Int J of Soc Robotics (2015) 7:601–616
DOI 10.1007/s12369-015-0315-x

A Reference Architecture for Social Head Gaze Generation in Social Robotics

Vasant Srinivasan[1] · **Robin R. Murphy**[1] · **Cindy L. Bethel**[2]

Accepted: 9 August 2015 / Published online: 23 August 2015

Abstract This article outlines a reference architecture for social head gaze generation in social robots. The architecture discussed here is grounded in human communication, based on behavioral robotics theory, and captures the commonalities, essence, and experience of 32 previous social robotics implementations of social head gaze. No such architecture currently exists, but such an architecture is needed to: (1) serve as a template for creating or re-engineering systems, (2) provide analyses and understanding of different systems, and (3) provide a common lexicon and taxonomy that facilitates communication across various communities. A constructed reference architecture and the Software Architecture Analysis Method (SAAM) are used to evaluate, improve, and re-engineer two existing head gaze system architectures (Human–Robot Collaboration architecture and Robot Behavior Toolkit architecture). SAAM shows that no existing architecture incorporated the summation of functionalities found in the 32 studies. SAAM suggests several architectural improvements so that the two existing architectures can better support adaptation to a new environment and extension of capability. The resulting reference architecture guides the implementation of social head gaze in a rescue robot for the purpose of victim management in urban search and rescue (US&R). Using the proposed reference architecture will benefit social robotics because it will simplify the principled implementations of head gaze generation and allow for comparisons between such implementations.

Keywords Reference architecture · Social head gaze · Behavioral robotics · Human–robot interaction

✉ Vasant Srinivasan
vasant.s@gmail.com

Robin R. Murphy
murphy@cse.tamu.edu

Cindy L. Bethel
cbethel@cse.msstate.edu

[1] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA

[2] Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762, USA

1 Introduction

Social head gaze is a complex system of coordinated head movements with the speech signal, in order to convey different communicative and nonverbal functions in an interaction [3]. The five main functions of social head gaze include three communicative functions—*look interested in human(s)*, *engage in a verbal conversation with human(s)*, and *gaze at objects in the environment*—and two nonverbal functions—*convey general liveliness and awareness* and *show various mental states*—are key components for an effective and natural social interaction between a robot and human(s). The importance of social head gaze to human–robot interaction is evident in it being proposed as a metric to evaluate the quality of different human–robot interactions [59].

A *reference architecture* captures the commonalities, essences, and experiences of various systems through a process of mining and generalization of their implementations [13,14]. It provides template solutions that guide specific implementations [14]. Currently, there is no reference architecture that provides a standard structure to aid in the understanding, analysis, and comparison of different head gaze generation systems. Existing architectures are typically *system architectures*, and describe the elements of a system and the relationships between those elements [13].

They do not provide broad coverage, templates, or a common lexicon and taxonomy specific to social head gaze generation. This article synthesizes a reference architecture for social head gaze generation from 32 studies by applying the methodology described in [23,35]. The reference architecture provides a computational mapping between different communicative and nonverbal functions of human social head gaze to the expression of one or more discrete robot head gaze actions (range, speed, and frequency). It embodies architectural best practices gathered from the design and development of various robotics applications that implement social head gaze, such as: healthcare [24,37], victim management [8,19], robot guides [7,38,50,62,70], entertainment [10,26,28,29,44,49], telepresence [39,63], and fundamental research [2,31–33,40,57,61,67].

The primary contribution of this article is a reference architecture for head gaze generation in robotic systems and the analyses of two existing architectures (Human–Robot Collaboration Architecture and Robot Behavior Toolkit) for *overall functionality*, *adaptation to a new environment*, and *extension of capability*. The proposed reference architecture is intended to provide guidance for the implementation of head gaze generation in social robotics. It provides a blueprint for the development of new social head gaze generation systems or for reverse-engineering existing systems. The templates provided by the blueprint help with system understanding and improve system reuse. They provide a broad coverage and wide applicability because they embody the knowledge gleaned from human head gaze research and 32 previous implementations of head gaze generation for social robotics. When used in combination with the five step Software Architecture Analysis Method (SAAM) [13] described in Sect. 4, it is a valuable tool for analyzing, comparing, and improving existing head gaze generation systems. Other software architecture analysis methods such as Tiny Architecture Review Approach (TARA) or Architecture Trade-Off Analysis Method (ATAM) can be used for evaluating the existing head gaze generation systems, however SAAM was chosen because it is the most mature and complete software architecture analysis method. The other methods are still young and are undergoing refinement and improvement [16]. Additionally, SAAM provides high validity and repeatability of results for different quality attributes like modifiability, performance, robustness, and portability in several different domains [16]. The synthesized reference architecture generalizes common functions and structures to provide a common lexicon and taxonomy that facilitates communication across diverse communities such as social scientists interested in understanding fundamental aspects of the social head gaze phenomena, or robot behavior designers/practitioners who need to implement head gaze elements in a specific application.

The scope of this work is to address social head gaze in robotic systems for dyadic interactions. The exclusion of eye gaze and support for multi-party interaction is due to the higher complexity introduced by computational models for eye gaze generation or multi-party interaction. Head gaze for dyadic scenarios has significant value without the need for eye gaze or support for multi-party interaction, and is additionally widely applicable. The extension of the reference architecture to include both head gaze and eye gaze generation in a multi-party scenario is a viable area for future work.

The article is organized as follows: Section 2 identifies the metrics for coverage in the reference architecture of the five main social head gaze functions from human communication and discusses two problems in implementing the models of these head gaze functions from human communication on robotic platforms. This section discerns six existing system architectures from 32 articles in human–robot interaction and shows that no existing system architecture meets the requirements to be characterized as a reference architecture. Section 3 discusses the derivation of a reference architecture using the lens of behavioral robotics theory [6,48], the knowledge from human head gaze research, and 32 articles on robot social head gaze discussed in Sect. 2. Section 4 introduces the Software Architecture Analysis Method (SAAM), which is the first documented and most popular software architecture analysis method used in software engineering, which uses a five-step approach to evaluate a system architecture [13]. This section illustrates how a combination of the proposed reference architecture and SAAM can be applied to analyze two existing architectures selected for their *overall functionality* and *modifiability*. Based on the analysis, Sect. 4 provides suggestions for architectural improvements so that the two existing architectures can better support *adaptation to a new environment* and *extension of capability*. Section 5 presents an instantiation of the reference architecture on a rescue robot for victim management. Section 6 discusses the evaluation of the reference architecture implementation in a large-scale user study. Section 7 provides a summary of the contributions of this research and a possible body of future work.

2 Related Work

The reference architecture serves as a standard for building head gaze systems for social robotics, hence it should be grounded in human communication, and consistent with Nass et al.'s Computers are Social Actors (CASA) model, where humans treat computers as if they were human [52]. Therefore, the reference architecture should provide broad coverage that captures the five main communicative and nonverbal functions of human social head gaze discussed in Sect. 2.1. Section 2.2 evaluates six existing system archi-

tectures identified from a review of 32 articles from the human–robot interaction literature [1,2,4,7,8,10,19,24,26,28–34,37–40,44,46,49,50,53,57,58,61–63,67,70] to determine if they can be characterized as a reference architecture. These 32 articles capture atleast.

2.1 Metrics for Coverage in the Reference Architecture of Social Head Gaze for Social Robots

The reference architecture must express the broad representation of communicative and nonverbal functions found in any human–human interaction. Therefore, the metrics for coverage in the reference architecture are the five main human social head gaze functions—*look interested in human(s)* [5,36], *engage in a verbal conversation with human(s)* [5,11,17,33,50,56], *gaze at objects in the environment* [20,43], *show various mental states* [18,36], and *convey general liveliness and awareness* [5,36]. The first three human social head gaze functions are communicative functions best understood computationally. The function *look interested in human(s)* is driven by salient stimuli, such as when another human enters the social zone [5] of the speaker, or if the other human shows initial interest. Head gaze supports speech in *engage in a verbal conversation with human(s)* function to communicate syntactic signals such as verbal utterances, accentuation, and emphasis [12,18,21,47]. For example, during situated communication among humans, a human might direct gaze not only at the face of the human with whom they are communicating, but also other humans in the environment [17,56]. Head gaze is used to direct *gaze at objects in the environment* if the stimuli is a salient object in the environment, such as a fast moving ball [36] or if the topic of the discussion is an object in the environment [20,43]. The latter case is known as *referential gaze* and the human fixates toward the object 800 m s to 1 s, before the utterance of the object's name [20,43].

The remaining two human social head gaze functions *show various mental states* [18,36] and *convey general liveliness and awareness* [5,36] convey nonverbal information and are critical to the performance of the reference architecture. However, their functions and dynamics have been studied much less than co-verbal, facial, and other non-verbal signals [3]. Head gaze patterns can *show various mental states*. Speakers tend to look at listeners more when they intend to be more persuasive, deceptive, ingratiating, or assertive [36,49]. Additionally, gestures such as head nods and head shakes have been found to be used for showing mental states, such as emotions [18] or intention [34,53]. The *convey general liveliness and awareness* function is employed for idle looking behaviors, i.e., when there are no tasks at hand or to interrupt tasks. Head movements used to scan the environment or slight head shakes and nods help indicate the person is alive [5,36].

