IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 6, 2018, accepted June 24, 2018, date of publication June 29, 2018, date of current version July 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2851841

A Brain-Inspired Multi-Modal Perceptual System for Social Robots: An Experimental Realization

MOHAMMAD K. AL-QADERI AND AHMAD B. RAD, (Senior Member, IEEE)
Autonomous and Intelligent Systems Laboratory, School of Mechatronic Systems Engineering, Simon Fraser University,
Surrey Campus, Surrey, BC V3T 0A3, Canada

Corresponding author: Ahmad B. Rad (arad@sfu.ca)

ABSTRACT We propose a multi-modal perceptual system that is inspired by the inner working of the human brain; in particular, the hierarchical structure of the sensory cortex and the spatial-temporal binding criteria. The system is context independent and can be applied to many on-going problems in social robotics, including but not limited to person recognition, emotion recognition, and multi-modal robot doctor to name a few. The system encapsulates the parallel distributed processing of real-world stimuli through different sensor modalities and encoding them into features vectors which in turn are processed via a number of dedicated processing units (DPUs) through hierarchical paths. DPUs are algorithmic realizations of the cell assemblies in neuroscience. A plausible and realistic perceptual system is presented via the integration of the outputs from these units by spiking neural networks. We will also discuss other components of the system including top–down influences and the integration of information through temporal binding with fading memory and suggest two alternatives to realize these criteria. Finally, we will demonstrate the implementation of this architecture on a hardware platform as a social robot and report experimental studies on the system.

INDEX TERMS Human-robot interaction, machine perception, multi-modal systems, social robots, spiking neural networks, top-down influences.

I. INTRODUCTION

Among the species, humans have this remarkable capacity to make sound and relatively quick decisions in diverse situations based on incomplete and at times vague information. Understanding the complicated dynamics of human decision making process has been linked to an underlying sophisticated perceptual system. The new findings in neuroscience and psychology have led scientists to form hypotheses and to conjecture models of the human perceptual process. Social robotics research as a fundamentally interdisciplinary endeavor, can significantly benefit from these discoveries. Social robotics researchers are inspired by such models and whereas they are aware that they cannot duplicate the entire process, they attempt to emulate such biological models closely and formulate computationally viable replicas through a reverse engineering process. As social robots are expected to function predominantly in human environments and social settings; it is most likely and desirable that these systems are patterned after humans and respond appropriately to emotional and social cues in social settings. The seamless integration of social robots into human social settings and their ultimate acceptance by humans will largely depend on how natural (in the human sense) such robots behave. Research by Niculescu *et al.*, in the context of human-robot interaction, suggests that people consider the level of closeness to human-like response from a social robot as a highly favorable attribute. The authors argued that in their interactions with a social robot, humans regard fast reaction time (within the range of their own) more important than the accuracy and the correctness of the response itself [1]. In related studies [2], [3], the importance of reading and expressing social cues, and real-time response at human interaction rate are noted to be important considerations that must be addressed by social robotic systems. These machines must be capable of sensing, perceiving, and interpreting the real-world environment similar or as close as possible to that of humans. Therefore, mimicking the way humans process stimuli, and synthesizing similar interpretations of those stimuli as of humans should be among the salient features of social robots.

As social robots are employed to assist [4], collaborate [5], provide physical, mental, and/or social support to humans [6], the need for close spatial and temporal association between the two in all aspects of their interactions is also important and is referred to as proximate interaction [7]. Similarly, the mixed-initiative interaction which takes the form of dialogue interaction is highlighted and regarded as relevant research in the area of social robotics. In this category of human-robot interaction, the role of the robot is shifted from being an operator of something to imitate and/or act, collaborate, or assist humans. Thus, an essential element that leverages the performance of the social robots in achieving mixed-initiative interaction is the availability of an effective perceptual system capable of perceiving the state of the real-world environment and having a capacity to grasp the social and emotional cues (i.e. providing social situation assessment) in a rather natural (in the human sense) and fast -close to average human reaction time.

Within the above context and along with many researchers, we advocate that not only social robots are desired to be equipped with an analogous sensory system as humans; they also ought to mimic a closely related architecture to human's perception process [8], [9]. In this manuscript, the authors propose and evaluate a context-independent robotic perceptual architecture on the premise of multi-modal stimuli. The design is essentially a simplified algorithmic interpretation of the human perceptual process.

The motivation of this study is to address the less researched problem of machine perception which is a central pursuit in artificial intelligence. The terms machine learning and machine perception are used interchangeably in literature. However, there are subtle differences between the two. The former addresses general learning form signals whereas the latter is concerned with perception and understanding of the machine from the world around it. This process is essentially an analogous process to the human perception. Inspired by the current understanding of the architecture of human perception, we propose a machine perception system that is multi-modal, and follows a close hierarchy as the biological counterpart. We borrow the notions of binding criteria [10], fading memory [11], top-down influences [12], cell assemblies [13], and convergence zone [14] from neuroscience and experimental psychology and suggest respective computational synthesis.

Perusing the literature in social robotics; one may note that they are mostly concerned with affective computing [9]. In a parallel note, reported studies in machine perception deal with unimodal sensory systems [8]. The rationale for this paper is to bridge the gap and design a perceptual system for social robots. The distinctive characteristics of the proposed architecture are: (a) it is inspired by the human sensory cortex and has a broad one-to-one analogy with its architecture, (b) the system responds within the range of the human reaction time, (c) it is multi-modal yet it does not require concurrent presence of all modalities in order to complete a perceptual task, (d) it is generally context independent implying that it can be configured to address different perceptual problems such as person recognition, object, and emotion recognition (affective computing), etc. A customized version of the basic architecture of the perceptual system was configured for the person recognition problem and its performance was studied via simulation studies [15]. In this paper, we present a thorough exposition of the whole system as well as its components and suggest two alternative temporal-based integration methods and demonstrate the effect of incorporating top-down influences on the performance of the system in real-time implementation. We also report the design of an in-house multi-modal social robot to test and demonstrate the properties of the perceptual system.

The rest of this paper is organized as follows. We will review the relevant literature in section II. In section III, we will then present an overview of the system, its relation to the primate sensory cortex and human perception process. We will also examine all the subsystems in details and integrate them into an elegant perceptual system. In section IV, the realization and evaluation of the system in real-world application is presented. Finally, the paper is concluded in section V with concluding remarks and potential refinements.

II. RELATED STUDIES

Perusal through the relevant and vast literature, it is interesting to note that most available methodologies do not take into consideration the fact that humans interact among themselves and with the environment through an efficient processing of available information from multisensory modalities and the integration of this information over a finite and short time [16]. For example, in the person recognition problem, humans use various modalities and different aspects of measures to facilitate their perception and response to stimuli within 150 to 200ms. [17]. Face recognition, person identification, and re-identification, and gesture recognition are generally classified as the most challenging tasks by a machine yet essential for human-robot interaction. However, to humans, these and similar tasks appear mundane, effortless, and fast. This motivates and inspires researchers to develop brain-inspired systems to achieve improved performance in similar perceptual tasks. Among such studies are convolutional deep neural networks, Hierarchical Max-pooling model (HMAX), and Stable Model. Convolutional deep neural networks have demonstrated impressive results for object recognition using the ImageNet Large Scale Recognition Challenge dataset [18]. HMAX model [19], which is an invariant object recognition system and inspired from primate visual cortex, shows relatively high performance on invariant single object recognition, multi-class categorization, and complex scene understanding tasks. Stable Model was proposed by Rajaei *et al.* [20] and can be considered as an extension of HMAX model. It uses the adaptive resonance theory (ART) mechanism for extracting intermediate visual descriptors that makes the model stable against forgetting previously learned patterns. It shares C2-like features with HMAX model, but in the training phase only highest active C2 units are selected as prototypes for the input image by mean of feedback projection. The aforementioned systems borrow their hierarchical architecture from primate visual cortex. It is worthwhile noting that these and many other reported architectures fall within the so-called mono modality systems; i.e. primarily vision modality. Visual perception is considered a crucial element in human perceptual system and plays a vital role in achieving many tasks such as person recognition, facial expression recognition, and object recognition to name a few [21]–[23]. However, other sensor modalities may provide complementary information or essential information (as in case of absence of visual modality) that enhance the outcome of the perception process.

Significant research effort has been devoted to understand processing of multisensory information in the sensory cortex. Neuroscientists, psychologists, and psychophysicists suggest models to explain how humans perceive the environment by integration of information from different sensory modalities. The notion of ''convergence-zone'' proposed by Damasio [14] is one such plausible model that attempts to explain the mechanism of multi-modal human perception process. In this model, the convergence-zone ensembles which are presumed to be located in the higher-order integrative cortical areas play a major role in integrating multiple aspects of external reality, perceiving, and recalling experiences. Damasio contends that the human perceptual process encodes all the attributes that belong to a certain object and available to sensor modalities as a set of feature vectors. These features vectors are then fed to various dedicated neural circuits for further processing. Then, the synergistic integration of the outputs of these dedicated neural networks at different levels using different binding criteria produces the final output of the perception process. The areas where the binding process is performed are referred to as ''convergence-zone'' ensembles. The central observation of Damasio's model is that the real-world entities are perceived by a synergistic process that uses information extracted from various sensor modalities and integrate this information through a hierarchical structure with feedback projection. Each sensory modality provides one or more of the entity's features and the hierarchical integration of this sensory information over time is the key factor that makes humans superior in perceiving and distinguishing real-world objects in diverse settings and scenarios. The question raised by this model is how the convergence-zone ensembles bind the outputs of these neural networks to create a perception or experience about a certain object or event respectively.

The feature integration theory (FIT) that was proposed by Terisman and Gelade suggests a solution for the binding problem [24]. FIT provides a mechanism for binding by location and shared features. In this theory, the spatial attention window provides access to the features in the associated locations in various feature maps and facilitates the integration of information from these locations to a single object file for further analysis and identification. One of the key principles of FIT is that various features of an object are ''registered early, automatically, and in parallel across the visual field'' whereas the object as a whole is identified separately and at later stage in the perception process [24]. One could argue that fast perception of humans in the presence of so many distracting stimuli is a process of selective attention that is explained by FIT.

Another prominent solution of binding problem was formulated by Milner, Grossberg, and von der Malsburg in [25]–[27] respectively. However, Gray and Singer were the first who experimentally demonstrated the role of synchrony in the binding process [28]. They coined the term ''Binding-by-synchrony'' and suggested that binding problem could be solved by temporal synchrony of population of neurons. The neurons that encode the same object are distinguished from other neurons by synchronous firing. In other words, the matching frequency firing of population of neurons indicates that these neurons encode the same object, while other frequency firing highlights other objects.

In contrast to classical theories of sensory processing that view brain as a stimulus-driven mechanism, the current findings consider the human perception process as an active as well as a proactive process [29], [30]. Within this framework, the processing of stimuli is controlled by the top-down influences that shape the outcome of the sensory processing by creating predication about the forthcoming sensory events and providing shortlisted candidates that might fit the pattern represented by the attended stimulus. The effect of the top-down influences is to alter or multiplex the function of neurons at the receptive field system according to the object or the task that is attended to [12]. In other words, the expectations, which are given as pre-knowledge or induced by the outcome of processing the feature vectors in the fastest processing routes, generate top-down influences by instructing the receptive field system to process only a set of appropriate feature vectors.

III. OVERVIEW OF THE MULTI-MODAL PERCEPTUAL SYSTEM

In this section, we present a thorough exposition to the proposed architecture and explain in details the key subsystems and the interrelationship among them and discuss the contribution of each module in the multi-modal perception process. In order to show the similarities with the biological system, the presentation in this section switches between the biological system and how it is ''replicated'' in this architecture. The system is essentially built around the current understanding of the architecture of the sensory cortex as shown in the left hand side of Fig. 1 (hereafter referred to Fig. 1−L). As shown, an attended stimulus is mapped by independent unimodal sensory processing routes (i.e., V1 and V2 in visual, A1 and A2 in auditory, and S1 and S2 in somatosensory) to a set of feature vectors that represent the stimulus's attributes. The sensory information undergoes modality-specific processing before it converges to form a perception about the stimulus [31], [32]. Recent research studies in neuroscience and psychophysics suggest that cross-modal interactions may happen early in perception process

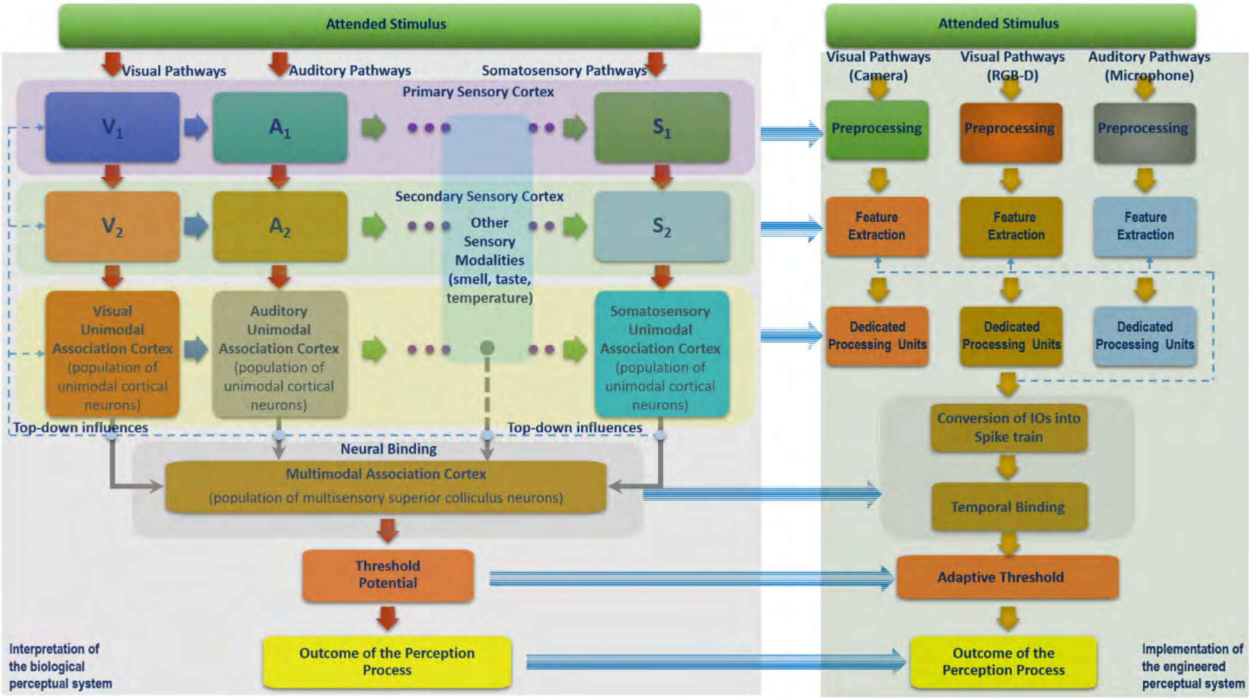

Attended Stimulus
Visual Pathways
Auditory Pathways
Somatosensory Pathways
Primary Sensory Cortex
V_1
A_1
S_1
Secondary Sensory Cortex
V_2
A_2
S_2
Other Sensory Modalities (smell, taste, temperature)
Visual Unimodal Association Cortex (population of unimodal cortical neurons)
Auditory Unimodal Association Cortex (population of unimodal cortical neurons)
Somatosensory Unimodal Association Cortex (population of unimodal cortical neurons)
Top-down influences
Top-down influences
Neural Binding
Multimodal Association Cortex (population of multisensory superior colliculus neurons)
Threshold Potential
Outcome of the Perception Process
Interpretation of the biological perceptual system
Attended Stimulus
Visual Pathways (Camera)
Visual Pathways (RGB-D)
Auditory Pathways (Microphone)
Preprocessing
Preprocessing
Preprocessing
Feature Extraction
Feature Extraction
Feature Extraction
Dedicated Processing Units
Dedicated Processing Units
Dedicated Processing Units
Conversion of IOs into Spike train
Temporal Binding
Adaptive Threshold
Outcome of the Perception Process
Implementation of the engineered perceptual system

FIGURE 1. An overview of biological perception system (left hand side) and a possible realization (right-hand side).

to provide new and essential information, that is unobtainable by unimodal sensory input in isolation [33], [34]. Each stimulus's attribute is mapped by population of neurons (cell assemblies) that have similar receptive field properties and each cell assembly maps one feature for an attended stimulus. Each sensor modality has its own receptive field system and generates neural responses representing its domain-specific characteristics. These neural responses are further processed by population of unimodal cortical neurons that respond selectively to various attributes within same sensor modality.

