sensors

MDPI

Article

Vision-Based Attentiveness Determination Using Scalable HMM Based on Relevance Theory

Prasertsak Tiawongsombat [1], Mun-Ho Jeong [2,*], Alongkorn Pirayawaraporn [3], Joong-Jae Lee [4] and Joo-Seop Yun [5]

1 Electronics Engineering Technology, College of Industrial Technology, King Mongkut's University of Technology North Bangkok, 1518 Pracharad 1 Rd., Wongsawang, Bangsue, Bangkok 10800, Thailand; prasertsak.t@cit.kmutnb.ac.th
2 Division of Robotics, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul 01897, Korea
3 Department of Control and Instrumentation Engineering, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul 01897, Korea; alongkorn.kmutnb@gmail.com
4 Center of Human-centered Interaction for Coexistence, 5, Hwarang-ro 14-gil, Seongbuk-gu, CHIC, Seoul 02792, Korea; arbitlee@chic.re.kr
5 Mechatronics Technology Convergence R&D Group, Korea Institute of Industrial Technology, 320 Techno Sunhwan-ro, Yuga-eup, Dalseong-gun, Daegu 42994, Korea; jsyun@kitech.re.kr
* Correspondence: mhjeong@kw.ac.kr; Tel.: +82-2-940-5625

Received: 4 October 2019; Accepted: 29 November 2019; Published: 3 December 2019

Abstract: Attention capability is an essential component of human–robot interaction. Several robot attention models have been proposed which aim to enable a robot to identify the attentiveness of the humans with which it communicates and gives them its attention accordingly. However, previous proposed models are often susceptible to noisy observations and result in the robot's frequent and undesired shifts in attention. Furthermore, most approaches have difficulty adapting to change in the number of participants. To address these limitations, a novel attentiveness determination algorithm is proposed for determining the most attentive person, as well as prioritizing people based on attentiveness. The proposed algorithm, which is based on relevance theory, is named the Scalable Hidden Markov Model (Scalable HMM). The Scalable HMM allows effective computation and contributes an adaptation approach for human attentiveness; unlike conventional HMMs, Scalable HMM has a scalable number of states and observations and online adaptability for state transition probabilities, in terms of changes in the current number of states, i.e., the number of participants in a robot's view. The proposed approach was successfully tested on image sequences (7567 frames) of individuals exhibiting a variety of actions (speaking, walking, turning head, and entering or leaving a robot's view). From these experimental results, Scalable HMM showed a detection rate of 76% in determining the most attentive person and over 75% in prioritizing people's attention with variation in the number of participants. Compared to recent attention approaches, Scalable HMM's performance in people attention prioritization presents an approximately 20% improvement.

Keywords: human–robot interaction; attention model; measure of attentiveness; relevance theory; Scalable Hidden Markov Model

1. Introduction

Attention is a process involving human factors. Human factors play a central role in the attentiveness determination process, especially when qualitative information and uncertainties are involved [1–5]. Intuitively, attention is an essential process for starting social interaction between human beings. To begin giving attention in a social interaction, a person with whom to communicate must first be identified. Most people perform this selection subconsciously, i.e., they identify who,

from those in a given room or group, is worthy of their attention. Likewise, an intelligent service robot has to select a person in the group before their bi-directional communication starts. Therefore, the robot is required to possess attention-selecting capabilities as a fundamental function based on human social expectations; when the robot is equipped with such capabilities, people can interact with it in the same way that they interact with other people [6–11]. However, humans often stay in groups instinctively for communication. Before starting a conversation, the speaker evaluates to their prospective communicators and selects one from among them with whom to communicate. For this reason, when the service robot communicates in a multi-person interaction, it also evaluates and prioritizes the attention of the prospective communicators members individually, in terms of their perceived attentiveness; this process is called attention prioritization. The robot then selects the person who has the highest attentiveness (i.e., the most attentive person) to be the person with whom it communicates.

Most attention systems are generally composed of two distinctive sections: (1) feature extraction and (2) an attention model. (1) Feature extraction extracts attention-related visual features (ostensive-stimuli) from an image sequence and/or audio features from a sound stream. Various visual features are often chosen to be used as stimuli for the attention system such as the distance between a robot and a person [12,13], the head direction of the people participating in an interaction [14–19], and/or visual speaking status detection [20–25]. When audio features are used for the attention model, the direction of a sound source and the distance to a sound source are usually adopted [26–28]. (2) The attention model evaluates the selected stimuli and computes the attentiveness of each person. Finally, the most attentive person, as well as the attention priority of the individuals, in terms of their attentiveness, are determined. Intuitively, a robot equipped with an attention system can be considered to be more flexible and effective in their interactions with humans compared to the robot without one.

Overall, most previous methods have employed either a set of event conditions, heuristic equations, or both, such that the methods operate under predefined parameters and rules. However, the heuristic approaches presented in the literature are often susceptible to noisy observations and may produce frequent undesired attention shifts by the robot. Furthermore, their performance also suffers when they must contend with changes in the number of persons and observations, as they have difficulty adapting the state numbers accordingly in real-time.

To overcome such difficulties, a novel attentiveness determination approach based on relevance theory [29] is introduced. The relevance theory describes how humans communicate with each other and how a person evaluates the attention of other people during interaction exchanges. Thus, this theory was applied and converted to a mathematical form that aims to determine the most attentive person and prioritize people according to their relative attentiveness. Thus, a model was developed which aims to determine the most attentive person and prioritize people according to their relative attentiveness. The proposed approach consists of (1) a Scalable Hidden Markov Model (Scalable HMM) for attentiveness determination and (2) a probabilistic approach to compute the relevance for stimuli. The Scalable HMM has a scalable number of states and observations, and online adaptability for state transition probabilities with respect to changes in the current number of states. To test the proposed approach, the Scalable HMM was applied to 10 image sequences (7567 frames) of individuals exhibiting a variety of actions (speaking, walking, turning head, and entering or leaving a robot's view). The detection rates achieved by the proposed approach, for both determination of the most attentive person and for people attention prioritization, were obtained and compared to those by recent robot attention model approaches.

The remainder of the paper is organized as follows. Section 2 reviews related research. Section 3 introduces a probabilistic stimuli-relevance computation approach based on relevance theory. The Scalable HMM-based attentiveness determination method is described in Section 4. Experimental results and the conclusions are discussed in Sections 5 and 7, respectively.

2. Related Work

In the past decade, several researchers have integrated psychological studies into robotics research. Such works have estimated the mental states of other people by observing their behaviors and aimed to design a robot with human-like attention capabilities [30–34]. Psycholinguistic studies revealed that speaking status plays an important role in attention, in that a listener's visual attention is driven by what they hear [35,36]. As a result, the speaking status is usually considered as a fundamental feature of a robot's attention system [37–42]. Robot attention models can be categorized into two groups: those which rely on fixed rules and those which rely on arithmetic equations. When fixed rules are employed based on a logical set of event conditions, the satisfaction of a given measure leads to the model selecting a person as the most attentive person. Use of adopted arithmetic equations involves computation of the attentiveness of each person; finally, people prioritization can be determined by a comparison of the computed attentiveness.

In an approach that utilized a set of event conditions based on locations of a sound source and human face [37], an attention system was proposed for receptionists and companion robots. The system operated under the assumption that there is a single sound source at a time. The rules for the selection of the most attentive person are defined as follows: (1) if the location difference between a located sound source and a detected human face in the robot's view is within $\pm 10^{\circ}$, the system associates the sound source with the human face. The person belonging to the associated face is determined as the most attentive person; (2) if the location difference changes such that it exceeds $\pm 30^{\circ}$ for three seconds, the system dissociates the sound source from that detected face, and the robot then loses its focus on the most attentive person; (3) step (1) and (2) are repeated.

A few years later, a focus of attention (FOA) system based on a detected speaking person was proposed [38]. Their method applied a multi-modal anchoring [39] for tracking a person of interest (POI). The only speaking person, who is facing the robot, can assume the role of POI. The event conditions are as follows: (1) a robot determines a speaking person as the POI (i.e., the most attentive person); (2) as long as the speech of the POI is anchored, other speaking people are ignored; (3) when the POI stops speaking for more than two seconds, the POI loses its speech anchor and another person can become the POI; (4) if no other person appearing in the robot's view is speaking, the previous POI remains the most attentive person.

POI selection based on the gazing direction of a human face and a sound source location [40] was presented. However, people attention prioritization cannot be achieved in this case. In short, the face direction validates detected sounds as voices and the robot only gives attention to a person facing the robot. The logical set of event conditions for selecting the most attentive person is as follows: (1) there is a person facing a robot; (2) a sound source is located and associated to the detected face; (3) the person associated with the detected face and sound becomes the speaking person and the POI.

A parameter of intimacy to determine the selection priority of an interactive partner using interaction distance was also proposed [41]. This proposed method is based on the concept of proxemics for communication between a robot and multiple people. Proxemics suggests that the more intimate the communication, the nearer the target person stands. Interaction distance is roughly classified into four groups: intimate distance, personal distance, social distance, and public distance. A person with the highest intimacy is determined as the most attentive person for an interaction, with parameters of the intimacy equation pre-defined heuristically.

A value representing the attentiveness of a person was presented by Bennewitz et al. [42]. This value is computed by a weighted sum of three multimodal factors where the weights are constant and heuristically decided. Three factors are: (1) the time when the person last spoke, (2) the distance of the person to the robot (estimated according to the size of a bounding box around a person's face), and (3) the person's location relative to the front of the robot. The person with the highest value is determined as the most attentive person and is given the robot's focused attention of attention of the robot. Attention prioritization is then simply achieved by sorting the magnitude of the computed values.

3. Introduction to Relevance Theory and Attention Model

This section introduces the concept of the relevance of observed features (i.e., stimuli-relevance) for attentiveness evaluation. Unlike previous attention approaches that employ heuristic parameters to calculate the attentiveness values of people, the stimuli-relevance values are computed probabilistically. Particularly, the proposed approach is derived based on a psychological theory of human communication methodology, called relevance theory. With this approach, a robot may evaluate attention as a person does during an interaction.

3.1. Relevance Theory in Multiple People-to-Robot Interactions

Relevance theory [29] explains a method of human communication that takes into account implicit inferences. Inferential communication not only intends to affect the thoughts of an audience but also seeks to elicit recognition from the audience that the communicator has an intention.

The method argues that individuals who engage in communication usually have the same notion of relevance in mind. To determine the most relevant communicator, an audience (in case of, a robot) searches for a certain meaning in any given communication situation and stops processing the situation when a meaning that fits the audience member's expectation of relevance is found (i.e., the communicator with maximum relevance is identified).

In human–human interaction, the ostensive-stimulus is an act by a human, produced during the interaction, which is appealing and mutually manifests between people. Ostensive-stimuli must satisfy two conditions: (1) they must attract the audience's attention, and (2) they must allow the attention of the audience to be focused on the communicator's intentions. In this work, the people-to-robot interaction (MPRI) can be illustrated as shown in Figure 1.

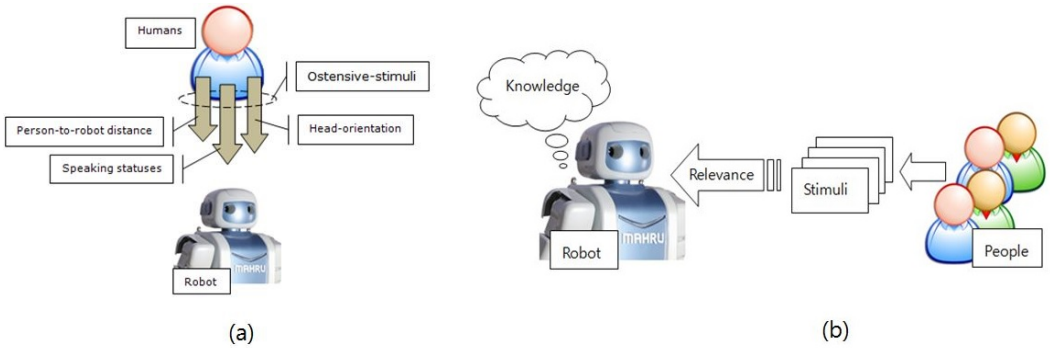

Figure 1. Ostensive-stimuli in people-to-robot interaction: (**a**) A robot and ostensive-stimuli of a single person; (**b**) relevance acquisition from observed ostensive-stimuli of multi-person with respect to robot's equipped knowledge.

Figure 1a depicts the robot perceiving ostensive-stimuli from a single person (i.e., the person-to-robot distance, the head-orientation, and the speaking statuses). Figure 1b shows an overview of the robot acquiring the relevance of ostensive-stimuli based on its equipped knowledge to understand people intentions. Particularly, let us clarify this situation as follows:

- A robot is the only audience.
- N persons possess N different levels of intention that should be understood by the robot.
- The intention of any person is to start communication with the robot and become the most attentive person.
- Human's ostensive stimuli are assumed to be confined to person-to-robot distance, head-orientation, and speaking statuses.
- The robot simultaneously receives N intentions of people, evaluated from all relevant ostensive stimuli, and searches for the person with the maximum relevance in terms of intention.