While the human communication literature provides metrics for coverage in the reference architecture, it does not provide direct guidance for practical implementation details for robots in two important areas. First, the mapping of human social head gaze functions to the expression of one or more discrete robot head gaze actions (range, speed, and frequency) will differ. This is because of hardware constraints in robots, such as (a) robots have fewer degrees of freedom, (b) upper and lower limits for acceleration and the speed of the robot's movements, and (c) lag in reactions to motor commands due to physical inertia and communication latency [55]. Second, there is a lack of methodologies in the human communication literature to integrate different communicative and nonverbal functions to resolve any resulting collisions. Therefore, a reference architecture is constructed in Sect. 3 using the implementation of these five important human social head gaze functions into human–robot interaction.

2.2 Characterization of Existing System Architectures in Human–Robot Interaction as a Social Head Gaze Reference Architecture

A review of 32 articles [1,2,4,7,8,10,19,24,26,28–33,37–40,44,46,49–51,53,57,58,61–63,67,70] from the human–robot interaction literature that addressed some aspect of social head gaze generation revealed that no existing system architecture can be established as a reference architecture. A two step process was followed in this analysis:

1. *Identify system architectures for head gaze generation*: The review revealed eight studies that describe six system architectures used for head gaze generation. The six system architectures described are—Linta-III [32], extension of the C5M architecture [10], an architecture to initiate and maintain engagement [62], an architecture for multi-modal interaction [7], the Human–Robot Collaboration architecture [26], and the Robot Behavior Toolkit architecture [28–30].
2. *Evaluate the system architecture to determine if it meets the coverage requirements to be characterized as a reference architecture (as listed in Sect. 2.1)*: As shown in Table 1, no system architecture competently integrates all five main human social head gaze functions into a single robotics implementation. Two system architectures—Linta III [32] and the extension of the C5M architecture [10]—do not generate head gaze for the *engage in a verbal conversation* human social head gaze function. The importance of implementation of this human social head gaze function in social robotics is well established for achieving rapport-related outcomes in measures such as Robot Likeability [49,50], Attentiveness of the Human to the Robot [37,62], Perceived Intelligence [49,50], Empa-

Table 1 The human social head gaze functions implemented by six existing system architectures in human–robot interaction

Studies	System architecture	LIH	EVCH	GOE	SVMS	CGLA
[32]	Linta III	X		X		
[10]	Extension of the C5M architecture	X		X	X	X
[7]	Architecture for multi-modal interaction	X	X	X		
[62]	Architecture to initiate and maintain engagement	X	X	X		
[26]	Human–Robot Collaboration Architecture	X	X	X		
[28–30]	Robot Behavior Toolkit Architecture	X	X	X		

LIH look interested in human(s), *EVCH* engage in verbal conversation with human(s), *GOE* gaze at objects in the environment, *SVMS* show various mental states, and *CGLA* convey general liveliness and awareness

thy [50,57], Groupness [50], and Positive Emotional State [50,62]. Five out of the six architectures—Linta-III [32], architecture to initiate and maintain engagement [62], architecture for multi-modal interaction [7], the Human–Robot Collaboration architecture [26], and the Robot Behavior Toolkit architecture [28–30]—do not generate head gaze for two important human social head gaze functions—*show various emotional states* and *convey general liveliness and awareness*. These two functions have been shown to be very important in human–robot interaction for positive outcomes in measures such as Effectiveness of Task [10], Human's Mental Model of the Robot [10], Attribution of Intentionality [24,34,53], and Social Presence [24].

The review identifies two gaps in the design of existing system architectures that the proposed reference architecture aims to address. First, no existing system architecture provides guidance on the integration and coordination of the five human social head gaze functions for social robotics. The maximum number of head gaze behaviors integrated by a system architecture in a single implementation is four [10]. This research synthesizes a reference architecture in Sect. 3.3 and provides specific recommendations on the integration and coordination of all five human social head gaze functions. Section 6 presents results from an implementation that competently integrates all five human social head gaze functions for a rescue robot used in victim management. Second, there are no common terminologies used for describing the human social head gaze functions in the six system architectures. Therefore, it becomes difficult to compare and share implementations. The reference architecture for social head gaze normalizes the different nomenclatures from the 32 human–robot interaction studies as a part of the Commonality Analysis step in Sect. 3.2.

3 Approach

Using the two step methodology outlined in [23,35] and described below, a reference architecture for social head gaze was constructed. First, *conceptual architectures* were derived for each of the 32 previous implementations using behavioral robotics theory [6,48] as the common framework. The constructed reference architecture is based on behavioral robotics theory because it has been shown to be capable of expressing ethological and robotic concepts, and it is consistent with good software engineering principles such as modularity and extensibility [6,48]. Conceptual architectures are abstract representations of subsystems and inter-subsystem relations, not specific procedures or variables [23]. Second, a *commonality analysis* [69] was employed to synthesize shared elements between the resulting conceptual architectures to form a reference architecture.

3.1 Construction of Conceptual Architectures

In order to apply a commonality analysis, each of the 32 prior implementations of social head gaze had to be expressed using a single framework. A behavioral robotics framework, also called programming by behavior, was selected. The key construct in behavioral robotics is a behavior b, which maps a percept s onto an act r [6]. An agent may have multiple behaviors active at the same time therefore, the combined observable response is given as $\rho = C(G \times B(S))$, where B is a vector of behaviors, S is a vector of sensed percepts, G is a vector of the gain functions, and C is the coordination function that determines the overall response ρ. The strength of act r may be modified by a gain G, which may amplify or reduce the contribution of an individual behavior to the overall behavior. Examples of gains in social head gaze are covariate factors such as culture, gender, or proxemics, which are identified to have a significant influence on head gaze [25,27,49,50,68]. Table 2 illustrates a social head gaze example using behavioral robotics terminologies; the terminologies from ethology are also included for readers familiar with that field. For example, while communicating social attention to a human, a robot should consider gender when determining the amount of time it will fixate [49]. Here the observable response is $\rho = C(G(gender) \times \text{COMMUNICATING SOCIAL ATTENTION } (human))$. If there are no behaviors other than

Table 2 Common terminology in the ethology and behavioral robotics communities [48] illustrated with a social head gaze example

	Common terms in the ethology community	Common terms in the behavioral robotics community	Example from social head gaze
b	Behavior	Behavior	Communicating social attention
s	Stimulus	Percept	Human shows initial interest
r	Response	Act	Fixate
G	Gain	Gain	Culture, gender, and proxemics
C	Coordination function	Coordination function	Arbitration by prioritization
ρ	Overall response	Overall response	Fixate

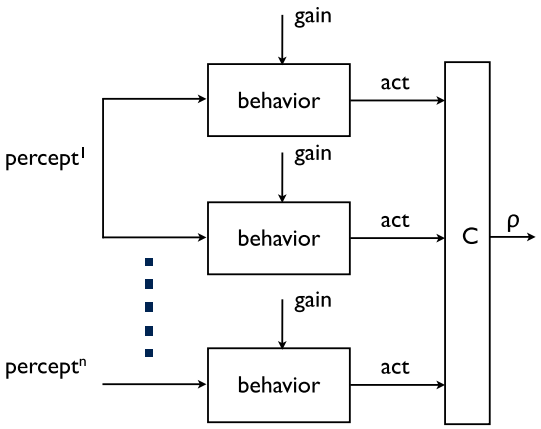

Fig. 1 Example of a conceptual architecture in which gains impact the behaviors and the resulting acts are passed through a coordination function

COMMUNICATING SOCIAL ATTENTION executing, then the observable overall response ρ will be *fixate*.

As the first step toward construction of the reference architecture, this work derived conceptual architectures for each of the 32 previous implementations of head gaze using the behavioral robotics notations described above. The conceptual architectures are comprised of behaviors, percepts, acts, gains, and coordination functions. Figure 1 provides a diagrammatic representation of the conceptual architectures.

3.2 Commonality Analysis

A commonality analysis is an analytical technique used to determine the components of an architecture [69]. It helps identify the domain concepts that represent common elements of the domain at its highest level of abstraction; it is also useful for normalizing existing notations produced by previous implementations.

The 32 conceptual architectures derived previously were iteratively analyzed to identify and create useful abstractions common to all conceptual architecture components. The nomenclature was then standardized. For example, *Avert* [2,61], *Look Away* [11,26,49,50], and *Avoid Gaze* [70] all corresponded to the same head gaze act; hence, that head gaze act was standardized to Avert. This step ensured that implementations with different overall functionality, environments, and robot types were taken into consideration and supported. The commonality analysis identified three types of percepts, six head gaze acts, five behaviors, and one coordination function based on prioritization. None of the 32 studies reported implementations using gain parameters such as gender or culture to actively influence the generation of head gaze. Henkel et al. [25] reports an implementation that uses gains based on proxemics to influence head gaze behaviors generated using the proposed reference architecture.

