What’s “up”? - Resolving interaction ambiguity through non-visual cues for a robotic dressing assistant

Greg Chance, Praminda Caleb-Solly, Aleksandar Jevtić, Sanja Dogramadzi

Abstract— **Robots that can assist in activities of daily living (ADL) such as dressing assistance, need to be capable of intuitive and safe interaction. Vision systems are often used to provide information on the position and movement of the robot and user. However, in a dressing context, technical complexity, occlusion and concerns over user privacy pushes research to investigate other approaches for human-robot interaction (HRI). We analysed verbal, proprioceptive and force feedback from 18 participants during a human-human dressing experiment where users received dressing assistance from a researcher mimicking robot behaviour. This paper investigates the occurrence of deictic speech in an assisted-dressing task and how any ambiguity could be resolved to ensure safe and reliable HRI. We focus on one of the most frequently occurring deictic words “up”, which was captured over 300 times during the experiments and is used as an example of an ambiguous command. We attempt to resolve the ambiguity of these commands through predictive models. These models were used to predict end effector choice and the direction in which the garment should move. The model for predicting end effector choice resulted in 70.4% accuracy based on the user’s head orientation. For predicting garment direction, the model used the angle of the user’s arm and resulted in 87.8% accuracy. We also found that additional categories such as the starting position of the user’s arms and end-effector height may improve the accuracy of a predictive model. We present suggestions on how these inputs may be attained through non-visual means, for example through haptic perception of end-effector position, proximity sensors and acoustic source localisation.**

I. INTRODUCTION

It is predicted that by 2050 many countries could have around 30 or 40% of their population aged over 65 years resulting in a huge social and economic impact, especially with regard to healthcare provision [1], [2]. This is likely to result in shortage of care workers [3], posing significant challenges to healthcare provision and additional pressure on informal carers [4]. To address these needs, promotion of wellbeing and independence for older people [5] is required, such as the long term care revolution [6]. Independence for older adults requires assistance with the ADL of which assistance with dressing is in the top 4 but has the lowest technological maturity and commercial development [7], possibly due to the inherent difficulties of close human-robot interaction in this area. The majority of older adults would prefer to stay at home for as long as possible and avoid the traumatic and costly procedure of moving to a care home [8]. General acceptance of this type of assistive technology must be specific to the user’s need and expectations for a specific situation [9]. The expected interaction may use many modalities, of which speech may be preferred. However, one of the problematic issues is the semantic ambiguity of the vocabulary used, especially deictic words whose meaning is dependent on the context in which they are used. This paper investigates the occurrence of deictic speech in an assisted dressing task, the level of variability, and how any ambiguity could be resolved to ensure safe and reliable We construct decision tree models to evaluate accuracy of correct interpretation of ‘up’ when other, non-visual cues are used as inputs.

II. BACKGROUND

Previous studies on robotic dressing have mainly focused on vision systems as the primary modality of interaction. A 2-arm robot has been used to dress a t-shirt onto a mannequin using reflection markers and MAC3D motion capture system [10]. Similarly, stereo cameras used in combination with garment markers have been used for dressing a mannequin with a t-shirt [11] and the group have also achieved similar results using depth (RGB-D) cameras [12]. Using a Baxter robot to put a hat on whilst compensating for variations in the user’s position was achieved using a Kinect sensor and open source code [13]. The Asus Xtion depth sensor has also been used to estimate garment [14] and user pose [15] for dressing assistance by humanoid robots.

Reliance on a single modality is clearly possible but the use of multiple modalities is more conductive to natural human-robot interaction. Initial work along these lines has made some progress in the area of robotic assisted dressing. For a trouser dressing task, one group integrated both vision and force feedback with a focus on dressing errors [14] where the force feedback was used to trigger failure identification. Gao et al. also used RGB-D data to estimate user pose and initial robot trajectory for dressing a sleeveless jacket using the Baxter robot [16]. The initial robot trajectory was optimized based on feedback from the limb endpoint force estimation.