Hence, Figure 1 can be represented as the diagram of MPRI, as shown in Figure 2, in which the people can be considered as the sources of ostensive-stimuli.

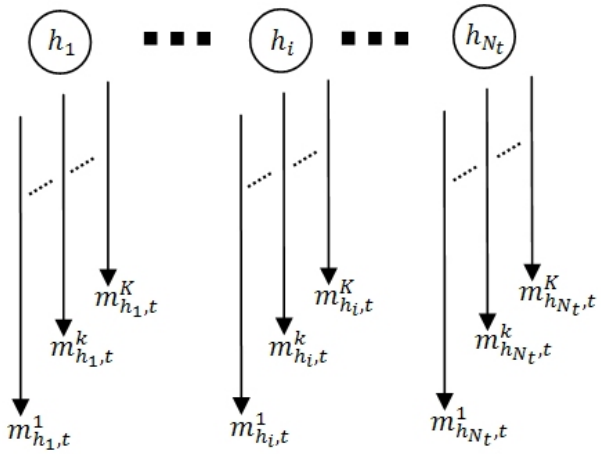

Figure 2. Participating people $\{h_i\}$ as sources of the ostensive-stimuli $m^k_{h_i,t}$, where $1 \leq i \leq N_t$ and $1 \leq k \leq K$. N_t implies the number of participants at time t, and K denotes the number of ostensive-stimuli.

Let us define a group of participants as $\{h_i\} = \{h_1, h_2, \ldots, h_{N_t}\}$, where N_t is the number of participants at time t and $1 \leq i \leq N_t$. Focusing on any person h_i, $m^k_{h_i,t}$ are the observed ostensive-stimuli, which are scalar, independent, and bounded with their observation ranges. $\mathbf{m}_{h_i,t} = [m^1_{h_i,t}, \ldots, m^k_{h_i,t}, \ldots, m^K_{h_i,t}]^T$ denotes an ostensive-stimuli vector of person h_i at t, where K is the number of ostensive-stimuli and $1 \leq k \leq K$. Hence, $\mathbf{o}_t = [\mathbf{m}^T_{h_1,t} \cdots \mathbf{m}^T_{h_i,t} \cdots \mathbf{m}^T_{h_{N_t},t}]^T$ becomes an ostensive-stimuli vector of a group of participants at t. Later, Section 4.1.1 describes probabilistic stimuli-relevance based on the relevance theory.

3.2. Attention Model's Structure

Attention model algorithms presented in past literature mostly involve algorithms with heuristic parameters for the determination of the most attentive person. As the use of fixed parameters can result in frequent undesired changes of states (i.e, undesired attention shifts) in situations with noisy observations. A probabilistic approach may be better suited to robot attention models.

Previously presented heuristic attention approaches simply define a detected speaking person as the most attentive person. Although it seems natural to employ the speaking status as the most influential stimulus, defining the speaker as the most attentive person over-emphasizes the importance of speaking, i.e., the selection of a person can be impacted by false detection of speaking status and should involve other stimuli, such as the person's head pan or the person's distance from the robot. The proposed approach differs from such heuristic approaches in that it considers multiple stimuli to order attentiveness, rather than solely speaking status.

The proposed probabilistic method focuses on improving the performance of determining the most attentive person and prioritizing people based on their attentiveness. To do so, both effective computation of attentiveness and adaptation to the changes in the number of participants (i.e., communicators) and observations (i.e., changes in person-to-robot distances, head pans, and speaking statuses) are taken into account. Three ostensive-stimuli are considered: person-to-robot distance, head pan of a person, and speaking status. Particularly, person-to-robot distance and the head pan of a person are treated as typical observations for computation of stimuli-relevance probabilities (Section 4.1.1). Speaking status is used for determining adaptable state transition probabilities in run-time (Section 4.1.2). Hence, with adjustable state transition probabilities, more flexible and efficient attentiveness determination, for a robot attention model, can be achieved.

The model's capability of coping with a change in the number of observations is useful for some situations, such as those with several associated observations, which may be occasionally inaccessible

or unimportant during operation. In such situations, by temporarily and effectively scaling down the number of observations, the proposed approach can still robustly compute the probabilities of observations. In this way, computational failure during run-time can be avoided. Furthermore, when the previously missing observations become available again, the approach simultaneously adapts its computation process of probabilities of observations according to the current number of observations and states. This issue is discussed in detail in Section 4.2.

4. Attentiveness Determination Using Scalable Hidden Markov Model

In this section, a Scalable HMM, based on the relevance theory, for attentiveness determination is described. The Scalable HMM recalls a similar approach [43]. Both the proposed model and the dynamic HMM are able to handle changes in the number of states during run-time. Differently, our proposed Scalable HMM is also capable of coping with changes in the number of observations attributed to changes in the number of states. Figure 3 depicts the main processes of the proposed approach in five parts. First, the probabilistic attentiveness computation based on stimuli-relevance (Section 4.1) presents the method by which the stimuli-relevance probabilities are computed using the three ostensive-stimuli (distance from person-to-robot, angle of a person's head pan and a person's speaking status). Next, an online probabilistic attentiveness analysis (Section 4.2) demonstrates the probabilistic computation between the previous and current states, for the case in which the number of detected persons changes. Section 4.3 explains how the most attentive person is selected and the attention prioritized in run-time. Finally, Section 4.4 describes how the Scalable HMM-based attention model is applied, using Particle Swarm Optimization (PSO), to the robot attention model.

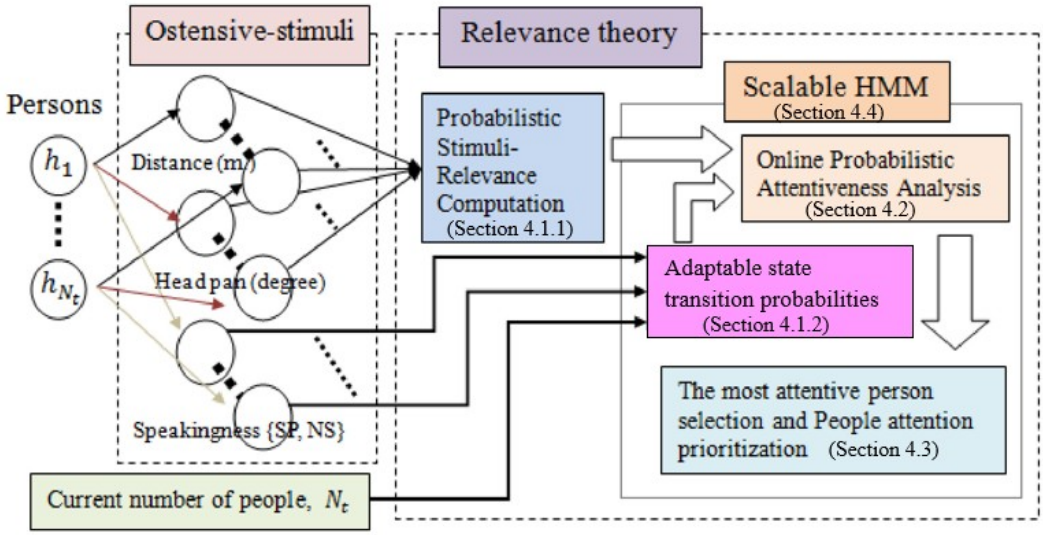

Figure 3. Attention model for the most attentive person selection and people attention prioritization based on relevance theory. HMM = Hidden Markov Model; NS = non-speaking status; SP = speaking status.

4.1. Probabilistic Acomputation Based on Stimuli-Relevance

Probabilistic attentiveness computation from three ostensive-stimuli (introduced in Section 3.2) is thoroughly explained in two parts: (1) Section 4.1.1 describes how stimuli-relevance probabilities from two stimuli (a person-to-robot distance and angle of a head pan) are obtained; (2) Section 4.1.2 depicts the adaptable state transition probabilities, which are used to flexibly consider state transition probabilities in run-time and thus increase efficient attentiveness determination for the robot attention model. In consideration of state transition probabilities, a person's speaking status, as well as the number of persons in camera view, are used to determine probabilistic attentiveness in run-time. Then, the probabilistic attentiveness is used to determine the most attentive person and arrange people attention prioritization, as discussed in the following sections.

4.1.1. Probabilistic Stimuli-Relevance Computation

As discussed in Section 3.1, an ostensive-stimulus both attracts the attention of a robot and tries to convey to the robot the meaning intended by the communicators. Because ostensive-stimuli are noisy, the probabilistic approach is considered as a potentially superior alternative to computing the stimuli-relevance of a given person.

In particular, human's relevance computation here consists of two fundamental properties. The first is the attraction of ostensive-stimuli, wherein a person's intention to communicate is conveyed through the emission of stimuli in the form of probability density function (pdf). The other is restraint of ostensive-stimuli, wherein a person has no intention to emit particular stimuli, and is also in the form of pdf. Considering $\mathbf{m}_{h_i,t}$ and the ostensive-stimuli vector of any person h_i, the probabilities of attraction, $P_i(\mathbf{m}_{h_i,t})$, and restraint, $\bar{P}_i(\mathbf{m}_{h_i,t})$, can be defined by

$$P_i(\mathbf{m}_{h_i,t}) = \prod_{k=1}^{K} c_k(m^k_{h_i,t}), \tag{1}$$

$$\bar{P}_i(\mathbf{m}_{h_i,t}) = \prod_{k=1}^{K} \bar{c}_k(m^k_{h_i,t}), \tag{2}$$

where $c_k(m^k_{h_i,t})$ and $\bar{c}_k(m^k_{h_i,t})$ denote the attraction and the restraint distributions of the kth ostensive-stimulus, respectively.

From cognitive psychology, attention is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information [44], whether deemed subjective or objective, while ignoring other perceivable information. Building on this, we define a state variable, $q_t \in \{s_1, \ldots, s_i, \ldots, s_{N_t}\}$, where a person h_i has an intention to start communication with the robot while others have no intention to do it. Consider $\{h_i\}$ as a group of participating people. Using Equations (1) and (2), the probability of relevance of the ostensive-stimuli of given a state s_i, $P(\mathbf{o}_t|q_t = s_i)$, is defined as follows:

$$P(\mathbf{o}_t|q_t = s_i) = P_i(\mathbf{m}_{h_i,t}) \prod_{j \neq i} \bar{P}_j(\mathbf{m}_{h_j,t}), \quad 1 \leq i, j \leq N_t,\ i \neq j. \tag{3}$$

For example, if the current number of participating people is N_t, Equation (3) becomes:

$$\begin{aligned}
P(\mathbf{o}_t|q_t = s_1) &= P_1(\mathbf{m}_{h_1,t})\bar{P}_2(\mathbf{m}_{h_2,t}) \cdots \bar{P}_{N_t-1}(\mathbf{m}_{h_{N_t-1},t})\bar{P}_{N_t}(\mathbf{m}_{h_{N_t},t}) \\
P(\mathbf{o}_t|q_t = s_2) &= \bar{P}_1(\mathbf{m}_{h_1,t})P_2(\mathbf{m}_{h_2,t}) \cdots \bar{P}_{N_t-1}(\mathbf{m}_{h_{N_t-1},t})\bar{P}_{N_t}(\mathbf{m}_{h_{N_t},t}) \\
&\vdots \\
P(\mathbf{o}_t|q_t = s_{N_t}) &= \bar{P}_1(\mathbf{m}_{h_1,t})\bar{P}_2(\mathbf{m}_{h_2,t}) \cdots \bar{P}_{N_t-1}(\mathbf{m}_{h_{N_t-1},t})P_{N_t}(\mathbf{m}_{h_{N_t},t}).
\end{aligned} \tag{4}$$

Note that Equation (1)−(3) illustrate good scalability in the sense that $P(\mathbf{o}_t|q_t = s_i)$ adapts efficiently with respect to the number of participating people and observations in run-time. As a result, effective computation of attentiveness of people can be achieved.

4.1.2. Online Adaptable State Transition Probabilities

This section introduces online adjustable state transition probabilities based on the speaking statuses of participants. Speaking statuses of persons are used as the conditional parameter in the model. Furthermore, the current number of participants are also taken into account. Hence, an effective and improved computation of attentiveness can be achieved in terms of sensitivity and adaptability with respect to speakers and changes in the number of participants.