Three types of percepts were identified in the commonality analysis: *external*, *linguistic*, and *internal*.

1. *External* percepts are visible states of the external world. They typically require inference and/or interpretation of sensor data. For example, such percepts include *Presence of Human* [8,10,19,24,26,28,29,31,37,38,40,44,45,50,57,61,62,70], *Human Shows Initial Interest* [8,10,19,24,26,28,29,31,37,38,40,44,45,50,57,61,62,70], *Listening to Human* [24,26,33,62,63], and *Presence of Object* [1,10,26,28,29,32,46,58,62,66,67].
2. *Linguistic* percepts occur in the robot's dialog (text or audio). For example, such percepts include *Start of Turn* [4,11,26,28–30,33,38,49,50,57,70], *Middle of Turn* [49,50], *End of Turn* [4,11,26,28–30,33,38,49,50,57,70], *First Word in Theme* [49,50], *First Word in Rheme* [49,50], and *Onset of Speech Utterance* [1,10,26,28,29,32,46,58,62,66,67]. The theme specifies the topic of a sentence, while the rheme specifies what is new or interesting about the topic [22].
3. *Internal* percepts are self-perception of an internal state of the robot. This is based on the robot's beliefs about people and objects in the world. For example, such percepts include *Internal State*$_{liveliness}$ [7,10,37], *Internal State*$_{mentalstate}$ [7,10,37], and *Internal State*$_{acknowledge}$ [24,26,33,62].

Head gaze acts are "head" movements used to generate a social head gaze. Of the six head gaze acts described below, three (*fixate*, *avert*, and *concurrence*) are considered computational primitives, and the others (*short glance*, *confusion*, and *scan*), are considered as compound head gaze acts. The compound head gaze acts consist of one or more primitive head gaze acts and are used to convey functions, which are different or cannot be captured by primitive head gaze act nomenclature.

1. *Fixate* is a head gaze that persists on a target person, object, or location in space (e.g., space between two humans for social attention [31]). If the person or object is moving, fixation tracks and maintains a gaze with the target [10,26,28,29,31,38,40,44,45,49,50,57,61,62,70].
2. *Avert* is a head gaze away from a person or a look away from the person toward the environment [11,26,28,29,33,38,49,50,57,70].
3. *Back-Channel Head Gaze Act* is any head gaze act used to signal concurrence or disagreement that follows the expression of opinions, evaluations, and planning [36]. *Concurrence* is a repetitive vertical movement of the head, which interrupts fixation [26,62]. Head nodding was used only in conjunction with fixation. *Disagreement* is indicated by turning the head from side to side [36]. However, this head gaze act has not been defined and used in the social robotics literature.
4. *Short Glance* is a fixation persisting for a short duration [7,50]. Short glances are often used in a multi-party situation when the robot needs to acknowledge the presence of bystanders or other conversants.
5. The *Mental State Head Gaze Act* is any head gaze act used to express mental states such as emotions or intentions. One example of a mental state head gaze act is *Confusion*, which is a series of rapid shifts back and forth accompanied by a roll of the head for amplification [10]. Other mental states (such as emotions like Happy, Sad, or Surprise) may require additional head gaze acts.
6. *Scan* is a short glance to a series of random points in space [62].

Five social head gaze behaviors were identified:

1. COMMUNICATING SOCIAL ATTENTION is a behavior where head gaze is used by robots to look interested in humans[8,10,19,24,26,28,29,31,37,38,40,44,45,50,57,61,62,70]. This behavior maps an *external* percept *Human Shows Initial Interest* on to the *fixate* head gaze act. This behavior is initiated at the beginning of an interaction or if the robot is not capable of speech.
2. REGULATING AN INTERACTION is a behavior where head gaze is used for engaging in a conversation[11,26,28,29,33,38,49,50,57,70]. This behavior maps combinations of *linguistic* and *external* percepts on to the *fixate*, *avert*, or *concurrence* head gaze acts. The linguistic percepts facilitate turn-taking and are as follows: *Start of Turn*, *Middle of Turn*, *End of Turn*, *First Word in Theme*, and *First Word in Rheme*. The external percept *Listening to Human* and internal percept $Internal\ State_{acknowledge}$ activate back-channeling.
3. MANIFESTING AN INTERACTION is a behavior where head gaze is used to direct attention towards objects in the environment [10,26,28,29,32,62,66,67]. The behavior maps *external* and *linguistic* percepts onto the *fixate* head gaze act. A combination of the *external* percept *Presence of Object* and the *linguistic* percept *Onset of Object Utterance* facilitates *referential gaze*.
4. PROJECTING MENTAL STATE is a behavior where head gaze is used for showing various mental states, such as emotions. This behavior maps an *internal* percept such as $Internal\ State_{mentalstate}$ onto the head gaze act for a mental state such as confusion [10], emotions such as happiness, sadness, surprise, etc [37], or to show the robot's intentions to humans such as head gaze at a location in space in which the robot is going to act [34,53]. The internal state of the robot can be set based on its beliefs about people and objects in the world.
5. ESTABLISHING AGENCY is a behavior where a head gaze is used to convey general liveliness and awareness [10,24,62]. This behavior maps an *internal* percept such as $Internal\ State_{liveliness}$ onto the *scan* head gaze act.

A coordination function fuses the responses of multiple active behaviors [6,48]. The existing system architectures [26,28] use a competitive arbitration method based on the prioritization of behaviors to select a single overall response.

3.3 Resulting Reference Architecture

Following the commonality analysis, a reference architecture was finalized as shown in Fig. 2. The components of the architecture are grouped into *Perceptual Schemas*, *Behaviors*, *Motor Schemas*, and a *Coordination Function*. This grouping of the components is as suggested by Murphy [48] and Arkin [6].

Perceptual Schemas have at least one method that takes sensor input and transforms it into a data structure called a percept [48]. Perceptual schemas are used to generate the *external*, *linguistic*, and *internal* percepts. Table 3 enumerates a list of nine possible perceptual schemas to generate the percepts. The perceptual schemas can share the same sensors. For example, the *detect_human* perceptual schema shares the sensor data from the webcam with the *detect_object* perceptual schema. Additionally, the computational processes in the behaviors can share the percepts created by the perceptual schemas. The *Presence of Human* percept is shared between COMMUNICATING SOCIAL ATTENTION and REGULATING AN INTERACTION behaviors.

The reference architecture consists of five behaviors: COMMUNICATING SOCIAL ATTENTION, REGULATING AN INTERACTION, MANIFESTING AN INTERACTION, P- ROJECTING MENTAL STATE, and ESTABLISHING AGENCY. These behaviors become active when their corresponding percept is detected. The behaviors are transformation units; they map the percept to an appropriate head gaze act. The