Predictive models for dressing a hospital gown have been proposed by [20], based on just the force feedback without the use of vision in a simple single arm experiment. Outside of dressing assistance the use of multimodal robot interaction has been well researched to include speech [17], gesture [18] and gaze [19] and their combination [20]. Other more subtle cues have been used to understand the user such as emotional

This research was supported by EPSRC and EU CHIST-ERA. This work was partially funded by the Spanish Ministry of Economy, Industry and Competitiveness, under the programme “Acciones de Programación Conjunta Internacional 2015”, project reference number: PCIN-2015-147

G. Chance, P. Caleb-Solly and S. Dogramadzi are with the Bristol Robotics Laboratory, University of the West of England, BS16 1QY, UK (e-mail: greg.chance@brl.ac.uk, praminda.caleb-solly@brl.ac.uk and sanja.dogramadzi@brl.ac.uk). A. Jevtić is with the Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Parc Tecnològic de Barcelona, C/ Llorens i Artigas 4-6, 08028 Barcelona, Spain (e-mail: ajevtic@iri.upc.edu).

state through facial expressions [21] or through physiological inputs such as heart rate and perspiration [22]. User attention [23] and user intention [24] may also be rich sources of information about the interaction scenario whilst also being some of the more difficult to capture.

In our previous work, we focused on trajectory planning for error handling using IMU sensors [25] and differentiating garments using force sensors [26] using the Baxter robot. In this work successful detection of dressing errors was obtained using gyroscope data and Support Vector Machines were found to give the highest garment prediction accuracy.

In many areas of robotics research the need for visual detection and tracking of the real world is needed, for example in the fields of robotic surgery and industrial processing. In our current work we look at the possibility to remove the vision modality in human-robot interaction due to concerns of user privacy. In addition, vision systems can be either very large installations suitable for laboratory settings only (e.g. Vicon) or single point devices which may suffer from occlusions. When using an RGB-D sensor mounted on the robot, occlusion may occur due to either the garment or the robot arm blocking the view of the user. Moreover, the processing of such data can be computationally expensive resulting in a loss of natural interaction and also be sensitive to lighting conditions. However, this may be alleviated to some extent with the advancement of computing technology and availability of greater computational power. Although vision has been removed from this work to focus on the robot proprioception, reintegrating vision into a multimodal system is planned for the final prototype. The work presented in this paper forms part of a larger research and development effort under the I-DRESS project[1], with the ultimate aim to provide dressing assistance through multimodal interaction.

A. Scope of Work

We have undertaken a human-human interaction (HHI) study at the Bristol Robotics Laboratory to investigate the modalities that are used in an assisted dressing task. We examine video footage from 18 participants (over 200 interactions) to identify interactions in the dressing scenario and cues that people use to indicate their intention. Our aim is to show that robot-assisted dressing could be achieved without the need for a visual input.

Section III explains the details of the HHI experiments, definition of the terms used, captured interaction data and their interpretation. We highlight potential ambiguity when using the utterance 'up' as an example of a deictic command due to its varied interpretation based on the context of the dressing situation. In section IV we demonstrate the results of using decision trees to analyze and interpret the use of 'up' in different dressing sequences. The results have shown high percentage of accuracy in predicting correct robot actions based on the non-visual inputs.

III. Human-Human Interaction Study

Staff and students from the University of the West of England and Bristol Robotics Laboratory were invited to participate in the HHI study where they would receive assistance to put on a jacket. The participants were introduced to the research and submitted their consent to being video and audio recorded for the purposes of the experiment. The testing took place in a quiet location screened-off from external observers and interruptions, see Figure 1. The webcams had integrated microphones and one of these was used to monitor the utterances of the participant. Other data about the interaction was captured but not used in this analysis, e.g. user pose from the Xsens suit shown in Figure 1, force and torque data from a hand-held device.

During the experiment the researcher "acting" as the robot had their eyes closed. This was done to promote verbal responses from the user and to simulate no visual input to the 'robot'. The participant was told to use speech to give commands to the researcher acting as "the robot" in order to get the jacket on. In the first part of the experiment the participant was allowed full mobility, able to move about as much as they found helpful to get the jacket on. The jacket was put on three times with the participant seated on a stool and three times whilst standing. The experiment was then repeated, but with an imagined restricted mobility issue: the participant was told not to bend either elbow. Again, the dressing task was repeated three times, for both seated and standing conditions.

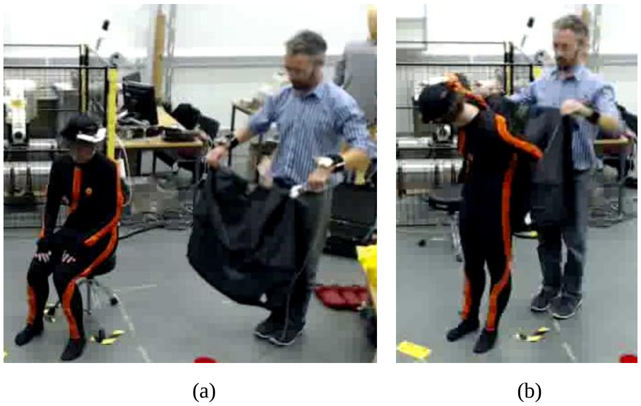

(a) (b)

Figure 1. Participant receiving assistance in putting on a jacket while sitting (a) and standing (b).