In an analogous fashion to the biological system, the proposed perceptual system attends to external stimuli through its multi-modal sensory system as depicted in Fig. 1−R (right hand side). As shown in Fig. 1−R and Fig. 2, each sensor modality (vision, voice, etc.) undergoes its specific preprocessing and feature extraction modules to generate a set of feature vectors representing its domain-specific characteristics. Each feature vector is then processed by its respective Dedicated Processing Unit (DPU) which in turn contributes to production of the intermediate outputs (IOs). The DPUs are the ''engineered replica'' of cell assemblies that are formed by unimodal cortical neurons at unimodal association cortex.

There is consensus among neuroscientists that superior colliculus neurons have a characteristic feature of integrating multi-modal sensory information. Even though there is no established computational model for these neurons, however there are reasonable models that bridge the gap between the physiological and psychophysics interpretations of multisensory integration information. We have adopted temporal binding mechanisms, which use spiking neurons as their core computational elements in the proposed perceptual architecture as criteria of how multi-modal sensory information are integrated at higher level. Therefore, the outputs of DPUs (intermediate outputs) will be transformed into temporal spikes to be processed by the temporal binding system that accommodates for one type of top-down influences (type 2). Top-down influences are depicted as feedback projection signals (type 1) and lateral connections (type 2) as shown in Fig. 2. We will discuss the realization of top-down influences in details in the next section and demonstrate its impact on the perceptual system in section IV. Once the intermediate outputs are transformed into temporal spikes, they will be introduced to temporal binding system. The generation of temporal spikes is central in the operation of the algorithm. We will describe the individual modules in the rest of this section.

A. GENERATION OF FEATURE VECTORS

Humans perceive the world via their senses each of which is heavily specialized through its distinctively hardwired circuitry to process sound and sight, etc. The perceptual system is also equipped with distinct sensors (camera, RGB-D sensors, microphones, etc.) to sense objects, people, or places. The pattern recognition literature is abundant of techniques and methods to process the raw unimodal sensory inputs and there are established methodologies for generation of feature vectors (feature extraction modules for vision and voice). The use of each methodology, is of course, application dependent. The key point is that each modality undergoes some preprocessing followed by specialized transformation to extract its

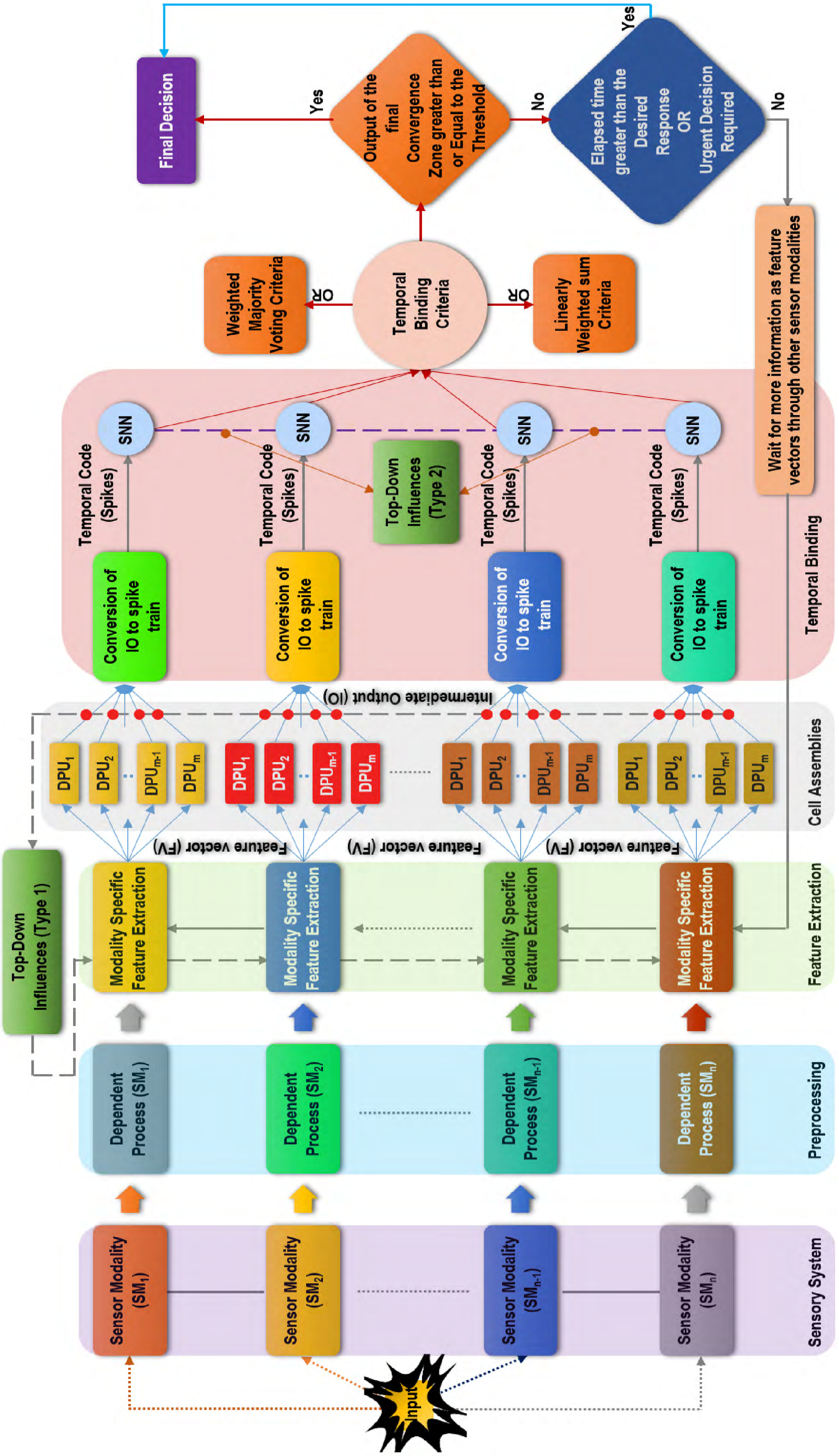

Input
Sensory System
Sensor Modality (SM$_1$)
Sensor Modality (SM$_2$)
Sensor Modality (SM$_{n-1}$)
Sensor Modality (SM$_n$)
Preprocessing
Dependent Process (SM$_1$)
Dependent Process (SM$_2$)
Dependent Process (SM$_{n-1}$)
Dependent Process (SM$_n$)
Top-Down Influences (Type 1)
Feature Extraction
Modality Specific Feature Extraction
Modality Specific Feature Extraction
Modality Specific Feature Extraction
Modality Specific Feature Extraction
Feature vector (FV)
Feature vector (FV)
Feature vector (FV)
Cell Assemblies
DPU$_1$
DPU$_2$
DPU$_{m-1}$
DPU$_m$
Intermediate Output (IO)
Temporal Binding
Conversion of IO to spike train
Conversion of IO to spike train
Conversion of IO to spike train
Conversion of IO to spike train
Temporal Code (Spikes)
Temporal Code (Spikes)
Temporal Code (Spikes)
Temporal Code (Spikes)
SNN
SNN
SNN
SNN
Top-Down Influences (Type 2)
Wait for more information as feature vectors through other sensor modalities
Temporal Binding Criteria
OR
OR
Weighted Majority Voting Criteria
Linearly Weighted sum Criteria
Output of the final Convergence Zone greater than or Equal to the Threshold
Yes
No
Final Decision
Elapsed time greater than the Desired Response OR Urgent Decision Required
Yes
No

FIGURE 2. The architecture of the proposed human-oriented perceptual system for social robots.

key features attributed to that specific modality (please refer to section IV for more information). Fig. 1−R and Fig. 2 show these modules respectively. The output from each feature extraction module is fed to its respective dedicated processing unit (DPU) which is analogous to unimodal association cortex whereby neurons in each cell assembly respond to a dedicated attribute of the attended stimulus. These feature vectors are generated by extracting meaning from the streams of data that are provided by different sensor modalities.

Since each feature vector is processed by its respective DPU, the error due to model bias is alleviated. In addition, using a set of feature vectors that represent the real-world stimulus from different and independent aspects of measures is expected to reduce the error variance. In other words, the built-in or hardwired property of the perceptual system, which is manifested by connecting each sensory modality to its respective DPUs, accommodates for the reduction in both types of error. As each DPU employs its own learning algorithm that infers specific information from the feature vectors, the error bias is reduced. On the other hand, integrating the outputs of these DPUs in a hierarchical manner using either spatial, temporal, or hybrid spatial-temporal binding mechanisms reduces the error variance.

B. DEDICATED PROCESSING UNITS (DPUs)

The information embedded in the data stream in each modality is decoded into its associated feature vectors which in turn are fed to its corresponding processing unit. DPUs can be considered as computational analogies of the cell assemblies in the biological system (Fig. 1−L). Machine learning and artificial intelligence literature offer many computational models for mimicking these cell assemblies. Multilayer feedforward neural networks, self-organized map, probabilistic neural networks, Gaussian mixture model, Hopfield network, Boltzmann machine, adaptive resonance network, support vector machine, linear and non-linear classifiers are possible computational model candidates for DPUs. Selection of an appropriate computational model is application and task dependent. Along the same argument, other classifiers that use various similarity and distance measures provide functional plausible models as well. For example, Mahalanobis distance considers the variance of each variable and covariance between variables in a feature vector. In other words, it provides a similarity measure that takes into account the scale of the data. The choice of a particular computational model is problem dependent.

The output of the DPUs are referred to as intermediate outputs (IO) (Fig. 2). The time required to generate the intermediate outputs depends on the type of the sensor modalities and its specific processing route. This implies that some of the processing units are faster and produce quicker intermediate outputs whereas others require longer time to generate their corresponding outputs. The variation of the processing time, that is required to produce the intermediate outputs, initiates the feedback projection from higher layer to lower layer in the hierarchical architecture which is essentially the same process referred as top-down influences in cognitive psychology (Fig. 2).

Another important contributing factor to the generation of the intermediate outputs is the presence or absence of a sensory modality. Consider the problem of person recognition; humans use distinctive cues to recognize a person in a short time and with high reliability. If the subject's face features are not clear due to occlusion, illumination, or distance; humans may use other cues such as body features, voice, or gait to identify the person [34]. The proposed framework incorporates this feature by including a feedback connection that compares the reliability of the perception outcome through an adaptive threshold. The perception process terminates if the reliability of the outcome is equal or greater than a preset threshold; otherwise it waits for more information from other sensor modalities to refine the perception process.

C. STIMULUS REPRESENTATION AND THE BINDING MECHANISMS

One of the main characteristics of the proposed perceptual system is that the classification of the real-world stimuli is performed by extracting a limited set of feature vectors from each modality in a parallel manner. Using a limited set of feature vectors to represent the real-world stimuli is inspired by the fact that human's channel capacity of processing information is limited as suggested in Miller's seminal paper [35]. These feature vectors, which represent various attributes of a real-world stimuli, are processed through separate processing channels (pathways) by DPUs (cell assembly).

According to [36], humans use two different routes for face recognition; configural and featural processing paths. The configural route is ''fast'' and facilitated by low spatial frequency and characterized as holistic, whereas the featural route is ''slow'' and uses the information available in high spatial frequency and characterized as accurate. In a parallel fashion, the system (section IV), employs five feature vectors to represent the attended subject. One feature vector is extracted from voice modality and the rest are extracted from visual modality. Vison-based feature vectors are purposely selected to represent both configural and featural information, respectively.

Humans use both routes through an integrated and dynamic manner by exploiting the information that is available first from the configural route to guide and direct the features route in order to refine the information and finalize the perception process. Though, these pathways operate in different time scales. This property is realized in the proposed architecture by means of top-down influences implemented as feedback signal and lateral connections. Feedback signal can be used for two functions: (1) they initiate the processing of appropriate feature vectors that are suitable for the task at hand. This function is depicted in Fig. 3 as a selection of specific feature vector, which is highlighted in red color, to be processed by its associated DPU; (2) they limit the scope of search for the best candidates; that is once the output from the fastest pathway

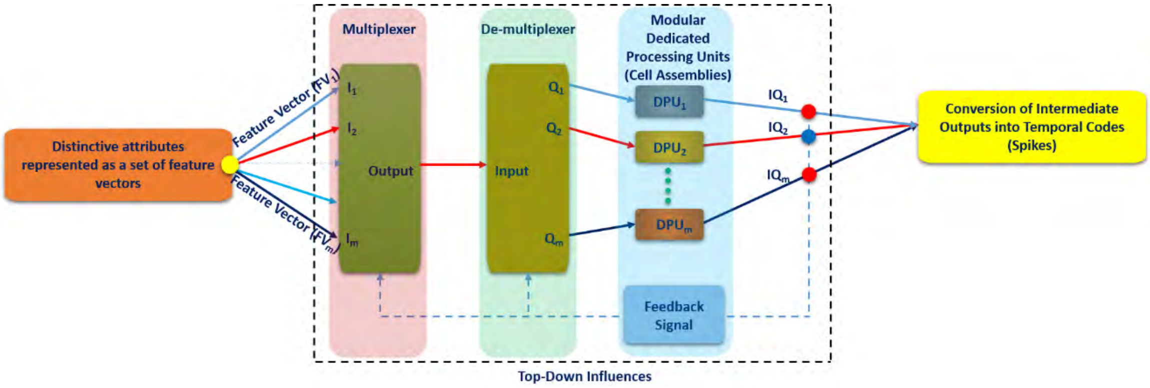

Distinctive attributes represented as a set of feature vectors
Feature Vector (FV1)
Feature Vector (FVm)
Multiplexer
I1
I2
Output
Im
De-multiplexer
Input
Q1
Q2
Qm
Modular Dedicated Processing Units (Cell Assemblies)
DPU1
DPU2
DPUm
IQ1
IQ2
IQm
Feedback Signal
Conversion of Intermediate Outputs into Temporal Codes (Spikes)
Top-Down Influences

FIGURE 3. Dedicated processing units and top-down influence (type 1).

