saini[1], *Sagar Choudhary*[2], *Aastha Verma*[3]

[1] M.Tech Student, Department of CSE, Quantum University, Roorkee, India
[2] Assistant Professor, Department of CSE, Quantum University, Roorkee, India
[3] B.Tech Student, Department of CSE, Quantum University, Roorkee, India

ABSTRACT

To develop a set of 16 dynamic gestures for real-time gesture identification, a custom dataset was created using convolutional neural networks and motion history images. At the same time, verbal input from human operators is processed by an improved open-source voice recognizer. A multi-threaded system design enables the simultaneous operation of gesture and speech recognition, sensor fusion, robot control, and user interface. The experimental evidence supports that multimodal communication increases usability, trust, and understandability, but we cannot ignore the potential negative implications, like overconfidence and distraction. A comprehensive review of the latest multi-modal fusion techniques for physical HRC (pHRI) indicates that accurate and real-time spatiotemporal data fusion across different sensory modalities is critical. This research highlights the importance of using multi-modal interaction to develop safe, usable, and efficient collaborative robotic systems in industry.

Keywords: Human–Robot Collaboration(HRC), Multi-Modal Interaction, Gesture and Speech Recognition, Physical Human–Robot Interaction (pHRI), Sensor Fusion Techniques

Introduction

Rapid developments in industrial automation and artificial intelligence have made human–robot collaboration (HRC) with improved interaction a major and expanding area of current research.. We present human-robot collaboration (HRC) design and development of a real-time, multi-modal HRC version that provides speech and gesture-based interaction. A dataset of the gestures has been created and openly shared. For gesture recognition, a convolutional neural network (CNN) was developed, making use of motion history images combined with advanced deep learning, and allowing real-time operation. An enhanced open-source speech recognition engine is also used to process human operator gestures and verbal commands.[1]

The proposed system includes an innovative integration method that combines gesture and speech input with a custom-designed software interface that supports visualization and user interaction. The system architecture includes multi-threading to facilitate simultaneous data acquisition, gesture and speech acquisition, data processing, input fusion, robot control, and user interface updates. This fully integrated system demonstrates the efficacy and feasibility of the proposed approaches, with experimental validation of the methods.

Industrial collaborative robots, often called "cobots," are designed to work with people within shared workspaces.[2] Human collaborators typically apply social behaviors learned while collaborating with other humans (e.g., applying speech, gestures, facial expressions, and gaze) when interacting with non-human agents (robots). Allowing cobots to reciprocate social behaviors with similar social signals will also facilitate intuitive and more efficient collaboration. Cobots can signal information via three fundamental modalities: visual, acoustic, and mechanical forms of communication. Visual channels may involve lights, text, and expressive gestures as forms of feedback. Acoustic signals (e.g., speech or audio cues) can potentially complement visual communications but are poorly suited to portraying social signals in noisy industrial environments.[19] Mechanical communication, such as providing haptic information, is available, but it requires specialized hardware. Any social signals from visual, acoustic, or mechanical channels must be easily detectable by human operators without extensive specialized training. Previous research on multi-modal signals in HRC has defined social cues for collaborative robots by concentrating on visual modalities and varying degrees of anthropomorphism.[3] Robots have been working on repetitive and high-load motions for a long time in manufacturing, thus providing them with access to inexpensive, user-friendly automation, and contributing to small and medium-sized business growth. Cobots have entered increasingly unstructured and dynamic environments, and the necessity of intelligent real-time collaboration is growing. HRC permits humans and robots to leverage their unique abilities (robot accuracy and endurance versus human flexibility, situational awareness, and cognitive skills), which augments productivity, flexibility, and overall efficiency.

However, effective collaboration necessitates that cobots operate in safe and intuitive ways in shared spaces. It requires a higher perceptual understanding for example knowing objects or human actions, and most of the papers and reviews focus on wearable robotics, generic HRC processes, or even systems based on just one modality such as solely vision or use of tactile, but little consideration has been given to the way systems can utilize information from multiple modalities in HRC.[4],[5]

Objective

The main goal of this research is to develop and implement a real-time, multi-modality human–robot collaboration (HRC) system that combines dynamic gesture recognition with speech recognition, to enable real-time communication and collaboration between humans and collaborative robots (cobots) in industrial contexts.[1] To achieve seamless gesture recognition, we created a set of 16 dynamic hand gestures and a new dataset. The gesture recognition uses Convolutional Neural Networks (CNNs) and motion history images to create a real-time, robust gesture recognition system, while an improved, open-source speech recognition engine is used to interpret verbal commands.

The use of multi-threading functionality available in the system is a distinctive aspect of this research plan, having gesture recognition, speech recognition, sensor data fusion, robotic actions and the GUI all simultaneously running, lends itself to much more natural interaction, effective workflows, and allows us to bypass some limitations associated with unimodal systems.[3] Furthermore, the research has outlined the necessity for multi-modal fusion methods for even greater impacts to safety, usability, and user trust in the often high-stakes and dynamic industrial settings. The performance of the system was verified in lab tests that demonstrated that verbal and visual modalities will rapidly improve the adaptability and reactivity of cobots. In conclusion, this work was a step towards the evolution of collaborative robotic systems that are safe, intelligent, and user-friendly.

Problem Statements:-

1. Human-robot communication (HRC) systems do not allow real-time integration of gesture and speech to facilitate effective information transfer.
2. Any single-sensor (single-modality) system cannot provide an accurate and reliable perception of the perceived actions in physical human-robot interactions.
3. Low-anthropomorphic industrial robots cannot properly express social cues and thus limit the understanding of humans.[2]
4. In physical human-robot interactions, multi-modal fusion methods are inconsistent and not standard across applications.[3]
5. Speech recognition systems do not work in a noisy industrial environment.[6]
6. There is no reliable mechanism to incorporate data from different types of sensors within HRC systems.

Literature Review

Human-Robot Collaboration (HRC) has become an important research area in robotics, mainly due to advancing the cognitive flexibility and variability offered by humans with the reproducible precision and productivity afforded by robotic systems. This approach, to human-robot learning and training, in industrial and service settings, is increasingly being applied in multi-modal communication systems, aspiring to achieve similar communication richness to human-human communication. Multi-model HRC systems combine all different modes of interaction in the form of gesture, speech, vision, audio, tactile, and physiological signals, to support human-robot interaction that is natural, organized, and responsive in real-time. New academic work provided by [1] explores complementary perspectives associated with the design, use, and future of multimodal human-robot collaboration systems. Each piece, while differing operationally, explored interrelated concepts of collaborative robots, ranging from real-time communication systems, social modeling and interaction, and sensor fusion approaches.

The development of a real-time, multi-modal communication system for industrial HRC, allowing for relative gesture and speech recognition to promote natural and effective human–robot interaction.[1] The developed system consists of 16 dynamically designed gestures inputted visually, recognized via a convolutional neural network (CNN), and received in conjunction with natural speech commands that are operated through a modified open-source speech recognizer. A particularly new aspect of the study is adopting motion history images (MHI) from some gesture sequences to extract spatio-temporal features, which improved its accuracy on real-time recognition of the gesture. The way they're able to integrate gesture recognition, speech processing, robot control, and interface visualization, even while being a multi-threaded paradigm (running concurrently), was a notable feature for painting a picture as to how the mitigation of various latency challenges is possible. The systemic characteristic of avoiding wearable sensors and tapping into just visual and auditory modalities on the operator in the HRC system encourages natural movements without encumbering mobility or comfort in dynamic environments. This research tackled key dynamics and challenges to using poor speech recognition systems in industrial settings with ambient noise, lighting, and complexity of dynamic sequences, moving toward improved opportunities for incorporating intuitive command systems.

By comparison,[2]focus on the socio-cognitive aspect of HRC and the multi-modal social cues: head-like gestures, light displays, and audio cues—to determine how multi-modal social cues can improve communication from robot to human and how combinations of social cues influence human perception, trust, enjoyment, and understanding of a given task, studying both light display and/or sounds along with visual displays of low to high anthropomorphic robots. Two experimental studies revealed that multi-modal social cues improved users' perceptions of the robot as being more friendly and indicated an intention to collaborate when visual gestures are paired with either sound or light. The study indicates that social cues are non-trivial, contribute toward establishing a mutual understanding of purpose in human–robot teams, although having excessive anthropomorphism can create a condition of excessive trust, when users assign unreasonable levels of intelligence or reliable behavior to a robot, that can put a potentially user in harms way in terms of safety or in certain contexts, e.g. Aviation. Thus, designers of socially interactive robots are challenged to strike a balance between social expressiveness and user awareness.