Let us denote $Y_{h_j,t-1} \in \{\text{NS}, \text{SP}\}$ as the speaking statuses of person h_j at time $t-1$, where NS is the non-speaking status and SP is the speaking status. The state transition probability distribution given a person's speaking status, $P(q_t = s_j | q_{t-1} = s_i, Y_{h_j,t-1})$, is denoted by

$$P(q_t = s_j | q_{t-1} = s_i, Y_{h_j,t-1}) = \begin{cases} \frac{\rho_{\text{ns}}}{\text{R}} & \text{if } i = j, Y_{h_j,t-1} = \text{NS}; \\ \frac{\rho_{\text{sp}}}{\text{R}} & \text{if } i = j, Y_{h_j,t-1} = \text{SP}; \\ \frac{1}{\text{R}} & \text{if } i \neq j, Y_{h_j,t-1} = \text{NS or SP}, \end{cases} \tag{5}$$

where ρ_{ns} and ρ_{sp} define the sensibility parameters of the attention model, influencing the sensitivity of the robot's attention shifts with regard to the speaking and non-speaking persons during run-time. Further, $N_{t-1} \times N_{t-1}$ is the dimension of the state transition matrix.

The state transition matrix is now designed such that the transition to the same person (state) is ρ_{ns} or ρ_{sp} times more likely than the transitions to other persons, conditioned according to the previous speaking status of that person, whether it was non-speaking or speaking, respectively. Note also that $1 \leq \rho_{\text{ns}} < \rho_{\text{sp}}$, and R is a normalizing constant used to ensure that each row of the state transition matrix sums to 1.

4.2. Online Probabilistic Attentiveness Analysis

Relevance theory states that a person retrieves relevance assumptions stored in their memory (knowledge of ostensive-stimuli with respect to a situation) and processes them with an inferential procedure to draw a conclusion.

To implement this procedure into the robot's attention model in a similar manner, the inferential procedure can be mathematically emulated by statistical inference. The inference is performed by using quantitative data. A greater informativeness quantity results in better accuracy of inference.

For the analysis of attentiveness of a person h_j until the current time t, we consider the probability of relevance of the partial observation sequence until time t, $\mathbf{O}_{1:t} = \{\mathbf{o}_1\ \mathbf{o}_2 \cdots \mathbf{o}_t\}$ and state $q_t = s_j$ given a robot's attention model λ. This yields:

$$\alpha_t(j) = P(\mathbf{o}_1\ \mathbf{o}_2 \cdots \mathbf{o}_t, q_t = s_j | \lambda). \tag{6}$$

Efficiently, we can solve for $\alpha_t(j)$ from Equation (6) inductively, as follows:

(1) Initialization

$$\alpha_1(i) = \pi_i\ P(\mathbf{o}_1 | q_1 = s_i), \quad \pi_i = \frac{1}{N_1},\ 1 \leq i \leq N_1. \tag{7}$$

(2) Induction

(2.1) Validation
- Checking the current number of participants.
- Iterating over states at $t-1$ and t for states comparison and validation.
Correcting state indexes, if required.

(2.2) Computation

case 1: $N_{t-1} = N_t$

$$\alpha_t(j) = \sum_{i=1}^{N_t} \left[\alpha_{t-1}(i) a_{ij}\right]\ P(\mathbf{o}_t | q_t = s_j), \quad 1 \leq j \leq N_t, \tag{8}$$

case 2: $N_t > N_{t-1}$

$$\alpha_t(j) = \begin{cases} \sum_{i=1}^{N_{t-1}} \left[\alpha_{t-1}(i)a_{ij}\right] P(\mathbf{o}_t|q_t = s_j), & 1 \le j \le N_{t-1}, \\ \pi_j P(\mathbf{o}_t|q_t = s_j), & \pi_j = \frac{1}{N_t},\ N_{t-1} < j \le N_t \end{cases} \tag{9}$$

case 3: $N_t < N_{t-1}$

$$\alpha_t(j) = \sum_{i=1}^{N_{t-1}} \left[\alpha_{t-1}(i)a_{ij}\right] P(\mathbf{o}_t|q_t = s_j), \quad 1 \le j \le N_t, \tag{10}$$

where $\pi = \{\pi_i\}$ is the initial state distribution of Scalable HMM for the attention model, and $a_{ij} = P(q_t = s_j|q_{t-1} = s_i, Y_{h_j,t-1})$ are the state transition probabilities.

Figure 4 illustrates the induction procedure, showing how state s_j can be reached at time t from the N_{t-1} possible states at previous time $t-1$. Prior to $\alpha_t(j)$ computation, the validation process (Step (2.1)) must be performed at each induction step. The process, which has $O(N^2K)$ computational complexity, validates the current number of states (the participants in a robot's view). In the case of a decrease in the number of participants, an index re-correction of the participants may be required, such that computation failure can be avoided in run-time.

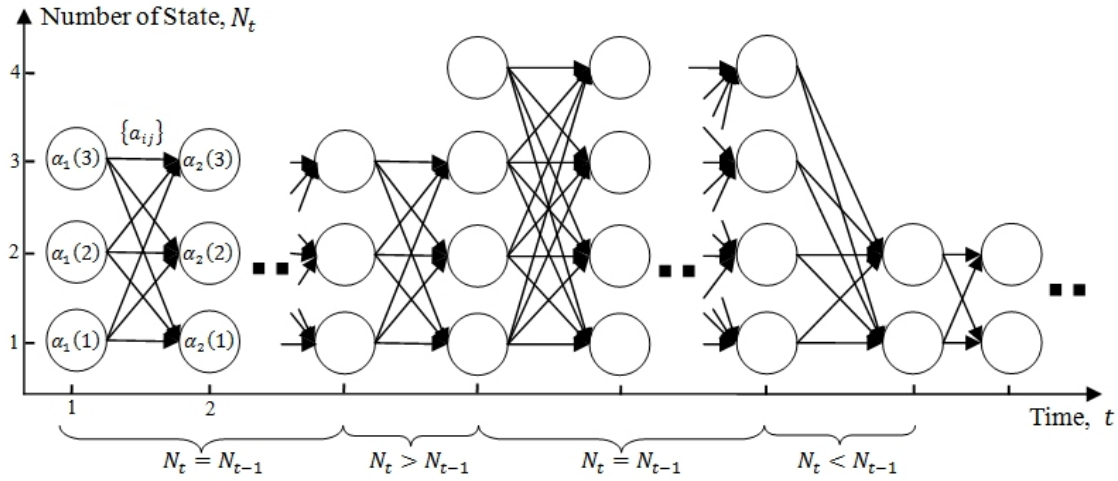

Figure 4. Illustration of the implementation for $\alpha_t(i)$ computation in terms of a lattice of observations and states during different cases: $N_t = N_{t-1}$, $N_t > N_{t-1}$, and $N_t < N_{t-1}$.

4.3. Online Most Attentive Person Selection and People Attention Prioritization

For selecting the most attentive person and prioritizing people based on their attentiveness, the proposed attention model first evaluates the probabilistic stimuli-relevance of participants. Next, the probabilistic attentiveness of each person, $\alpha_t(j)$, is computed.

Finally, the most attentive person, denoted by q_t^*, is determined as the person with the maximum attentiveness. Hence, q_t^* is denoted by

$$\begin{aligned} q_t^* &= \underset{1\le j\le N_t}{\text{argmax}}[P(q_t = s_j|\mathbf{O}_{1:t})] \\ &= \underset{1\le j\le N_t}{\text{argmax}}[P(\mathbf{O}_{1:t},\ q_t = s_j)/P(\mathbf{O}_{1:t})] \\ &= \underset{1\le j\le N_t}{\text{argmax}}[P(\mathbf{O}_{1:t},\ q_t = s_j)] \\ &= \underset{1\le j\le N_t}{\text{argmax}}[\alpha_t(j)]. \end{aligned} \tag{11}$$

By comparing $\{\alpha_t(j)\}$ for current participants at t, the prioritization of participants with respect to their attentiveness can be achieved.

4.4. Learning Approach for Scalable HMM-Based Attention Model Using Particle Swarm Optimization

In general, the HMM parameters are estimated using the Baum–Welch algorithm [45]. However, it is well known that the Baum–Welch algorithm easily converges to local optimum solutions. To find the global solution or better optimum solutions, estimating HMM parameters using Particle Swarm Optimization (PSO) [25,46,47] has been an alternative method, showing superior results compared to the conventional Baum–Welch method. Further, the PSO algorithm also provides a simple method for solving complex optimization problems. Therefore, we apply a training approach based on PSO for our robot attention model.

The PSO-based learning approach is briefly introduced in this section. Let us denote $\psi = \{\rho_{\text{ns}}, \rho_{\text{sp}}, \{\mu_k, \bar{\mu}_k\}, \{\sigma_k^2, \bar{\sigma}_k^2\}\}$ as a vector of system parameters to be estimated, where $\{\mu_k, \bar{\mu}_k\}$ and $\{\sigma_k^2, \bar{\sigma}_k^2\}$ are the means and variances of attraction and restraint distributions of ostensive-stimuli, $1 \leq k \leq K$, respectively.

In the PSO-based learning approach, the model is encoded into a string of real numbers. The vector ψ acts as a particle, representing the position vector $\mathbf{x}_i$. With each position vector $\mathbf{x}_i$, there is an associate velocity vector $\mathbf{v}_i$, modeling the capacity of the particle to move from a given position $\mathbf{x}_i^z$ at the zth iteration to another position $\mathbf{x}_i^{z+1}$ in a successive iteration of the space solution sampling process.

The initial positions $\mathbf{X}^0 = \{\mathbf{x}_i^0; i = 1, 2, \ldots, N_p\}$ and their velocities $\mathbf{V}^0 = \{\mathbf{v}_i^0; i = 1, 2, \ldots, N_p\}$ of N_p particles of the swarm can be randomly generated [48]. The ranges can differ for different dimensions of particles.

The degree of optimality of each particle is evaluated at the zth iteration by computing its $\log P(\mathbf{O}|\psi)$. The fitness function is defined as follows:

$$P(\mathbf{O}^m|\psi) = \sum_{j=1}^{N} \alpha_T^m(j) \tag{12}$$

$$f(x_i) = \log P(\mathbf{O}|\psi) = \frac{1}{M} \sum_{m=1}^{M} \log P(\mathbf{O}^m|\psi), \tag{13}$$

where $\mathbf{O}^m = \{\mathbf{o}_1^m \ \mathbf{o}_2^m \cdots \mathbf{o}_T^m\}$ is the m^{th} observation sequence. The previous best particle, **pbest**, storing the best position that has been reached up to now by the ith particle, is found by $\mathbf{pbest}_i^z = \text{argmax}_{1 \leq h \leq z} \left[f(x_i^h) \right]$. Next, the global best particle, **gbest**, which is the optimum position in the overall swarm, can be computed by $\mathbf{gbest}^z = \text{argmax}_{1 \leq i \leq N_p} \left[f(x_i^z) \right]$.

The velocity of the d^{th} of each particle is updated with dynamic inertia as follows:

$$\mathbf{v}_{i,d}^{z+1} = w^z \mathbf{v}_{i,d}^z + \phi_1 r_1 (\mathbf{pbest}_{i,d}^z - \mathbf{x}_{i,d}^z) + \phi_2 r_2 (\mathbf{gbest}_d^z - \mathbf{x}_{i,d}^z) \tag{14}$$

$$w^z = w_{\max} - \frac{w_{\max} - w_{\min}}{\text{iter}_{\max}} z, \tag{15}$$

where r_1 and r_2 are two uniformly distributed random positive numbers, used to provide the stochastic weighting. w is the inertia weight, affecting the influence of the old velocity on the new velocity. ϕ_1 and ϕ_2 are constants, called cognition and social acceleration, respectively. Next, the particle position is then updated as follows:

$$\mathbf{x}_{i,d}^{z+1} = \mathbf{x}_{i,d}^z + \mathbf{v}_{i,d}^{z+1}. \tag{16}$$

The velocity and position updating, and the optimization process are stopped when the condition of termination is satisfied. Finally, **gbest** is assumed to be the optimum solution for the model.

The terminating condition is that the maximum number of iterations, $\text{iter}_{\max}$, is reached ($z = \text{iter}_{\max}$) or the increase of the optimum fitness is below a given threshold (i.e., $|f(\mathbf{gbest}^z)| - |f(\mathbf{gbest}^{z-1})| < threshold$).

5. Experiments

To evaluate the performance, the proposed method was tested with 10 image sequences consisting of a total of 7567 frames, displaying individuals who were speaking, walking, turning their head, and entering or leaving a robot's view. None of the 10 image sequences were used during the training phase. The performance results were compared to the performance of a human and two widely used attention approaches: a set of event conditions [38] and a heuristic algorithm [42].