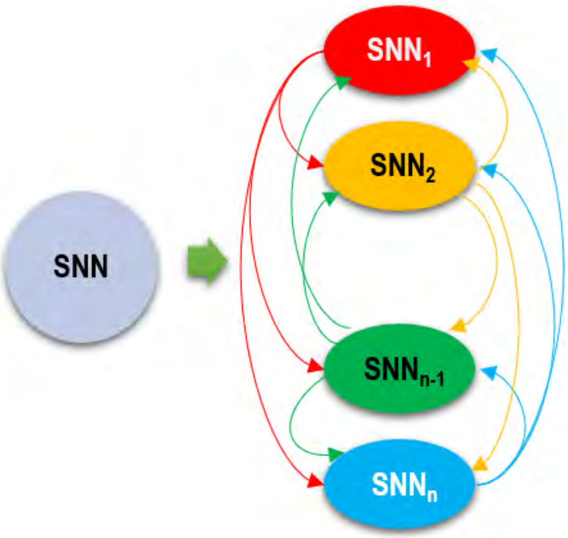


FIGURE 4. Lateral connection between spiking neurons (SN) in spiking neural networks (SNN).

is available, it will be used to direct the outputs of other routes by providing shortlisted candidate for the attended stimulus.

The spatial information is processed by the specialized DPUs and is transformed into temporal spikes that will be fed to population of spiking neurons in order to bias top-ranked spiking neurons to be the shortlisted candidates for the attended stimulus. The function of lateral connection is to force the competitive behavior among spiking neurons (SN) in integrating the intermediate outputs of DPUs. As shown in Fig. 4, each spiking neuron is connected laterally to all other spiking neurons that receive input from the DPUs participating in the mapping of the attended stimulus. Once the membrane potential of one of these spiking neurons reaches the threshold value, the rest of spiking neurons within the circuit are reset to a resting potential value. This pattern of lateral connection essentially mimics the effect of the top-down influences and significantly reduces the computational cost of the perceptual system.

D. INTEGRATION OF INFORMATION AND FADING MEMORY

Maass *et al.* [11], [37], [38] proposed the so-called "generic cortical microcircuits" to explain the neural computation at the micro-level as a means of interpretation of the cognitive processing in large-scale neural systems. The salient point is that integration of information over time with fading memory is a fundamental computational operation of these generic cortical microcircuits. The concept of fading memory implies that the influence of the input stimulus is present for a limited time after which it would gradually fade away if it is not rehearsed or refreshed. In addition to this property, these models have universal approximation and separation properties that make them excellent candidates for perceptual tasks such classification, recognition, and localization.

The intermediate outputs (from DPUs) are available at different times (due to difference in duration of processing from each modality). This is consistent with the temporal coding of information in spiking neurons and can also be observed in human's visual perception [17], [39], [40]. Therefore, the integration of these outputs over the time of perception process and incorporation of the fading memory is practical and compatible with biological interpretations. Here, we consider two alternative methods to implement the temporal integration with fading memory, both of them employ spiking neurons as their core element. These two models are referred to as *temporal integration with fading memory via liquid state machine* and *temporal integration with fading memory via leaky integrate-and-fire* neuron (LIF), respectively.

1) TEMPORAL INTEGRATION WITH FADING MEMORY VIA LIQUID STATE MACHINE

The first method that can be used to implement the adopted hypothesis is to use one of the two well-known computational models of reservoir computing; liquid state machine or echo state network. In this methodology, each intermediate output (from DPUs) is converted to a segment of spike train using one of the methods suggested in [41] and [42]. Then, all of these segments are configured as one spike train that

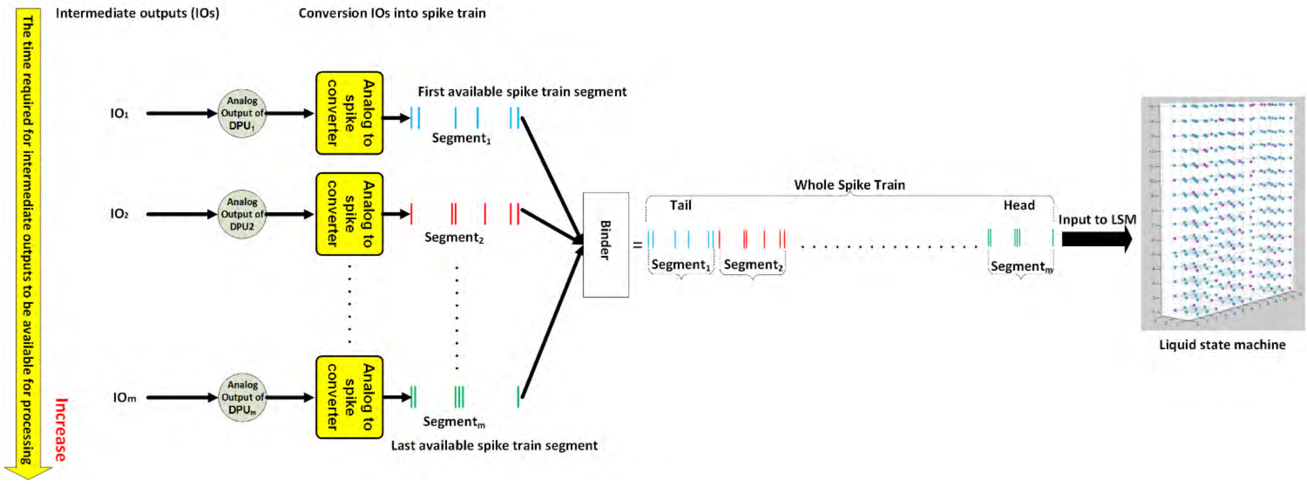

Intermediate outputs (IOs)
Conversion IOs into spike train
The time required for intermediate outputs to be available for processing
Increase
IO_1
Analog Output of DPU_1
Analog to spike converter
First available spike train segment
$Segment_1$
IO_2
Analog Output of DPU2
Analog to spike converter
$Segment_2$
IO_m
Analog Output of DPU_m
Analog to spike converter
$Segment_m$
Last available spike train segment
Binder
=
Whole Spike Train
Tail
Head
$Segment_1$
$Segment_2$
$Segment_m$
Input to LSM
Liquid state machine

FIGURE 5. Implementing temporal integration with fading memory via liquid state machine.

is fed to a liquid state machine. The liquid state machine is an excellent computational candidate that can be used to classify each spike segment. In that sense, each input spike segment represents one attribute of an attended stimulus and the whole spike train represents the complete feature map of that attended stimulus for any classification tasks such as facial-expression recognition, person recognition, or object recognition. The formation of the whole spike train considers the times that the outputs from the DPUs (IOs) are available in the shape of a stack memory. In other words, the first available output is converted to segment of spike train and will be considered as the base of the whole spike train. The next available output from these DPUs will be converted to segment of spike train in the same way and is attached to the front of first segment, and so on. The process of conversion from intermediate outputs into spike train is depicted in Fig. 5. Since the formation of the whole spike train that represents a real-world stimulus is performed as described above, we adopted the weighted majority voting criterion to finalize the outcome of the perception process such that more weight is given to the recent available information represented as the head segment in Fig. 5. This is compatible with both criteria of encoding the output of DPUs and the performance of liquid state machine in recognizing each spike segment in the whole spike train. The performance of the liquid state machine in classifying the segments in the whole spike train increases from tail segment (early available information) towards head segment (recent available information) as demonstrated in [37]. The capability of the liquid state machine to classify the most recent available information with higher performance than that of earlier information fits our hypothesis of the importance of considering fading memory property in integrating information over the perception time. Also, the recent available information or the last available output from the DPUs is encoded as head segment in our criterion that used to form the whole spike train as shown in Fig. 5. This provides accurate information but slow processing such as the case of information available in high spatial frequency during face perception [43], [44]. Moreover, the most recent available information is derived in the light of top-down influences (i.e., limiting the search scope looking for best candidates of attended stimulus). Hence, most often, this represents the most accurate information and refined version for the best candidate of the attended stimulus. The last output from DPUs (most recent available information) that will be considered in forming the whole spike train is highly affected by the natural perception time that allowed for an attended task. We suggest that the allowed perception time for an attended stimulus is affected by the specific scenario of social human-robot interaction (HRI); i.e., the emotional of the human and the type of sensor modalities that participate in encoding the attended stimulus.

2) TEMPORAL INTEGRATION WITH FADING MEMORY VIA LEAKY INTEGRATE-AND-FIRE NEURON

The second method of forcing temporal integration with fading memory in integrating the information generated by the DPUs is to use leaky integrate-and-fire neuron (LIF) model. In this method, each output of the DPU is converted into spike firing time that is inversely proportional to the associated output score as shown in Fig. 6. In this conversion, the high output score is equivalent to early firing time. Hence, the neuron fired first represents the best candidate of the attended stimulus. These spikes will be fed to LIF neurons by means of pre-synaptic inputs at their prescribed firing times, the membrane potential U of LIF neurons will increase due to pre-synaptic inputs. Whenever the membrane potential U of one of these neurons reaches a threshold V, the neuron fires a spike and its potential is reset to a resting potential U_r. During no activity period, i.e. no spikes at the pre-synaptic neuron, the neuron will start leaking its potential U until a resting potential Ur is reached. Therefore, these models

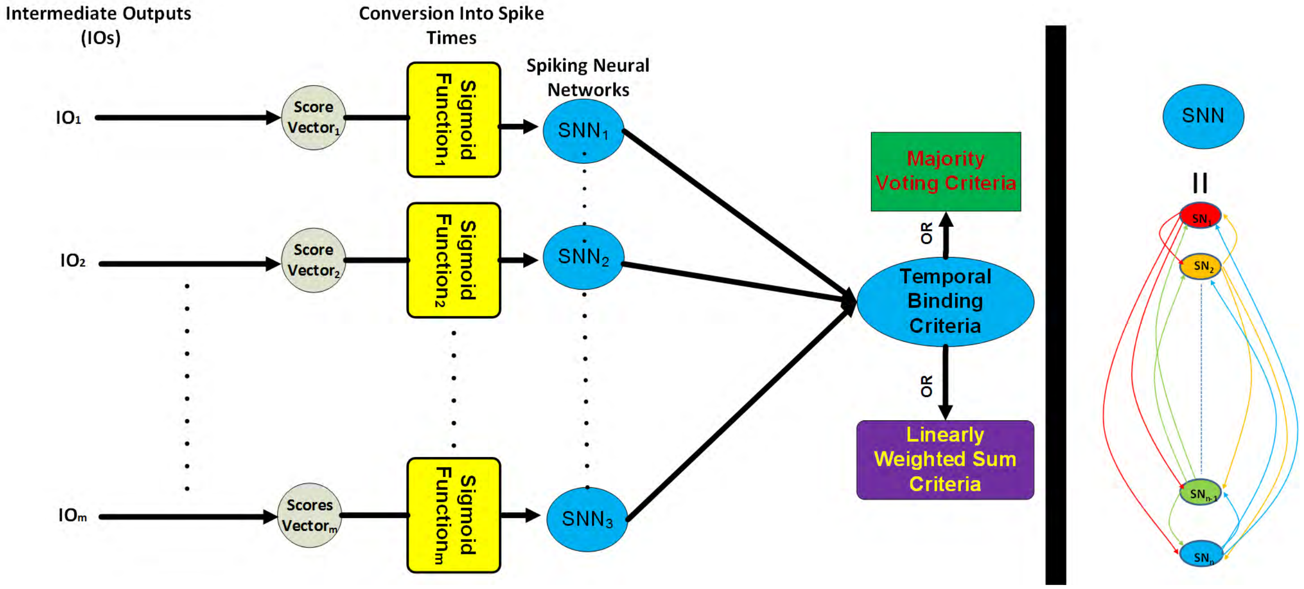


FIGURE 6. Implementing temporal integration information with fading memory via LIF.

are also referred to as forgetful, but the term leaky is more popular [45]. Since this neuron model does not fire a spike until the threshold is reached, the threshold value controls the allowed perception time for the attended task. In our model, increasing the threshold value is corresponding to giving more time for the neurons to integrate more information that is available in sensor modalities through the DPUs' pathways. We suggest that the threshold value should be adjusted in an adaptive manner to take into account the reliability of the perception outcomes, the natural real-time response, and the emotional state of the social robot.

These spiking neurons are connected to each other in lateral way as shown in Fig. 4. The main purpose of this connection is to rest the potential fields of all connected neurons within the spiking neural network whenever one of them reaches the threshold value and fire a spike. In other words, introducing the lateral connection between the spiking neurons, which are used to integrate the information over the perception time, reduces the computational cost by means of competitive behavior of the spiking neurons. As shown in Fig. 2, the integration of intermediate outputs from different DPUs within the same modality is performed at early stage in the hierarchical structure, while the integration of the finalized outputs from different sensory modality pathways happened at later stage in same hierarchical structure. We adopted this way of integrating the intermediate outputs because the time needed by each modality pathways to finalize their outputs has different time scales [46]. In most cases, the temporal window of integrating information from different sensor modality pathways is wider than that of integrating information from DPUs within the same sensor modality. This implies that the information available at the same time scale and from the same modality should be integrated together before being aggregated to the next level in the hierarchical structure in order to reinforce each other. In some social HRI scenarios, the outputs from sensor modality pathways are available at different time scales such that the effect generated from the output of one or more sensory modality pathways are completely faded out before being reinforced by the outputs of other sensory modality pathways that participate in the perception process of the attended stimulus. In this case, we adopted the linearly weighted sum criterion for integrating the outputs of DPUs at the last stage of the hierarchical architecture in addition to the temporal binding that used for early stages. In many perceptual tasks, the perception outcomes are produced by a single modality pathway. For example, humans recognize individuals who have odd features faster [47]. In this case, the most distinctive feature, which is the odd feature, is processed through the fastest path and its intermediate output is enough to drive the membrane potential U of LIF neuron to its threshold value and consequently fire a spike which means completion of the perceptual process of the attended stimulus.