Provide a comprehensive systematic review of the applications of multi-modal fusion techniques in physical HRC (pHRI). pHRI is defined as 'Where a human and a robot work collaboratively in the same physical space', and bi-modal fusion typically refers to using multiple sensory modalities, such as visual, EMG, EEG, proprioception (force/torque), tactile sensors, or Inertial Measurement Units (IMUs). For their review, categorize sensor data as either feedforward communications or feedback communications.[3] They highlight differences in how these differ from intent recognition, motion prediction, environmental awareness, and interaction with safe contact. Traditional means of multi-modal fusion, such as Kalman filtering, Bayesian inference, rule-based systems, etc., are compared to newer models for multi-modal data fusion with base-case computational learning, such as deep neural networks, probabilistic graphical models, etc. The authors also emphasize the need for spatiotemporal modeling concerning pHRI, where a robot has to disentangle a stream of continuously variable signals to infer the intent of a human participant and change their behavior, while remaining in safe contact. They believe that learning-based fusion may be more adaptive to dynamic environments and across differences in users, but are also aware that fusional learning models are likely to have greater computational requirements, sensor correlations, and challenges associated with the interpretability of users' decisions or thinking processes.

As a collective, these three studies provide an overall perspective for multi-modal HRC. Together, [1] provides evidence of the operational viability of gesture and speech for real-time control in HRC via the combination of the advancements in multi-modal human robot interaction (HRI), and in particular, multi-threaded capabilities to ensure responsiveness and usability. Cao et al. focused on the human emotional/psychological dimensions of interaction. They found that the use of multi-modal social cues increased user engagement and perceptions of robot intelligence, but that when designing multi-modal cues, careful consideration should be given to possible overdependence or false attribution. Lastly, [3] offers a broad framework of sensor fusion and signal integration, clearly delineating the constituents of HRC that are necessary for effective interaction in a physical context.

Even with such advancements, there are still potential issues to resolve. Real-time multimodal systems need to pay particular attention to the amount of computational infrastructure they require so that there are no consequences for scalability or responsiveness in crowds and high-demand contexts. Models driven by deep learning may be powerful, but can be brittle when they encounter occlusions, sensor noise, or unfamiliar contexts. In addition, gesture and speech vocabularies can fail to transfer across cultures, industries, or tasks in which significant retraining or adaptation will be required. Ultimately, considerations of human social cues in robotic behavior must also be balanced with what the impact of this may be psychologically, where facilitating likability or anthropomorphism does not decrease the user's situational awareness or safety.

Methodology

Design of Dynamic Gestures

Dynamic gestures are made through coordinated movements encompassing both the upper arm and forearm, with no movement of the fingers. Dynamic gestures, unlike static gestures, make use of additional temporal characteristics (e.g., motion trajectory, orientation, and speed), which implies that dynamic gestures have much greater variability and richer information than static gestures [1].In human-robot collaboration (HRC) systems, effectively designing gestures must adhere to three fundamental principles: simplicity, social correctness, and minimal cognitive affordance on the user. Based on the classification outlined, we identify four gestures:

1. Iconic gestures explicitly represent an action or the physical presence of an object,

2. Metaphoric gestures represent an abstract idea,

3. Deictic gestures, normally involving a hand or limb that physically points towards an object, and

4. Beats, which are hand movements focused on rhythmically supporting speech that has no semantic [25] relevance.

Given the nature of dynamic gestures about clarity and semantic value, we produced 16 dynamic gestures in both iconic and deictic form for use in HRC system development (see Fig. 2). The gestures were delineated into two different functional groups: calibration and operation. The calibration gestures (Gestures 1 - 4) assist the robot to commence standard robot behaviours: clockwise (CW) or counterclockwise (CCW).

Start: Starts kinematic calibration - finding a set of initial values in the robot's motion structure and executing the position for task execution.

Stop: Stops all robot motion and maintains the current pose.

Go Home: Puts the robot back in its default position, or home position (for example, full arm extension in the upright position).

Disengage: Used when the robot hits its motion limits, this gesture allows the robot to move and retract safely back into (only) a functional workspace. The remaining operation gestures (Gestures 5-16) control specific robot actions (e.g., directing the end-effector movement, speed, gripping and opening/closing the gripper, rotating the end-effector CW and CCW).

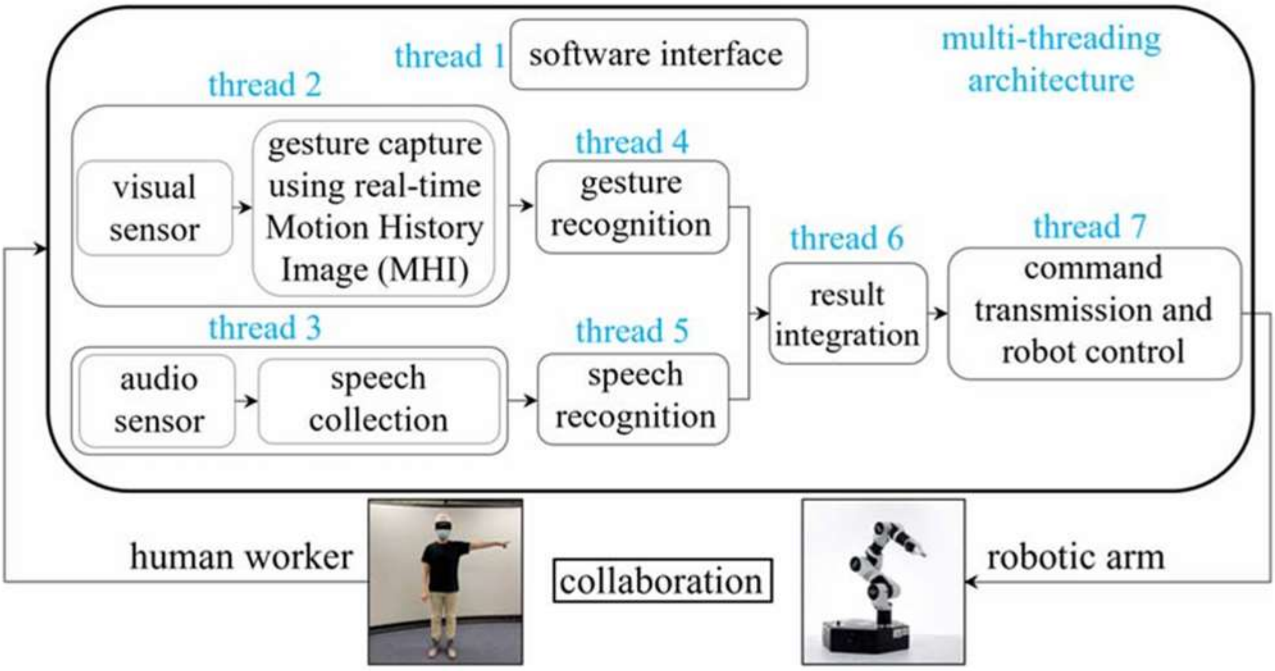

Fig. 1 System overview

Speech Recognition

To make the proposed human–robot collaboration (HRC) system more resistant to disruption and more flexible, we combine speech input from the human operator and real-time gesture recognition to enhance the overall accuracy of command interpretation. In addition to using log-Mel filterbanks to generate low-dimensional representations of log-spectral magnitude vectors, the system makes use of the Google open-source voice recognition software, which supports over 120 languages. This preprocessing step significantly increases overall speech recognition accuracy, especially for deep neural networks that have a larger number of parameters.

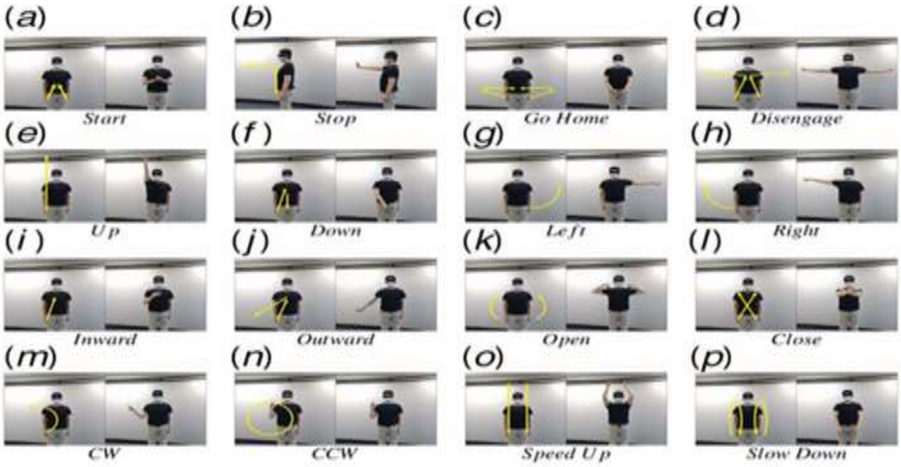

Fig. 2 Illustration of the designed 16 dynamic gestures (CW, clockwise; CCW, counterclockwise): (*a*) gesture 1, (*b*) gesture 2, (*c*) gesture 3, (*d*) gesture 4, (*e*) gesture 5, (*f*) gesture 6, (*g*) gesture 7, (*h*) gesture 8, (*i*) gesture 9, (*j*) gesture 10, (*k*) gesture 11, (*l*) gesture 12, (*m*) gesture 13, (*n*) gesture 14, (*o*) gesture 15, and (*p*) gesture 16

To test the recogniser in an industrial context, we ran speech recognition experiments under ten different background noise conditions. The danger condition samples are shown in Figure 10 and were acquired from freely accessible public information sources on the internet. All noise conditions were recorded at around 60 to 70 decibels (dB) or similar to a functioning manufacturing facility. Any noise sources above this level (>70 dB) were not included because typical human speech occurs at around 60 dB, and the loud noise from the environment would likely drown out the spoken command.