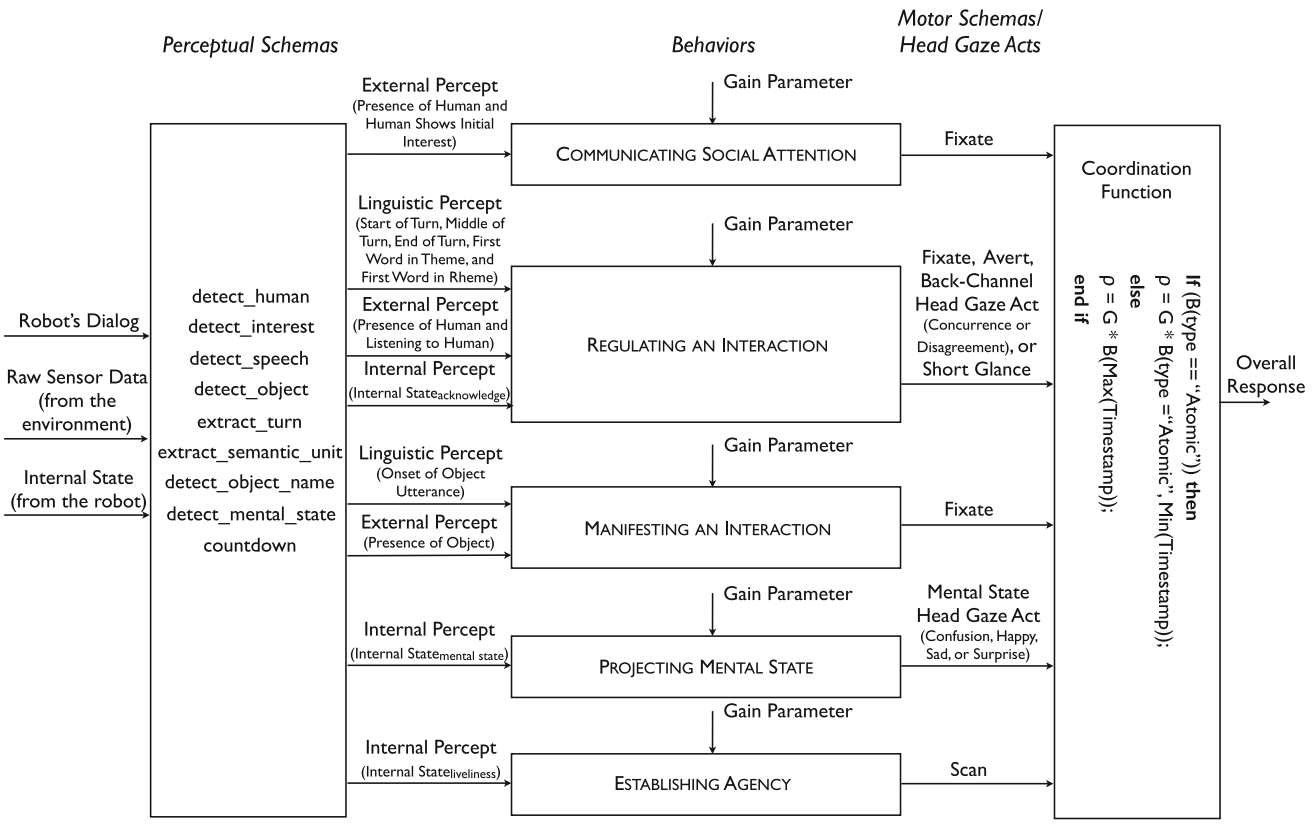

Fig. 2 Notational view of the possible behaviors in the reference architecture for social head gaze. The range of possible perceptual schemas, percepts, and head gaze acts have been enumerated

Table 3 The nine perpetual schemas and corresponding percepts used by the resulting Reference Architecture

Perceptual Schema	Percepts
detect_human	Presence of human
detect_interest	Human shows initial interest
detect_speech	Listening to human
extract_turn	Start of turn, middle of turn, and end of turn
extract_semantic_unit	start of rheme and start of theme
detect_object	Presence of object
detect_object_name	Onset of speech utterance
detect_mental_ state (for example, joy, happiness, and confusion)	Internal state$_{mentalstate}$
countdown (timer event-based)	Internal state$_{liveliness}$ and Internal state$_{acknowledge}$

behaviors themselves can employ different computational mechanisms (models based on probability or learning). The gains are shown as parameters that apply to each of the individual behaviors. The gain parameters can be used to modify range, speed, and frequency based on covariate factors that influence head gaze, such as the culture and gender [49,50,68] of the interaction partner or proxemics [25,27,45].

The motor schema represents the template for physical activities and is connected to the actuators [48]. The reference architecture currently supports six head gaze acts: *Fixate, Avert, Back-Channel Head Gaze Act, Short Glance, Mental State Head Gaze Act,* and *Scan.*

The coordination function is used to coordinate the responses of multiple active behaviors [6,48]. The coordination function ensures that the robot is sensitive to the current context and conveys the appropriate meaning. The prioritization rules used by existing architectures [26,28] are ad hoc [26], lack implementation details [28], and are limited to two or three behaviors. Hence, a coordination scheme based on timestamps (higher priority for the most recent behavior) and the nature of behavior (atomic or non-atomic) is proposed in Sect. 5. Atomic behaviors run to completion without interruption, where as non-atomic behaviors can be interrupted by the most recent behavior. The algorithm in Fig. 2 describes this function. While any other appropriate coordination methods such as action selection methods, fuzzy logic, or voting can be used, the coordination method using timestamps and the nature of the behavior was the first scheme to be implemented for the five behaviors.

Table 4 Re-characterization of Human–Robot collaboration architecture [26] and Robot Behavior Toolkit architecture [28] to the Reference Architecture based on Functionality

Reference architecture		Human robot collaboration architecture	Robot behavior toolkit architecture
Category	Component	Component	Component
Sensor processing	Raw sensor data	Raw sensor data	Raw sensor data
	Robot's dialog	Collaboration manager	Activity model
	Internal state	–	–
Perception	Perceptual schemas	Collaboration manager	Perceptual system
		Behavior recognition	Cognitive system
Behaviors	Communicating social attention	Response policy	Behavior selection system and knowledge base
	Regulating an interaction	Turn policy	
	Manifesting an interaction	Reference policy	
	Projecting mental state	–	–
	Establishing agency	–	–
Coordination	Coordination function	Maintenance policy	Behavior coordination system
		Collaboration manager	
Action execution	Overall response	BML realizer	Behavior generator

4 Evaluation of Two Representative Existing Architectures Using the Reference Architecture

A reference architecture is an important tool to evaluate, improve, or re-engineer existing architectures. When used in combination with the Software Architecture Analysis Method (SAAM), it gives insight into the quality of existing architectures and aids a robot designer in determining if a particular architecture meets their requirements. While any of the six system architectures identified in Sect. 2 could be used to illustrate SAAM, the Human–Robot Collaboration architecture [26] and the Robot Behavior Toolkit architecture [28–30] were considered. This was because only these two system architectures used a social science model for head gaze generation, which ensured that the head gaze was of high quality, elicited human acceptance [15], was repeatable, and consistent. SAAM is a five step process [13] which includes the following:

1. *Characterize a reference architecture of the domain*. For this analyses we used the reference architecture described in Sect. 3.3.
2. *Describe the existing architecture in terms of the reference architecture*. Section 4.1 describes the structural decomposition of the two systems architectures, which are mapped on to the reference architecture, followed by an allocation of functionality to the structure (Table 4).
3. *Choose a set of quality attributes with which to assess the architecture*. The two system architectures are formally evaluated for *overall functionality* [13] and two types of modifiability that are common in software engineering and also constitute a significant percentage of modifications for social head gaze: *adaptation to a new environment* and *extension of capability* [13]. The attribute *overall functionality* helps evaluate the existing architecture, whereas *adaptation to a new environment* and *extension of capability* provides insight to improve or re-engineer the existing architecture.
4. *Choose a set of concrete tasks which test the desired quality attributes*. *Overall functionality* is the number of head gaze behaviors supported by the architecture. For the attribute *adaptation to a new environment* potential factors that can change when operating in a new environment (e.g. search and rescue) are the robot and dialog. Similarly for *extension of capability*, extension of the architectures to a multi-party scenario is considered, which would require the addition of new behaviors and production rules.
5. *Evaluate the degree to which each architecture provides support for each task*. To *architecturally support* each task, subsystems that are responsible for the functionality should be (a) isolated in architectural description, that is the subsystem should be isolated from the rest of the architecture, and (b) non-monolithic. There should be support for subdivision of functionality within the subsystem.

4.1 Re-characterization of Existing Architectures to the Reference Architecture

This section uses the reference architecture to re-characterize two existing architectures—the Human–Robot Collaboration architecture and the Robot Behavior Toolkit architecture (Table 4). Re-characterization allocates the components of the existing architectures to the reference architecture based on Functionality. This information is then used in Sect. 4.2 to evaluate architectural support for the quality attributes and provide suggestions for architectural improvements.

4.1.1 Human–Robot Collaboration Architecture

The Human–Robot Collaboration architecture supports engagement between a human and a humanoid robot by generating head gaze behaviors [26]. The system consists of seven subsystems that provide structure to head gaze. Specifications from human–human communication for conversational turn taking [17,56] and engagement [54] are used in the architecture. The implementation of the architecture has been validated for a humanoid robot "Melvin" in a tangram game task.

The re-characterization of the Human–Robot Collaboration architecture is shown in Table 4. The Raw Sensor Data component of the Human–Robot Collaboration architecture is assigned to the Raw Sensor Data component of the Reference Architecture. Two subsystems, Collaboration Manager and Behavior Recognition are allocated to the Perceptual Schemas. The Collaboration Manager subsystem contains dialog annotated with turn status, hence it is also assigned to the Robot's Dialog component. The functionality of the Behavior Recognition subsystem is to perceive behavior indicators such as when a human initiates a connection. Three subsystems—Response Policy, Turn Policy, and Reference Policy—are assigned to the Behaviors. The Turn Policy subsystem generates the head gaze required for engaging in a conversation. This subsystem performs the function of the REGULATING AN INTERACTION subsystem. The Response Policy subsystem generates the head gaze necessary for looking interested in humans, which is the function of the COMMUNICATING SOCIAL ATTENTION subsystem. The Reference Policy subsystem generates referential head gazes for looking at objects in the environment. This subsystem captures the functionality of the MANIFESTING AN INTERACTION subsystem. Two subsystems, Maintenance Policy and Collaboration Manager, are allocated to the Coordination Function. The role of the Maintenance Policy subsystem is to prioritize the head gaze policy. The Collaboration Manager described above has one additional function. This subsystem is responsible for inhibiting turn or pointing gestures. Both of these subsystems perform the function of the Coordination Function of the reference architecture. The BML Realizer subsystem is allocated to the action execution of the overall response.