Remark: The reader should have already realized that the multi-modals systems designed based on "fusion" are conceptually and operationally different from the proposed architecture. The idea of fusion is to integrate the effect of several modalities with a view that each modality by its own is not able to contribute to a correct perception; as such the signals are fused together to enhance the perception. The proposed system is designed based on the idea of convergence zone (as the term is used in neuroscience). This is further elaborated in Fig. 1−R and Fig. 1−L. The modules "Conversion of IOs to spiking networks" and "Temporal binding" are analogous to "Multi-modal Association Cortex". As such, the process is essentially different from fusion. One of the attractive properties of such a system is that it does not require all sensor modalities concurrently. The perception process is facilitated by any modality that is rich in information and first becomes available.

E. DESIGN GUIDELINES FOR GENERAL PURPOSE HUMAN-ORIENTED PERCEPTUAL SYSTEM

In this section, we outline some design guidelines for the proposed system:

1) A real-world stimulus is represented by a limited set of features that are generated by different sensor modalities and provide different types of information. Adopting limited set of features to represent a real-world stimulus is inspired by the fact that humans use limited channel capacity of processing the data generated by the sensory system.
2) The set of features vectors that represents the attended stimulus are then processed by DPU's. These DPU's have dedicated architectures, use different learning methods, and superior in extracting a specific type of information available in these feature vectors.
3) The human-oriented perceptual system should be able to finalize the decision making process in finite time by efficient processing of the available information according to specific HRI scenarios.
4) The final outcome of human-oriented perceptual system is generated by integrating the outputs from different DPUS at different levels, which are called ‘‘convergences zones’’, in a hierarchical manner.
5) The integration or binding of information criterion should consider the following:
6) The importance of natural real-time response at human rate for social robots that proposed to interact with human in social settings.
7) Most often, a real-world stimulus is composed of a set of features that vary in their relative salience on the perception outcome and complement each other.
8) Humans interact with the real-world environment by means of what information is available in the streams of sensory systems (in some scenarios, sensor modalities are not available simultaneously).

A pseudo-code representation of the proposed platform is provided in Fig. 7 to facilitate the implementation of the proposed system. It should be noted that processing of streams of data from sensor modalities runs in parallel through independent pathways.

We would also like to highlight the inherent difference between the proposed perceptual system and the notion of sensor fusion. The key feature of a sensor fusion system is that the output of such systems requires the presence and significance (contribution) of all the modalities at all times. In many applications, it is very difficult to know a priori the actual contribution of a particular sensor (modality). Also, the presence of all sensory modalities at the same time is not practical in social HRI scenarios. However, as discussed in previous section, the proposed system is principally different from sensor fusion.

IV. SYSTEM IMPLEMENTATION

In order to assess the performance of the perceptual system, we configured it for the problem of person recognition (Fig. 8) on an in-house designed social robot. In this section, we will present the implementation details as well as experimental studies

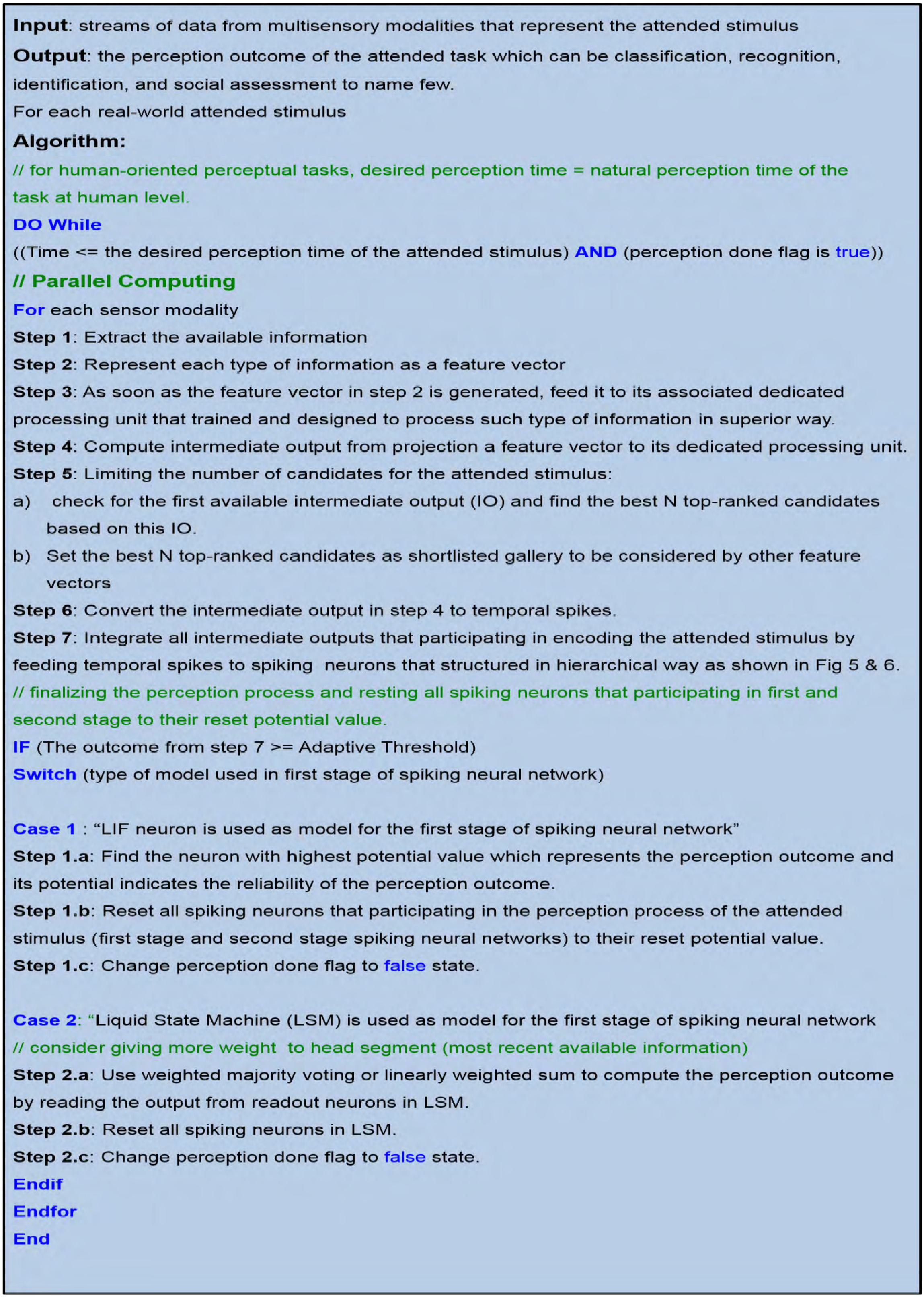
A. SOCIAL ROBOT PLATFORM

Let us first go through the design of the social robot. The base of the social robot is a Pioneer 3DX (P3-DX) – a popular research mobile robot platform. The robot has a two-wheel differential drive system with rear caster and 500-tick encoder. The body and the base of the robot are made from aluminum with dimension (51×38×22) with 19.5cm diameter drive foam filled wheel, it can move at speed of 1.6 meter per second on flat floor with capability of carrying payload up to 23 kg. The Pioneer robot system is extensively retrofitted. The height of the robot is extended to 150 cm by mounting additional stand on the top of the robot's base. The additional stand is designed such that it does not block other sensors required for navigation (i.e., SEEK Laser). The system is equipped with a RGB-D sensor (Kinect sensor), RGB camera (USB3 Flea3 machine vision camera), Directional microphone (Rode VideoMic Pro), SICK LMS-200 laser rangefinder, 16 sonar sensors are distributed around the robot to provide a 360−degree coverage with read ranges from 15cm to approximately 500 cm. The Kinect sensor has a multi-array microphone distributed around it. The data stream from this multi-array microphone can be used for acoustic source localization and ambient noise suppression. The social robot is also equipped with Lenovo W530 laptop as on board PC to provide the required computational power. The hardware may be augmented with additional modalities such as olfactory or tactile sensors in future. However, for the purpose of person recognition in social settings, the existing sensors modalities are completely sufficient.

The social human-robot interaction is facilitated by voice and an interactive 3D avatar capable of showing different emotional states using facial morphing, body movement, and voice. The virtual avatar character is displayed on a touchscreen mounted on the top of the stand. We used the virtual avatar system developed by Charamel Inc. The avatar system is rendered in variety of characters (face, shape, gender) that are selected by the nature of application and the tasks that the social robot is expected to perform. Moreover, the avatar system is augmented with a text-to-speech engine and lip synchronization (vocalizer) features to facilitate a realistic social human-robot interaction.

B. EXPERIMENTAL SETUP

Fig. 8 shows the configuration of the architecture of the perceptual system for the problem of person recognition problem in social settings. The biometric information relating to a subject's identity is represented as a set of feature vectors. These feature vectors may not be available



```
Input: streams of data from multisensory modalities that represent the attended stimulus
Output: the perception outcome of the attended task which can be classification, recognition,
identification, and social assessment to name few.
For each real-world attended stimulus
Algorithm:
// for human-oriented perceptual tasks, desired perception time = natural perception time of the
task at human level.
DO While
((Time <= the desired perception time of the attended stimulus) AND (perception done flag is true))
// Parallel Computing
For each sensor modality
Step 1: Extract the available information
Step 2: Represent each type of information as a feature vector
Step 3: As soon as the feature vector in step 2 is generated, feed it to its associated dedicated
processing unit that trained and designed to process such type of information in superior way.
Step 4: Compute intermediate output from projection a feature vector to its dedicated processing unit.
Step 5: Limiting the number of candidates for the attended stimulus:
a)  check for the first available intermediate output (IO) and find the best N top-ranked candidates
    based on this IO.
b)  Set the best N top-ranked candidates as shortlisted gallery to be considered by other feature
    vectors
Step 6: Convert the intermediate output in step 4 to temporal spikes.
Step 7: Integrate all intermediate outputs that participating in encoding the attended stimulus by
feeding temporal spikes to spiking  neurons that structured in hierarchical way as shown in Fig 5 & 6.
// finalizing the perception process and resting all spiking neurons that participating in first and
second stage to their reset potential value.
IF (The outcome from step 7 >= Adaptive Threshold)
Switch (type of model used in first stage of spiking neural network)

Case 1 : "LIF neuron is used as model for the first stage of spiking neural network"
Step 1.a: Find the neuron with highest potential value which represents the perception outcome and
its potential indicates the reliability of the perception outcome.
Step 1.b: Reset all spiking neurons that participating in the perception process of the attended
stimulus (first stage and second stage spiking neural networks) to their reset potential value.
Step 1.c: Change perception done flag to false state.

Case 2: "Liquid State Machine (LSM) is used as model for the first stage of spiking neural network
// consider giving more weight  to head segment (most recent available information)
Step 2.a: Use weighted majority voting or linearly weighted sum to compute the perception outcome
by reading the output from readout neurons in LSM.
Step 2.b: Reset all spiking neurons in LSM.
Step 2.c: Change perception done flag to false state.
Endif
Endfor
End
```

FIGURE 7. Pseudo-code summarizing implementation steps of the proposed framework of developing human-oriented perceptual system.

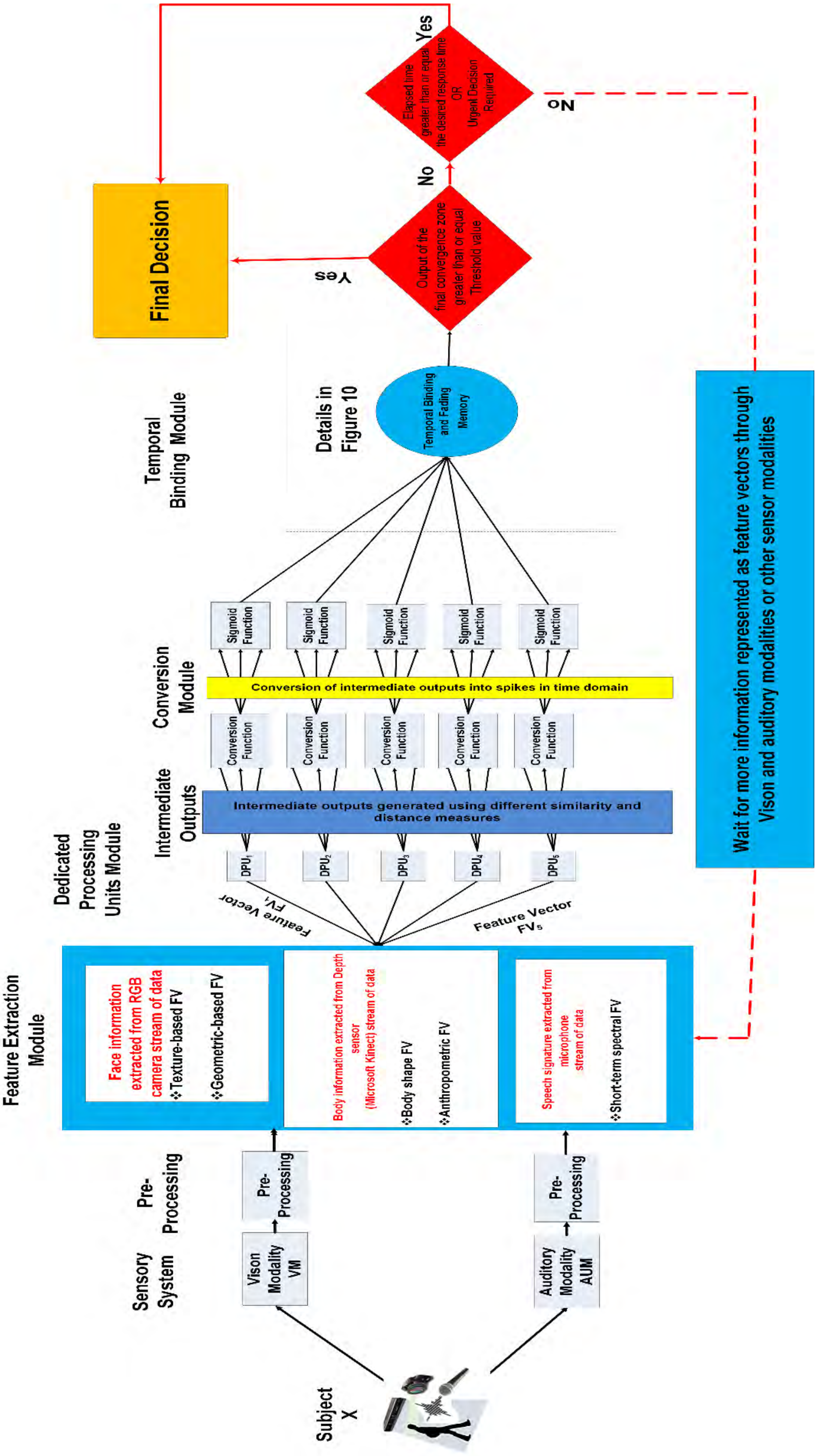

Subject X
Sensory System
Pre-Processing
Feature Extraction Module
Dedicated Processing Units Module
Intermediate Outputs
Conversion Module
Temporal Binding Module
Vison Modality VM
Auditory Modality AUM
Pre-Processing
Pre-Processing
Face information extracted from RGB camera stream of data
❖Texture-based FV
❖Geometric-based FV
Body information extracted from Depth sensor (Microsoft Kinect) stream of data
❖Body shape FV
❖Anthropometric FV
Speech signature extracted from microphone stream of data
❖Short-term spectral FV
Feature Vector FV1
Feature Vector FV5
DPU1
DPU2
DPU3
DPU4
DPU5
Intermediate outputs generated using different similarity and distance measures
Conversion Function
Conversion Function
Conversion Function
Conversion Function
Conversion Function
Conversion of intermediate outputs into spikes in time domain
Sigmoid Function
Sigmoid Function
Sigmoid Function
Sigmoid Function
Sigmoid Function
Temporal Binding and Fading Memory
Details in Figure 10
Output of the final convergence zone greater than or equal Threshold value
Yes
No
Elapsed time greater than or equal the desired response time OR Urgent Decision Required
Yes
No
Final Decision
Wait for more information represented as feature vectors through Vison and auditory modalities or other sensor modalities