In Figure 3, we present speech command waveforms collected in an isolated and quiet condition. Figure 12 presents cases for (1) the "Start" command without background noise; (2) only drilling noise in empty air; and (3) the "Start" command and drilling noise.

A spectral subtraction method is used to minimize the degradation of speech recognition performance due to ambient noise. The spectral subtraction method measures the spectrum of noise and speech signal, then takes the noise spectrum away from the speech, converting the speech into a clearer representation. The first 0.5 seconds of each audio recording, which describes only background noise, is used for noise modeling and to create a noise profile, while the remainder is actual speech. This lead time allows for the collection of ambient noise before processing the gestures and speech inputs from the user.

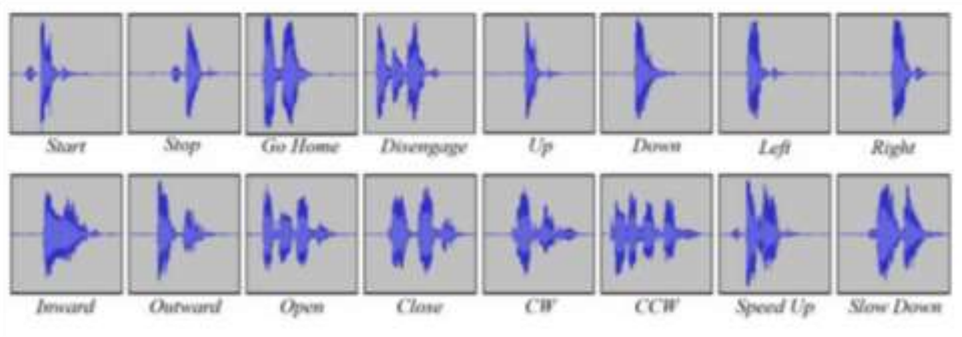

Fig.3 Waveform samples of the 16 commands

The audio dataset consists of 16 different speech commands and recordings made by two native English speakers (one male and one female) in ten separate noise conditions, leading to around 1,600 audio recordings. The accuracy of speech recognition models, for a given command class (C), is defined as the number of times the command class was recognized divided by the total number of samples in that command class. Figure 4 indicates that the majority of command sounds had recognition accuracy of more than 90%. On the other hand, two commands had notably lower recognition accuracies, which were 'speed up' and 'inward'. The "speed up" command had an accuracy of around 75%, and "inward" had an accuracy below 15%.

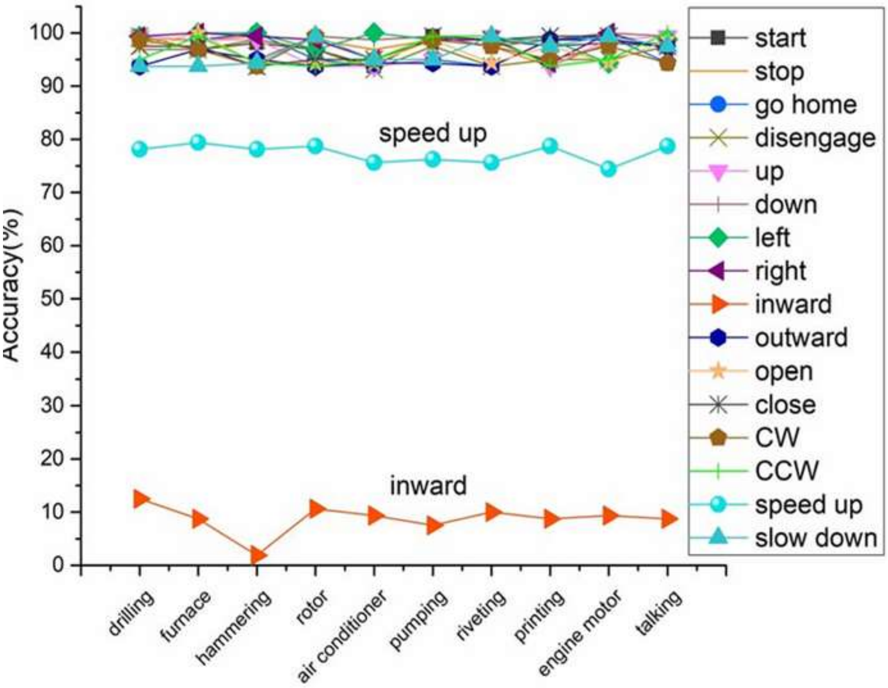

Fig. 4: Performance (%) of the speech recognizer on our speech data set

Table 1 shows the recognition transcripts for the two commands that had low recognition accuracy. During the speech recognition process, the Google open-source speech recognizer runs through multiple candidate transcripts for each input. Each candidate transcript has a confidence score associated with it (0 - 100%). The final output transcript is the one with the highest confidence. For example, for the first sample of the "inward command", the possible transcripts were "uworld", "in word", "inward", "keyword", and "in world". The final transcript selected was "uworld" because it had the highest level of confidence available at 69.70%.

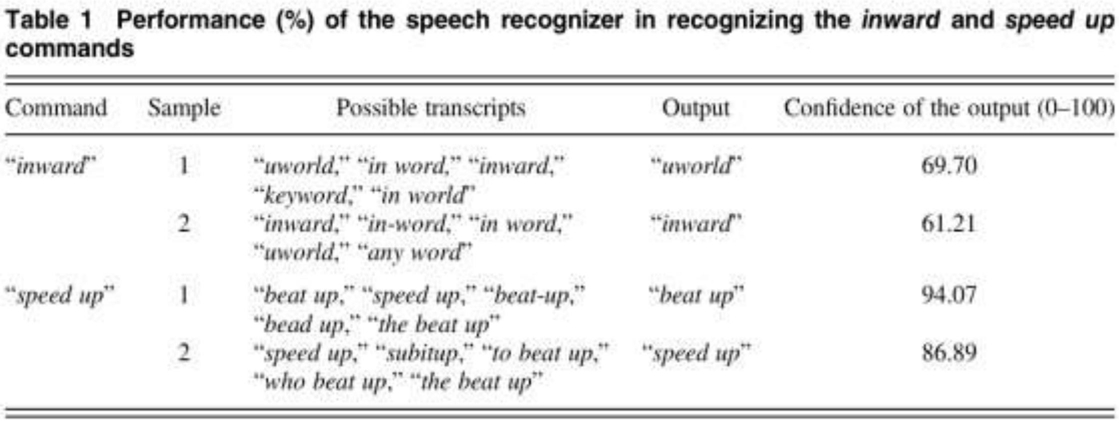

Table 1 Performance (%) of the speech recognizer in recognizing the *inward* and *speed up* commands

Command	Sample	Possible transcripts	Output	Confidence of the output (0–100)
"inward"	1	*"uworld," "in word," "inward," "keyword," "in world"*	*"uworld"*	69.70
	2	*"inward," "in-word," "in word," "uworld," "any word"*	*"inward"*	61.21
"speed up"	1	*"beat up," "speed up," "beat-up," "bead up," "the beat up"*	*"beat up"*	94.07
	2	*"speed up," "subitup," "to beat up," "who beat up," "the beat up"*	*"speed up"*	86.89

Note: The recognizer outputs the transcription with the highest confidence, and the lowercase/capitalization does not affect the recognition result.