5.1. Experimental Setup

This section gives an overview of the experiment setup and experimental scenarios used to verify the performance of the proposed robot attention model. Multiple persons stood in front of the robot and far away from the robot, by 2–3 m. In our experiment, three persons were a suitable number for our devices to be able to observe them fully in the camera's view while they were still in the range of sound communication. For all image sequences, three attention-related features (speaking status, the distance between a person and a robot, and a person's head pan angle), were determined visually and automatically using approaches described in Appendix A. The proposed attentiveness computation model and all feature detection algorithms have been implemented using a PC equipped with a 3GHz Pentium 4 CPU, with 1GB of RAM, and an NVIDIA GeForce 6600.

Figure 5 illustrates the experiment setup; participating people stand in front of MAHRU-M, which has a Bumblebee2 stereo camera (www.ptgrey.com) attached on its head. MAHRU-M is a mobile humanoid robot platform based on a dual-network control system and coordinated task execution [49].

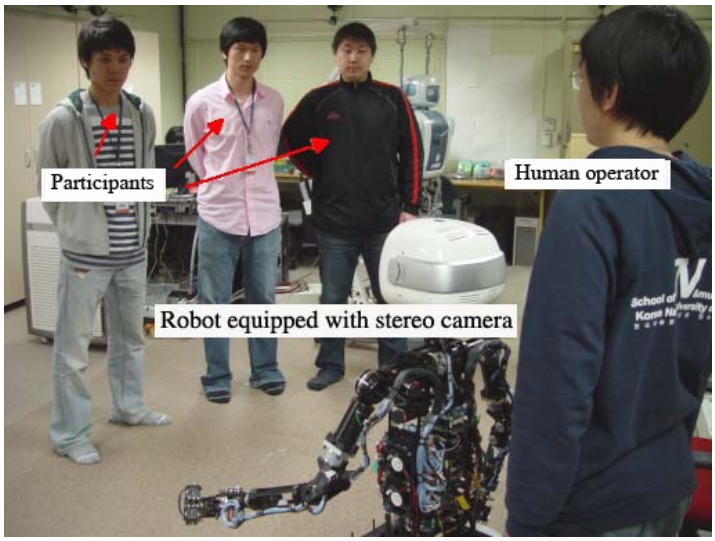

Figure 5. Our experimental setup.

Figure 6 illustrates various situations with participants randomly performing actions (speaking, head turning, walking toward or away from a robot, and entering or leaving a robot's view). The first row (Figure 6a) shows an image sequence of three participants randomly turning their heads, exhibiting speaking or non-speaking intervals, and walking towards or away from the robot. The second and the third rows (Figures 6b and 6c, respectively) illustrate the image sequences, showing an individual entering or leaving the robot's view.

The illumination conditions were natural and not controlled in any of the collected image sequences. As a result, the face size and the lighting conditions for the persons in different locations in

the robot's view differed, as shown in Figure 7. The illumination can be calculated and represented by the luma component (Y′) in Y′UV color space which is the weighted sum of RGB components of a color image. In Figure 7, the brightness and its variation of a face image of each person is demonstrated by the mean of luma ($\bar{Y}'$) and its standard deviation ($\sigma_{Y'}$).

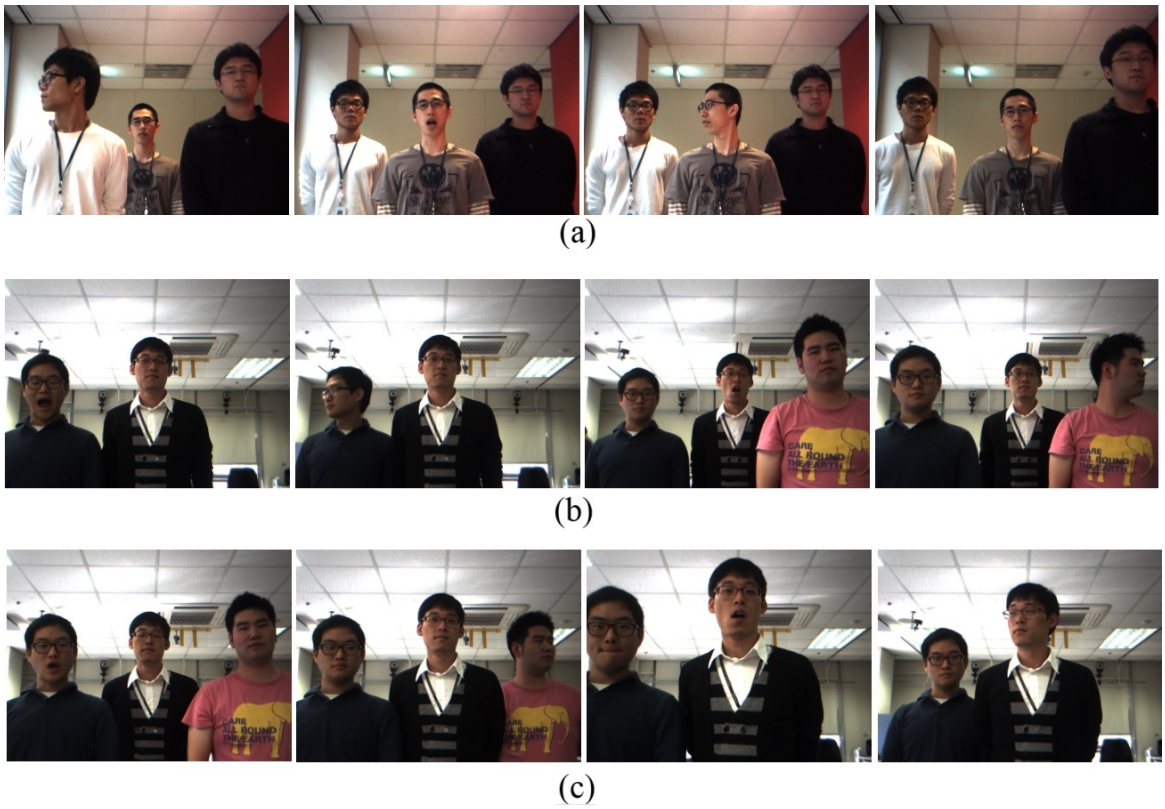

Figure 6. Sample images displaying individuals who exhibited speaking, walking, turning head, and entering/leaving a robot's view.

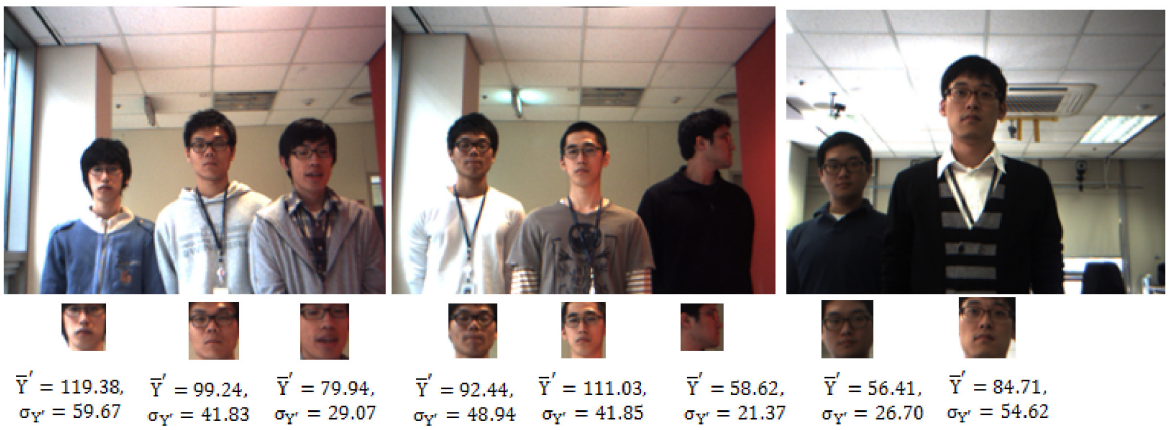

Figure 7. Face images of participants in different locations, sizes, and lighting conditions, in which ($\bar{Y}'$) is the mean of luma and its standard deviation ($\sigma_{Y'}$).

5.2. Attraction and Restraint Distributions of Ostensive-Stimuli

According to Figure 3, two ostensive-stimuli are involved in the computation of probabilistic stimuli-relevance in our attention model. Hence, the corresponding number of ostensive-stimuli is two (i.e., $K = 2$; $k = 1$ corresponds to the person-to-robot distance, and $k = 2$ corresponds to head

orientation (or head pan)). To design attraction and restraint distributions for each ostensive-stimulus, the stimuli must be analyzed with regard to their contributions to attentiveness.

Hence, we examined how much attention a person might give to someone who stands at different distances and looks in different directions. Intuitively, we assume that a person is likely to pay less attention to people who are further away or looking away, compared to those who are closer distances or looking directly at the person.

The attraction and restraint distributions for both the person-to-robot distance and the head pan angle can be reasonably designed by a folded normal distribution [50]. Conveniently, the folded normal distribution can be tuned such that it satisfactorily delivers the interpreted distribution's characteristics of both stimuli, using the mean μ and variance σ^2 as parameters.

Figure 8 illustrates $\{c_k, \bar{c}_k\}$, designed by the folded normal distributions, of the kth ostensive-stimulus of person h_i, in which $[\mu_k, \bar{\mu}_k]$ denote the measuring scope of the kth ostensive-stimulus. Here, the measuring scopes of the person-to-robot distance and head-pan angles in our attention model were set to $[\mu_1 = 0.5\text{ m}, \bar{\mu}_1 = 2.0\text{ m}]$, and $[\mu_2 = 0°, \bar{\mu}_2 = 90°]$, respectively.

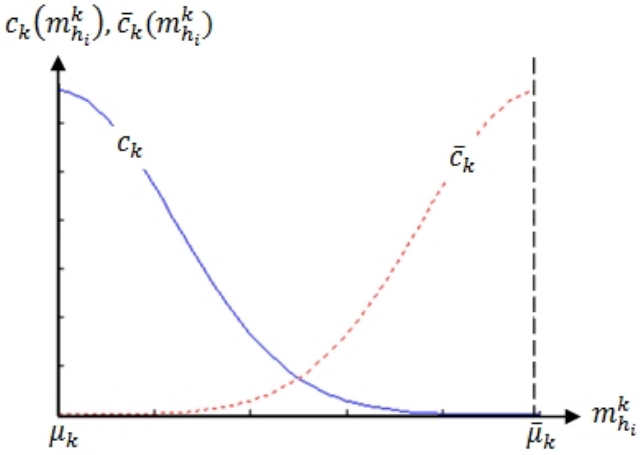

Figure 8. Probability distributions of attraction, c_k, and restraint, $\bar{c}_k$, of k^{th} ostensive-stimulus of person h_i ($k = 1$ indicates distance of person-to-robot and $k = 2$ indicates head pan).

6. Results

The human mind is a complex entity that represents a particular characteristic of people [51]. Every person has an individual opinion regarding who is the most attentive person during an interaction. Because of this, the common ground truth for determining the selection criteria for the most attentive person is difficult to practically determine.

Hence, to evaluate the proposed attention model, attention evaluation experiments were conducted with people. In these human evaluations of the most attentive person including attention prioritization were obtained. Ten users participated with the experiments on the same 10 image sequences used for testing the proposed attention model with the robot as the attention evaluator. They were asked to watch videos of image sequences of interacting people (see Figure 6), to evaluate who is the most attentive person, and prioritize people based on attentiveness. The users were also asked to consider only speaking statuses, distance, and head pan for the evaluation of attentiveness. Finally, for each image sequence, the most likely outcomes of the human evaluation were obtained by finding the maximum among user decisions.

The probabilities of detection (P_d), false alarm (P_f), and attention shift during intervals (P_s) were used as performance indicators. P_d indicates the attention model's performance regarding the detection of the most attentive person and the detection of people prioritization with respect to attentiveness. Let us denote $P_{d,\text{m}}$ as the ratio of the frames where the most attentive person is correctly detected to the

total number of frames with the most attentive person. $P_{d,\mathrm{p}}$ is the ratio of the frames where people's attention is correctly prioritized to the total number of frames.

$P_{f,\mathrm{m}}$ is the ratio of the frames where a person who is not the most attentive person is incorrectly detected as the most attentive person to the total number of frames where the given person is not the most attentive person. Finally, P_s is calculated as follows:

$$\begin{aligned} P^i_{s,\mathrm{m}} &= \frac{\text{number of transitions from MP state to } \neg\text{MP state in MP intervals}}{\text{number of MP frames}} \\ P^i_{s,\neg\mathrm{m}} &= \frac{\text{number of transitions from } \neg\text{MP state to MP state in } \neg\text{MP intervals}}{\text{number of } \neg\text{MP frames}} \\ P_s &= \sum_{i=1}^{N} \left[\frac{P^i_{s,\mathrm{m}} + P^i_{s,\neg\mathrm{m}}}{2} \right], \end{aligned} \tag{17}$$

where N is the total number of participants. MP and ¬MP refer to "the most attentive person" and "not the most attentive person," respectively.