FIGURE 8. The perceptual system configured for person recognition task in social settings.

concurrently due to many reasons including sensors limitations, occlusion, and variation in the processing time for pre-processing and feature extraction process. The features vectors that are used to represent the biometric features in order to identify a subject in different settings and postures are short-term spectral feature, facial texture-based feature vector, facial geometry-based feature vector, anthropometric feature vector, and body shape feature vector. Some of these feature vectors are computationally expensive such as the texture-based feature which is based on Gabor filter (for vision modality) which uses convolution operator. Since this kind of features need more time and available later in the perception process, the information carried by this type of features may not contribute to the final outcome when time-critical response is required. In contrast, other feature vectors such as anthropometric feature vector (RGB-D sensor), which is based on Euclidian distance computed between pairs of selected skeleton joints, needs less computational cost and can be detected and extracted from longer distance compared to other feature vectors. Hence, it is available early in the perception process and consequently contributes to the person recognition process. In order to generate the intermediate outputs, these features vectors are processed by dedicated processing units (DPU) as shown in Fig. 8. These DPUs are application dependent. In this real-time implementation, we adopted support vector machine (SVM), K-NN classifiers that used various distance and similarity measures, and Gaussian mixture model (GMM) as computational model candidates for DPUs. These computational models represent generative (e.g. GMM) and discriminant (e.g. SVM) models that estimate the distribution within each class and the boundary between classes, respectively. In addition, the outputs of the classifiers that use different similarity and distance measures represent how much a test subject is close or similar to each candidate in the gallery set from different measure perspectives. Using various similarity and distance measures in the classifier design reduce the error variance. On other hand, adopting SVM and GMM that are superior in processing such feature vectors lessen the error bias of the system. Since the temporal binding via liquid state machine has been already evaluated for classification tasks [48], the temporal binding via LIF neuron was adopted in this implementation (section III-C.2). The role of top-down influences, which is one of the main subsystems of the perceptual system, in reducing the computational cost, improve the recognition rate, and facilitating real-time response of social HRI is highlighted in this paper. The top-down influence is implemented in this paper by means of expectation which limits the scope of search for the best candidate of the attended subject.

1) REGISTRATION OF THE SUBJECTS

Like humans that are introduced to each other in social occasions, the social robot also needs to be "introduced" to the people in its social circle. We refer to the introduction phase as "registration". As there are no multi-modal datasets, we constructed a database for 79 subjects by combining three well known datasets namely, TIDIGITS corpus [49], FERET database [50], and RGB-D database [51]. Therefore, a gallery set was constructed from the combined dataset and the biometric data captured from 8 members of our research laboratory. Thus, the number of subjects in the gallery set increased to 87. However, the 8 members had to go through the process of registration so that the robot obtains speech, facial views, and body features for these individuals. The gathered information is the same as those available in the combined database. Six groups of data were collected using Kinect sensor, Flea 3 USB3 camera, and Rode shotgun microphone. Four of these groups of data, which represented depth information of subjects in various scenarios, were collected using Kinect sensor in an indoor environment. In the first four groups (depth information), we required capturing from each subject in the group the synchronized information of 1) skeleton joints, 2) body's point cloud. The first group of data was captured by recording each individual walking slowly with frontal view and stretched arm. The second group of data was obtained where a subject walking normally without stretching his/her arm. The third group of data represents back view of each subject walking away from the robot. The last group of data in depth information category was a side view of each subject (90 degrees from sensor view) walking normally. In all the four scenarios, the data was captured by Kinect sensor where a subject was at least 2 meters away from the robot. The fifth group of data, which represented frontal facial image of subjects, were collected

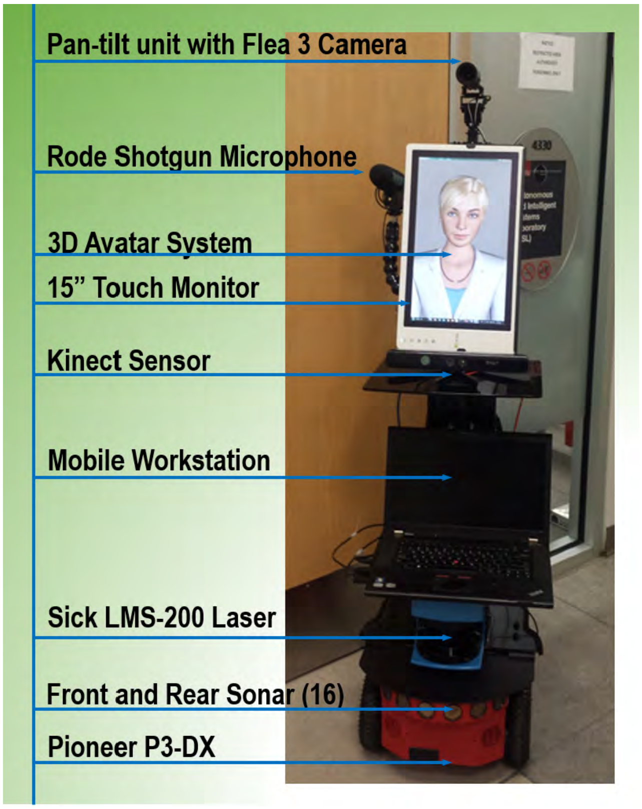


FIGURE 9. Pioneer 3DX reconfigured as social robot platform.

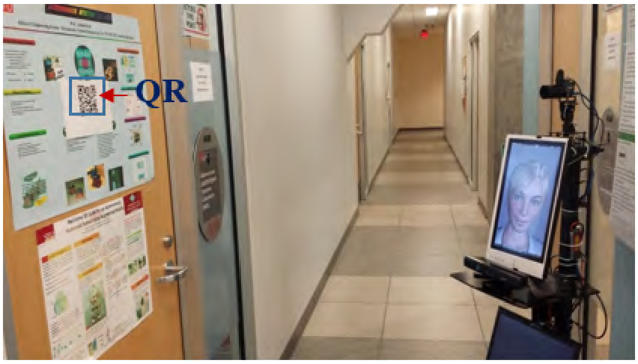


FIGURE 10. The social robot platform detecting one QR posted on room's door.

using Flea3 USB3 camera where a subject is 2 meter away from the camera. The last group of data represented speech data of each individual pronouncing 77 digit sequences as in TIDIGITS corpus. As shown in Fig. 10, the robot infers its approximate location (room identification) by detecting and decoding the QR code posted on room's door. We adopted ZXing (''zebra crossing'') which is an open-source multi-format 1D/2D barcode image processing library implemented in different programming languages. The nature of person identification for social HRI imposes some requirements on social robot; robot needs to respond in real-time fashion, robot needs to identify a subject in different postures and various lighting condition, the robot should infer the identity of a subject based on ''incomplete'' but what available information. The combination of the depth sensor and RGB camera is a plausible choice for the long-term person identification task that considered one of the primary requirements of social HRI. Depth sensor information is not affected by changing lighting conditions or by changing clothes which is one of primary challenges in long-term person identification. On the other hand, RGB camera provides important information such as appearance-based features which carry high discriminant power. Kinect sensor and Flea3 camera work in a synergetic manner and provide information about a subject in different scenarios related to the social HRI. For example, the Kinect sensor is capable of detecting a subject in range of 0.8 to 4.2 meters regardless of a subject's posture. In our implementation, whenever a subject is detected by Kinect sensor, his/her body's point cloud and skeleton joint are extracted and fed to the associated pre-processing unit to generate anthropometric and body shape feature vectors. Also, whenever a subject is close enough to the camera so that his/her face is detected in one frame of the camera video stream, a cropped facial image is fed to the associated pre-processing unit to construct texture-based and geometry-based facial feature vectors. However, in some scenarios of the social HRI the information available in visual sensory streams is useless for the attended task (person recognition). To accommodate these scenarios where the information in visual sensory streams does not carry discriminant information to recognize a subject such as a subject's skeleton is not detected by Kinect sensor due to distance or substantial occlusion, or subject's face is not detected by Flea 3 camera due to distance or camera viewing angle and head orientation, the system exploits the speech data that available in auditory modality (i.e. Rode shotgun microphone) to extract voiceprint and recognize a subject.

In the temporal binding via LIF neuron, the number of spiking neuron participating in recognition process is a monotonically increasing function (function of number of subject in gallery set). This number of neurons creates computational burden that cause difficulty to meet real-time response requirements when the system is implemented on an average personal computer with modest computational power. The expectation solves this problem by grasping simple and distinctive cues from robot's environment which in turn helps robot to infer useful information about the environment such as type of room, location of room, people probably work in that room. This useful information reduces the computational cost by two ways; the robot expects to meet specific group with limited number of people which reduces the computational cost needed for the classifiers, also the number of active spiking neurons needed to construct the binding circuit is dramatically reduced. Hence, real-time response can be achieved easily.

2) TOP-DOWN INFLUENCES

A distinctive feature of the system is the utilization of top-down influences to significantly speed up the perception process. We adopted quick response (QR) codes as a sensible and straightforward solution to infer crucial information about the environment. The cues from QR supplies complementary features to simplify the person recognition task. QR codes were posted at different locations, in particular on the front door of different research laboratories. Each QR is assigned with a string to identify the associated laboratory and the individuals who normally work therein. Once the QR codes are detected and decoded by the robot's vision system, the social robot gets prior knowledge on the specific group of individuals expected to be in that area. Hence the search space is dramatically reduced leading to less computational burden as well as faster perception speed. We used the open-source code to detect and decode the QR codes in this implementation. Fig. 10 shows the robot detecting and decoding one of QR codes that was posted on the front door of our research laboratory. As the robot passes by the door, the robot's vision system detects and decodes the posted QR code. The robot uses the decoded information assigned to the QR code (containing important cues about this particular environment), to provide prior information about the group of people who normally work in this place for the purpose of person recognition, thus the search space (from all subjects in the gallery set) is reduced to only few people.

Reducing the number of candidates in the gallery set (search space) will alleviate the computation burden of the system as less number of feature vectors need to be generated. These feature vectors are further processed by

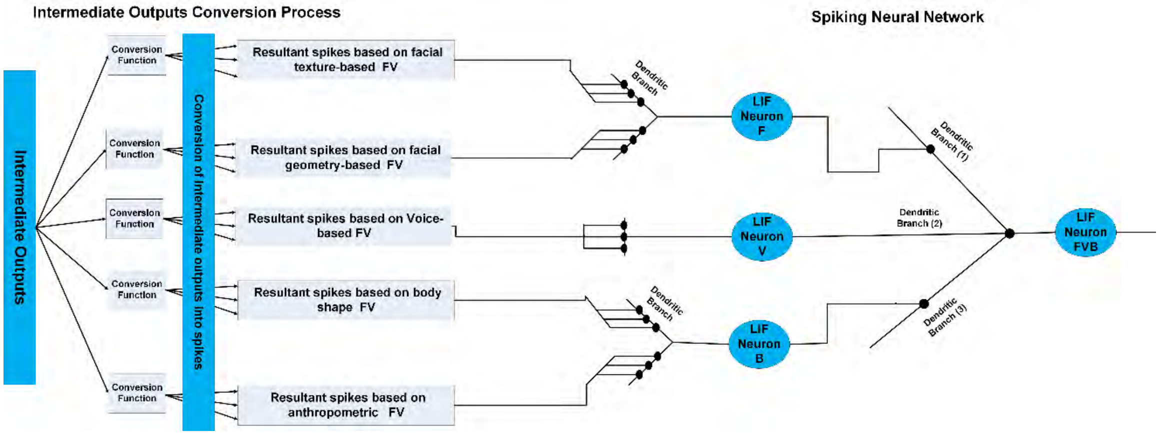

Intermediate Outputs Conversion Process
Spiking Neural Network
Intermediate Outputs
Conversion Function
Conversion Function
Conversion Function
Conversion Function
Conversion Function
Conversion of intermediate outputs into spikes
Resultant spikes based on facial texture-based FV
Resultant spikes based on facial geometry-based FV
Resultant spikes based on Voice-based FV
Resultant spikes based on body shape FV
Resultant spikes based on anthropometric FV
Dendritic Branch
Dendritic Branch
LIF Neuron F
LIF Neuron V
LIF Neuron B
Dendritic Branch (1)
Dendritic Branch (2)
Dendritic Branch (3)
LIF Neuron FVB

FIGURE 11. One block of the SNN circuit that are used to construct the overall SNN.

their respective DPUs (GMM, SVM, and K-NN classifiers). Their outputs (the intermediate outputs) are then integrated via the spiking neural networks as shown in Fig. 11. As the number of spiking neurons that will compete among each other to represent the best candidate for the attended subject is a function of the number of individuals in the gallery set, the computation cost of the system is considerably reduced. The number of spiking neurons to construct the spiking neural network can be expressed as N=4×n, where N is the overall number of spiking neurons, and n is the number of subjects in gallery set. In this implementation, the top-down influence (QR codes) limits the number of spiking neurons participating in the recognition of the subject by biasing group of spiking neurons (the number of spiking neurons is reduced from 384 neurons to 32 neurons). If the detected person is not a member of the identified group, the system utilizes this problem by checking the threshold value at the final stage of multi-modal perceptual system. In a case where the final output of the system does not satisfy the threshold value, the system will expand the scope of the search by adding more individuals to the galley set. One of the key features of the multi-modal perceptual system is that it is capable to finalize the perceptual task using the available information within the allowed time for the perception process. In order to finalize the process within the specified time, the system processes the data of multiple sensory modalities in parallel. For example, as shown in Fig.11 when some biometric features such as face information is not available due to occlusion, or limitation of sensors, the system exploits other biometric features such as speech, anthropometric, and body-shape information in order to identify a person and responds within the desired time in real-time fashion.