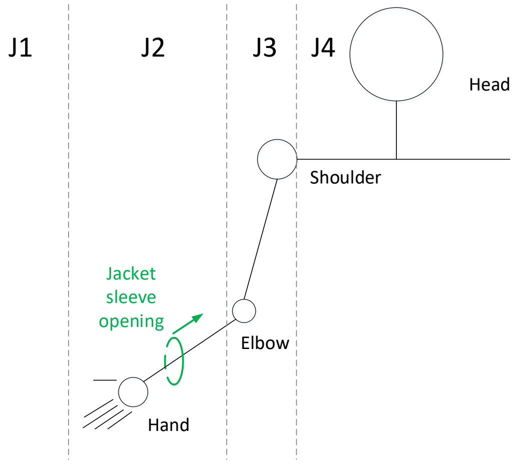


Figure 2. Jacket dressing task is segmented by the position of the jacket sleeve opening relative to the user which is shown by the green ring travelling up the arm. J1= approach to hand, J2=hand to elbow, J3=elbow to shoulder, J4=finish. In this example the Jacket is said to be in segment J2.

[1] https://www.i-dress-project.eu/

A. *Definitions*

Within this paper the following definitions are used to refer to people or objects. The *participant* is the experimental subject and this is used interchangeably with *user* to refer more specifically to the *end-user* of the system. No robot was used in this testing but a researcher was posing as a robot for the dressing testing, is referred to here as the *robot*. The terms *word*, *utterance* or *phrase* are all associated with the spoken words from the user whilst receiving dressing assistance from the *robot*. The term *end-effector* is used here to mean the hand of the researcher. As the researcher is trying to mimic a robot, this terminology is suitable. The term *dressing segment* refers to the progress of one arm of the jacket up the arm of the user (J1, J2, J3) which are defined in Figure 2. The participant in Figure 1 is in J1 for (a) and J2 for (b). The *arm angle* is defined as the angle that a straight arm makes with the normal to the floor. The term *sitting* is referring to the participant when they were seated on a stool during the dressing test. The term *mobility constraint* indicates when the participant was not allowed to bend either arm at the elbow. The term *head orientation* refers to the general direction that the participant was looking in, for example this would be *left* in Figure 1 (a) and *right* in (b). The head orientation *straight ahead* was also captured. The *dressing strategy* is the sequence of actions taken by the user to get the jacket on: strategy a) one arm at a time or strategy b) both arms into the jacket at the same time.

B. *User Utterances*

Several user utterances were observed that may be ambiguous in isolation or without contextual information. These are known as deictic phrases and are researched within the field of speech recognition [27]. The words "left" and "right" have an explicit meaning but depend on the frame of reference in which they are used. Interestingly, in most cases the participant would phrase the command to suit the 'robot'. The phrase "pull" usually infers an action of drawing an object towards oneself, but in these tests the phrase has a myriad of interpretations relating to movement in all directions, some of which were directly away from the robot. Also phrases such as "move down a little" or "left a bit" require some tacit knowledge of the situation in order to interpret exactly the distance implied by 'a little' and 'a bit'. The phrase "ok" was also used to mean "ok, let's start" and equally "ok stop, I'm finished" and when used in isolation, which was observed in many instances, would need contextual information for a correct interpretation.

For this analysis we have chosen the word "up" due to its ambiguity and proliferation throughout the experiment. The word "up" was the 2nd most used word throughout the test (7.4% of all words) after "stop" (9.7%) and for comparison "ok" was used 2.7% of the time. In TABLE I we define different ways that the phrase "up" could be interpreted by the robot. These definitions can be partitioned along the lines of i) direction and ii) end effector selection. The identifier is prefixed with the end effector selection (L=left, R=right, B=both) and is suffixed with the direction the end effector should move: up the arm[2] (A) or, up the z-axis (Z)[3].