We treat the transcripts in Table 1 as the correct recognitions for the speech database. We use the correct recognition approach since the available command phrases are very distinct and not likely to be confused with one another in the factory floor HRC cases. We then achieved a real-time keyword extraction technique during recognition for the 16 speech commands. Specifically, the system could filter necessary keywords from short spoken phrases—for example, reducing "speed up the process" to the "speedup" keyword, or recognizing "go inward" as the "inward" keyword. Experimental results are presented in Section 5 and show that this method offered a gain since the accuracy of both the "inward" and "speed up" commands was increased to over 90%.

System Integration:-

Integration of Gesture and Speech Recognition Results:-

The strategy of integrating the system is to combine gestures and speech recognition outputs to produce dependable command interpretation in a natural human-robot collaborative environment. The system uses a confidence score between 0% and 100% to compute a recognition score and confirm the recognition outcome, with 0% being that what was identified was not the command intended through the gesture or speech, and 100% being confident that what was identified is a correct gesture or speech command. The system will accept a command using either or both outputs (gestures and speech) only if the confidence scores indicate a signal of 90% or above. If the confidence score from either or both outputs produces a result scoring below 90%, the results will be invalidated, the system will sound a beep for the user, and the system will start awaiting input for a suitable command - in this way, false positives and false negatives can be decreased[1].

The integration logic is intended to be able to accommodate the scenarios described in Algorithm 2. First of all, the program checks the ambient noise level to determine if it is appropriate to consider speech commands; if the noise level is above 70 dB, the system will not utilize any speech commands. Second, the system processes gesture and speech outputs as well as the accompanying recognition confidence assessments. The program handles 5 possible integration cases:

1. When both gesture and speech results are the same and valid (i.e., one of the 16 command labels), the output will be the same command label.
2. When gesture and speech results are different but are both valid, then the output will be determined based on the confidence values by taking the command with the highest confidence value.
3. When only a gesture result is valid, then the gesture will be output as the final command.
4. If the speech command is valid but the gesture is invalid, then the speech outcome is selected.
5. Both gesture and speech inputs provided by a user are invalid, and "invalid" will be displayed, informing that the system is waiting and to provide more input. The system remains active, collecting speech and gesture data continuously, only waiting to be activated until an input is considered valid. Where background noise exceeds 70dB, the system does not use speech input due to the level of chance misinterpretations, and only utilizes gesture identification. In this case, there are two more cases to be assessed.
6. It is where the gesture is valid and is accepted as the final command.
7. The gesture is also invalid; no command is issued to the robot, and the system will remain on standby, continuously collecting new observation events and information until a valid input is detected.

Multi-Threading Architecture of the HR system:-

The HRC system is built on a multi-threading configuration to allow multitasking, necessary for high-speed performance with real-time responsiveness. The next two paragraphs describe the seven operational threads in detail. Figure 5 shows the high-level organization of the thread design for operational tasking and the management of the dedicated threads. The system initializes and organizes seven operational threads: user interface thread, gesture capture/thread, speech capture/thread, gesture recognition thread, speech recognition thread, command/feedback result integration thread, and executable robot actuators thread. By separating these operations into different threads, they can be performed in parallel, which is crucial when executing and communicating between humans.[1]

Fig.14 Multi-threading model of the proposed real-time HRC system with an example

Robot in adjusted timing activities with low latency. For optimal system performance, all threads are designed to run continuously in a collaborative multi-threading environment with the goal of meeting or exceeding time constraints imposed by the adaptive real-time collaborative environments with robots and humans working together.

To guarantee smooth and continuous real-time operation, the real-time capabilities and functions in the system are isolated from other processes and assigned to independent concurrent threads. They are autonomous threads that receive and send information from and to queues (to simplify implementation, the queues are linear queues shown in Fig. 5), which will simply store the data in a FIFO-first-in, first-out process.

The system is implemented using the Python language, developed in the PyCharm environment. The performance hardware used for implementation is a 24-core CPU, a dual Nvidia GeForce GTX 1080 Ti Graphics Processing Units (GPU), and 64 GB of RAM. The gesture recognition model processes one Motion History Image at an average of 0.026 seconds per MHI, which essentially processes at 30 fps (0.0333 seconds per frame) and thus supports real-time gesture processing. The speech commands are recognized using the Google Speech Recognition API at an average of 0.218 seconds for speech commands that average three seconds in duration, which is within the threshold for real-time. Therefore, the CNN-based gesture recognition and speech recognition modules effectively support real-time performance without delay.

Thread 1 is running continuously while the system is running to manage the user interface. Threads 2 and 3 read the gesture and speech inputs from the RGB camera and microphone, respectively. The input devices are automatically shut off at the end of each data collection cycle, then turned back on when Threads 6 and 7 have completed processing. The data collection window (known as "execution time") is fixed at 2.5 seconds based on the maximum time needed for a user to complete a gesture or speech command from the collected dataset. During this time window, the camera and microphone, recording gesture and speech input respectively, store input data into separate gesture and speech data queues.

Fig.5 Multi-threading model of the proposed real-time HRC system with an example

The MHI is rendered in real time on the interface, and the microphone is turned on 0.5 seconds earlier than the camera to allow the ambient noise to be collected to denoise speech inputs.

The study

Threads 4 and 5 handle recognition, processing MHIs, and speech inputs as they are dequeued. The recognition results are labeled and displayed on the interface, as well as stored in separate label queues. Thread 6 takes the labeled results and fuses the gesture and speech outputs into a single command and stores that final decision in the result queue. Finally, Thread 7 continuously reads from the result queue to communicate the interpreted result.

To the best of our knowledge, prior work with industrial robot arms does not fully investigate the interactions between head-like gestures and combinations of integrated light and sound. In the study, we are considering the gap of examining the effects of multi-modal social cue behaviors in a robotic arm that produces three communication modalities together: head gestures (gaze cues, nodding, and head shaking) with an improvisational "breather," visual feedback with LED lights, and auditory feedback with an informative sound cue. Rather than looking at these modalities separately, we will compare a condition where all three modalities are enabled (i.e., full condition), to other conditions where either the light/or sound is not enabled. This way of comparing a multi-modal level condition to another condition assumes these modalities have possible coupling effects, looking at the head-like gestures, building off the findings of Tatarian et al. (2021). All conditions will be considered to have the same contextual information if everything is enabled. Based on this design, the following hypotheses were formulated.

Hypothesis 1 (H1): Expressing multi-modal social cues can influence people’s perception of industrial robot arms.[20]

Prior work has demonstrated that expressing head gestures can influence people’s perception of industrial robot arms using five scales of the God Speed questionnaire, except perceived safety.[20] The study demonstrates that a higher rating in Anthropomorphism can offer a lower rating in perceived safety. The exploratory study suggests that light and sound modalities can affect making people feel more comfortable with robot motion. Additionally, alarm lights and sounds have been used to raise concerns about safety in industry, and especially human-robot collaboration in industrial setups. Hence, we hypothesize that the use of combining head gestures, light, and sound modalities will influence the ratings of all five scales of people’s perceptions of the robots.

H2 (Hypothesis 2)

The display of multi-modal social cues can positively influence users’ perceived enjoyment, perceived usefulness, and intention to use industrial robotic arms. For example, robot head gestures have been shown to positively influence perceived enjoyment, perceived usefulness, and intention to use within high and low anthropomorphic robot contexts in some studies.[46],[54] We forecast, exploratory study, that the positive influence of head gestures, given light and sound multimodal. Social cues will persist in perceived enjoyment, perceived usefulness, and intention to use.

Hypothesis 3 (H3):

The addition of light and sound will further enhance the legibility of social cues provided through head gestures.

The combination of visual (e.g., light) and auditory (e.g., sound) cues provides additional layers of information that can better the understanding of humans about the behavior of robots, increasing the effect of human-robot interactions. Using prior studies as a basis, we hypothesize that the social cues (e.g., head gestures) will be easier to recognize and interpret with both light and sound, compared to head gestures alone. Furthermore, we examine if the combination of both modalities produces better understood results than either modality (light or sound) alone. [21]

To test our hypotheses, we completed two user studies online. In both studies, users watched videos of a human collaborating with a Franka Emika robot arm and were asked to evaluate their impressions of the robot, as well as interpret the action of the robot. Study I was focused on evaluating overall impression of the robot (H1), as well as perceived enjoyment and intention to use the robot (H2).[43]

I. MULTI-MODAL FUSION METHODS:-

Multi-modal fusion has been used in many areas (content analysis, fault detection, aerial navigation), and many review studies attempt to categorize these processes into various techniques.[1] provided a widely cited organization, sorted by sensor configuration, into three types: complementary, competitive, and cooperative. They also discussed fusion techniques, which exist at both the signal and decision levels. Similarly, on wearable robotics and recommend four unique fusion approaches: single fusion algorithm, unimodal switching, multimodal switching, and mixing.