In the experiments with our proposed attention model, the model parameters were estimated with the PSO approach described in Section 4.4 using the training data (sequences of the observed ostensive-stimuli of people: speaking statuses, person-to-robot distance, and head-pan angle).

In the experiments with the two other attention methods, i.e., the set of event conditions [38] and the heuristic equations [42], the respective parameters of each approach were determined according to recommendations described in their studies. The same observations used in our proposed method were also used in these two approaches. A brief description of each approach is provided in Section 1.

Figure 9 demonstrates the observed ostensive-stimuli of a single person (person 1) from one of the image sequences. The first three graphs from the top depict the detected speaking status of person 1, the person's estimated distance from the robot in meters, and the estimated head-pan angle in degrees over time, respectively. The fourth graph illustrates the probabilistic attentiveness result of person 1, computed by the proposed attentiveness computation model. Sample images on the right side of the figure illustrate the action sequences of people in this image sequence, in which the people are labeled as person no. 1, person no. 2, and person no. 3.

In frame 25, person no. 1 was approximately 1.9 m away from the robot and was looking relatively straight at the robot. He had a respectively low attentiveness of 0.1185. Next, in frame 122, his attentiveness became higher ($\approx$ 0.98) because he came closer to the robot ($\approx$ 1.4 m away). At frame 480, person no. 1 had a very low attentiveness of 0.012 because he was very far away from the robot ($\approx$ 2 m), even though he looked straight at the robot. However, his attentiveness rose gradually as he spoke, and his attentiveness became 0.7 at frame 579.

Figure 10 illustrates the attention outcomes of one image sequence of a situation of two participants and the observed ostensive-stimuli of each person. The sample images of the sequence are also shown at the bottom of the figure. The most attentive person is indicated by a rectangle, and the attentiveness ranking is labeled by a number above each person.

At frame 50, person no. 1 was detected as a speaking person. His attentiveness became larger than the attentiveness of person no. 2. Examining frames 100 to 120, the proposed attention model chose person no. 2 as the most attentive person, while human evaluation considered person no. 1 as the most attentive person. The outcomes were different because both participants were located at similar distances, making it difficult for people to determine whether person no. 1 or person no. 2 was closer. Consequently, the most attentive person selection based on distance becomes critical. However, in the case of attentiveness quantification by a robot, the distance is computed as a real number, so determining the closest person among participants is an easy task.

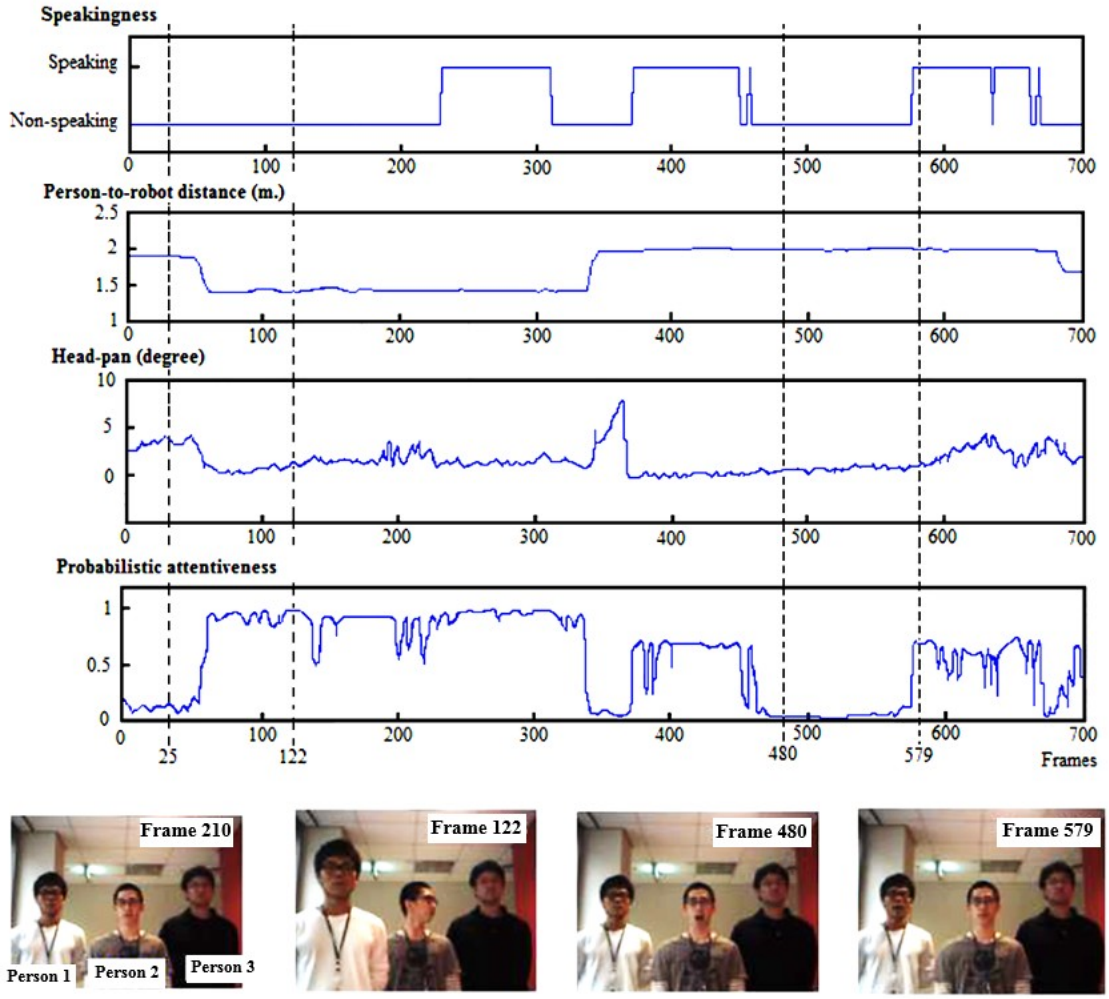

Figure 9. Observed attention-related features (ostensive-stimuli) of a single person (person 1) from one of the image sequences and his probabilistic attentiveness result.

At frames 129 and 210, the participant (person no. 1) turned his head and looked away from the robot. This resulted in a decrease in his attentiveness, and made another participant become more interesting to the robot. As a result, the other participant's attentiveness significantly increased. Frame 250 demonstrates a possible error in the selection of the most attentive person caused by consecutive false speaking status detection.

Figure 11 depicts the attention outcomes of one image sequence of a situation of three participants. The middle graph shows the selection of the most attentive person over time compared to the human evaluation, which is illustrated in the top graph. The bottom graph depicts the computed probabilistic attentiveness of participants in the robot's view.

In frames 350–370 (Figure 11), the proposed attention approach chose person no. 2 as the most attentive person instead of person no. 1, and there were several undesired attention shifts. These were caused by continuous errors in the estimated observations. These errors were due to consecutive errors in the estimation of the head-pan angles of person no. 2. Hence, despite the effective probabilistic attentiveness computation for the most attentive person selection and people attention prioritization, our approach cannot withstand extreme error in observations if the error occurs continuously for a long period of time.

The proposed approach performed well, as expected in terms of determining the most attentive person (Figure 10 at frames 50, 129, and 210, and Figure 11 at frame 60, 140, and 663). Even when there was no speaking person present, the proposed approach was able to determine the most attentive

person, such as in frames 70–300 in Figure 10 and in frames 140 and 663. In the absence of a speaking person, the transition probabilities of the robot's FOA become equally well-distributed according to the adaptable state transition probabilities described in Section 4.1.2. As a result, the selection procedure of the most attentive person is efficiently conducted and altered by other visual features.

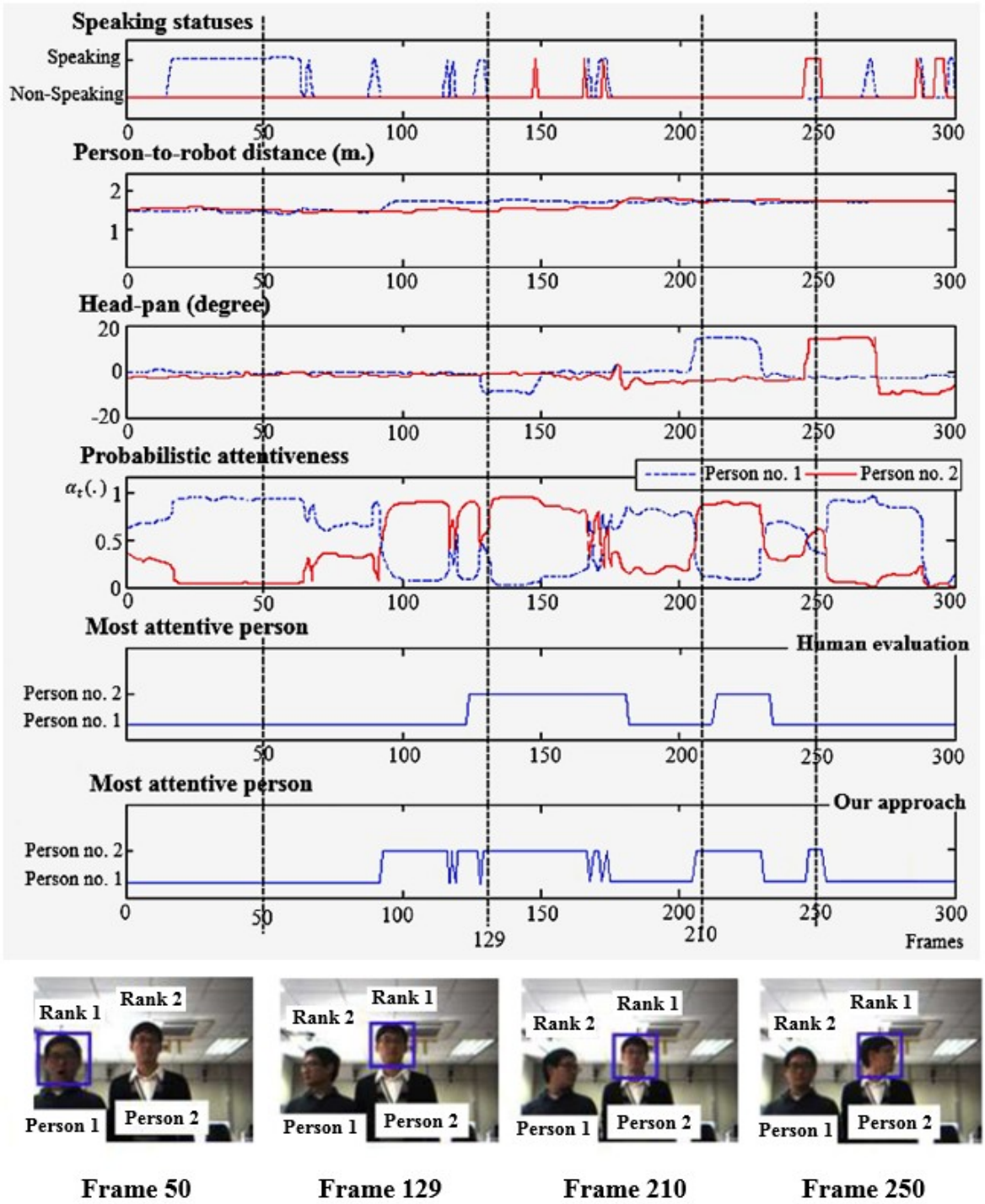

Figure 10. Robot's attention results, from one of the image sequences, showing observed attention-related visual features, the most attentive person selection, and people attention prioritization outcomes. The most attentive person is indicated by a rectangle and the attentiveness ranking is indicated by the number on the top of each person.