C. *Data Capture*

Video footage of the experiments was reviewed for every instance of the utterance "up" using the Nvivo platform. At each instance of the utterance some basic information was captured pertaining to the user at that time or within a 1s window, see TABLE II. X_1 defines the arm angle of the user which is estimated at the time of the utterance and is categorized into 4 groups. X_2 is the dressing segment as defined in Figure 2 categorized into 3 groups. X_3 is true if the participant was sitting at the time of the utterance and false for standing. X_4 is true if the participant was imitating the mobility constraint (locked elbows) and false for normal mobility. The general head orientation of the participant, X_5, is put into three categories: looking left, right and straight ahead (forward). This table also shows the classification definition for moving the end effector either up the z-axis, outcome Y_0, or moving up the arm, outcome Y_1. Additionally, when the arm is perpendicular to the floor (angle=0°) then both outcomes are true. In this case we set Y_0=1 and Y_1=0 to maintain a binary outcome.

TABLE I INTERPRETATIONS OF "UP"

Identifier	
LZ	Move the left end effector up in the z-axis.
RZ	Move the right end effector up in the z-axis.
BZ	Move both end effectors up in the z-axis.
LA	Move the left end effector up the arm.
RA	Move the right end effector up the arm.
BA	Move both end effectors up the arm.
Z	Up z-axis / towards the ceiling
A	Up the arm / towards the shoulder

TABLE II ATTRIBUTES & CLASS DEFINITIONS

	Attribute	Values
X_1	Arm angle	0, 20, 45, 90 (degrees)
X_2	Dressing segment	J1, J2, J3
X_3	Sitting	True/false (binary)
X_4	Mobility constraint	True/false (binary)
X_5	Head orientation	Left=1, right=2 or forward=0
	Classifier	**Values**
Y	Move end effector	Y_0 = Up the z-axis / Y_1 = up the arm
Z	End effector choice	Z_0 = Left / Z_1 = Right / Z_2 = Both

D. *Capturing Interaction Attributes from Video*

The research pertaining to feature extraction and behavior recognition in the field of robotics is expansive. Automatic proxemic feature extraction has been implemented in a social robotic interaction study [28]. By exploiting proxemic cues combined with motion data, some authors have gained information about interactions including the intentionality of a specific interaction [29]. Efficient robot-human interaction has been achieved through implementation of RGB-D cameras and a feature detector by applying separated filters to the 3D spatial dimensions [30]. Lourens et al. propose the analysis of interaction should include information beyond

[2] This can also be thought of as moving towards the shoulder if the arm is straight. If the arm is not straight the movement will comprise two vectors, the second of which will terminate at the shoulder.

[3] The XY plane is taken as the plane of the floor and the positive z direction is taken as towards the ceiling.

conscious expression which might include emotions, intentions and mental states [31].

In this study we use video footage and manually review the interactions to extract the features required, automatic detection was not required. Many of the required features are simple to interpret from the video footage, such as sitting, dressing segment and head orientation. The mobility constraint tests were separated from the other parts of the test so easily determined and codified within the video. The angle of the arm is estimated based on the arm position relative to the shoulder of the user, see Figure 3.

E. Utterance Interpretation

This section deals with the important issue of how the outcome classifier was determined, i.e. which of the commands in TABLE I should be used when the user said "up". This is split into two parts, the first is how the command was interpreted based on human perceptions during the experiment. The second part is how a robot could arrive at the same interpretation based on the same inputs.

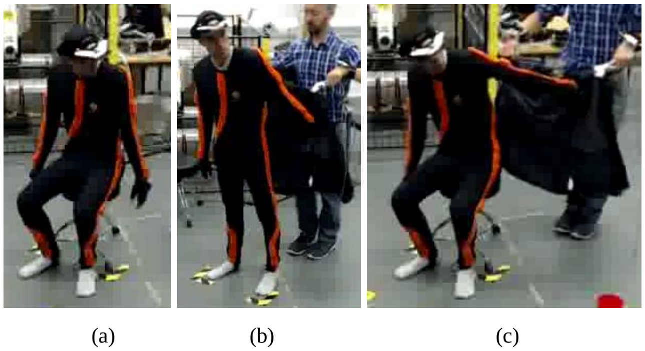


Figure 3. Arm angle of participant at approximately a) 0, b) 45, and c) 90 degrees.

1) Human Interpretation

During the HHI experiment the researcher had to interpret the spoken command from the participant and move the jacket with their eyes closed. We focus here only on the word "up" and what cues were available to interpret this word into an action. This interpretation was based on the information available at the time of testing. As the researcher had their eyes closed they had no visual cues about the situation and therefore used other non-visual information available, see TABLE III.