While these are general categories, this paper focuses on the specific area of pHRI. Based on the idea of a two-phase stable grasping concept, we describe pHRI as two distinct phases: global space perception and local contact behaviors. Each of these phases entails different sensor modalities and different fusion strategies. By aligning the fusion strategy with the phase of interaction, we offer a more context-responsive and effective design framework for achieving real-time, responsive, and ultimately safe, physical collaboration in human–robot systems.

Global Space Perception Phase

Global Space Perception Phase. Corresponding to the indirect contact signals, the global space perception phase is the first phase of pHRI. The big targets of observation can be split into three aspects: environments, task objects, and humans. Various methods for multi-modal fusion concerning these three aspects are shown. The perception of 3D environments is critical for robots to have coarse cognition about the environment and humans. To improve the capturing of the environment's information,[3]fused orientation information from inertial sensors, images from cameras, and depth information from ToF sensors, using a probabilistic framework to synergistically get robust reconstructions. The use of probabilistic models has an inherent advantage in the treatment of heterogeneous information, and the Bayesian equation uses prior knowledge of previous states. Concerning specific task objects in the global workspace, there are lots of interesting architectures proposed. used a weight-shared strategy and parameter-free correlation layer to fuse RGB and depth information. They designed a complex feature learning architecture, which is made up of two modality-specific and one modality-correlated deep feature learning networks. In the correlated detection net, after passing through 3 layers. Due to the inherent differences between sensor signals, the first step was to use a series of filters to process the raw input data during feature construction. This included the Sobel operator (with a window size of 3), Canny edge detector, Laplacian of Gaussian (LoG), and Gabor filters, each of which helped in pattern identification of the signals. After feature extraction and classification, the most informative features were found and selected for further refinement during feature selection. For modeling temporal sequences, Long Short-Term Memory (LSTM) networks were trained independently on the features derived from each modality for early prediction of turn-taking behavior. Finally, the Dempster–Shafer Theory (DST) was used during the decision-level fusion stage to merge the information from multiple sensor channels, providing a more robust and complete understanding of turn-taking intent.

Local Contact Operation Phase

After the global sensing phase, in which robots have a basic understanding of the workspace, task objects, and nearby humans, one must address how to support proximal human–robot collaboration to carry out complex, fine-grained tasks. This entails a dynamic long-distance interaction in a closely-coupled system involving direct physical contact; thus, this section will focus on fusion methods for interpreting multiple modalities during local interaction with contact.[3]

The salient aspect of this phase of interaction is the continuous, kinesthetic physical contact between human and robot. While the previous phase of interaction (global detection) had a touch and force/torque modality, the context of local contact means that one can engage with more active modalities of touch and force/torque, which preconditions the modalities towards object recognition and promotes precise real-time control over the motion of the robot. In the context of this close interaction, a tactile-image fusion framework for recognizing a target object engaged with physical movements. Specifically, the authors chose to model their tactile data as a multivariate time series, characterize the empirical equations, and generate covariance descriptors for the tactile images. The classification was undertaken with a k-nearest neighbors (K-NN) algorithm from a distance map modulated by references created with Dynamic Time Warping.[3]

To improve robotic perceptibility of object properties like compliance, texture, and thermal response, a Bayesian exploration model enhanced with reinforcement learning, allowing robots to make exploratory movements that reduce ambiguity concerning the candidate objects by building a confusion probability matrix, thereby improving recognition accuracy across a range of object properties.

While haptic sensing contributes valuably during physical contact in human–robot interaction, visual inputs in HRI play an equally critical role, as visual input gives more comprehensive awareness of the surrounding environment and object context. To give an example, collected tactile data from a Barrett Hand (BH8-280) and visual data from a Kinect sensor. For the two sensory modalities, raw sensor inputs are first processed to extract feature vectors that represent the tactile-visual informational content, and then conformed to the same representation structure and size using Bag-of-Systems (BoS). The sensory modalities are combined in a linear representation model, where each feature vector is represented as a linear combination of learned training dictionaries. The distribution of weights by each modality corresponds to classification accuracy during training, and the dictionary is optimized to improve object classification.[3]

II. MULTI-MODAL FUSION CLASSIFICATION:-

The process of multi-modal fusion in physical human–robot interaction (pHRI), combines the many sensing modes used to perceive the environment, the human actions, and object interactions. Coordinating a pHRI effect from multi-modal fusion can occur either during global space perception or local contact perception (Section III). All strategies in multi-modal fusion maintain the three integral features of pHRI:

(1) Shared spatio-temporal space:-Spatio-Temporal Sharing: pHRI occurs in a shared physical space over time, requiring fusion techniques that account for both space and time. Depending on whether the input is instantaneous or includes a time series, we can categorize the fusion approaches: transient fusion (one-time snapshot); sequential fusion (based on time-series data).

(2) Multi-modality:-Spatio-Temporal Sharing: pHRI occurs in a shared physical space over time, requiring fusion techniques that account for both space and time. Depending on whether the input is instantaneous or includes a time series, we can categorize the fusion approaches: transient fusion (one-time snapshot); sequential fusion (based on time-series data).

(3) Multi-task processing. Here, we provide an organized account of fusion techniques using these three features::-Multi-Task Processing: pHRI tasks are dynamic and inherently result from multiple tasks performed simultaneously. Each of these tasks will vary in its many characteristics.

Spatio-Temporal Sharing

Physical human–robot interaction (pHRI) typically includes how perception processes occur through space and time effectively within a common environment. In some examples or situations, the system may need only to perceive the instance of the elements that are supporting or in the interaction, for instance: the state of sensor data (i.e., an RGB image or depth image) would be sufficient to perceive with visual detection--and consequently ascertain the state relative to the moment, whilst making allied inference relative to global spatial perception. The process to capture all of the perception processes can be initiated for example, through using the moment in time of the interacting elements states or the physical human robot interactions for evaluating appropriateness; but there can be situations such as a robot estimating the future trajectory of a human where including previous observations is warranted to substantiate the transitions and temporal change which is significantly more pronounced for more dynamic or predictive tasks requiring incorporated fused state information. Temporal sequences of data that is representative of an interaction provide key contextual information and features important to inference and conclusions about future state.

To be more precise, fusion approaches that capitalize on this spatio-temporal characteristic are generally organized into two primary ways of fusion forms: transient fusion, which captures some snapshot of a single point.

Transient Fusion (TRA)

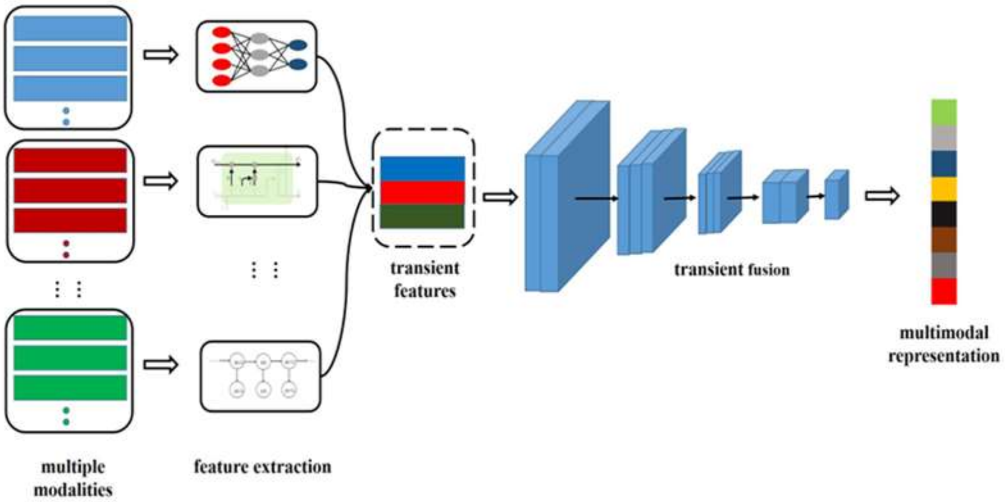

Fig.5 Transient fusion (TRA)

Transient fusion refers to taking representations of sensor data at one time point and processing and filtering the data. For our purposes, we want to represent a multi-modal representation of the current state of the system (depicted in Fig. 5).

In the transient fusion process, multiple modal data points (e.g., vision, geometry or skeletal tracking) occur concurrently from different sensors. In the first step of the process, the raw data are filtered into associated relative features that point to some state. Then, each modality feature is presented via a transient fusion module to give multi-modal representations for engaging with (i.e., decision making) later.

Transient fusion is well established for modal representations that can be captured and represented while the information is captured at one time (RGB images, geometry features, skeletal tracking, object affordances etc.). In general, transient fusion is common in situations where a user's state needs to be evaluated in real-time while eliminating motion or transitional actions between states. Classic examples of transient fusion include 3D reconstruction of an environment, real-time human detection or object detection, or immediate gesture grounding.