The re-characterization of Human–Robot Collaboration architecture reveals four points of interest:

(a) The description of the Collaboration Manager subsystem is monolithic; hence, it does not lend itself to a subdivision of functionality. This is because there is limited structural separation between perception (for example turn status), content of dialog, and behavior arbitration. The Collaboration Manager must provide the dialog, identify the turn events, and provide conflict resolution.
(b) The coordination mechanisms exist in both the Collaboration Manager and Maintenance Policy subsystems and their interactions are not fully defined and isolated. For example, the system does not have mechanisms to handle situations in which two rules have the same priority.
(c) The output of the Turn Policy subsystem feeds into the Reference Policy subsystem. As a result, the flow of data is serial between the two subsystems. The serial flow of data between these subsystems can affect real-time performance if the computational routines are more complex in the Turn Policy subsystem (e.g. multi-party interaction).
(d) In its current form, the architecture does not include mechanisms for PROJECTING MENTAL STATE and ESTABLISHING AGENCY, which have been shown to be important nonverbal head gaze functions as noted in Sect. 2.

4.1.2 Robot Behavior Toolkit Architecture

The Robot Behavior Toolkit architecture provides a framework for generating head gaze in human-like robots [28,29]. The architecture consists of eight subsystems that provide structure to head gaze. Specifications from human–human communication for conversational turn taking [17,56] and referential gaze [20,43] are used in the architecture. The implementation of the architecture has been validated for two robots, a simulated PR2 and a physical humanoid robot "Wakamaru". The tasks used in the studies are storytelling [28,29] and collaborative game tasks [28,29].

The re-characterization of Robot Behavior Toolkit architecture is shown in Table 4. The Raw Sensor Data component of the Robot Behavior Toolkit architecture is assigned to the Raw Sensor Data component of the Reference Architecture. Two subsystems, the Perceptual System and the Cognitive System, are allocated to the Perceptual Schemas. The Perceptual System transforms stimuli into a percept. The Cognitive System provides internal and external percepts based on the information from the Perceptual System and the current action prescribed by the Activity Model. The Activity Model contains dialog annotated with turn status, hence it is allo-

cated to the Robot's Dialog component. Two subsystems, the Knowledge Base and the Behavior Selection System, are assigned to the Behaviors. The Knowledge Base is a collection of behavioral specifications in XML. The Behavior Selection System queries the Knowledge Base for an appropriate behavior based on the percept. Both of these subsystems are responsible for the generation of head gaze and perform the function of three reference architecture subsystems: COMMUNICATING SOCIAL ATTENTION, REGULATING AN INTERACTION, and MANIFESTING AN INTERACTION. The Behavior Coordination subsystem is assigned to the Coordination Function. The role of the Behavior Coordination subsystem is to resolve conflicts and overlaps among behaviors by prioritization. This subsystem performs the function of the Coordination Function of the reference architecture. The Behavior Generator subsystem is allocated to action execution of the overall response of the robot.

There are two points of interest to note in this re-characterization of the Robot Behavior Toolkit architecture:

(a) The Knowledge Base subsystem is a collection of behavioral specifications in XML. The description of this subsystem is monolithic and not isolated. The use of the same subsystem for different behaviors reduces the maintainability, reusability, parallelizability, and robustness of the system.
(b) The architecture does not support the following two components: PROJECTING MENTAL STATE and ESTABLISHING AGENCY. These two nonverbal functions are important for achieving positive outcomes, as seen in Sect. 2.

4.2 Re-engineer or Improve Existing Architectures Using the Reference Architecture

The re-characterization of the two existing architectures in terms of the reference architecture reveals the degree of architectural support for *overall functionality*, *adaptation to a new environment*, or *extension of capability* and provides insights that can be used to make architectural improvements. The goal of the analysis using the reference architecture and SAAM is not to criticize or commend particular architectures, but to provide a method for determining which architecture supports a researcher's or robot behavior designer's needs.

The *overall functionality* of the Human–Robot Collaboration architecture and the Robot Behavior Toolkit architecture is limited to the communicative functions of social head gaze, since they are only capable of generating head gaze for COMMUNICATING SOCIAL ATTENTION, REGULATING AN INTERACTION, and MANIFESTING AN INTERACTION.

The architectural support for *adaptation to a new environment* and/or *extension of capability* in the Human–Robot Collaboration architecture is inadequate and can be improved. First, the functionalities supported by the Collaboration Manager such as perception, dialog management, and arbitration need to be upgraded for working in a new environment (e.g. search and rescue) or new capability (e.g. additional behaviors for multi-party scenario). In order to *architecturally support* the necessary modifications these functionalities should be fully defined, individually upgradable, and isolated from the Collaboration Manager. This can be achieved by transferring the coordination mechanism functions completely to the Maintenance Policy subsystem and perception functions to the Behavior Recognition subsystem. Second, the Turn Policy subsystem should be made independent of the Reference Policy subsystem. This change will improve the real-time performance of an implementation of the architecture and will reflect how the corresponding head gaze functions *engage in a verbal conversation with human(s)* and *gaze at objects in the environment* occur in human communication.

The Robot Behavior Toolkit offers adequate support for re-engineering to new environments. This is because the subsystems that support adaptation to different robots and dialog such as the Perceptual System, Cognitive System, Behavior Realizer, Activity Model, Behavior Generator are isolated in architectural description and non-monolithic. However, the architectural support for the *extension of capability* could be further improved in two areas. First, the Knowledge Base subsystem, which would incorporate a new behavior is monolithic. It should be made non-monolithic and subdivided into independent self contained components. These components should represent the individual social head gaze functions, as seen in the Human–Robot Collaboration architecture [26] or the proposed reference architecture. This will help in maintainability, re-usability, parallelizability, and robustness of the system. Second, the subsystems should not be solely developed for a rule-based system. This is because adding another behavior, such as extension to a multi-party scenario may involve computations that cannot be captured using simple rules in XML and might require advanced techniques such as learning.

5 Reference Architecture Implementation for Victim Management

In order to illustrate the reference architecture, it is helpful to consider how it is applied to guide the implementation of social head gaze for victim management in urban search and rescue (US&R). At a disaster such as a building collapse, it can take responders up to ten hours to safely extricate a trapped victim after they are discovered [8,48]. Throughout this critical period, it is important that the robot interact with the victim in a socially appropriate way in order to

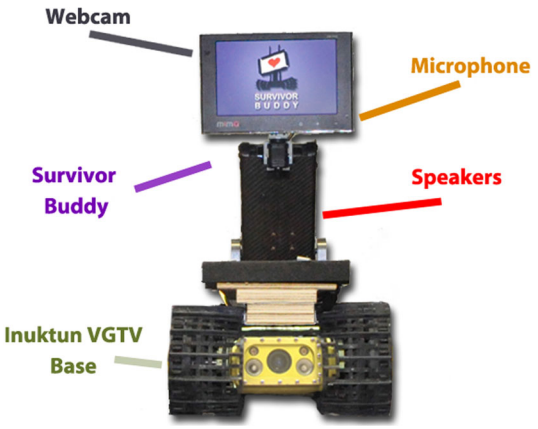

Fig. 3 Survivor Buddy Robot

reduce stress levels and prevent shock, as well as keep the victim calm, comforted, at ease, and engaged until assistance arrives. Survivor Buddy is a four-degree of freedom affective multimedia head mounted on an Inuktun Extreme-VGTV robot (see Fig. 3) used for urban search and rescue responses. Survivor Buddy's "head" is a small, 7-inch MIMO 740 touchscreen monitor, which also contains a webcam and microphone. A speaker system is mounted in Survivor Buddy's neck.

The Perceptual Schemas transform the information from the sensors into percepts. The *Presence of Human* is indicated by the *detect_human* perceptual schema if the human head is within the field of view of the robot's webcam and is detected by the FaceAPI algorithm [41]. The *detect_interest* perceptual schema records the initial distance and angle to the face of the human. If subsequent readings show a decrease in distance and angle, then it is determined that *Human Shows Initial Interest*. The robot is *Listening to Human* if human speech is detected by the *detect_speech* perceptual schema using an internally developed voice recognition toolkit [65]. The percept *Presence of Object* is generated by the *detect_object* perceptual schema if an object is within the field of view of the robot's webcam and is detected by the ROS object recognition stack [60]. The stack is comprised of models of regularly used objects such as cups, soda cans, bottles, etc. However, new objects that were not previously in the library can also be captured [60]. The *extract_turn* perceptual schema extracts the *Start of Turn*, *Middle of Turn*, *End of Turn* turn events from the robot's current turn. The *Start of Turn* is defined as the first word of the new turn. *Middle of Turn* is defined as the word after 50 % of the words in the turn. *End of Turn* is defined as the last word ± one word. The *Start of Rheme* and *Start of Theme* semantic units are computed by the *extract_semantic_unit* perceptual schema. The "rheme" represents the contribution to the pool of knowledge in the conversation. The first word of this semantic unit is the *First Word in Rheme*. The "theme" represents what the utterance is about, what links it to previous utterances. The first word of this semantic unit is the *First Word in Theme*. In the current implementation the *First Word in Rheme* and *First Word in Theme* are marked *a-priori* by manual identification. This is standard practice and has been used extensively for the generation of head gaze [26,28,29,49,62]. The *detect_mental_state* perceptual schema is used to detect the robot's mental state, for this example, "confusion." The $Internal\ State_{confusion}$ is returned when the robot does not understand the human because of failures in speech recognition. Finally the *countdown* perceptual schema sets the $Internal\ State_{liveliness}$ and $Internal\ State_{acknowledge}$ based on a timer event. The $Internal\ State_{liveliness}$ is activated if the robot is idle for more than 15 seconds and the $Internal\ State_{acknowledge}$ is set if the elapsed listening time is more than 6 seconds. Since the existing literature does not provide guidance on specific values for *elapsed idle time* [10] or *elapsed listening time* [26,62], these suitable timeouts were estimated by the researchers to communicate effectively the corresponding function of head gaze.