Speech: The user may give additional utterances which gives more context to the command or state explicitly the intention, for example "*move your right hand up to my right shoulder*".

Proprioception & Prior Knowledge: The researcher will have knowledge about dressing and what the participant is likely to say and which movements are unlikely to help in completing the task. The researcher also had their eyes closed for the test but prior to this they will have subconsciously noted the height of the user relative to themselves. This gives the researcher information about the approximate height of the shoulder relative to themselves which would infer where the trajectory path should terminate, i.e. the jacket should not be moved above the shoulder.

Force: The researcher also experienced a resistive force when moving the garment depending on the position and movement of the user giving information about the position of the arm. For example, if the jacket is half way up the arm but cannot be moved 'up' the z-axis then the command was interpreted as going 'up' the arm.

End Effector Position: In combination with proprioception and voice intensity, the position of the end effectors with reference to the perceived position of the user gives an indication of the dressing segment and the arm position and angle. In addition, if one end effector was higher than the other, i.e. one at the shoulder and the other near the waist, then interpretation of "up" would more likely be related to the end effector that is lower down.

Voice: The strength or intensity of the user's voice was used to gauge distance to the user. Also the direction of the user's head while talking can give an indication of the direction they are looking in (head orientation).

TABLE III INTERPRETATION CUES

	Interpretation Cues and Information
Speech	The user gives other explicit commands that clearly identify the outcome.
Prior knowledge	The researcher already knows how to put on the garment and which movements are likely to be asked for.
Proprioception	Approximate height of user and the general position of the user are known and subconsciously remembered prior to closing eyes.
Touch / Force	Experiencing resistance through the hands when moving the garment gives an indication of task segment, i.e. more resistance as jacket approaches the shoulder.
End effector / hands position	If the position of the hands relative to the user are close to the height of the shoulder then the task is nearly completed. This also gave an indication of the user arm angle knowing the user was in a fixed position. When the hands are at different heights (z-axis) this may indicate moving one end effector over the other.
Direction & loudness of voice	The head orientation of the user could be recognised by the direction the voice was coming from. Also the voice intensity could be used to judge distance to the user.
Dressing strategy	Force feedback and prior movement actions gives information about the dressing strategy (see below).

Dressing Strategy: The types of dressing strategy could be but are not limited to: a) both arms into the jacket at the same time and b) one arm at a time. The strategy can be inferred from the first few commands given by the user, usually indicated by the jacket being brought around to one side of the user (strategy b) or directly behind and at waist height (strategy a). Knowing the strategy gives information about the likely subsequent commands.

These cues gave the researcher a good indication of how to interpret the commands from the participant and dress the user without use of vision. The decision processes that were used to interpret the utterance "up" can be described by a series of flow charts, see TABLE IV.

2) Robot Interpretation

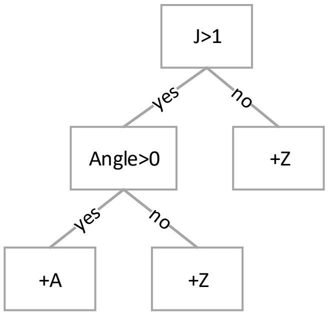
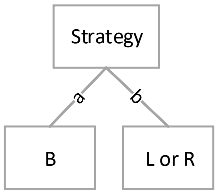
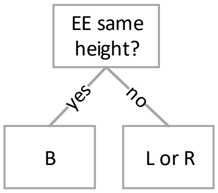
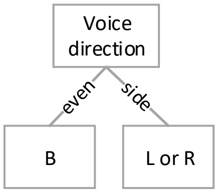
We have shown how a human can use non-visual cues during a dressing interaction to determine the user's intentions if the situation is ambiguous. Now we propose how these attributes could be attained in the robotic system without vision. We recognise from the human interpretation that there are 7 cues used to determine the 5 attributes. We propose how the robot may learn about each attribute using these cues or an alternative where it is not available.

Dressing Segment: It may be possible to determine when the dressing task is in segment J2 by sensing when the user is within reach of the robot, possibly using proximity sensors, and combining this with force feedback from load cells mounted at the end effectors. This could be used to determine when the hand (or hands) are first detected entering the garment.

Sitting: This information could be pre-defined to the robot prior to dressing along with the user's height and other fixed variables such as arm span. Furthermore, if the user has a preference on dressing strategy this can be added to a personal profile of the user. This data will give the robot a starting point for the trajectory.