Sequential Fusion (SEQ):

Sequential fusion merges temporal data into models that maintain some memory (e.g., Recurrent Neural Networks (RNNs) and/or hidden Markov Models (HMMs)). These models then provide a multimodal representation that maintains the value of temporal order (see Fig. 6).

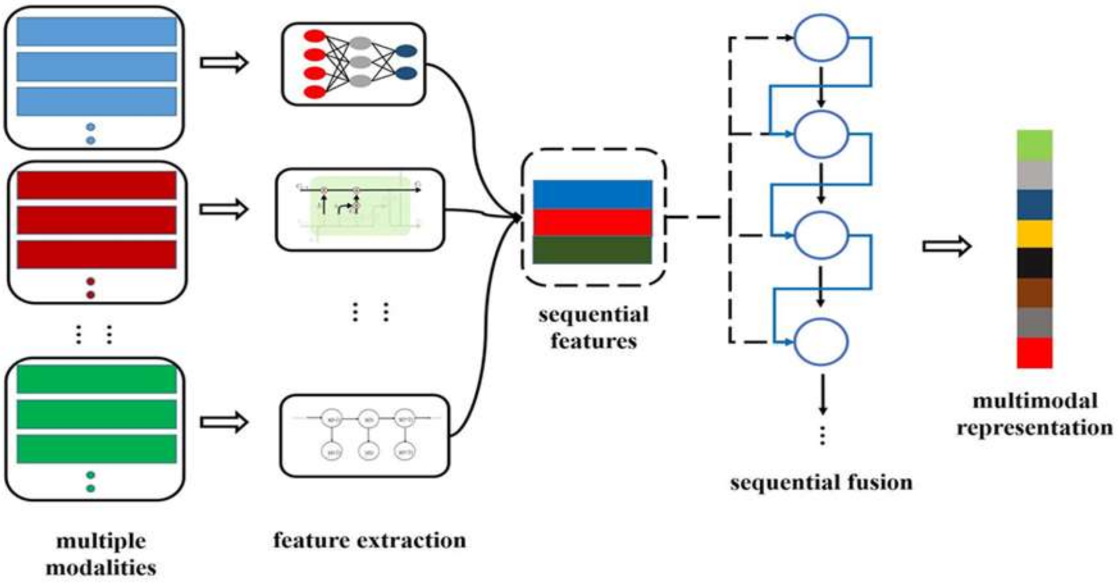

Fig.6 Sequential Fusion

This architecture gathers data over a time period continuously across multiple modalities. The raw data is passed through feature extractors, and the resulting sequential representations for each modality are sent together to a sequential fusion module. The fusion module remembers the time-series data, returns back a time-oriented and multi-modal representation, thus allowing for the perception to better represent the continuity of experience. [3]

Time-based sequential fusion is critically important for PHRI applications, where perception and decision making rely on a continuum and temporal evolution governed by sensor observations. It is most successfully integrated with modalities with natural temporal features (e.g. audio, center of pressure or haptic data, electroencephalogram (EEG) signals). The architecture has been deployed successfully for human tracking systems, detection tasks for tactile events, and action recognition tasks.

suggested a self-learning model for detecting human activity using a three-layer Long Short-Term Memory (LSTM) network to encode long motion sequences of a skeleton, which is of critical importance for dynamic human–robot interaction. Likewise, a deep neural network–hidden Markov model approach for speech recognition that recognized clean speech, acoustic responses, and background noise to ensure robustness and improve validity.

B. Multi-Modal

The modalities in a multi-modal system can be grouped into two categories, based on the type of raw data: homologous modalities (with similar types of data) and heterogeneous modalities (with different types of data). There are only two general fusion architectures to combine either one of the different categories of modalities.

1. The first one is a simply fusion method, where the modalities are directly fed into the fusion module with no dedicated preprocessing of any modality.

2. The second has recognized that each modality has different qualities as well as types of data. In this case, the first stage is to extract the features individually from each modality, and to respect the properties of the data type. These features could be brought together—in the case had more than one classifier— present as a thorough, multi-modal representation.

1) Swallowing Fusion (SWA)

Swallowing Fusion is a simple method where instead of preprocessing (feature extraction), we simply put multiple modalities directly into the fusion module (see Fig. 7)

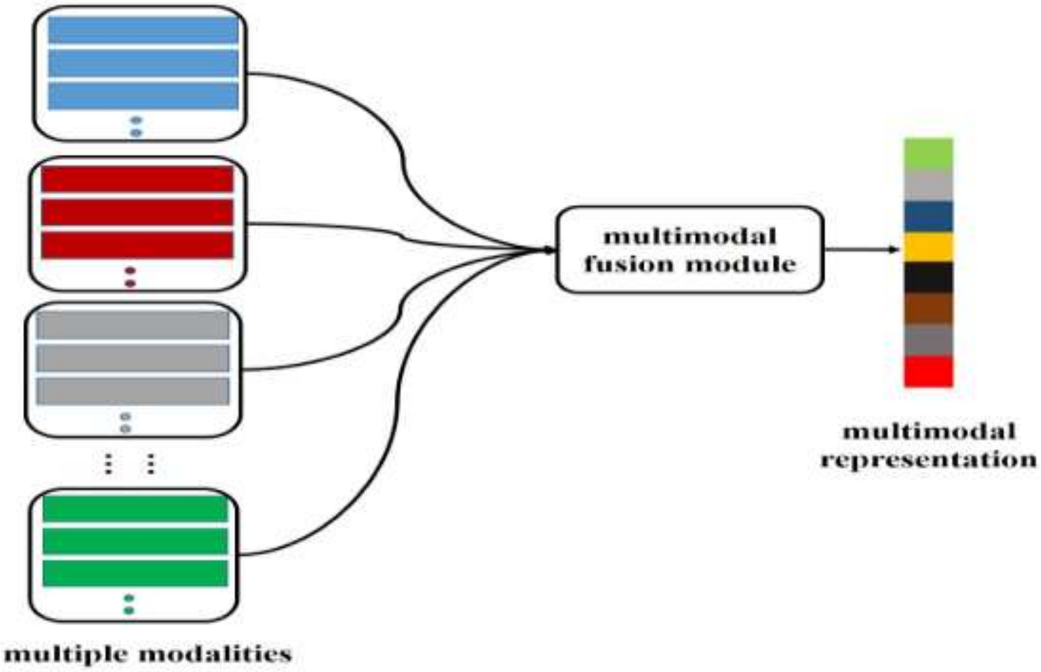

Fig. 7 Swallowing fusion (SWA)

In this architecture, homologous data is put simultaneously into the multi-modal fusion module, where it is used to provide a single representation. This approach is typically used to combine only a few modalities that have a similar data structure, for example, RGB, depth and skeleton data force and encoder outputs, and physiological data like ECG and SpO_2.

Recently, technologies like GelSight have made it possible to represent tactile information visually, so a combination of vision and haptic data could all be fed into the same convolutional neural network (CNN). For instance, reported improved grasping performance when both RGB images and GelSight tactile images were input into a deep CNN.

Swallowing Fusion architecture may be considered the simplest means to integrate multiple modalities, although it applies best in instances when the modalities have compatible or the same data representation.

2) Mixing Fusion (MIX)

Mixing Fusion uses independent feature extraction within each modality and parallel fusion processes. Produces multiple classifiers for fusion – only one classifier per modality or a combination of modalities. Each classifier's output is combined in a single representation (See Fig. 8).