3) FEATURE EXTRACTION

The face-based training model was trained using geometry-based and texture-based facial feature vectors. Six to eight frontal facial images at different facial expressions and different illuminations were employed as the facial gallery set. The body-based training model was trained using body-shape and anthropometric feature vectors. For each of the four scenario described in section IV-C, three out of five samples from depth information (i.e. skeleton joints, body's point cloud, and the estimated floor) were selected for construction of the body gallery set. Four feature vectors were extracted from each sample in the facial galley set and body galley set. Two of them represent facial information including facial geometry-based and texture-based feature vectors. The rest of the feature vectors, namely the body shape and anthropometric feature vectors represent the body information of the attended subject. The facial geometry-based feature vector consisted of combination of Euclidian distances among selected facial points and facial feature ratios, computed as shown in Table 1 (Appendix A). The combination of Euclidian distances among selected skeleton joints, shown in Fig. 12, is used to generate the anthropometric feature vector as described in Table 2 (Appendix A). The combination of geodesic distances among the projection of selected skeleton joints on the three-dimensional body point cloud were used to construct body shape feature vector. First, skeleton joints are detected by Kinect sensor and then projected on 3D body mesh generated from point cloud.

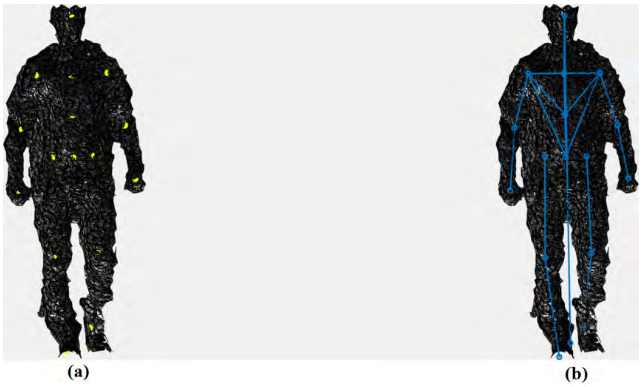


FIGURE 12. Selected skeleton joints and Euclidean distance among them; a) Projection of skeleton joints on the three-dimensional body point cloud, b) Euclidian distance of selected skeleton segments.

A good approximation of the shortest path among the projected skeleton joints are extracted by the fast-marching algorithm, implemented using the Dijkstra algorithm [36]. The complete list of geodesic distances that used to construct the body shape feature vector are described in Table 3 (Appendix A).

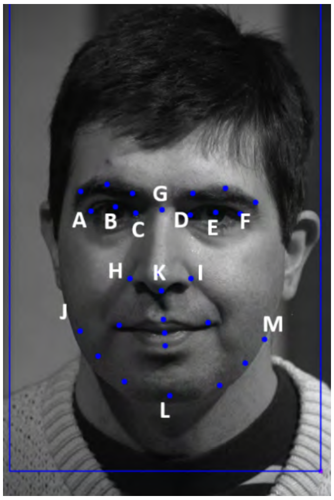


FIGURE 13. Selected facial points on a frontal facial image from FERET database.

The facial texture-based feature vector is represented as a set of multi-scale and multi-orientation Gabor filter coefficients extracted from the face image at facial points shown in Fig. 13. Since the Gabor filter uses convolution integral, which is computationally expensive, local facial texture patterns at the facial points are extracted by Gabor to speed up the computation. The 2D Gabor filter can be expressed as in (1):

$$G(x, y, \xi_x, \xi_y, \sigma_x, \sigma_y, \theta) = \frac{1}{\sqrt{\pi \sigma_x \sigma_y}} e^{-\frac{1}{2}[(\frac{R_1}{\sigma_x})^2 + (\frac{R_2}{\sigma_y})^2]} e^{j(\xi_x x + \xi_y y)} \tag{1}$$

Where $R_1 = xcos\theta + ysin\theta$ and $R_2 = -xsin\theta + ycos\theta$, ξ_x and ξ_y are spatial frequencies, σ_x and σ_y are the standard deviation of an elliptical Gaussian along the x and y axes, and θ represents the orientation.

Given an input image I, the response image of the Gabor filter can be computed using the convolution operation defined as in (2). We convolve the image I with every Gabor filter kernel in the Gabor filter banks centered at the pixels specified by the facial points.

$$z = \sum\nolimits_x \sum\nolimits_y I\,(x, y) G\left(x' - x, y' - y, \omega, \sigma, r, \theta\right) \tag{2}$$

Where $G\left(x' - x, y' - y, \omega, \sigma, r, \theta\right)$ is Gabor filter kernel centered at (x', y'). $I\,(x, y)$ is the intensity value of the image I at (x, y) location. The performance of the Gabor filter response in the face recognition and classification tasks is highly affected by the parameters used to construct the Gabor Kernel bank. In this study, we adopted the Gabor filter parameters suggested by [52]. The suggested parameters are: 8 orientations, 6 frequencies with Gaussian width ($\sigma_x = \sigma_y = 1$). The Gabor filter bank responses given in (2) consist of real and imaginary parts that can be represented as magnitudes and phases components. Since the magnitudes vary slowly with the position of facial points on the face, whereas the phases are very sensitive to them, we used only the magnitudes of the Gabor filter responses to generate the texture-based feature vector. Hence, we have 48 Gabor coefficients for each facial point on the face.

The voice signature is extracted based on the short-term spectral, specifically, the Mel-frequency cepstral coefficients (MFCCs). MFCCs is plausible choice for real-time speaker identification for many reasons: (1) MFCCs need low computing power and require short utterance to be extracted, and (2) MFCCs characterize a person's vocal tract. Hence it is text and language independent. Since a natural speech is analog signal, it must be sampled and digitized in order to be fed to feature extraction module. A natural speech signal composed of various segments including; voiced, unvoiced, silence, and other non-speech segments. In general, voiced segments characterized as high energy and periodic signal. In contrast, unvoiced segments have lower energy level and non-periodic in nature. Voiced segments carry the information that distinguishes voiceprint of one speaker from another; hence these segments must be localized and segmented from input speech signal first before feeding it to the feature extraction module. Voice activity detector (VAD) has been proposed by speech and speaker recognition researchers to divide input audio signal into voiced, silence, and non-speech segments. One of the readily used VAD approach in real-time application is energy based approach which assume that voiced segments have higher energy that other non-speech segments. A VAD algorithm based on signal energy and spectral centroid has been implemented in this study. The result from VAD is pre-emphasized and then windowing by applying hamming windows as in (3).

$$W(n) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n < N - 1 \tag{3}$$

The resultant time-domain signal is converted to frequency domain by applying the well-known Fast Fourier Transform (FFT). Finally, The MFCCs are obtained by applying logarithmic compression and discrete cosine transform as in (4). The discrete cosine transform converts log Mel spectrum into time domain.

$$C_n = \sum\nolimits_m^M \left[\log S(m)\right] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \tag{4}$$

Where $S\,(m)\,, m = 1, 2 \ldots., M$ is output of an M-channel filter-bank, n is the index of the cepstral coefficient. In this study, we retained the 12 lowest C_n excluding 0th coefficient.

4) COMPUTATIONAL MODELS FOR DPUs

Each of the aforementioned feature vectors in section IV-D is further processed by selected classification model that heuristically proved to provide satisfactory performance with such kind of feature vector. For facial geometric-based, body shape, and anthropometric feature vectors SVM with radial basis function kernel is adopted as a DPU. Facial geometric-based, body shape, and anthropometric feature vectors are composed of limited number of features (i.e. low-dimensional feature vector). It is sensible to adopt SVM with radial basis function kernel as DPU for these feature vectors in order to map them to high dimensional space and find a separating hyperplane with the maximal margin. 5-fold cross-validation method is used to find the best parameter for the kernel function. The probability estimate output from SVM were used to represent the IOs, the LIBSVM was adopted to train SVM [53]. The texture-based feature vector is processed by $7-NN$ that uses *L2norm*, *Mahalanobis distance*, and *Cosine Similarity*. It is worth to noting that these distance and similarity measures are applied to each facial point and then summed for each subject to measure the total closeness and similarity of all facial points of the attended subject as in (5) and (6). The Gaussian mixture model with universal background model (GMM-UBM) method is popular and considered as baseline to evaluate speaker recognition system. Here, the GMM-UBM is adopted as DPU for voice-based feature vector. We used 64-compnenent GMM to build the UBM. Then, for each subject a speaker-dependent GMM parameters is estimated by adapting the well-trained parameters in the UBM to fit specific speaker model. Finally, the IOs from this classifier is calculated using posteriori probability of all observations in probe set against all speakers' models in gallery set.

$$Similarity^{i}_{a}(J, J') = \frac{\sum_{k=1}^{N} a_{ki}a'_{ki}}{\sqrt{\sum_{k=1}^{N} a_{ki}^{2} \sum_{k=1}^{N} a_{ki}'^{2}}} \tag{5}$$

$$Similarity_{face} = \sum_{i=1}^{L} Similarity^{i}_{a}(J, J') \tag{6}$$

Where $Similarity^{i}_{a}\left(J, J'\right)$ is the similarity between two jets, J and J$'$ associated with i^{th} facial points on the face of the subject, a_{ki} is the amplitude of kth Gabor coefficient at ith facial points. N is the number of wavelet kernels. $Similarity_{face}$ represents the total similarity between the two faces as the sum of the similarities over all facial points as expressed in (6). Equations (5) and (6) can be modified to express the closeness of all facial points by replacing *Cosine Similarity* with either *L2norm* or *Mahalanobis distance*.

5) INTEGRATION OF INFORMATION VIA LIF

The temporal binding is implemented through the leaky integrate-and-fire neuron (LIF) model to manifest the integration of IOs generated from various classifier. One block of the spiking neural network (SNN) that is used to perform the integration process is shown in Fig. 11. The overall SNN that is used to integrate the information from various biometric modalities is constructed by laterally connecting n blocks from this circuit where n represents number of subjects in gallery sets. IOs generated based on the selected feature vectors were converted into spike times and normalized to range from *zero to* 150 *ms* prior to being fed to LIF neurons. The dynamic of LIF neuron is governed by (7); for more details about this model see section III-D.2. The integration process is expressed as the total response of postsynaptic potential due to different presynaptic inputs within-branch and between-branch (7) to (9). Adopting *Softmax* to integrate IOs generated from same feature vector as in (8) (within-branch) reduces error variance by not exaggerating the effect of one aspect of measure at the expense of other measures in deriving the final outcome. On the other hand, weighted-linear summation of IOs generated from different feature vector as in (9) (between-branch) reduces error bias by means of exploiting various attributes in deriving the final outcome.

$$\tau_m \frac{dV_m}{dt} = -(V_m - V_{resting}) + R_m \cdot (I_{syn}(t) + I_{noise} \tag{7}$$

$$Syn\_Integrate\_WB = \sum_{i=1}^{K} Softmax(IO_i) \tag{8}$$

$$Syn\_Integrate\_BB = \sum_{i=1}^{K} \alpha_i IO_i \tag{9}$$

Where IO_i represents the total input to the ith dendritic branch, α_i is the ith dendritic branch weight, K is the number of dendritic branches.

The SNN was constructed by neural circuit (CSIM) simulator [38]. The parameters of LIF neurons were set as follows. The weight synapses of neuron F, neuron V, and neuron B were equal and set to 1500×10^{-9}. The weight synapses of neuron FVB was set as follows. The weight synapse of dendritic branch one was set to 4500×10^{-9} and weight synapses of dendritic branch two and three were set to 3000×10^{-9} and 2000×10^{-9} respectively. $V_{thresh=}0.15$, $V_{reset} = -0.067$, $V_{reseting} = 0$, $C_m = 5 \times 10^{-8}$, $V_{init} = 0.08$, $R_m = 1 \times 10^{-6}$, $T_{refact} = 0.0025$, $I_{noise} = 50 \times 10^{-9}$, I_{sys}(t) represents the converted IOs as spike times. These input spike times were set in the range from *zero to* 150 *ms*. This selection is empirically derived and compatible with the natural perception time of human. As shown in Fig. 11, the spiking neuros in SNN circuit were labeled with F, V, B, or FVB to indicate that the tagged neuron integrate facial information, body information, voice information, or all of them respectively. The first neuron from the list of spiking neurons labeled with FVB that fires a spike represents the best candidate of the attended subject x from the gallery set. The recognition rates can be calculated based on the firing time of the three neurons (i.e. neuron F, neuron V, neuron B, and neuron FB) which corresponds to calculating recognition rate based on facial information, body information, voice information, or all of them respectively. Neuron F may use face geometry, face texture, or both of them to fire a spike. The same applies to neuron B which may use body shape, anthropometric, or both in order to drive its potential to threshold and consequently evoke a spike. Neuron V receives its input from one feature vector, hence it uses it to build up its potential and fires a spike. The overall recognition rates were calculated based on neuron FVB, which may use facial information, body information, voice information, or all of them. Cumulative match curves (CMCs) show the probability

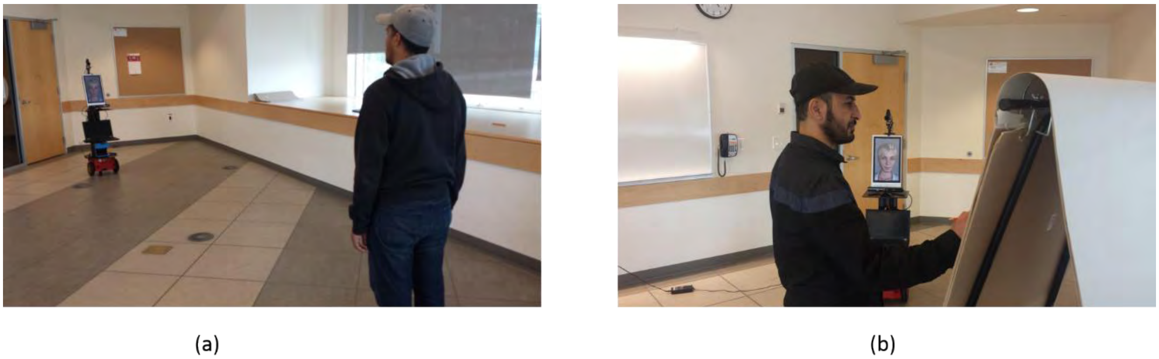


FIGURE 14. examples of scenarios that used to evaluate the perceptual system; a) face is not detectable due to distance, b) face information is not available due to head angle relative to camera view.

that the correct match of classification is found in the N most likely candidates, where N (the rank) is plotted on the x-axis. The recognition result was averaged over ten runs, the cumulative match curves (CMCs) were plotted for these recognition results.