Mobility Constraint: Mobility constraint may dictate the dressing strategy, i.e. if the user has difficulty moving their arms they may always adopt a dressing strategy where they put the garment onto both arms at the same time. This information could also be pre-defined to the robot.

TABLE IV HUMAN DECISION FLOW CHARTS

	Is the hand in the jacket already? If not then move +Z. If hand is already in jacket, is the angle greater than 0? If yes then go up arm, if no then +Z.
	What strategy is the user adopting for putting the jacket on? These are: a) both hands in at once b) one arm at a time. For a) move both end effectors at the same time, for b) use independently.
	Is one end effector higher in the z-axis than the other? If no, then use independent control, if yes move both end effectors simultaneously.
	How to choose which end effector to move? If the command is not explicitly stated then both is usually assumed especially in J1. However for segment >J1 the user may wish to manipulate a single EE and this also depends on dressing strategy.

Head orientation: The user head orientation could be determined using acoustic source localisation [32], [33] through implementation of multiple microphones (assuming that the user makes a vocal request).

Arm Angle: Detecting the angle of the arm may be possible based on the position of the end effectors, the height of the user, knowing if they are sitting and their dressing strategy. Feedback from a force sensor would also be necessary to make real-time adjustments based on the user movement. This combination of inputs maybe akin to the proprioception that the researcher used when dressing the participant with their eyes closed. This leads a potential solution towards a form of haptic perception from the robot's end effector as suggested by Kapusta et al. [34].

A. *Data Capture Cost*

It is also interesting at this point to discuss the cost of obtaining each of these attributes based on the above proposal. This may be defined as hardware or software costs and complexity of implementation. This can be used to rank the attributes on the approximate cost of obtaining them. A lower cost is associated with data that is simple to obtain, see TABLE V. Sitting and mobility constraints have the lowest costs as these can be simply obtained from the user prior to dressing. Implementing acoustic source localisation to determine the head orientation may have moderate hardware and implementation costs. Dressing segment would use proximity sensors and load cell technology with medium to high implementation complexity. Determining arm angle would have to be the largest of these costs as this relies on all the other attributes. The rank of cost low-to-high for each attribute is: 3, 4, 5, 2, 1.

TABLE V ESTIMATED DATA CAPTURE COSTS

Attribute		***Hard/Software***	***Implementation***
X_1	Arm angle	Med-High	Med - High
X_2	Dressing segment	Med	Med - High
X_3	Sitting – ask user	n/a	Low
X_4	Mobility constraint – ask user	n/a	Low
X_5	Head orientation – acoustic source	Low – Med	Med

IV. DATA ANALYSIS & DISCUSSION

The number of times the utterance "up" was observed during the jacket dressing task is shown in TABLE VI. This has been divided into the dressing segments showing that J2 (jacket between hand and elbow) and J3 (jacket between elbow and shoulder) are where the utterance is observed the most. In total 325 observations were made.

TABLE VI OBSERVATIONS OF UTTERANCE "UP"

Dressing Segment	Y_0	Y_1	**Sub-Total**
J1	20 (6%)	12 (4%)	32 (10%)
J2	53 (16%)	84 (26%)	137 (42%)
J3	22 (7%)	134 (41%)	156 (48%)
		***n* =**	**325**

A. *Regression Analysis*

Using a univariate logit regression, the outcome of *Y* (up in z direction or up the arm) is analysed separately against each of the attribute values *X*, see TABLE VII. Attributes with a higher positive coefficient give a higher probability of being related to the outcome Y_1 = up the arm and negative coefficient to Y_0 = up in z-axis. The 95% confidence interval of the coefficient is also shown in parentheses along with the p-value showing the result of the null hypothesis test. The r-squared value represents a measure of the variance in *Y* that is explained by the variable *X*.

Arm angle was treated as a continuous variable and showed statistical significance (P<0.001) and good model fit (R^2=0.24). The coefficient predicts that for every increase in arm angle by 1 degree there is an increased probability of outcome Y_1 of 0.078. Dressing segment J2 and J3 were analysed with respect to reference category J1. Of these only J3 was significant with p<0.001 indicating that Y_1 (up the arm) was 2.7x more likely in J3 compared to J1. This also has a reasonable R^2 value meaning it has a moderate fit to the regression model. Mobility constraint is also statistically significance and indicates that when the user imitates the mobility constraint they were 3.1x more likely to want the jacket to go up the arm. Information about head orientation and sitting are not statistically significant.