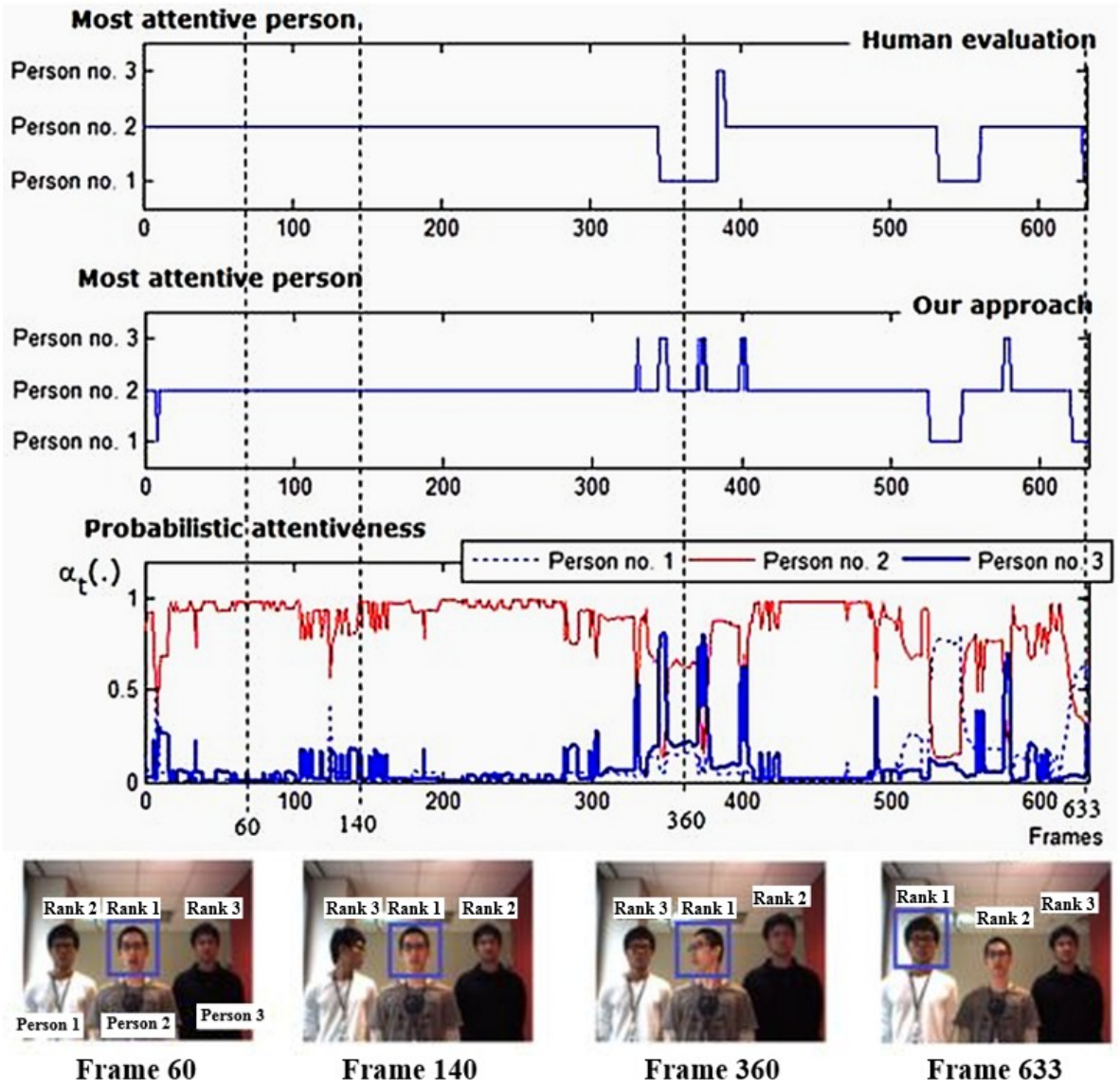

Figure 11. Robot's attention results, showing a performance of our attention approach in the sense of constant in the number of participants in a robot's view (three persons situation).

Figure 12 depicts the attention outcomes of one image sequence of a situation, in which there is a change in the number of participants. At the beginning, there were two participants in the robot's view. Our proposed attention approach prioritized these two persons with respect to the computed attentiveness. Starting from frame 289, a new person appeared and stayed in the robot's view. The current number of people in a robot's view became three. That person was included automatically and seamlessly into the attention model, and his attentiveness was calculated and compared with those of the other participants. This confirms the scalability of our proposed attention model based on the Scalable HMM in terms of the change in the number of people and observations. The model scalability also applies to situations with a decreasing number of participants.

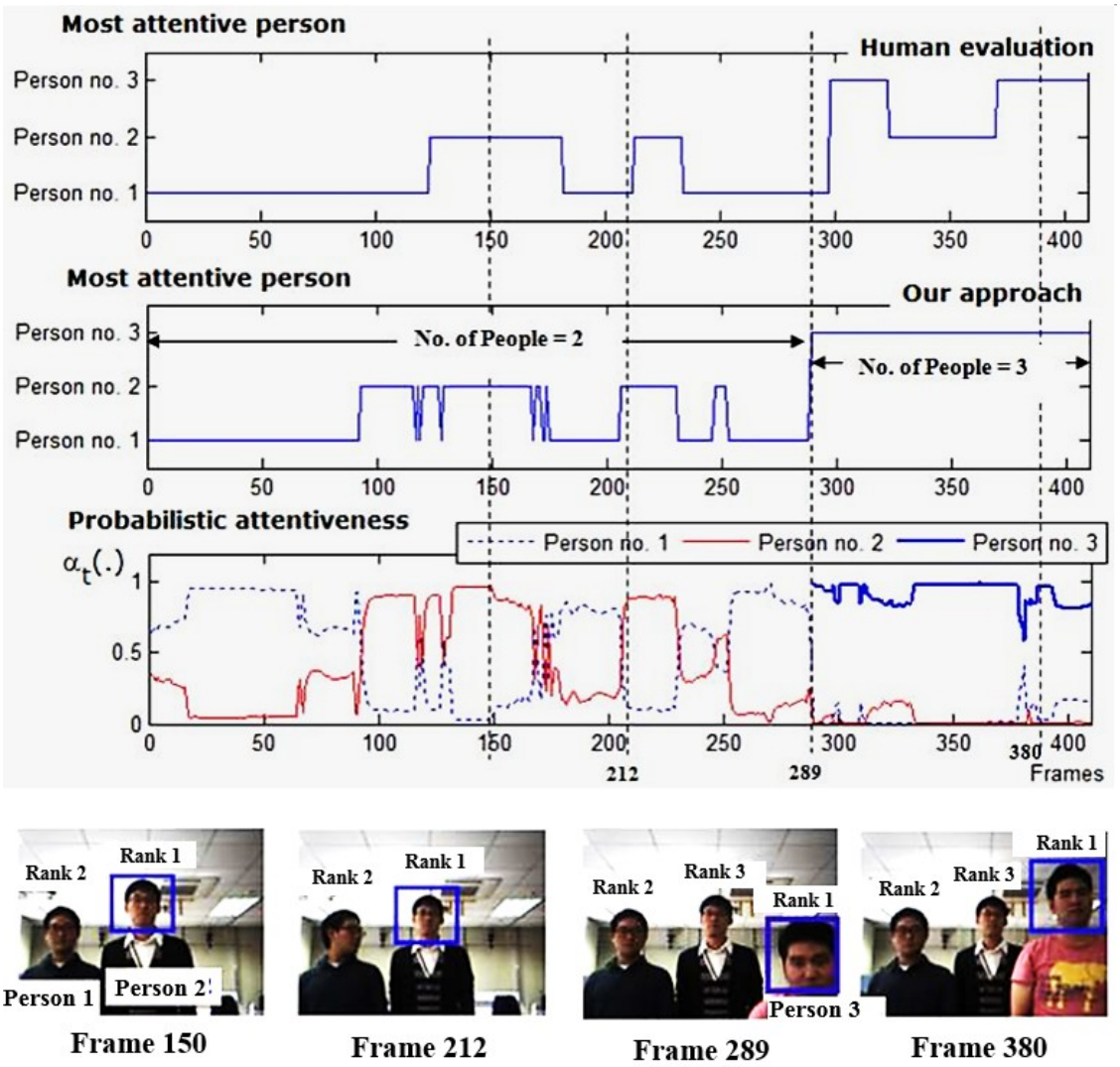

Figure 12. Robot's attention results, showing a performance of our attention approach in the sense of change in the number of participants in a robot's view (two persons to three persons situation).

The set of event conditions [38] for the determination of a robot's FOA are listed as follows:

- If the robot detects a speaking person, the speaker becomes the most attentive person.
- As long as the person is speaking, the speaking of other people is ignored.
- When the attentive person stops talking for more than two seconds, the robot loses its anchor on that person as being the most attentive person.
- Only a speaking person can take over the role of the most attentive person.

For the heuristic approach, the weighted sum approach [42] is tested. Five sets of pre-defined weights (Figure 13) for the three ostensive-stimuli were investigated to explore the approach's performance, where $w = [w_d, w_h, w_s]$ is a set of weights, and w_d, w_h, and w_s are weights for distance, head-pan angle, and speaking status, respectively.

The performance comparison in terms of the most attentive person selection using the receiver operating characteristic (ROC) space is shown in Figure 13a,b. The ROC curve is a graphical plot which illustrates the performance of a system via the comparison of two relative operating characteristics [52,53]. Figure 13c shows a performance comparison with respect to both the most attentive person selection and people attention prioritization. The plots show that our proposed robot attention model outperforms these two heuristic attention approaches.

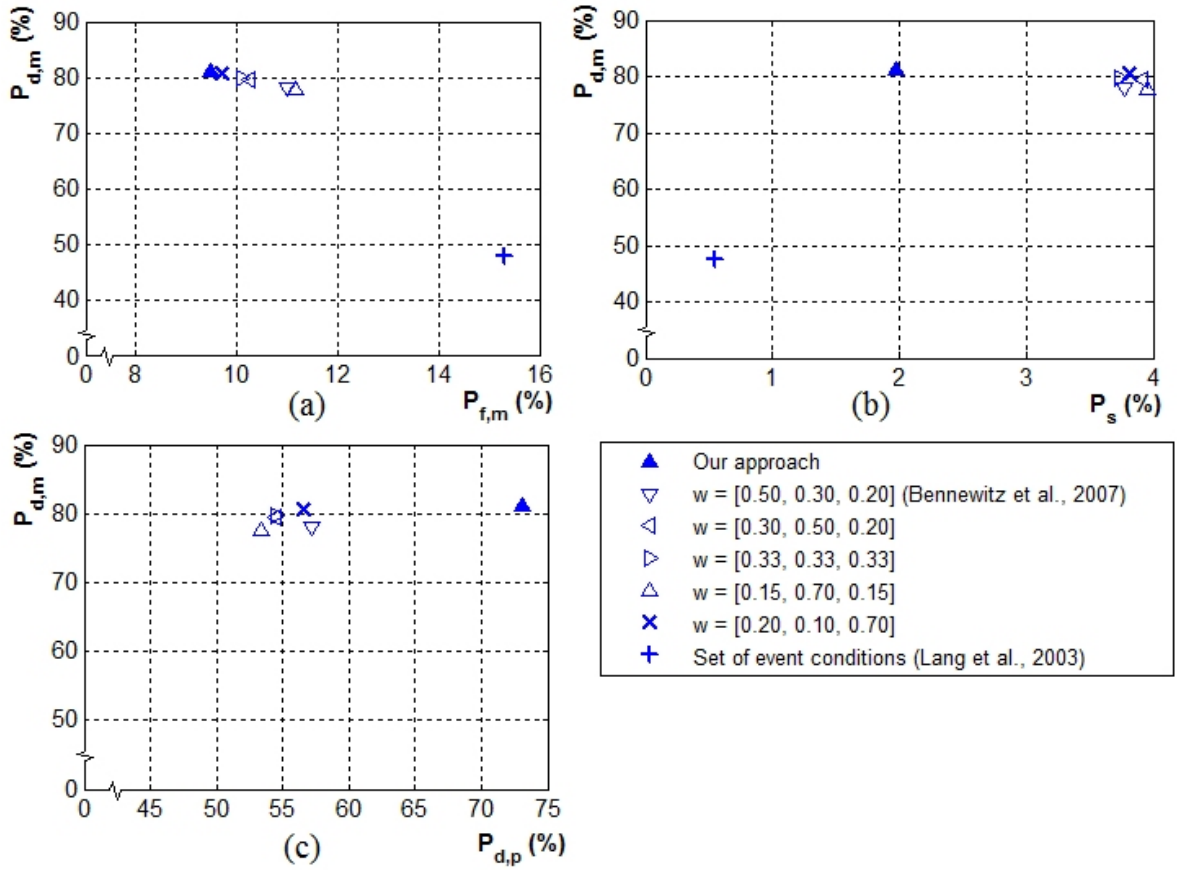

Figure 13. Performance comparison: (**a**,**b**) receiver operating characteristic (ROC) most attentive person detection space and plots of the proposed approach and two other attention approaches; (**c**) probability plots of the most attentive person detection $P_{d,\mathrm{m}}$ vs. the people attention prioritization $P_{d,\mathrm{p}}$.

Our attention model succeeded in obtaining a high detection rate of the most attentive person (≈76% of $P_{d,\mathrm{m}}$) and the highest detection rate of people attention prioritization (≈75% of $P_{d,\mathrm{p}}$) compared to the other two approaches. The proposed approach also achieved a small rate of attention shift during intervals, P_s, which was only (≈2%). Additionally, $P_{d,\mathrm{m}}$ was improved by over 30% compared to Lang et al.'s approach (the approach using the set of event conditions) and by almost 3% compared to Bennewitz et al.'s weighted sum approach. Compared to the other two attention approaches, the proposed approach had significant improvement of almost 20% regarding $P_{d,\mathrm{p}}$ and approximately 2% regarding P_s.