The behaviors are implemented using production rules (see Table 5). The production rules are if-then statements where if a percept is perceived, the head gaze act is called. The production rules directly correspond to the five behaviors, and a single behavior may comprise one or more production rules (see Table 5). For the user study described in this article, the gain parameters for each behavior were set to 1.

The Coordination Function is implemented using a prioritization scheme based on timestamps (highest priority for the most recent behavior) and the nature of the behavior (atomic or non-atomic). This implementation coordinates head movements simultaneously for five behaviors. Prior work [26,28] focuses on the coordination of just two or three behaviors, not all five together. The PROJECTING A MENTAL STATE and MANIFESTING AN INTERACTION behaviors are atomic and run to completion without interruption. This is because any interruptions from other head gaze acts may detract from the goals of the robot, perceived understanding of the robot's actions, and social competence of the robot. These behaviors are typically uninterrupted in human–human communication. PROJECTING A MENTAL STATE, such as anger makes any detected interruption, such as an aversion of head for turn-taking, of secondary importance. Expressing the emotion fully is more important than a filler gaze act like an aversion of the head that can be used to build rapport. Referential head gaze in humans is used to draw the attention of the human to an object. The purpose of the MANIFESTING AN INTERACTION behavior will not be accomplished if it is interrupted midway through the process. All other behaviors are non-atomic and may be interrupted by the most recent behavior. In terms of implementation, this translates to production rules 8 and 9, shown in Table 5, running to completion without interruption. All other production rules (1–7) can be

Table 5 Nine production rules

	Behavior	Production Rule
1	Communicating Social Attention [10,26,28,29,31,38,40,44,50,57,61,62,70]	IF **Presence of Human** and **Human Shows Initial Interest**, THEN **Fixate** toward the human for an undetermined duration until interruption occurs at a velocity of 33°/s
2		IF **Presence of Human, First word in Theme**, and **Start of Turn**, THEN **Avert** from the human with a ±7°/s simultaneous horizontal and vertical movement for an undetermined duration until interruption occurs at a velocity of 33°/s
3		IF **Presence of Human, First word in Theme**, and **Middle of Turn**, THEN **Avert (p = .73)** from the human with a ±7°/s simultaneous horizontal and vertical movement for an undetermined duration until interruption occurs at a velocity of 33°/s
4	Regulating an Interaction [42], [11,26,28,29,33,38,49,50,57,70]	IF **Presence of Human**, **First word in Rheme**, and **Middle of Turn**, THEN **Fixate (p = .7)** toward the human for an undetermined duration until interruption occurs at a velocity of 33°/s
5		IF **Presence of Human**, **First word in Rheme**, and **End of Turn**, THEN **Fixate** toward the human for an undetermined duration until interruption occurs at a velocity of 33°/s
6		IF **Listening to Human** and **Internal State**$_{acknowledge}$ THEN **Concurrence** toward the human with repetitive vertical head movement of ±10° every 3 s at a velocity of 33°/s
7	Establishing Agency [10,62]	IF **Internal State**$_{liveliness}$, THEN **Scan** three random points in the environment at a velocity of 33°/s
8	Projecting Mental States [10]	IF **Internal State**$_{confused}$, THEN **Confusion** toward the human with a head roll of ±20° and return to the fixation point at a velocity of 33°/s
9	Manifesting an Interaction [10,26,28,29,62,67]	IF **Presence of Object** and **Utterance of Object**, THEN **Fixate** toward the object in the environment at a velocity of 33°/s

interrupted by the most recent production rule that is activated.

The overall response after coordination is one of the following five out of six head gaze acts: *Fixate, Avert, Concurrence, Confusion,* and/or *Scan.* Some of the gaze acts are probabilistic in nature, ($p = value$) indicating the probability of activation of the gaze act. The *Short Glance* gaze act was not implemented in the user study because the architecture was instantiated for a dyadic conversation and not a multi-party conversation scenario. The head gaze acts vary with the robot, as each robot would have a different implementation of head gaze based on its degrees of freedom and motor characteristics. The implementation specifics for each of the head gaze acts on the Survivor buddy robot are shown in Table 5. The parameters such as duration and range matched known values used by earlier implementations of head gaze in the literature [10,26,28,31,38,44,45,49,50,57,61,62,70].

6 Evaluation of the Reference Architecture Implementation

The head gaze generated using the reference architecture implementation was evaluated with the Survivor Buddy robot in a high fidelity simulated parking structure collapse scenario. In the semi-wizard of oz study (the robot was autonomous, but required the operator to activate the robot's turn), 93 participants played the role of trapped "victims" and interacted with the robot in one of three conditions: Loosely Synchronized Head Gaze-Speech (LSHG-S), Tightly Synchronized Head Gaze-Speech (TSHG-S), or No Head Gaze-Speech (NHG-S). These conditions were chosen to investigate head gaze-speech synchronization for human–robot interaction. TSHG-S requires precise timing between the speech utterance and activation of the corresponding head gaze act. The head gaze generation for TSHG-S uses semantic content of the dialog, such as First Word in Theme, First Word in Rheme, etc which was similar to gaze behaviors exhibited in human–human conversation. These were identified by manual inspection using definitions described by Halliday [22], and marked with the corresponding head gaze acts on a pre-recorded audio file. The Microsoft speech system triggers the head gaze act when it encounters a head gaze marker. In LSHG-S, the timing between the speech utterance and the activation of the corresponding head gaze act is flexible; that is, the activation of the head gaze act can lead, lag, or occur at onset of the speech utterance. The generation of LSHG-S is based on sentence structure (for example, *Initial Word*, *Word following Punctuation : . ! ?*, *After 75 % of Words between Punctuation : . ! ?*, and *Carriage Return*) and time delays (for example, *Elapsed Listening Time* and *Elapsed Idle Time*). The NHG-S was the control condition, with the robot looking directly at the participant throughout the interaction and without displaying any head gaze acts, and using only speech to interact. Both the LSHG-S condition and TSHG-S condition were implemented using the reference architecture.

Table 6 Statistically significant results from the conducted user study (ANOVA and Tukey's HSD) [64]

Attribute	LSHG-S v NHG-S	TSHG-S v NHG-S	LSHG-S (Md) (IQR)	TSHG-S (Md) (IQR)	NHG-S (Md) (IQR)
SAM: arousal	$t(90) = 3.63$	$t(90) = 3.48$	8	8	6
$F(2, 90) = 8.43$	$p = .001$	$p = .002$	2	2	3
$p < .001$					
Robot likeability	$t(90) = 3.05$	$t(90) = 3.33$	4.91	5	4.4
$F(2, 90) = 6.75$	$p = .008$	$p = .004$	1.2	1.4	1.4
$p = .002$					
Human-like behavior	$t(90) = 4.03$	$t(90) = 3.10$	5	5	3
$F(2, 90) = 8.9$	$p < .001$	$p = .007$	2	2	3
$p < .001$					
Understanding robot behavior	$t(90) = 4.63$	$t(90) = 5.29$	5.33	6.16	3
$F(2, 90) = 18.09$	$p < .001$	$p < .001$	1.67	2.75	1.58
$p < .001$					
Gaze-speech synchronization	$t(90) = 8.66$	$t(90) = 8.28$	6	6	3
$F(2, 90) = 47.87$	$p < .001$	$p < .001$	2	2	2
$p < .001$					
Look at objects at appropriate times	$t(90) = 4.82$	$t(90) = 5.12$	6	6	3
$F(2, 90) = 14.6$	$p < .001$	$p < .001$	1	1.25	2.25
$p < .001$					
Natural movement	$t(90) = 4.79$	$t(90) = 5.19$	5	5	3
$F(2, 90) = 16.69$	$p < .001$	$p < .001$	2	2	3
$p < .001$					

Md median and *IQR* interquartile range

The robot interacted with the participant at a distance of 1.22 m, which was within the participant's personal zone [5]. The screen displayed only a Survivor Buddy logo so that the only social cues were voice and head gaze acts. The interaction with the participant lasted approximately 15 min, and the robot followed a predefined script consisting of questions and simple directions. The robot supervisor (hidden from view) would activate the text for the robot's turns in the dialog. The robot used all the five head gaze behaviors and five out of the six identified head gaze acts during the interaction. It activated the COMMUNICATING SOCIAL ATTENTION behavior to gain the participant's attention and convey that it was interested and ready for an interaction. The robot then used the REGULATING AN INTERACTION behavior for effective human-like turn-taking and back-channeling during a dialog with the victim, based on 911 dispatch and triage protocols. The dialog focused on assessing the participant's physical health and gaining information about the location and nature of the event. The robot posed questions like: *"Can you move your fingers?"* and *"Do you see anyone else with you?"* The robot also monitored the area surrounding the participant and used the MANIFESTING AN INTERACTION behavior to point toward objects of interest like a fire extinguisher or hazardous objects. The robot activated the PROJECTING MENTAL STATE behavior to indicate confusion and provide feedback to the participant that it did not understand. The robot used the ESTABLISHING AGENCY behavior to convey liveliness and let the participant know that it was functioning properly.