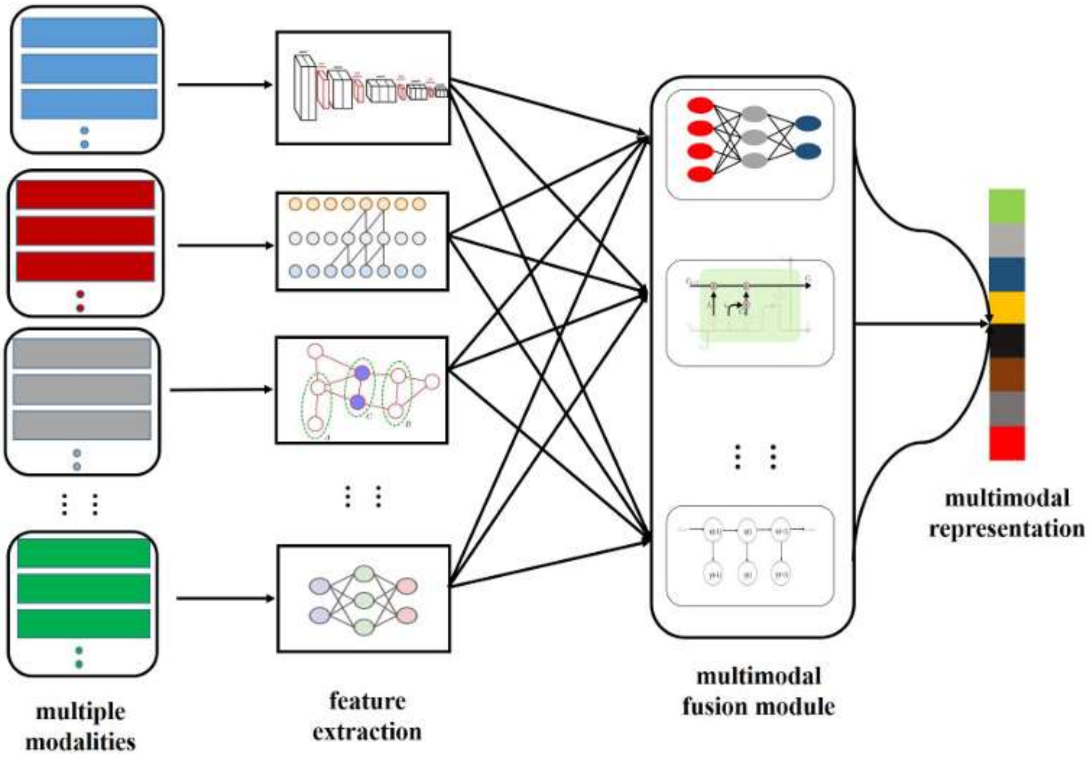

Fig. 8 Mixing fusion (MIX).

Heterogeneous data are pre-processed with individual feature extractors that confine and exploit the properties of each modality. The features of each modality are put into multi-modal fusion, and the final output is a representation. Mixing Fusion is useful for heterogeneous data types as heterogeneous data may differ within spatial, temporal, and structural meanwhile processing R and I-HRI in anthropomorphic humanoid robotic forms. Mixing Fusion is one of the most popular methods of fusion for incorporating diverse types of information and accommodating the resources of the various information. Especially in challenging human-robot interactions with complex tasks, and contact-rich actions where the data were collected at different time phases and spatial contexts.

Analysis

The first study highlights the innovative approach between people and robots by utilizing gestures and spoken language to control an industrial robot arm. The first study uses a unique gesture dataset and analyzes with Motion History Images from video plus a convolutional neural network (CNN) type of Artificial Intelligence. The authors achieved greater than 98% accuracy on recognizing gestures. The study also implements speech recognition using an open-source platform and noise noise-reduction techniques, achieving greater than 95% accuracy. Because the whole system is able to handle multiple inputs, the user may be able to interact with the robotic arms quickly and smoothly. The authors performed multiple experiments with a robotic arm performing pick-and-place tasks to assess the system's reliability and effectiveness. The most important contribution of this work is the combination of inputs to improve recognition and letting humans establish the intentions of their actions with the actions of the robot in the real-world context.[1]

The second study looked at how different types of social signals, such as gestures or sounds produced by a robotic arm, would influence people's perception or feeling towards interacting or collaborating. By using a small, robotic arm with various patterns of lights and sound to communicate actions, they demonstrated the head gesture plus lights and sounds led to participants perceiving the robot was friendlier and smarter.The results suggested that participants were easier able to understand the robot's actions and had relatively more positive perceptions of their interaction when multiple means of communication were offered instead of one. In addition, the studies were conducted online, and only applied in one task, they indicated how significant these multiple signals could be at making robots more engaging and also easier to interact with.[2],[21]

The next study provided a comprehensive description of the feasibility of using varied manners of weaving together signals for future human-robot interactions, where humans and robots were also physically working side by side. Authors separated types of Sensory signal inputs into indirect sensory inputs like EEG and cameras, and direct sensory inputs like tactile sensors. They also explored types of signal combinations on three distinct dimensions: timing; overlapping or simply mixing data; and how complex are the tasks. The authors indicated several limitations of their framework, including: timing can mismatch with sensory data, how to approach choosing which input to utilize, and also what can be accomplished with pre-trained models was not explored. Additionally, the authors noted possible directions of future work, such as improving tactile sensors, different ways to dynamically weave together multiple signals overtime, more flexibility concerning which input to select, and improvement in learning efficiency. Interestingly, the authors hope this review will provide insights in future directions for increasing robot's intelligence and interactivity.[3]

Conclusion

This research represents a significant contribution to the progress of collaborative human–robot interaction (HRC) systems through the development of a multimodal system incorporating gesture and speech recognition. The development and testing of a viable real-time interaction system capable of utilizing convolutional neural networks and motion history images to identify gestures, along with an improved and open-source speech recognition engine for noisy industrial environments, not only demonstrates the real-world applicability of multi-modal systems, it also reduces the margin for error. The capacity to combine human gestures, speech, and the function of the robot, as well as facilitate a prompt feedback loop depends on the integration of multi-threading functionality; thus, low-latency two-way interaction is achievable. In addition, the design of a multimodal system that enhances the effectiveness of the human-robot teaming process is a high priority. The three modalities were tested to provide evidence of the situational benefits associated with multimodality within HRC; analysis of the psychological and operational benefits associated with modality co-occurrence—benefits that affect user perception, trust, and engagement—also augment HRC systems. The discussion of multimodal fusion in detail contributes to the research on sociocultural signals in HRC at both transitory and sustained levels of structural development and application in the fields of robotics, interaction design, social cue investigation, and collaborative systems research. Although the work accomplished in this project presents obstacles (e.g.g, speech ambiguity associated with noise, computation burden associated with gesture and speech recognition), the integrated HRC approach further establishes the usability of robots making them more intuitive, more intelligent, and able to engage as peers in a two-way multimodal manner with humans during multimodal collaborative interaction.

References

[1] H. Chen, M. C. Leu, W. Tao, and Z. Yin, "Design of a Real-Time Human–Robot Collaboration System Using Dynamic Gestures," in ASME Int. Mech. Eng. Congr. Expo., Virtual Conf., Nov. 16–19, 2020

[2]H.-L. Cao, G. Van de Perre, J. Kennedy, et al., "A personalized and platform-independent behavior control system for social robots in therapy: Development and applications," IEEE Trans. Cogn. Dev. Syst., vol. 11, no. 3, pp. 334–346, 2018.

[3]Xue, T., Wang, W., Ma, J., Liu, W., Pan, Z., & Han, M. (2020). Progress and Prospects of Multimodal Fusion Methods in Physical Human–Robot Interaction: A review. IEEE Sensors Journal, 20(18), 10355–10370. https://doi.org/10.1109/jsen.2020.2995271

[4] J. F. Arinez, Q. Chang, R. X. Gao, C. Xu, and J. Zhang, "Artificial Intelligence in Advanced Manufacturing: Current Status and Future Outlook," ASME J. Manuf. Sci. Eng., vol. 142, no. 11, p. 110804, 2020.

[5] X. V. Wang and L. Wang, "A Literature Survey of the Robotic Technologies During the COVID-19 Pandemic," J. Manuf. Syst., vol. 60, pp. 823–836, 2021.

[6] K. Zinchenko, C.-Y. Wu, and K.-T. Song, "A Study on Speech Recognition Control for a Surgical Robot," IEEE Trans. Ind. Inf., vol. 13, no. 2, pp. 607–615, 2016.

[7] M. C. Bingol and O. Aydogmus, "Performing Predefined Tasks Using the Human–Robot Interaction on Speech Recognition for an Industrial Robot," Eng. Appl. Artif. Intell., vol. 95, p. 103903, 2020.

[8] M. Kuhn, K. Pollmann, and J. Papadopoulos, "I'm Your Partner-I'm Your Boss: Framing Human–Robot Collaboration With Conceptual Metaphors," in Proc. ACM/IEEE Int. Conf. Human–Robot Interaction, Virtual Conf., Mar. 24–26, 2020, pp. 322–324.

[9] E. Coupeté, F. Moutarde, and S. Manitsaris, "A User-Adaptive Gesture Recognition System Applied to Human–Robot Collaboration in Factories," in Proc. 3rd Int. Symp. Movement and Computing, Thessaloniki, Greece, Jul. 5–6, 2016, pp. 1–7.

[10] V. V. Unhelkar et al., "Human-Aware Robotic Assistant for Collaborative Assembly: Integrating Human Motion Prediction With Planning in Time," IEEE Rob. Autom. Lett., vol. 3, no. 3, pp. 2394–2401, 2018.