For the approach using the set of event conditions [38], although it achieved a very small rate of P_s (≈1%), an extremely low rate of $P_{d,\mathrm{m}}$ (≈47%) and a rather high rate of $P_{f,\mathrm{m}}$ (≈16%) resulted. $P_{d,\mathrm{p}}$ could not be calculated in this case because the designed event conditions were too simplified and did not cover the issue of people attention prioritization.

Considering the performance of the weighted sum approach [42], for all five weight sets, there were similar performances despite the differences in weight sets (≈73% of $P_{d,\mathrm{m}}$, ≈54 of $P_{d,\mathrm{p}}$, 14% of $P_{f,\mathrm{m}}$, and 4% of P_s). This indicates that a fixed weight distribution of stimuli did not guarantee the optimum performance of the attention model. The approach with one fixed weight set might result in a high detection rate of the most attentive person, but may deliver a low detection rate of people attention prioritization, with high susceptibility to frequent undesired attention shifts, or vice versa. This implies that with the heuristic equation approach, the determination of suitable parameters that ensure the optimum trade-off between the hit rate and false alarm rate is critical.

Figure 14 depicts the most attentive person selection outcomes of Lang et al.'s approach (the set of event conditions), Bennewitz et al.'s approach (the weighted sum approach), and our proposed approach tested on one image sequence of the three-person situation, compared to human evaluation.

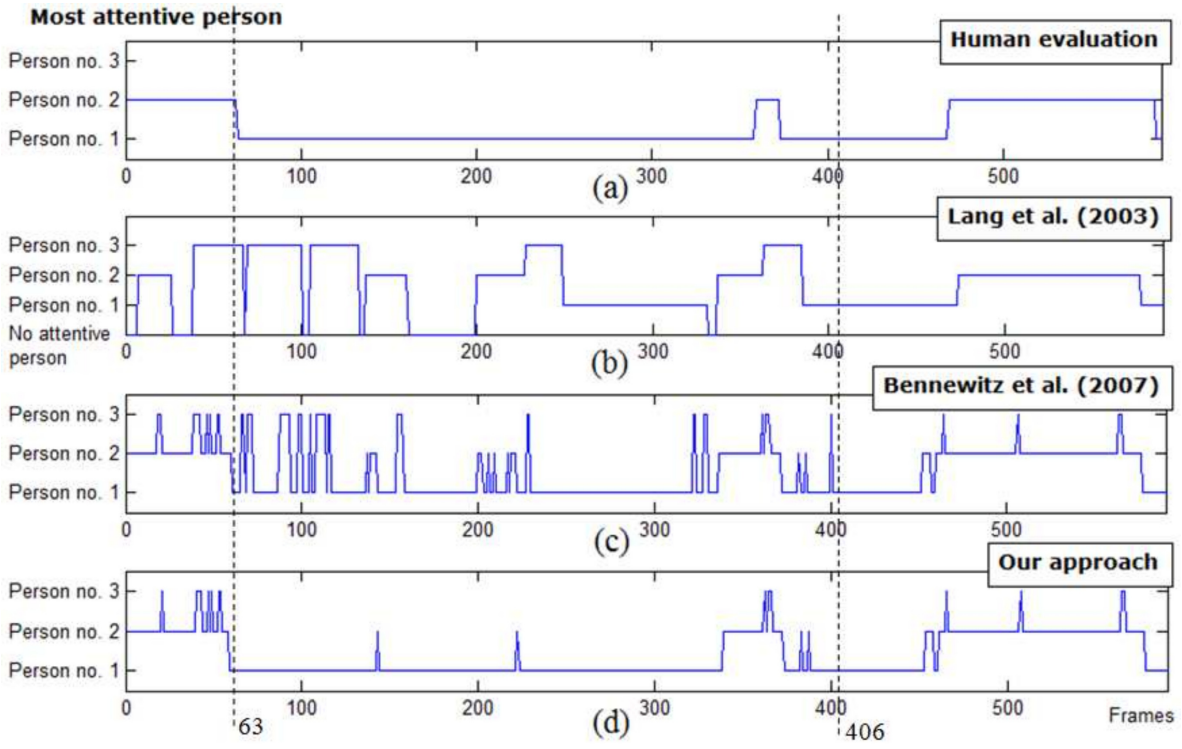

Figure 14. The most attentive person selection comparison between our attention approach and two other approaches, showing some false detection intervals of the most attentive person and improvements in terms of undesired attention shift moderation.

The first graph (Figure 14a) illustrates human evaluation of the most attentive person. The second graph (Figure 14b) depicts the result of Lang et. al.'s approach. Note that, compared to Figure 14c,d, there were several noticeable undetermined intervals of the most attentive person in Figure 14b (i.e., losing attention from the most attentive person). Specifically, these occurred during intervals in which no speaking person was detected. As a result, the attentive person was not unable to be determined by the proposed method.

The third graph (Figure 14c) depicts the result of Bennewitz et al.'s approach ($w = [w_d = 0.20, w_h = 0.10, w_s = 0.70]$). Next, the fourth graph (Figure 14d) shows the result of our proposed method. Considering frames 63–406 in Figure 14b–d, as expected from the probabilistic approach, several undesired attention shifts were moderated while correct detections of the most attentive person were maintained.

7. Conclusions

A novel vision-based attentiveness determination method has been presented to improve a robot attention model's performance in determining the most attentive person and prioritizing people based on attentiveness. Additionally, the effective computation of attentiveness and adaptation to changes in the number of participants and observations was accounted for in the proposed method. The proposed approach is based on relevance theory, a human communication methodology that explains how people evaluate the attention of other people during interactions.

The proposed approach consists of a computation method for probabilistic stimuli-relevance and Scalable HMM, an attentiveness determination model for most attentive person selection and people attention prioritization. Unlike the conventional HMM, the Scalable HMM has a scalable number of

states and observations, and online adaptability of the state transition probabilities with respect to changes in the number of states. Furthermore, for effective attentiveness determination, the speaking status of people was employed as conditional parameter for adaptable state transition probabilities, unlike most previous attention approaches. A better, more robust attentiveness determination was achieved, wherein the selection of the most attentive person could be conducted even in situations where no speaking person was detected. By employing online forward analysis, the probabilistic attentiveness of each person can be determined in real-time with low computation cost. A comparison of the computed attentiveness of people yielded the most attentive person selection and the people prioritization based on their attentiveness. The parameters of our proposed approach can be efficiently and conveniently learned based on PSO, such that good resistance to noisy observations and a good performance rate was achieved. The approach was successfully tested on 10 image sequences (7567 frames) with encouraging experimental results (≈76% accuracy in most attentive person detection and more than 75% accuracy for people attention prioritization). Additionally, the proposed method works robustly online in various lighting conditions and with changes in the number of participants. Compared to the two other more conventional attention approaches, improvements of nearly 20% in people attention prioritization and 2% in resisting undesired attention shifts were achieved. Overall, the most optimal performance was presented with the proposed method.

Despite being sufficiently robust in lighting variation, low-resolution images, and noisy observations, the proposed vision-based attentiveness determination for the most attentive person selection and people attention prioritization cannot operate well under extremely poor lighting conditions; in such conditions, image noise is high, resulting in extreme error in observations and a poor performance rate with the proposed model. Hence, in order to improve the model's efficiency, additional ostensive-stimuli could be included into the attentiveness determination model, such as hand gestures, facial expression, and natural language that can be trained using various deep learning architectures in [54,55]. Furthermore, to succeed in imitating a more human-like the attention system, a robot's head-eye control system and audio information could also be incorporated with visual information. In this manner, the robot could consider both vision and sound for its decision-making process, as humans do, thus improving its human likeness.

Author Contributions: Conceptualization, P.T. and M.-H.J.; Methodology, P.T., J.-J.L. and M.-H.J.; Software & Validation, P.T. and A.P.; Investigation, J.-S.Y.; Writing—Original Draft Preparation, P.T.; Writing—Review & Editing, J.-J.L., J.-S.Y. and M.-H.J.; Project Administration & Funding Acquisition, M.-H.J.

Funding: This research has been conducted by the Research Grant of Kwangwoon University in 2019, and supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2017R1D1A1B03035700).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Vision-Based Detection Methods for Distance of Person-to-Robot, Head-Orientation, and Speaking Statuses

In this section, vision-based detection approaches for three attention-related visual features (i.e., the distance of person-to-robot, head-orientation, and speaking statuses of a person) which are commonly observed and employed in the robot's attention model, are briefly described.

Appendix A.1. Distance of Person-to-Robot

Referring to psychological studies in personal space zones [56,57], humans regard their surrounding region (personal space) such that any invasion of personal space is always brought into attention. Table A1 summarizes the personal space zones between humans, so-called Hall's distances.

Table A1. Human–human personal space zones (Hall's distances).

Personal Space Zone	Range	Situation
Close Intimate Zone	0 m. to 0.15 m.	Lover of close friend touching
Intimate Zone	0.15 m. to 0.45 m.	Lover of close friend only
Personal Zone	0.45 m. to 1.2 m.	Conversation between friends
Social Zone	1.2 m. to 3.6 m.	Conversation to non-friends
Public Zone	over 3.6 m.	Public speech making

Walters et al. [58] presented a study in human–robot social space zones. The study shows that the human–robot interpersonal distances are comparable to those found in human–human interpersonal distances (see Table A1). Thus, the concept of Hall's distances can also apply to robots.

Here, to acknowledge people's whereabouts in a robot's view, we visually detect a person-to-robot distance using a BumbleBee2 stereo camera. By employing a stereo camera, a disparity image is computed and is readily accessible in real-time. To estimate a person-to-robot distance, first, a person's face should be located in the image. This can be done by a face detector algorithm. The face detector used in our system is described in several studies [59–61]. This face detector is one of the most powerful and widely used systems described in, and proposed by, the literature. Next, by averaging the intensity values of all existing pixels inside the located face window in the disparity image, the person-to-robot distance can be effectively estimated.

Appendix A.2. Head Orientation

In general, during an interaction between humans, the gaze direction is an important factor which notifies the object/subject of interest (the robot) of a gazer's interest (i.e., that of a human) at a particular time. Typically, the gaze can be determined from a combination of two basic actions: head-orientation and eyeball-motion.

Particularly, for accurate eyeball-motion detection and tracking, an adequate—or even high-resolution—face image sequence is necessary. However, as is known in MPRI, humans should be far enough away such that they can all fit in the robot's view. As a result, high-resolution face image sequences become a luxury difficult to obtain. Unavoidably, this raises difficulties in eyeball-motion detection and tracking. Hence, for simplicity in improving the accuracy with which a gaze is detected accuracy, we operate under the assumption that a human's looking-direction is the same direction as a human's facing-direction. In this way, the head orientation simply becomes a representation of the gaze direction.

Specifically, a structure of human's head has three degrees of freedom (3 DOF), with one rotational DOF around the neck (head pan) and two DOF at the neck (i.e., head tilt: looking up/down and tilting head left/right). Furthermore, to make the estimation less complex, this study considers only the head pan as a respective head movement for attracting a robot's attention. Therefore, in this case, the other DOFs at the neck (i.e., head tilt vertically and horizontally) are ignored and treated as noise motions.

To estimate the person's head pan, a coarse head-pose estimation algorithm is considered in this study. This approach can still efficiently estimate the person's head orientation in low-resolution face images. Brown and Tian explored the merits of two robust coarse head-pose estimation approaches [62].

To choose the applicable approach for our application, we tested the two approaches on the same dataset and found that the NN model approach was more robust and reliable than the probabilistic model approach. Hence, in this study, the NN-based coarse head pose estimation, similar to Zhao et al. [63], is employed to estimate a person's head pan angle.

Appendix A.3. Speaking Statuses of a Person

An act of speaking (i.e., calling a name, talking, shouting, etc.) is commonly considered an influential factor involved in attracting another person's attention. Two significant speaking statuses are concerned in this paper: speaking and non-speaking statuses.

Particularly, the speaking status detection approach should efficiently work online with low-resolution mouth images and in various lighting conditions because, in the case of MPRI, the participants should be located far enough away from the robot such that they can be in the robot's narrow view. Furthermore, the approach should be robust enough to withstand frequent undesired speaking state changes, which are often caused by noisy observation.

To overcome these requirements, a probabilistic approach described in Tiawongsombat et al. [25] is employed herein. The approach is a bi-level HMM, which analyzes lip activity energy on a mouth image to detect speaking statuses of a person in real-time. Specifically, the speaking status detection method is based on lip movement and speaking assumptions, thus embracing two essential consecutive procedures (post-processing and classification procedures) into a single model. Further, a bi-level HMM is an HMM with two state variables in different levels, where state occurrence in a lower level conditionally depends on the state in an upper level. Refer to [25] for more details.