As seen in Table 6 participants' rated the robot more positively with regards to measures of *Self Assessment Manikin (SAM):Arousal* [9], *Robot Likeability, Human-Like Behavior, Understanding Robot Behavior, Gaze-Speech Synchronization, Look at Objects at Appropriate Times,* and *Natural Movement* in both the LSHG-S condition and TSHG-S condition, when compared to the NHG-S condition. This suggests that following the reference architecture results in an implementation that competently integrates all five behaviors of head gaze. A detailed discussion of the results and implications of these findings can be found in Srinivasan et al. [64].

7 Summary

This article outlines the design and implementation of a reference architecture for the generation of social head gaze in social robotics which is grounded in human–human communication and based on 32 existing robotic implementations. The reference architecture is constructed following the procedure by [23,35], and using techniques from behavioral robotics theory [6,48]. It consists of—*Perceptual Schemas*, *Behaviors*, *Motor Schemas*, and a *Coordination Function*. This article is the first to integrate five social head gaze behaviors and use a coordination scheme based on timestamps and the nature of the behavior to resolve conflicts and overlaps in the behaviors. The reference architecture was validated in two ways. First, an evaluation of the *overall functionality*, *adaptation to a new environment*, and *extension of capability* of two existing system architectures was conducted using the reference architecture and SAAM. Second, the implementation of the reference architecture for use in victim management in a US&R scenario, which resulted in a high level of social acceptance.

The reference architecture provides two immediate benefits for researchers in the social robotics community. First, the reference architecture simplifies principled implementations of social head gaze. This is because it can be used as a blueprint when implementing applications that utilize social head gaze. The reference architecture defines important aspects of head gaze required for high quality, repeatable, and consistent implementations. It also provides guidance for integrating multiple behaviors into a single implementation. Second, a reference architecture is an important tool used to evaluate, improve, or re-engineer existing architectures. It gives a common vocabulary and taxonomy for making architectural comparisons and understanding systems. Comparisons of different architectures are more challenging when using various architectural representations and claims.

Four directions for future work have been identified. First, the reference architecture should be expanded to support both head and eye gaze. Second, the reference architecture should be extended to support multi-party interactions. Currently, head gaze is predominantly a feature of dyadic [1,2,4,8,10,19,24,26,28–33,37–40,44,46,49,51,53,57,58,61–63,66,67,70] or triadic situations [7,50]. However, it is reasonable to assume that robots will encounter multi-party situations often in the real world. Thus, it is also worth asking whether a multi-party situation can be approximated to several dyadic encounters. This would involve updating the implementation of REGULATING AN INTERACTION behavior to support the short glance gaze act and design of new behaviors and upgraded coordination system to engage bystanders or interrupt conversations. Third, the implementation of the reference architecture should be upgraded to support other emotional expression head gaze acts like *Happy*, *Angry*, or *Sad*. Four, the reference architecture needs to be extended to accommodate additional behavioral channels, such as arm postures and arm gestures.

Acknowledgments This work was supported in part by NSF IIS-0905485 "The Social Medium is the Message", RESPOND-R mobile, distributed test instrument created through NSF Grant CNS-0923203, and Microsoft External Research.

References

1. Admoni H, Dragan A, Srinivasa SS, Scassellati B (2014) Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, HRI '14, pp 49–56. ACM, New York
2. Admoni H, Hayes B, Feil-Seifer D, Ullman D, Scassellati B (2013) Are you looking at me? Perception of robot attention is mediated by gaze type and group size. In: Proceedings of the 8th ACM/IEEE international conference on human-robot interaction, pp 389–396. Tokyo
3. Al Moubayed S (2012) Bringing the avatar to life: studies and developments in facial communication for virtual agents and robots
4. Andrist S, Tan XZ, Gleicher M, Mutlu B (2014) Conversational gaze aversion for humanlike robots. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, pp 25–32. ACM, New York
5. Argyle M, Cook M (1976) Gaze and mutual gaze. Cambridge University Press, Cambridge
6. Arkin RC (1998) Behavior-based robotics. The MIT Press, Cambridge
7. Bennewitz M, Faber F, Joho D, Schreiber M, Behnke S (2005) Towards a humanoid museum guide robot that interacts with multiple persons. In: Proceedings of the IEEE/RSJ international conference on humanoid robots
8. Bethel C, Murphy R (2010) Non-facial and non-verbal affective expression for appearance-constrained robots used in victim management. Paladyn J Behav Robot 1:219–230
9. Bradley M (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25(1):49–59
10. Breazeal C, Kidd CD, Thomaz AL, Hoffman G, Berlin M (2003) Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In: Proceedings of IROS
11. Cassell J, Torres O, Prevost S (1998) Turn taking vs. discourse structure: how best to model multimodal conversation. In: Machine conversations. Kluwer, The Hague
12. Chovil N (1991) Discourse-oriented facial displays in conversation. Res Lang Soc Interact 25(1–4):163–194
13. Clements PC (2002) Software architecture in practice. Ph.D. thesis, Carnegie Mellon University
14. Cloutier R, Muller G, Verma D, Nilchiani R, Hole E, Bone M (2010) The concept of reference architectures. Syst Eng 13(1):14–27
15. Dautenhahn K (1999) Robots as social actors: Aurora and the case of autism. In: Proceedings of the third cognitive technology conference
16. Dobrica L, Niemela E (2002) A survey on software architecture analysis methods. IEEE Trans Softw Eng 28(7):638–653
17. Duncan S (1972) Some signals and rules for taking speaking turns in conversations. J Personal Soc Psychol 23(2):283–292
18. Exline RV, Winters LC (1965) Affective relations and mutual glances in dyads. In: Affect, cognition, and personality 1(5). Springer, New York