TABLE VII OUTCOME Y REGRESSION RESULTS

Attribute	***Coefficient / 95%CI range***	***p-value***	***R^2***
Angle (cont.)	**0.078 (0.057,0.10)**	**<0.001**	0.24
J_2 (ref. to J1)	0.617 (-0.159,1.392)	0.119	0.16
J_3 (ref. to J1)	**2.704 (1.778,3.630)**	**<0.001**	
Sitting	-0.229 (-0.751, 0.293)	0.390	0.0020
MobCon	**3.139 (1.959, 4.319)**	**<0.001**	0.17
Look left (ref. fwd)	-0.091 (-0.778, 0.596)	0.794	0.0003
Look right	-0.008 (-0.700, 0.684)	0.981	

B. *Decision Trees*

A categorical decision tree was trained using Matlab (2016b), using the *Y* outcome variable and the attributes X_1-X_5. Categorical trees were chosen over regression trees as the variables are non-continuous. Matlab uses the CART method for splitting prediction of the data, [35]. We quote two metrics for an indication of the classification error based on re-substitution and validation. The weighted average classification loss, *L*, is shown in equation (1), where *n* is the sample size, Y_j is the observed class, w_j is the weight of observation *j*, and I{*Y*} is the indicator function. The generalisation error based on validation testing, *V*, is calculated from the percentage of incorrectly predicted outcomes from a validation set. The data is split 70:30% for training and validation and the error is based on the 30% of data used in the model. Errors were calculated from 5 sets of 70:30 splits and averaged using different seed values for a random selector to split the data.

$$L = \sum_{j=1}^{n} w_j I\{\hat{Y} \neq Y_j\} \qquad (1)$$

1) *Classifier Y*

A decision tree was trained using a random 70% of the experimental data for outcome *Y* (up z-axis or up arm) using attributes X_1-X_5, see Figure 4. This decision tree has 25 nodes and is up to 7 levels deep. The resubstitution error *L*=0.071 and the validation error is a little higher at *V*=0.138. This difference is expected given resubstitution errors tend to be more optimistic about the model than actual validation testing, see TABLE VIII.

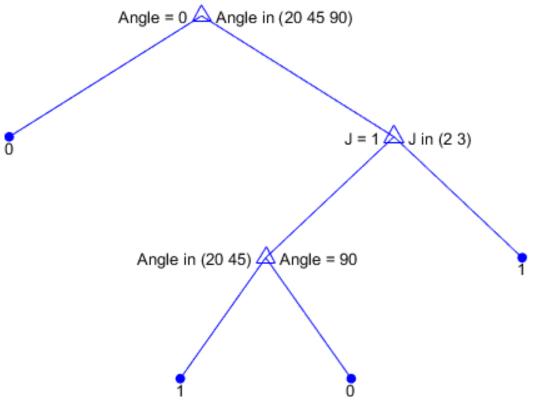


Figure 4. Decision tree for the outcome Y where the leaf node terminates in the outcome with the highest probability. Tree shown is pruned to 7 nodes for one of the five training sets.

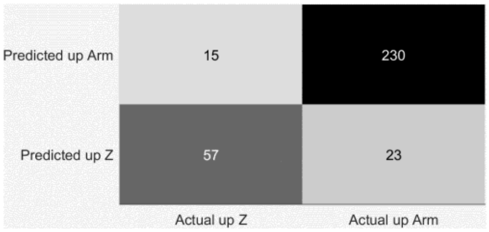


Figure 5. Confusion matrix for outcome *Y* showing the binary output either up the arm or up in the z-axis direction based on the decision tree. In 14 cases 'up arm' was incorrectly interpreted as 'up z' and in 18 cases 'up z' was incorrectly interpreted as 'up arm'.

The first branch uses the attribute X_1 (arm angle) and branches left with angle=0, giving probabilities of P(Y_0)=0.789 and P(Y_1)=0.211. Branching right for angles 20, 45 and 90 degrees gives probabilities of P(Y_0)=0.076 and P(Y_1)=0.924. The first level of the tree indicates quite high probabilities for predicting the correct outcome based on arm angle alone. The next branch splits with the X_2 attribute (dressing segment). For subsequent splits the number of nodes is reduced to less than 10% of *n* and over-fitting becomes a possible issue. A confusion matrix for this model shows where the going up the "arm" is mistaken for going up in "z", see Figure 5.