In the next section, we will assess the performance of the of the multimodal social robot in different scenarios whereby one or more modalities are not available.

C. EXPERIMENT 1

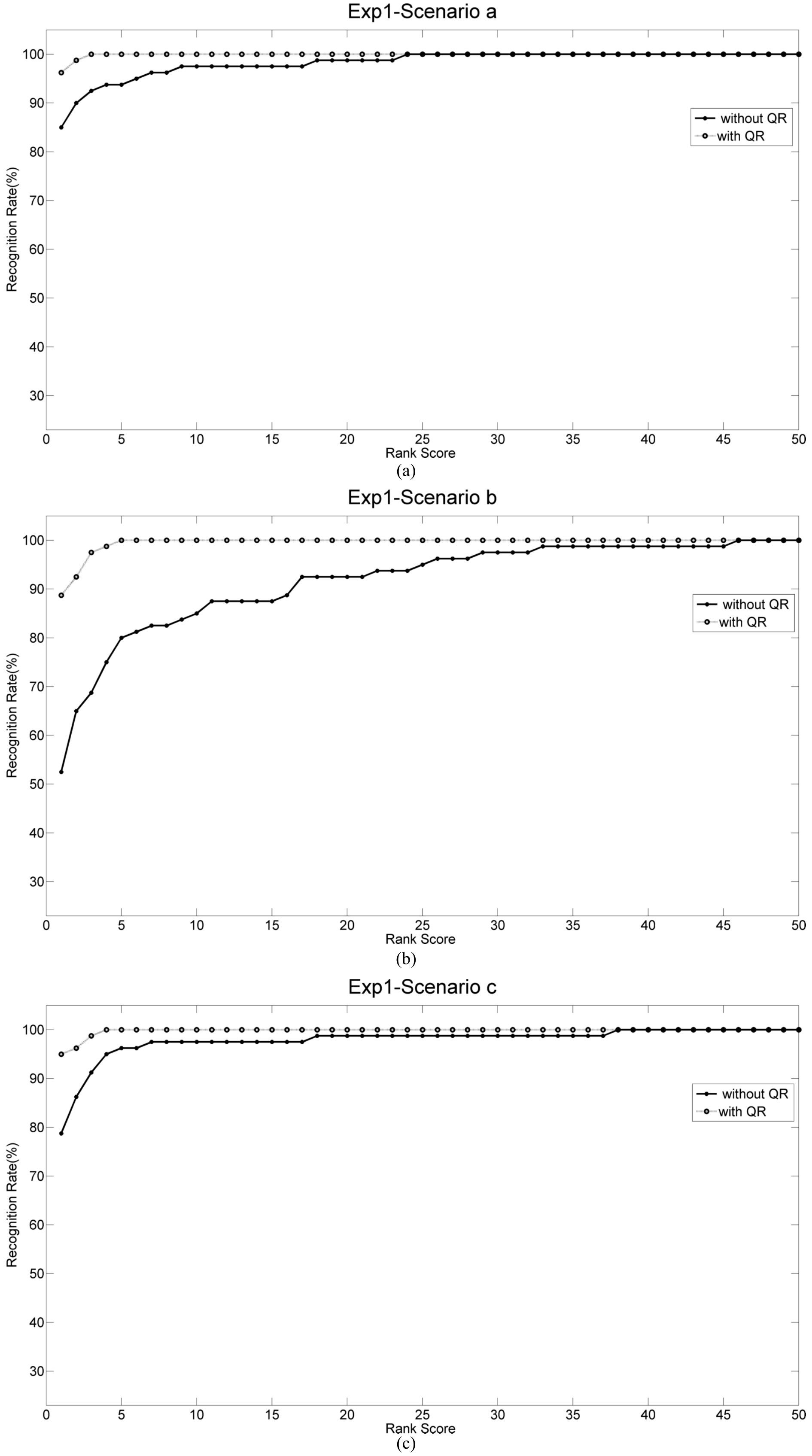
The system was evaluated in three scenarios; in the first two scenarios, the subject's facial information was absent due to difficulty of detection of the face in RGB image but the body information was available. Fig. 14-a depicts the first scenario (scenario a) where the subject was 4 meters away from the robot such that a small area of RGB image was reserved to face and consequently the face was not detected by the face detection algorithm. The second scenario (scenario b) is shown in Fig. 14-b, where the subject's face was extremely angled relative to camera view and reasonably far from the robot such that even a state-of-art face detection algorithm fails to detect it. In both scenarios, the body information could be detected and extracted from the data stream of the depth sensor (Microsoft Kinect). The third scenario (scenario c), the subject's facial and body information were absent due to substantial occlusion of the subject but the speech data could be captured by the microphone.

To demonstrate the effect of using top-down influences via expectations through QR code, we calculated the cumulative match curve for aforementioned scenarios in two cases; in the first case, the QR codes were not posted on doors and the robot's approximate location was not available while in the second case, the QR codes were posted on doors and the robot inferred its approximate location by identifying QR code. For each subject (eight individuals), we performed ten trials for each scenario. Cumulative match curves (CMCs) show the probability that the correct match of classification is found in the N most likely candidates, where N (the rank) is plotted on the x-axis. The recognition result was averaged over ten runs, the cumulative match curves (CMCs) were plotted for scenarios a, b, and c as shown in Figs. 15-a, 15-b, and 15-c, respectively.

D. EXPERIMENT 2

The system was evaluated in various scenarios where facial and body information were available but at different times (i.e. biometric features were not available simultaneously). The biometric features were available at different times due to the limited range of sensors including occlusion of some biometric features due to human activities, the processing time needed for feature extraction module was variable and the biometric feature was dependent. In such scenarios, the person identification system may use all or some of these biometric features to finalize the recognition process. We designed two scenarios to replicate these situations: in the first scenario, the subject approached the robot from distance of 4 meters such that the body information was available before the facial information Fig. 16-a. The second scenario was designed to have facial information available before the body information. Hence, the subject was sitting on a chair close to the robot such that most of his skeleton was not detectable by the Kinect sensor but when he stood up, his skeleton become detectable as shown in Fig. 16-b. Ten trials for each member in the research lab (eight individuals) was performed in both scenarios. Cumulative match curves (CMCs) show the probability that the correct match of classification is found in the N most likely candidates, where N (the rank) is plotted on the x-axis. The recognition result was averaged over ten runs, the cumulative match curves (CMCs) were plotted for the two scenarios Fig. 17.

The system uses the threshold value as a compromise between the reliability of the recognition outcome and real-time response of the robot. Increasing the threshold value corresponds to giving more time for the spiking neurons to integrate more information encoded as spikes at their presynaptic inputs. The more time is allowed for recognition process; the more reliable outcome is achieved in recognition process. Therefore, in experiment 2, the threshold value of spiking neurons was increased to make sure both of facial and body information will contribute to the recognition outcome.



(a)

(b)

(c)

FIGURE 15. (a) CMC for scenario a. (b) CMC for scenario b. (c) CMC for scenario c.

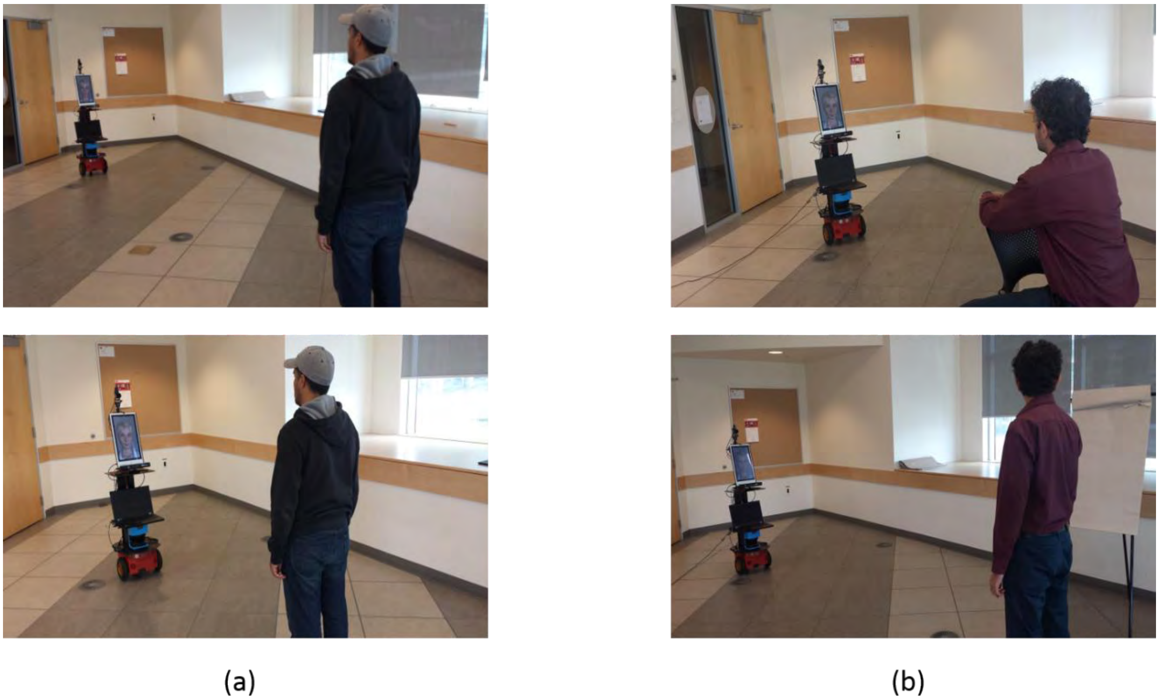


FIGURE 16. Examples of scenarios that used to evaluate the perceptual system; a) body information available before the face information, b) face information is available first then body information become available.

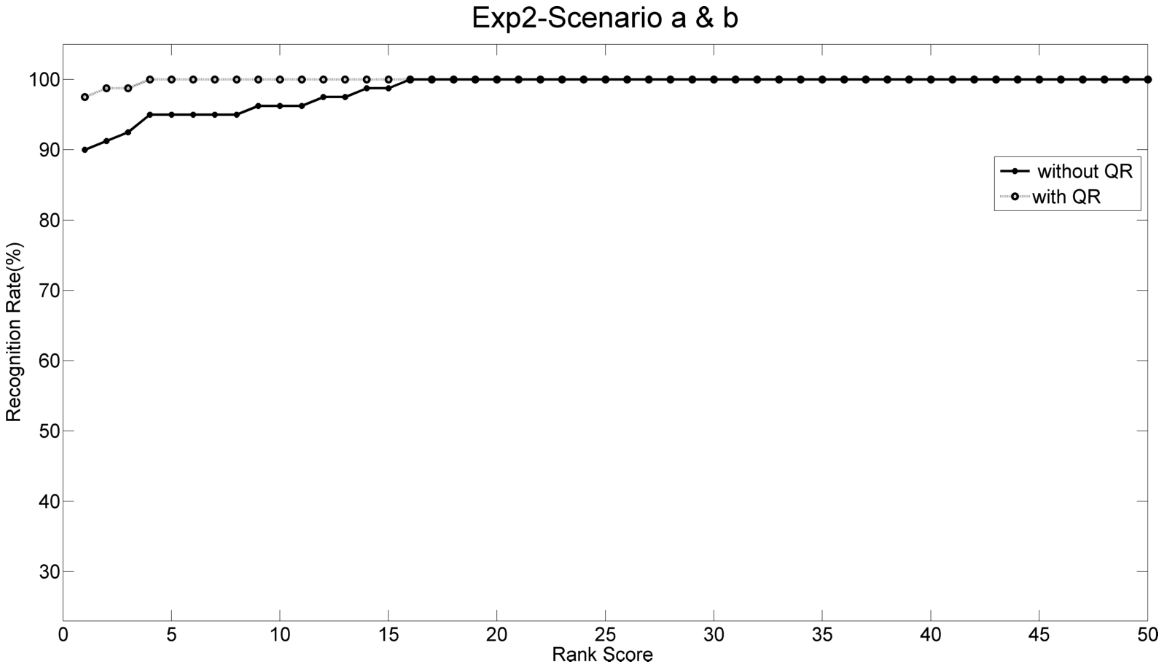


FIGURE 17. CMC for scenarios a and b.

E. DISCUSSION

The experimental studies suggest that the perceptual system recognizes a person in different social scenarios. The efficient processing is manifested as 1) exploiting the pre-knowledge and easy captured cues (i.e. QR code) about real-world environment to limit the search scope of the attended perceptual task and to emphasize the real-time fashion of social HRI, 2) using the outcome of processing a set of feature vectors, which is easy to be processed and available early in perception process (i.e. geometry-based feature vector), to bias limited number of spiking neurons and consequently reduce the computational cost, and 3) using various limited number

TABLE 1. Facial geometry-based feature vector.

$$Ratio1 = \frac{Nose\ \ width}{Nose\ \ height} = \frac{distance\ between\ H\ and\ I}{distance\ between\ G\ and\ K}$$

$$Ratio2 = \frac{eyes\ \ to\ \ nose\ \ distance}{eyes\ to\ chin\ \ distance} = \frac{distance\ between\ G\ and\ K}{distance\ between\ G\ and\ L}$$

$$Ratio3 = \frac{Face\ Width}{distance\ between\ \ the\ inner\_corner\ of\ the\ eyes} = \frac{distance\ between\ J\ and\ M}{distance\ between\ C\ and\ D}$$

$$Ratio4 = \frac{distance\ between\ \ cetners\ of\ eyes}{distance\ from\ \ midpoint\ between\ \ eyes\ \ to\ \ nose\ \ tip} = \frac{distance\ between\ E\ and\ B}{distance\ between\ G\ and\ K}$$

$$Ratio5 = \frac{Face\ \ \ width}{distance\ from\ the\ inner\ edge\ of\ \ the\ left\ eye\ to\ \ nose\ tip} = \frac{distance\ between\ J\ and\ M}{distance\ between\ C\ and\ K}$$

$$Ratio6 = \frac{distance\ between\ the\ inner\_corner\ of\ the\ eyes}{Nose\ width} = \frac{distance\ between\ C\ and\ D}{distance\ between\ H\ and\ I}$$

$$Ratio7 = \frac{The\ distance\ between\ the\ inner\_corner\ of\ the\ eyes}{The\ distance\ between\ the\ outer\_corner\ of\ the\ eyes} = \frac{distance\ between\ C\ and\ D}{distance\ between\ A\ and\ F}$$

$$Ratio8 = \frac{distance\ from\ the\ inner\ edge\ of\ the\ right\ eye\ to\ the\ nose\ tip}{distance\ from\ the\ outer\ edge\ of\ the\ right\ eye\ to\ the\ nose\ tip} = \frac{distance\ between\ K\ and\ D}{distance\ between\ K\ and\ F}$$

$$Ratio9 = \frac{distance\ from\ the\ inner\ edge\ of\ the\ right\ eye\ to\ the\ chin\ tip}{distance\ from\ the\ outer\ edge\ of\ the\ right\ eye\ to\ the\ chin\ tip} = \frac{distance\ between\ L\ and\ D}{distance\ between\ L\ and\ F}$$

Where $A, B, C, D, E, F, G, H, I, J, K, L, M$ are the selected facial points on a face image, as shown in Fig. 13.