19. Fincannon T, Barnes L, Murphy R, Riddle D (2004) Evidence of the need for social intelligence in rescue robots. In: Proceedings of the international conference on intelligent robots and systems (IROS), vol 2, pp 1089–1095
20. Griffin ZM (2001) Gaze durations during speech reflect word selection and phonological encoding. Cognition 82(1):B1–B14
21. Hadar U, Steiner T, Grant E, Rose FC (1983) Kinematics of head movements accompanying speech during conversation. Hum Mov Sci 2(1):35–46
22. Halliday M (1967) Intonation and grammar in British English. Janua linguarum: Series practica. Mouton, The Hague
23. Hassan AE, Holt RC (2000) A reference architecture for web servers. In: Proceedings of the 7th working conference on reverse engineering, pp 150–159. IEEE, Washington, DC
24. Heerink M, Kröse B, Evers V, Wielinga B (2010) Relating conversational expressiveness to social presence and acceptance of an assistive social robot. Virtual Real 14(1):77–84
25. Henkel Z, Bethel C, Murphy R, Srinivasan V (2014) Evaluation of proxemic scaling functions for social robotics. IEEE Trans Hum-Mach Syst 44(3):374–385
26. Holroyd A, Rich C, Sidner CL, Ponsler B (2011) Generating connection events for human-robot collaboration. In: 20th IEEE international workshop on robot and human interactive communication, RO-MAN, pp 24–246
27. Holthaus P, Pitsch K, Wachsmuth S (2011) How can I help? Int J Soc Robot 3(4):383–393
28. Huang C, Mutlu B (2012) Robot behavior toolkit: generating effective social behaviors for robots. In: Proceedings of the 7th ACM/IEEE conference on human-robot interaction
29. Huang CM, Mutlu B (2013) The repertoire of robot behavior: designing social behaviors to support human-robot joint activity. J Hum-Robot Interact 2(2):80–102
30. Huang CM, Mutlu B (2014) Learning-based modeling of multimodal behaviors for humanlike robots. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, HRI '14, pp 57–64. ACM, New York
31. Imai M, Kanda T, Ono T, Ishiguro H, Mase K (2002) Robot mediated round table: analysis of the effect of robot's gaze. In: Proceedings of 11th IEEE international workshop on robot and human interactive communication, pp 411–416
32. Imai M, Ono T, Ishiguro H (2001) Physical relation and expression: joint attention for human-robot interaction. In: Proceedings of 10th IEEE international workshop on robot and human interactive communication, pp 512–517
33. Ishi CT, Liu C, Ishiguro H, Hagita N (2010) Head motions during dialogue speech and nod timing control in humanoid robots. In: Proceedings of the 5th ACM/IEEE international conference on human-robot interaction, pp 293–300. IEEE Press, Piscataway
34. Jamone L, Brandao M, Natale L, Hashimoto K, Sandini G, Takanishi A (2014) Autonomous online generation of a motor representation of the workspace for intelligent whole-body reaching. Robot Auton Syst 62(4):556–567
35. Kazman R, Bass L, Webb M, Abowd G (1994) Saam: a method for analyzing the properties of software architectures. In: Proceedings of the 16th international conference on software engineering, pp 81–90. IEEE Computer Society Press, Los Alamitos
36. Kendon A (1967) Some functions of gaze-direction in social interaction. Acta Psychol 26:22–63
37. Kozima H, Nakagawa C, Yasuda Y (2005) Interactive robots for communication-care: a case-study in autism therapy. In: 14th IEEE international workshop on robot and human interactive communication (RO-MAN), pp 341–346
38. Kuno Y, Sadazuka K, Kawashima M, Yamazaki K, Yamazaki A, Kuzuoka H (2007) Museum guide robot based on sociological interaction analysis. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1191–1194. ACM, New York
39. Liu C, Ishi CT, Ishiguro H, Hagita N (2012) Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In: Proceedings of the 7th annual ACM/IEEE international conference on human-robot interaction, pp 285–292. ACM, New York
40. MacDorman K, Minato T, Shimada M, Itakura S, Cowley S, Ishiguro H (2005) Assessing human likeness by eye contact in an android testbed. In: Proceedings of the XXVII annual meeting of the Cognitive Science Society, pp 21–23
41. Machines S (2009) Faceapi. http://www.seeingmachines.com/product/faceapi
42. Matsusaka Y, Fujie S, Kobayashi T (2001) Modeling of conversational strategy for the robot participating in the group conversation. In: Interspeech'01, pp 2173–2176
43. Meyer AS, Sleiderink AM, Levelt WJ (1998) Viewing and naming objects: eye movements during noun phrase production. Cognition 66(2):B25–B33
44. Minato T, Shimada M, Ishiguro H, Itakura S (2004) Development of an android robot for human-robot interaction. Innovations in applied artificial intelligence, pp 424–434
45. Mitsunaga N, Smith C, Kanda T, Ishiguro H, Hagita N (2008) Adapting robot behavior for human-robot interaction. IEEE Trans Robot 24(4):911–916
46. Moon A, Troniak DM, Gleeson B, Pan MK, Zeng M, Blumer BA, MacLean K, Croft EA (2014) Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, HRI '14, pp 334–341. ACM, New York
47. Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E (2004) Visual prosody and speech intelligibility head movement improves auditory speech perception. Psychol Sci 15(2):133–137
48. Murphy RR (2000) Introduction to AI robotics. The MIT Press, Cambridge
49. Mutlu B, Forlizzi J, Hodgins J (2006) A storytelling robot: modeling and evaluation of human-like gaze behavior. In: Proceedings of the international conference on humanoid robots. IEEE, Piscataway
50. Mutlu B, Shiwa TKT, Ishiguro H, Hagita N (2009) Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In: Proceedings of the 4th ACM/IEEE international conference on human robot interaction, pp 61–68. ACM, New York
51. Mutlu B, Yamaoka F, Kanda T, Ishiguro H, Hagita N (2009) Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In: HRI '09: Proceedings of the 4th ACM/IEEE international conference on human robot interaction, pp 69–76. ACM, New York
52. Nass C, Steuer J, Tauber ER (1994) Computers are social actors. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 72–78. ACM, New York
53. Pitsch K, Vollmer AL, Muhlig M (2013) Robot feedback shapes the tutor's presentation how a robot's online gaze strategies lead to micro-adaptation of the human's conduct. Interact Stud 14(2):268–296
54. Rich C, Ponsleur B, Holroyd A, Sidner CL (2010) Recognizing engagement in human-robot interaction. In: Proceeding of the 5th ACM/IEEE international conference on human-robot interaction, pp 375–382. ACM, New York
55. Ruhland K, Andrist S, Badler J, Peters C, Badler N, Gleicher M, Mutlu B, McDonnell R (2014) Look me in the eyes: a survey of eye and gaze animation for virtual agents and artificial systems. In: Eurographics 2014—State of the Art Reports, pp 69–91. The Eurographics Association, Aire-la-Ville

56. Sacks H, Schegloff EA, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50(4):696–735
57. Sakamoto D, Kanda T, Ono T, Kamashima M, Imai M, Ishiguro H (2004) Cooperative embodied communication emerged by interactive humanoid robots. In: 13th IEEE international workshop on robot and human interactive communication, RO-MAN, pp 443–448
58. Sauppé A, Mutlu B (2014) Robot deictics: how gesture and context shape referential communication. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, HRI '14, pp 342–349. ACM, New York
59. Sciutti A, Bisio A, Nori F, Metta G, Fadiga L, Sandini G (2013) Robots can be perceived as goal-oriented agents. Interact Stud 14(3):329–350
60. Seib V (2010) Ros object recognition stack. http://wiki.ros.org/object_recognition
61. Shimada M, Yoshikawa Y, Asada M, Saiwaki N, Ishiguro H (2011) Effects of observing eye contact between a robot and another person. Int J Soc Robot 3:143–154
62. Sidner CL, Lee C, Kidd CD, Lesh N, Rich C (2005) Explorations in engagement for humans and robots. Artif Intell 166:10–1016
63. Sirkin D, Ju W, Cutkosky M (2012) Communicating meaning and role in distributed design collaboration: how crowdsourced users help inform the design of telepresence robotics. In: Design thinking research, pp 173–187. Springer, New York
64. Srinivasan V, Bethel C, Murphy R (2014) Evaluation of head gaze loosely synchronized with real-time synthetic speech for social robots. IEEE Trans Hum-Mach Syst 44(6):767–778
65. Srinivasan V, Murphy R, Henkel Z, Groom V, Nass C (2011) A toolkit for exploring the role of voice in human-robot interaction. In: Proceedings of the 6th international conference on human-robot interaction, HRI '11, pp 255–256. ACM, New York
66. Staudte M, Crocker M (2009) The effect of robot gaze on processing robot utterances. In: Proceedings of the 31th annual conference of the Cognitive Science Society. Cognitive Science Society, Amsterdam
67. Staudte M, Crocker MW (2009) Visual attention in spoken human-robot interaction. In: HRI '09: Proceedings of the 4th ACM/IEEE international conference on human robot interaction, pp 77–84. ACM, New York
68. Trovato G, Zecca M, Sessa S, Jamone L, Ham J, Hashimoto K, Takanishi A (2013) Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by Egyptians and Japanese. Paladyn J Behav Robot 4(2):83–93
69. Weiss DM (1998) Commonality analysis: a systematic process for defining families. In: Development and evolution of software architectures for product families, pp 214–222. Springer, Heidelberg
70. Yamazaki A, Yamazaki K, Kuno Y, Burdelski M, Kawashima M, Kuzuoka H (2008) Precision timing in human-robot interaction: coordination of head movement and utterance. In: Proceeding of the twenty-sixth annual SIGCHI conference on human factors in computing systems, pp 131–140. ACM, New York

Vasant Srinivasan is a Development Engineer at Sprinklr. He graduated in 2014 with his Ph.D. in Computer Engineering from the Computer Science and Engineering Department, Texas A&M University.

Robin R. Murphy (IEEE Fellow) is the Raytheon Professor of Computer Science and Engineering at Texas A&M, Director of the Center for Robot-Assisted Search and Rescue, and the Center for Emergency Informatics. She received a B.M.E. in mechanical engineering, a M.S. and Ph.D in computer science in 1980, 1989, and 1992, respectively, from Georgia Tech where she was a Rockwell International Fellow. She has over 150 publications on artificial intelligence, human–robot interaction, and robotics including two textbooks, Introduction to AI Robotics, and Disaster Robotics.

Cindy L. Bethel is an Assistant Professor in the Computer Science and Engineering Department at Mississippi State University (MSU). She is the Director of the Social, Therapeutic, and Robotic Systems (STaRS) lab and a Research Fellow with the MSU Center for Advanced Vehicular Systems Human Performance Group. She was a NSF/CRA/CCC Computing Innovation Postdoctoral Fellow in the Social Robotics Laboratory at Yale University. She graduated in 2009 with her Ph.D. in Computer Science and Engineering from the University of South Florida. Her research interests are in human–robot interaction, affective computing, robotics, human–computer interaction and interface design, artificial intelligence, and psychology. Her research focuses on applications associated with the use of robots for therapeutic support, law enforcement, search and rescue, and military.