[11] R. F. Pinto, C. D. Borges, A. Almeida, and I. C. Paula, "Static Hand Gesture Recognition Based on Convolutional Neural Networks," J. Electr. Comput. Eng., vol. 2019, 2019.

[12] J. Li, X. Liu, M. Zhang, and D. Wang, "Spatio-Temporal Deformable 3D Convnets With Attention for Action Recognition," Pattern Recognit., vol. 98, p. 107037, 2020.

[13] W. Tao, Z.-H. Lai, M. C. Leu, and Z. Yin, "Worker Activity Recognition in Smart Manufacturing Using IMU and SEMG Signals With Convolutional Neural Networks," Procedia Manuf., vol. 26, pp. 1159–1166, 2018.

[14] B. Treussart, F. Geffard, N. Vignais, and F. Marin, "Controlling an Upper-Limb Exoskeleton by EMG Signal While Carrying Unknown Load," in IEEE Int. Conf. Robot. Autom. (ICRA), Virtual Conf., May 31–Aug. 31, 2020, pp. 9107–9113.

[15] A. Ajoudani et al., "Progress and Prospects of the Human–Robot Collaboration," Auton. Rob., vol. 42, no. 5, pp. 957–975, 2018.

[16] D. Yongda, L. Fang, and X. Huang, "Research on Multimodal Human–Robot Interaction Based on Speech and Gesture," Comput. Electr. Eng., vol. 72, pp. 443–454, 2018.

[17] Lin, K., Li, Y., Sun, J., Zhou, D., and Zhang, Q., 2020, “Multi-sensor Fusion for

Body Sensor Network in Medical Human–Robot Interaction Scenario,” Inf. Fusion, 57, pp. 15–26.

[18] A. Banh, D. J. Rea, J. E. Young, et al., "Inspector Baxter: The social aspects of integrating a robot as a quality inspector in an assembly line," in Proc. 3rd Int. Conf. Human-Agent Interaction, 2015, pp. 19–26.

[19] K. Baraka and M. M. Veloso, "Mobile service robot state revealing through expressive lights: Formalism, design, and evaluation," Int. J. Social Robotics, vol. 10, no. 1, pp. 65–92, 2018.

[20] C. Bartneck, D. Kulić, E. Croft, et al., "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," Int. J. Social Robotics, vol. 1, no. 1, pp. 71–81, 2009.

[21] G. Bolano, A. Roennau, and R. Dillmann, "Transparent robot behavior by adding intuitive visual and acoustic feedback to motion replanning," in Proc. IEEE Int. Symp. Robot and Human Interactive Communication (RO-MAN), 2018, pp. 1075–1080.

[23] A. J. Brammer and C. Laroche, "Noise and communication: A three-year update," Noise and Health, vol. 14, no. 61, pp. 281, 2012.

[24] British Standards Institute, BS EN 60073:2002: Basic and safety principles for man-machine interface, marking and identification, 2002.

[25] E. Cha, M. Matarić, and T. Fong, "Nonverbal signaling for non-humanoid robots during human-robot collaboration," in Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), 2016, pp. 601–602.

[28] J. E. Colgate, W. Wannasuphoprasit, and M. A. Peshkin, "Cobots: Robots for collaboration with human operators," in Proc. ASME Int. Mech. Eng. Congr. Expo., 1996, pp. 433–439.

[29] I El Makrini, S. A. Elprama, J. Van den Bergh, et al., "Working with Walt: How a cobot was developed and inserted on an auto assembly line," IEEE Robot. Autom. Mag., vol. 25, no. 2, pp. 51–58, 2018.

[30] S. Elprama, I. El Makrini, B. Vanderborght, et al., "Acceptance of collaborative robots by factory workers: A pilot study on the importance of social cues of anthropomorphic robots," in Int. Symp. Robot and Human Interactive Communication, 2016, pp. 919–924.

[31] S. Embgen, M. Luber, C. Becker-Asano, et al., "Robot-specific social cues in emotional body language," in Proc. IEEE RO-MAN, 2012, pp. 1019–1025.

[32] T. Ende, S. Haddadin, S. Parusel, et al., "A human-centered approach to robot gesture-based communication within collaborative working processes," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2011, pp. 3367–3374.

[33] H. Erel, D. Trayman, C. Levy, et al., "Enhancing emotional support: The effect of a robotic object on human–human support quality," Int. J. Social Robotics, pp. 1–20, 2021.

[34] T. Faibish, A. Kshirsagar, G. Hoffman, et al., "Human preferences for robot eye gaze in human-to-robot handovers," Int. J. Social Robotics, pp. 1–18, 2022.

[35] F. Faul, E. Erdfelder, A.-G. Lang, et al., "G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," Behavior Res. Methods, vol. 39, no. 2, pp. 175–191, 2007.

[36] K. Fischer, "Why collaborative robots must be social (and even emotional) actors," Techné: Res. Philos. Technol., vol. 23, no. 3, pp. 270–289, 2019.

[37] K. Fischer, L. C. Jensen, F. Kirstein, et al., "The effects of social gaze in human-robot collaborative assembly," in Proc. Int. Conf. Social Robotics, 2015, pp. 204–213.

[38] B. Gleeson, K. MacLean, A. Haddadi, et al., "Gestures for industry: Intuitive human-robot communication from human observation," in Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), 2013, pp. 349–356.

[39] S. Grushko, A. Vysockỳ, D. Heczko, et al., "Intuitive spatial tactile feedback for better awareness about robot trajectory during human–robot collaboration," Sensors, vol. 21, no. 17, p. 5748, 2021.

[40] M. Heerink, B. Kröse, V. Evers, et al., "Assessing acceptance of assistive social agent technology by older adults: The Almere model," Int. J. Social Robotics, vol. 2, no. 4, pp. 361–375, 2010.

[41] S. K. Kundu, S. Kumagai, and M. Sasaki, "A wearable capacitive sensor for monitoring human respiratory rate," Jpn. J. Appl. Phys., vol. 52, p. 04CL05, 2013.

[42] N. Lazzeri, D. Mazzei, A. Zaraki, et al., "Towards a believable social robot," in Conf. Biomimetic and Biohybrid Systems, Springer, 2013, pp. 393–395.

[43] L. Onnasch, X. Maier, and T. Jürgensohn, "Mensch-Roboter-Interaktion-Eine Taxonomie für alle Anwendungsfälle," Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Dortmund, 2016.

[44] H. Park, D. Park, and J. Lee, "How important alarm types for situation awareness at the smart factory?," in Int. Conf. Human-Computer Interaction, Springer, 2019, pp. 113–118.

[45] T. Ribeiro and A. Paiva, "The illusion of robotic life: Principles and practices of animation for robots," in Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), 2012, pp. 383–390.

[46] J. Saldien, B. Vanderborght, K. Goris, et al., "A motion system for social and animated robots," Int. J. Adv. Robot. Syst., vol. 11, no. 5, p. 72, 2014.

[47] V. Sauer, A. Sauer, and A. Mertens, "Zoomorphic gestures for communicating cobot states," IEEE Robot. Autom. Lett., vol. 6, no. 2, pp. 2179–2185, 2021.

[48] A. Sauppé and B. Mutlu, "How social cues shape task coordination and communication," in Proc. 17th ACM Conf. Computer Supported Cooperative Work & Social Computing, 2014, pp. 97–108.

[49] A. Sauppé and B. Mutlu, "The social impact of a robot co-worker in industrial settings," in Proc. 33rd Annu. ACM Conf. Human Factors in Computing Systems, 2015, pp. 3613–3622.

[50] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," Trends Cogn. Sci., vol. 10, no. 2, pp. 70–76, 2006.

[51] S. Sheikholeslami, A. Moon, and E. A. Croft, "Cooperative gestures for industry: Exploring the efficacy of robot hand configurations in expression of instructional gestures for human–robot interaction," Int. J. Robot. Res., vol. 36, no. 5-7, pp. 699–720, 2017.

[52] S. Song and S. Yamada, "Designing expressive lights and in-situ motions for robots to express emotions," in Proc. 6th Int. Conf. Human-Agent Interaction, 2018, pp. 222–228.

[53] G. Tang, P. Webb, and J. Thrower, "The development and evaluation of robot light skin: A novel robot signalling system to improve communication in industrial human–robot collaboration," Robot. Comput.-Integr. Manuf., vol. 56, pp. 85–94, 2019.

[54] K. Tatarian, R. Stower, D. Rudaz, et al., "How does modality matter? Investigating the synthesis and effects of multi-modal robot behavior on social intelligence," Int. J. Social Robotics, pp. 1–19, 2021.

[55] Y. Terzioğlu, B. Mutlu, and E. Şahin, "Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration," in Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), 2020, pp. 343–357.