1 Supplementary: The components and effects for an CATR's framework

Table 1.1 The components and their attributes with descriptions and examples (bold* are the preferred types)

Component	Attribute	Description	Example
Input interface	Master-slave	The operator manipulates a primary system. The system will acquire the information and send them to a remote and identical system. It then executes the similar motion.	For MeBot [1], the operator had a pair of three DOF manipulators. He could express his desired gestures on his manipulators, and the remote system would imitate the motion.
	Natural interface*	The operator can express his nonverbal cues by acting them naturally. The system will acquire the information and send them to the remote system. The remote system then executes the motion given the information.	For Hasegawa [2], a KINECT sensor acquired the operator's poses information. The system sent that information the remote-system. The remote system would imitate the operator by moving its joints base on the pose information.
	Traditional input devices	Operator needs to use traditional input devices to operate the various nonverbal cues.	In a traditional telepresence robot, the operator needed to select the intended gestures from the keyboard. For instance, each key or a combination of keys could represent a distinct gesture.
Representation	**Auto*** or Manual	Auto represents a set of features that are extracted by an unsupervised algorithm.	For Lau's model [3], a probabilistic model, dynamic Bayesian network, was trained to model a segment of an input sequence. The parameters or probability distributions were the features for this model.
		Manual represents a set of features that are selected by a human.	In Park's work [4], the study used 13 distinct gestures, and they included lift-right, lift-left, and lift both.
	Distributed* or local	Distributed is a set of features that represent the underlying structure of the information.	For Hartmann's model [5], Hartmann identified six distributed parameters for each prototypical gesture. The parameters included spatial extent, temporal extent, fluidity, and more. The six parameters redefined the prototypical gestures, and they are not mutually exclusive to one another.
		Local are a set of features that are mutually exclusive to one another.	For Park's work [4], the 13 distinct gestures were mutually exclusive from one another. Lift-right and lift-left could not occur at the same time. If the operator lifted both his arms, then lift-both would be selected.

Continue from previous table

Component	Attribute	Description	Example
Representation (continue)	Spatial or **spatiotemporal***	Spatial features do not capture any temporal information	For MeBot [1], the transmitted data is the joints information, which does not contain any temporal information.
		Spatiotemporal features encompass the temporal information.	For Park's model [4], a gesture recognition module would classify the operator's motion to a predefined action class, e.g. lift-right.
Encoding	**Auto***	An automated system encodes the intended information before the information is transmitted.	For Lau's model [3], the probabilistic model automatically generated a distributed output given a sequential input.
	Direct	The system directly transmitted intended information to the remote system without any transformation.	For MeBot [1] and Hasegawa [2], both the systems directly transmitted the joints information to the remote system without transformation.
	Manual	The operator must manually encode the intended information.	For Hartmann's model [5], the operator must select the value for the six distributed parameters and a prototypical gesture for every intended gesture.
Decoding	**Auto***	A remote automated system decodes the encoded information before the remote system can execute the command.	For Park's model [4], a motion generator decodes the encoded data. In this case, the motion generator had a set of predefine motions, and it would execute one of the motions depend on the encoded data.
	Direct	The remote system can directly execute the received information without any transformation.	For MeBot [1] and Hasegawa [2], both remote systems executed their motion based on the received joints information.
Associating	**Auto***	For multimodal situation, the system automatically maps the relationship between different modalities.	Not available in the existing interfaces
	Manual	For multimodal situation, the operator will map or indicate the relationships between different modalities.	For Neff's model [6], a human coder annotated and aligned the gestures behavior and speech structure to create the rules for the system database.
	None	The system deals with each modality independently.	For instance, MeBot [1] has multimodal like neck, arms, video, and audio. However, each modality worked independently from one another.

Table 1.2 The effects and its attributes with descriptions and examples (bold* are the preferred types)

Effect	Attribute	Description	Example
Expressivity	**High***	A high degree of diversity within a modality.	For Hartmann's model [5], the operator had six distributed parameters to create a personal gesture for each prototypical gesture.
	Low	A low degree of variation within a modality.	For Park's model [4] , it had 13 gestures only. It was not appropriate to be an operator interface for telepresence robot because the remote system should express personal gestures and not generic gestures.
Cognitive load	High	The system requires a high amount of cognitive load to execute the nonverbal cues.	For instance, Hartmann's model [5] required the operator to select a prototypical gesture and fill in the intensity, [-1, 1], for the six distributed parameters during the encoding process.
	Low*	The system requires a low amount of cognitive load to execute the nonverbal cues.	In Lau's model [3], the operator just needed to express his intended gestures, and the system would automatically handle the rest of the processes.
	Low (require remapping)	The system requires a low amount of cognitive load, but the system still consume some degree of operator's cognitive load for remapping.	For MeBot [1], the operator must remap his intended gestures onto a primary system before he could execute his action. This process might filter the unconscious gestures.
Decoupling	No	The system will continue to send the nonverbal information even when the nonverbal information might be undesirable	For Hasegawa's bot [2], the remote system would execute any gestures the operator's gestures. Undesirable gestures, e.g. typing on the keyboard, might not be congruent to the going conversation.
	Yes (idle)	The system can stop any undesirable nonverbal information from transmitting, but the system will be in an idling state	For MeBot [1], the operator could stop his gesturing by not moving the primary system. At this point, the operator could continue with his remote-task, but the remote system would be in an idling state.
	Yes* (associate)	The system can stop transmitting any undesirable nonverbal signal. It also can conceal the missing signal with a coherence signal.	For Neff [6], the system could automatically generate the intended gestures that were congruent with the structure of the speech.
Obstacle avoidance	**Deliberative***	A deliberative planner generates motion and keeps its overall intentions. It tends to be slower when compares with a reactive planner.	For Lau's model [3], the output could be a variance of the input, which preserved the overall shape of the input.
	Reactive*	A reactive planner is a reflexive behavior system that responds to the real-time external stimulus.	The MeBot [1] had range sensors for obstacle avoidance. It directly transmitted the joints information, so it could only spontaneously avoid the obstacle when the range sensors detected the obstacle.

Reference

[1] S. O. Adalgeirsson and C. Breazeal, “MeBot : A robotic platform for socially embodied telepresence,” in *ACM/IEEE International Conference on Human-Robot Interaction*, 2010, pp. 15–22.

[2] K. Hasegawa and Y. Nakauchi, “Preliminary Evaluation of a Telepresence Robot Conveying Pre-motions for Avoiding Speech Collisions,” in *HAI-Conference.net*, 2013, pp. 5–8.

[3] M. Lau, Z. Bar-Joseph, and J. Kuffner, “Modeling spatial and temporal variation in motion data,” *ACM Transactions on Graphics*, vol. 28, no. 5, p. 171, Dec. 2009.

[4] H. Park, E. Kim, S. Jang, and S. Park, “HMM-based gesture recognition for robot control,” in *Iberian Conference on Pattern Recognition and Image Analysis*, 2005, pp. 607–614.

[5] B. Hartmann, M. Mancini, and C. Pelachaud, “Implementing expressive gesture synthesis for embodied conversational agents,” in *Gesture in human-Computer Interaction and Simulation*, 2006, pp. 188–199.

[6] M. Neff, M. Kipp, I. Albrecht, and H. Seidel, “Gesture modeling and animation based on a probabilistic re-creation of speaker style,” *ACM Transactions on Graphics*, vol. 27, no. 1, Mar. 2008.