References

1. Governatori, G.; Iannella, R. A modeling and reasoning framework for social networks polices. *Enterp. Inf. Syst.* **2011**, *5*, 144–167. [CrossRef]
2. Bruckner, D.; Zeilinger, H.; Dietrich, D. Cognitive automation—Survey of novel artificial general intelligence methods for the automation of human technical environment. *IEEE Trans. Ind. Inf.* **2012**, *8*, 206–215. [CrossRef]
3. Lam, C.; Ip, W.H. An improved spanning tree approach for the reliability analysis of supply chain collaborative network. *Enterp. Inf. Syst.* **2012**, *6*, 405–418. [CrossRef]
4. Yang, L.; Xu, L.; Shi, Z. An enhanced dynamic hash TRIE algorithm for lexicon search. *Enterp. Inf. Syst.* **2012**, *6*, 419–432. [CrossRef]
5. Wang, C. Editorial advances in information integration infrastructures supporting multidisciplinary design optimization. *Enterp. Inf. Syst.* **2012**, *6*, 265. [CrossRef]
6. Reeves, B.; Nass, C.I. *The Media Equation: how People Treat Computers, Television, and New Media Like Real People and Places*; American Psychological Association: Washington, DC, USA, 1996.
7. Kopp, L.; Gardenfors, P. Attention as minimal criterion of intentionality in robotics. *Lund Univ. Cogn. Stud.* **2001**, *89*, 1–10.
8. Fong, T.; Nourbakhsh, I.; Dautenhahn, K. A Survey of socially interactive robots. *Robot. Auton. Syst.* **2003**, *42*, 143–166. [CrossRef]
9. Raquel, V.A.; Rebeca, M.; Jose, M.P.-L.; Juan, P.B.; Adrian, R.G.; Pedro, R.L. Audio-Visual Perception System for a Humanoid Robotic Head. *Sensors* **2014**, *14*, 9522–9545.
10. Nuovo, A.D.; Conti, D.; Trubia, G.; Buono, S.; Nuovo, S.D. Deep Learning Systems for Estimating Visual Attention in Robot-Assisted Therapy of Children with Autism and Intellectual Disability. *Robotics* **2018**, *7*, 25. [CrossRef]
11. Li, K.; Wu, J.; Zhao, X.; Tan, M. Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-Mimicking Decision Mechanism. In Proceedings of the IEEE Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), Tianjin, China, 19–23 July 2018; pp. 498–503.
12. Alonso-Martín, F.; Gorostiza, J.F.; Malfaz, M.; Salichs, M.A. User Localization During Human-Robot Interaction. *Sensors* **2012**, *12*, 9913–9935. [CrossRef]
13. Pathi, S.K.; Kiselev, A.; Kristoffersson, A.; Repsilber, D.; Loutfi, A. A Novel Method for Estimating Distances from a Robot to Humans Using Egocentric RGB Camera. *Sensors* **2019**, *19*, 3142. [CrossRef]
14. Stiefelhagen, R.; Yang, J.; Waibel, A. Tracking focus of attention for human-robot communication. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots, Tokyo, Japan, 22–24 November 2001.

15. Michalowski, M.P.; Sabanovic, S.; Simmons, R. A spatial model of engagement for a social robot. In Proceedings of the 9th International Workshop on Advanced Motion Control, Istanbul, Turkey, 27–29 March 2006.
16. Massé, B.; Ba, S.; Horaud, R. Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2711–2724. [CrossRef]
17. Lemaignan, S.; Garcia, F.; Jacq, A.; Dllenbourg, P. From Real-time Attention Assessment to "With-me-ness" in Human-Robot Interaction. In Proceedings of the IEEE ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7–10 May 2016; pp. 157–164.
18. Sheikhia, S.; Odobeza, J.M. Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognit. Lett.* **2015**, *66*, 81–90. [CrossRef]
19. Das, D.; Rashed, M.G.; Kobayashi, Y.; Kuno, Y. Supporting Human–Robot Interaction Based on the Level of Visual Focus of Attention. *IEEE Trans. Hum. -Mach. Syst.* **2015**, *45*, 664–675. [CrossRef]
20. Yau, W.C.; Kumar, D.K.; Weghorn, H. Visual speech recognition using motion features and Hidden Markov Models. *Lect. Notes Comput. Sci.* **2007**, *4673*, 832–839.
21. Aubrey, A.; Rivet, B.; Hicks, Y.; Girin, L.; Chambers, J.; Jutten, C. Two novel visual activity detectors based on appearance models and retinal filtering. In Proceedings of the European Signal Processing Conference (EUSIPCO), Poznań, Poland, 3–7 September 2007; pp. 2409–2413.
22. Rivet, B.; Girin, L.; Jutten, C. Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Speech Commun.* **2007**, *49*, 667–677. [CrossRef]
23. Libal, V.; Connell, J.; Potamianos, C.; Marcheret, E. An embedded system of in-vehicle visual speech activity detection. In Proceedings of the IEEE International Workshop on Multimedia Signal Processing, Chania, Greece, 1–3 October 2007.
24. Siatras, S.; Nikolaidis, N.; Krinidis, M.; Pitas, I. Visual lip activity detection and speaker detection using mouth region intensities. *IEEE Trans. Circuits Syst. Video Technol.* **2009**, *19*, 133–137. [CrossRef]
25. Tiawongsombat, P.; Jeong, M.-H.; Yun, J.-S.; You, B.-J.; Oh, S.-R. Robust visual speakingness detection using Bi-level HMM. *Pattern Recognit.* **2012**, *45*, 783–793. [CrossRef]
26. Otsuka, K.; Takemae, Y.; Yamoto, J.; Murase, H. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In Proceedings of the 7th International Conference on Multimodal Interface (ICMI), Trento, Italy, 4–6 October 2005.
27. Schauerte, B.; Fink, G.A. Focusing computational visual attention in multi-modal human-robot interaction. In Proceedings of the International Conference on Multimodal Interfaces (ICMI), Beijing, China, 8–12 November 2010.
28. Ba, S.O.; Odobez, J.M. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 101–116. [CrossRef]
29. Sperber, D. Wilson, D. *Relevance: Communication and Cognition*, Harvard University Press: Cambridge, MA, USA, 1986; pp. 1–279.
30. Kelley, R.; Tavakkoli, A.; King, C.; Nicolescu, M.; Bebis, G. Undertanding human intentions via Hidden Markov Models in autonomous mobile robots. In Proceedings of the 3rd ACM/IEEE International Conference on Human robot interaction, Amsterdam, The Netherlands, 12–15 March 2008.
31. Kooijmans, T.; Kanda, T.; Bartneck, C.; Ishiguro, H.; Hagita, N. Accelerating robot development through integral analysis of human-robot interaction. *IEEE Trans. Robot.* **2007**, *23*, 1001–1012. [CrossRef]
32. Ito, A.; Terada, K. The importance of human stance in reading machine's mind (Intention). In Proceedings of the Conference on Human interface: Part I, Beijing, China, 22–27 July 2007; pp. 795–803.
33. Saadi, A.; Sahnoun, Z. Towards intentional agents to manipulate belief, desire, and commitment degrees. In Proceedings of the IEEE International Conference on Computer Systems and Application, Dubai/Sharjah, UAE, 8–11 March 2006.
34. Ono, T.; Imai, M. Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism. *Adv. Robot.* **2000**, *14*, 142–148. [CrossRef]
35. Tanenhaus, M.K.; Spivey-Knolton, M.; Eberhard, K.; Sedivy, J. Integration of visual and liguistic information in spoken language comprehension. *Science* **1995**, *268*, 1632–1634. [CrossRef]
36. Griffin, M.Z.; Bock, K. What the eyes say about speaking. *Psychol. Sci.* **2000**, *11*, 274–279. [CrossRef]
37. Okuno, H.G.; Nakadai, K.; Kitano, H. Social interaction of humanoid robot based on audio-visual tracking. In Proceedings of the International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE), Budapest, Hungary, 4–7 June 2001.

38. Lang, S.; Kleinehagenbrock, M.; Hohenner, S.; Fritsch, J.; Fink, G.A.; Sagerer, G. Providing the basic for human-robot-interaction: A multi-modal attention system for a mobile robot. In Proceedings of the International Conference on Multimodal Interfaces, Vancouver, BC, Canada, 5–7 November 2003.
39. Fritsch, J.; Kleinehagenbrock, M.; Plotz, T.; Fink, G.A.; Sagerer, G. Multi-model anchoring for human-robot interaction. *Rob. Autom. Syst.* **2003**, *43*, 133–147. [CrossRef]
40. Spexard, T.; Haasch, A.; Fritsch, J.; Sagerer, G. Human-like person tracking with an anthropomorphic robot. In Proceedings of the IEEE International Conference on Robotics & Automation, Orlando, FL, USA, 15–19 May 2006.
41. Tasaki, T.; Matsumoto, S.; Ohba, H.; Toba, M.; Komatani, K.; Ogata, T.; Okuno, H.G. Dynamic communication of humanoid robot with multiple people based on interaction distance. In Proceedings of the 2nd International Workshop on Man-Machine Symbolic System, Kurashiki, Okayama, Japan, 20–22 September 2004.
42. Bennewitz, M.; Faber, F.; Joho, D.; Behnke, S. Fritz—A humanoid communication robot. In Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication (ROMAN), Jeju, Korea, 26–29 August 2007.
43. Kohlmorgen, J.; Lemm, S. A dynamic hmm for on-line segmentation of sequential data. *Proc. NIPS* **2001**, *14*, 739–800.
44. Anderson, J.R. *Cognitive Psychology and Its Implication*, 6th ed.; Worth: New York, NY, USA, 2005; pp. 1–503.
45. Baum, L.E.; Sell, G.R. Growth functions for transformations on manifolds. *Pac. J. Math.* **1968**, *27*, 211–227. [CrossRef]
46. Xue, L.; Yin, J.; Ji, Z.; Jiang, L. A particle swarm optimization for hidden Markov model training. In Proceedings of the 8th International Conference on Signal Processing, Guilin, China, 16–20 November 2006.
47. Somnuk, P.A. Estimating HMM parameters using particle swarm optimization. *Lect. Notes Comput. Sci.* **2009**, *5484*, 625–634.
48. Kennedy, J.; Eberhart, R.C. Particle swarm optimization. In Proceedings of the IEEE International Conference Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; pp. 1942–1948.
49. Cha, Y.S.; Kim, K.G.; Lee, J.Y.; Lee, J.J.; Choi, M.J.; Jeong, M.H.; Kim, C.H.; You, B.J.; Oh, S.R. MARHU-M: A mobile humanoid robot platform based on a dual-network control system and coordinated task execution. *Robot. Auton. Syst.* **2011**, *59*, 354–366. [CrossRef]
50. Leone, F.C.; Nottingham, R.B.; Nelson, L.S. The folded normal distribution. *Technometrics* **1961**, *3*, 543–550. [CrossRef]
51. Jung, C.G. *The Development of Personality: Papers on Child Psychology, Education, and Related Subjects*; Princeton University Press: Princeton, NJ, USA, 1981; pp. 1–223.
52. Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*; OUP: Oxford, UK, 2003; pp. 1–318.
53. Obuchowski, N.A. Receiver operating characteristic curves and their use in radiology. *Radiology* **2003**, *229*, 3–8. [CrossRef]
54. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
55. Voulodimos, A.; Doulamis, N. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 1–13. [CrossRef]
56. Hall, E.T. *The Hidden Dimension: Man's Use of Space in Public and Private*; Bodley Head Ltd.: London, UK, 1966.
57. Hall, E.T.; Ray, L.B.; Bernhard, B.; Paul, B.; Diebold, A.R.; Marshall, D.; Edmonson, M.S.; Fischer, J.L.; Dell, H.; Kimball, S.T.; et al. Proxemics [and Comments and Replies]. *Curr. Anthropol.* **1968**, *9*, 83–108. [CrossRef]
58. Walters, M.L.; Dautenhahn, K.; Koay, K.L.; Kaouri, C.; Boekhorst, R.T.; Nehaniv, C.; Werry, I.; Lee, D. Close encounters: Spatial distances between people and a robot of mechanistic appearance. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots, Tsukuba, Japan, 5–7 December 2005; pp. 450–455.
59. Viola, P.; Jones, M. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
60. Lienhart, R.; Maydt, J. An extended set of Haar-like features for rapid object detection. In Proceedings of the IEEE ICIP, Rochester, NY, USA, 22–25 September 2002; pp. 900–903.
61. OpenCV on Sourceforge. Available online: http://sourceforge.net/projects/opencvlibrary (accessed on 28 June 2018).

62. Brown, L.M.; Tian, Y.L. Comparative study of coarse head pose estimation. *Workshop Motion Video Comput.* **2002**, *4*, 125–130.
63. Zhao, L.; Pingali, G. Carlbom, I. Real-time head orientation estimation using neural networks. In Proceedings of the International Conference on Image Processing, Rochester, NY, USA, 22–25 September 2002.