TABLE 2. Anthropometric feature vector.

-
- Euclidean distance between floor and head.
- Euclidean distance between floor and neck.
- Euclidean distance between floor and left hip.
- Euclidean distance between floor and right hip.
- Mean of Euclidean distances of floor to right hip and floor to left hip.
- Euclidean distance between neck and left shoulder.
- Euclidean distance between neck and right shoulder.
- Mean of Euclidean distances of neck to left shoulder and neck to right shoulder.
- Ratio between torso and legs.
- Euclidean distance between torso and left shoulder.
- Euclidean distance between torso and right shoulder.
- Euclidean distance between torso and mid hip.
- Euclidean distance between torso and neck.
- Euclidean distance between left hip and left knee.
- Euclidean distance between right hip and right knee.
- Euclidean distance between left knee and left foot.
- Euclidean distance between right knee and right foot.
- Left leg length.
- Right leg length.
- Euclidean distance between left shoulder and left elbow.
- Euclidean distance between right shoulder and right elbow.
- Euclidean distance between left elbow and left hand.
- Euclidean distance between right elbow and right hand.
- Left arm length.
- Right arm length.
- Torso length.
- Height estimate.
- Euclidean distance between hip center and right shoulder.
- Euclidean distance between hip center and left shoulder.

of feature vectors that possesses complementary information to increase the reliability of the recognition rate. In some scenarios of social interaction, not all the biometric modalities of an attended subject are available as shown in experiment 1. Nevertheless, the system overcomes this challenge by reducing the threshold value and exploiting the available modality (body information) to derive the potential of one neuron to the set threshold. Consequently, a spike is fired and the recognition task is completed successfully. It can be noted from Fig. 15-a and Fig. 15-b that the recognition rate

TABLE 3. Body shape feature vector.

- Geodesic distance between left hip and left knee.
- Geodesic distance between right hip and right knee.
- Geodesic distance between torso center and left shoulder.
- Geodesic distance between torso center and right shoulder.
- Geodesic distance between torso center and left hip.
- Geodesic distance between torso center and right hip.
- Geodesic distance between right shoulder and left shoulder.
- Geodesic distance between left hip and left knee.
- Geodesic distance between right hip and right knee.
- Geodesic distance between torso center and left shoulder.
- Geodesic distance between torso center and right shoulder.
- Geodesic distance between torso center and left hip.
- Geodesic distance between torso center and right hip.
- Geodesic distance between right shoulder and left shoulder.

is degraded when a subject's body is angled relative to the camera view as in scenario b. This degradation in recognition rate is due to the fact that the body shape and anthropometric feature vector, which extracted from side view of a subject in scenario b, contain lower discriminant level. In a case when higher reliability is needed, but the current available modality is not sufficient to drive the system to the required threshold to finalize the identification process, the system waits for more sensory information to be available and to be combined with other modalities to satisfy the required threshold (experiment 2). The results show that the system achieves high recognition rate in real-time fashion despite the fact that not all biometric modalities are concurrently available.

V. CONCLUSION

We have proposed a perceptual system inspired by the human brain. It is a modest and crude attempt to reverse engineering a neuroscience model of human perceptual system. The proposed architecture and its sub-systems are explained and viable computing models are introduced. The main contributions can be summarized as: 1) Introducing a large-scale end-to-end social robotic perceptual architecture as depicted in Fig.2. The architecture is inspired by exploiting the structure of multisensory information processing through the hierarchical structure of sensory cortex and the intermediate representations of DPUs' outputs, 2) suggesting a unified and functional plausible methodology for multisensory information processing by integrating the outputs of DPUs over the perception time using temporal binding with fading memory, 3) presenting a top-down influence on multisensory information processing, which is depicted in two types of connections; the feedback projection from fast processing routes to slower ones in order to refine the scope of search for the best candidates and the lateral connection among spiking neurons at different hierarchical stages, as a significant factor in reducing the computational cost of the proposed model and in limiting the search scope for stimulus candidate, 4). The key findings in neuroscience, psychology, and psychophysics about human perceptual process and human channel capacity of processing information are compiled and reformed as design guidelines for general purpose human-oriented perceptual system. The top-down influences feature, which is essential feature of the multi-modal perceptual system was implemented as a means of improving the perception speed and hence facilitating social human-robot interaction. We also investigated the effectiveness of using QR codes as identification signs containing implicit knowledge about the environment. The architecture was implemented on an in-house designed social robots and we reported extensive experimental studies.

APPENDIX

See Tables 1–3.

REFERENCES

[1] A. Niculescu, B. Van Dijk, A. Nijholt, and D. K. Limbu, "Socializing with Olivia, the youngest robot receptionist outside the lab," in *Proc. Int. Conf. Soc. Robot.*, 2010, pp. 50–62.

[2] R. Q. Stafford, B. A. MacDonald, C. Jayawardena, D. M. Wegner, and E. Broadbent, "Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot," *Int. J. Soc. Robot.*, vol. 6, no. 1, pp. 17–32, Apr. 2013.

[3] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "How quickly should a communication robot respond? Delaying strategies and habituation effects," *Int. J. Soc. Robot.*, vol. 1, no. 2, pp. 141–155, Feb. 2009.

[4] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: A review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, Jun. 2009.

[5] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Hum.-Comput. Interact.*, vol. 19, no. 1, pp. 61–84, Jun. 2004.

[6] F. Michaud *et al.*, "Socially interactive robots for real life use," in *Proc. Workshop Mobile Robot Competition, Amer. Assoc. Artif. Intell. Conf.*, 2006, pp. 45–52.

[7] M. Goodrich and A. Schultz, "Human-robot interaction: A survey," *Found. Trends Human Comput. Interact.*, vol. 1, no. 3, pp. 203–275, 2008.

[8] H. Yan, M. H. Ang, Jr., and A. N. Poo, "A survey on perception methods for human–robot interaction in social robots," *Int. J. Soc. Robot.*, vol. 6, no. 1, pp. 85–119, Jul. 2013.

[9] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auto. Syst.*, vol. 42, no. 3, pp. 143–166, 2003.

[10] A. Treisman, "The binding problem," *Current Opinion Neurobiol.*, vol. 6, no. 2, pp. 171–178, 1996.

[11] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Comput.*, vol. 14, no. 11, pp. 2531–2560, Nov. 2002.

[12] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Rev. Neurosci.*, vol. 14, no. 5, pp. 350–363, 2013.

[13] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. Mahwah, NJ, USA: Psychology Press, 2002.

[14] A. R. Damasio, "The brain binds entities and events by multiregional activation from convergence zones," *Neural Comput.*, vol. 1, no. 1, pp. 123–132, 1989.

[15] M. Al-Qaderi and A. Rad, "A multi-modal person recognition system for social robots," *Appl. Sci.*, vol. 8, no. 3, p. 387, 2018.

[16] S. Jahfari, K. R. Ridderinkhof, and H. S. Scholte, "Spatial frequency information modulates response inhibition and decision-making processes," *PLoS ONE*, vol. 8, no. 10, p. e76467, Jan. 2013.

[17] K. Amano, N. Goda, S. Nishida, Y. Ejima, T. Takeda, and Y. Ohtani, "Estimation of the timing of human visual perception from magnetoencephalography," *J. Neurosci.*, vol. 26, no. 15, pp. 3981–3991, Apr. 2006.

[18] C. F. Cadieu *et al.*, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS Comput. Biol.*, vol. 10, no. 12, p. e1003963, Aug. 2014.

[19] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[20] K. Rajaei, S.-M. Khaligh-Razavi, M. Ghodrati, R. Ebrahimpour, and M. E. S. A. Abadi, "A stable biologically motivated learning mechanism for visual feature extraction to handle facial categorization," *PLoS ONE*, vol. 7, no. 6, p. e38478, 2012.

[21] R. Chellappa, P. Sinha, and P. J. Phillips, "Face recognition by computers and humans," *Computer*, vol. 43, no. 2, pp. 46–55, Feb. 2010.

[22] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003.

[23] J. M. Susskind, G. Littlewort, M. S. Bartlett, J. Movellan, and A. K. Anderson, "Human and computer recognition of facial expressions of emotion," *Neuropsychologia*, vol. 45, no. 1, pp. 152–162, Jan. 2007.

[24] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[25] P. M. Milner, "A model for visual shape recognition," *Psychol. Rev.*, vol. 81, no. 6, pp. 521–535, Nov. 1974.

[26] S. Grossberg, "Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions," *Biol. Cybern.*, vol. 23, no. 4, pp. 187–202, 1976.

[27] C. von der Malsburg, "The correlation theory of brain function," in *Models of Neural Networks*. New York, NY, USA: Springer, 1994, pp. 95–119.

[28] C. M. Gray and W. Singer, "Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex," *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 5, pp. 1698–1702, Mar. 1989.

[29] W. Singer, "Neuronal synchrony: A versatile code for the definition of relations?" *Neuron*, vol. 24, no. 1, pp. 49–65, Sep. 1999.

[30] A. K. Engel, P. Fries, and W. Singer, "Dynamic predictions: Oscillations and synchrony in top–down processing," *Nature Rev.*, vol. 2, no. 10, pp. 704–716, 2001.

[31] H. R. Clemo, L. P. Keniston, and M. A. Meredith, "Structural basis of multisensory processing: Convergence," in *The Neural Bases of Multisensory Processes*. Boca Raton, FL, USA: CRC Press, 2011, pp. 3–14.

[32] L. M. Romanski, "Convergence of auditory, visual, and somatosensory information in ventral prefrontal cortex," in *The Neural Bases of Multisensory Processes*. Boca Raton, FL, USA: CRC Press, 2011, pp. 667–682.

[33] U. Noppeney, "Characterization of multisensory integration with fMRI: Experimental design, statistical analysis, and interpretation," in *The Neural Bases of Multisensory Processes*. Boca Raton, FL, USA: CRC Press, 2011, pp. 233–252.

[34] T. W. James and R. A. Stevenson, "The use of fMRI to assess multisensory integration," in *The Neural Bases of Multisensory Processes*. Boca Raton, FL, USA: CRC Press, 2011, pp. 131–146.

[35] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, no. 2, pp. 81–97, Mar. 1956.

[36] C. Wallraven, A. Schwaninger, and H. H. Bülthoff, "Learning from humans: Computational modeling of face recognition," *Netw. Comput. Neural Syst.*, vol. 16, no. 4, pp. 401–418, Jun. 2005.

[37] W. Maass and H. Markram, "On the computational power of circuits of spiking neurons," *J. Comput. Syst. Sci.*, vol. 69, no. 4, pp. 593–616, Dec. 2004.

[38] W. Maass, T. Natschläger, and H. Markram, "A model for real-time computation in generic neural microcircuits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 229–236.

[39] R. VanRullen and S. J. Thorpe, "The time course of visual processing: From early perception to decision-making," *J. Cogn. Neurosci.*, vol. 13, no. 4, pp. 454–461, May 2001.

[40] D. Nikolić, S. Häusler, W. Singer, and W. Maass, "Distributed fading memory for stimulus properties in the primary visual cortex," *PLoS Biol.*, vol. 7, no. 12, p. e1000260, Dec. 2009.

[41] A. Belatreche, L. P. Maguire, and M. McGinnity, "Advances in design and application of spiking neural networks," *Soft Comput.*, vol. 11, no. 3, pp. 239–248, Feb. 2007.

[42] S. Schliebs and N. Kasabov, "Evolving spiking neural network—A survey," *Evol. Syst.*, vol. 4, no. 2, pp. 87–98, Jun. 2013.

[43] V. Goffaux, B. Hault, C. Michel, Q. C. Vuong, and B. Rossion, "The respective role of low and high spatial frequencies in supporting configural and featural processing of faces," *Perception*, vol. 34, no. 1, pp. 77–86, 2005.

[44] H. Halit, M. de Haan, P. G. Schyns, and M. H. Johnson, "Is high-spatial frequency information used in the early stages of face detection?" *Brain Res.*, vol. 1117, no. 1, pp. 154–161, Jun. 2006.

[45] S. Schliebs and N. Kasabov, "Computational modeling with spiking neural networks," in *Springer Handbook of Bio-/Neuroinformatics*. Berlin, Germany: Springer, 2014, pp. 625–646.

[46] A. J. King, "Multisensory integration: Strategies for synchronization," *Current Biol.*, vol. 15, no. 9, pp. R339–R341, 2005.

[47] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–741, May 1995.

[48] W. Maass, T. Natschläger, and H. Markram, "Computational models for generic cortical microcircuits," in *Computational Neuroscience: A Comprehensive Approach*, vol. 18. London, U.K.: Chapman & Hall, 2004, pp. 575–605.

[49] R. G. Leonard and G. R. Doddington, *TIDIGITS LDC93S10*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[50] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[51] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *Proc. 1st Int. Workshop Re-Identificat. (Re-Id)*, Florence, Italy, 2012, pp. 433–442. [Online]. Available: https://www.iit.it/research/lines/pattern-analysis-and-computer-vision/pavis-datasets/534-rgb-d-person-re-identification-dataset

[52] Á. Serrano, I. M. de Diego, C. Conde, and E. Cabello, "Analysis of variance of Gabor filter banks parameters for optimal face recognition," *Pattern Recognit. Lett.*, vol. 32, no. 15, pp. 1998–2008, Nov. 2011.

[53] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

MOHAMMAD K. AL-QADERI received the B.Sc. and M.Sc. degrees in mechanical/mechatronics engineering from the Jordan University of Science and Technology. He is currently pursuing the Ph.D. degree in mechatronics engineering with Simon Fraser University, Canada. He was with many institutes as a research assistant and an instructor in the area of mechanical/mechatronics engineering. He is conducting his research at the Autonomous and Intelligent Systems Laboratory. His research goal is to enhance the ability of social robot to interact socially with its environments. His research interest is brain-inspired perceptual system for social robots, social human–robot interaction, and social robotics.

AHMAD B. RAD (M'99–SM'02) is currently a Professor with the School of Mechatronic Systems Engineering, Simon Fraser University, Surrey Campus, Surrey, BC, Canada. His current research interests include autonomous systems, robotics, machine perception and learning, and applications of soft computing in modeling and control.