To determine the model error as a function of the number of nodes, 5 trees were produced with limits on the maximum number of branches. The results show that arm angle and dressing segment alone can be used to create an fairly accurate model (87.80%) and that adding in additional attributes will not improve the model accuracy by more than around 1%.

TABLE VIII MODEL ERROR FOR OUTCOME Y

Max. Nodes	**Attributes Used**	***L***	***V***
25	1-5	0.071	0.138
15	1-5	0.098	0.122
9	1-5	0.105	0.127
7	1,2	0.117	0.135
5	1,2	0.117	0.129

Referring back to the attribute cost analysis, the arm angle and dressing segment were estimated as the most difficult or costly inputs to the system. In this case, a choice for a lower cost attribute cannot be made as most of the variance in the model is accounted for by variables X_1 and X_2.

2) *Classifier Z*

A second model was trained using a decision tree based on the Z outcome classifier to determine which end effector to move. The possible outcomes are: left end effector only, right end effector only or both simultaneously. A random sample of 70% of the data was used to train the model. The errors are L=0.175 and V=0.245 indicating that this model has only a 75% accuracy. This may signify that there is only a limited association between outcome Z and the attributes chosen or that the categorisation of the data does not have sufficient resolution. There is also the fact that the sample size is smaller when distributed over more outcomes.

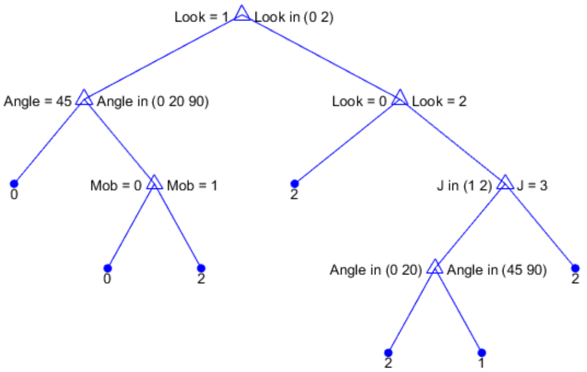
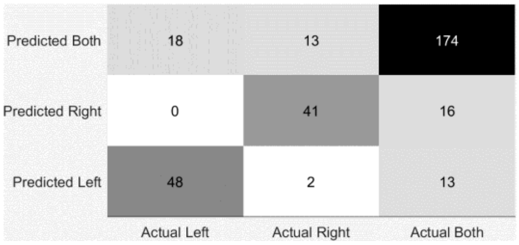


Figure 6. Decision tree for determining which end effector to move (outcome Z). The leaf node indicates which end effector to move: left = 0, right = 1 or both = 2. Tree shown is pruned to 13 nodes.

A confusion matrix for this decision tree is shown in Figure 7. This shows that needing just the left or right end effector is rarely confused with the individual end effector on the wrong side (left = 0, right =2) but is often confused with selecting both end effectors.



	Actual Left	Actual Right	Actual Both
Predicted Both	18	13	174
Predicted Right	0	41	16
Predicted Left	48	2	13

Figure 7. Confusion matrix for outcome Z showing which end effector to move based on the 37 node decision tree.

The ideal model would be able to predict which end effector to move based on the fewest attributes that are the easiest and most reliable to detect. TABLE IX shows the error rates for the model with increasing levels of pruning. The resubstitution error increases with a reduction in nodes whereas the validation error is minimised between 15 and 30 nodes (shown in bold). Approaching the minimum number of nodes indicates attributes 5, 1 and 2 (head orientation, arm angle and segment) give the highest model gains with 70.41% model accuracy. Referring to the attribute cost, head orientation was ranked moderately but by itself would only give 58.16% accuracy. Including arm angle (highest cost) improves the accuracy to 70.41%.

C. *Human vs. Machine*

Referring back to the decision flow charts in TABLE IV we can compare the initial human decision logic to the predictive models. We see similarities in the decision making process for the choice of direction, although the human flow chart begins with the proximity to the user (J>1) whereas the model used the arm angle to initially branch the decisions. However, in both cases arm angle and dressing segment were the most significant attributes.

TABLE IX MODEL ERROR FOR OUTCOME Z

Number Nodes	Attributes Used	L	V
37	1-5	0.1754	0.2449
29	1-5	0.2000	**0.2347**
23	1-5	0.2062	**0.2347**
15	1-5	0.2185	**0.2347**
13	1,2,4,5	0.2277	0.2499
11	1,2,4,5	0.2646	0.2959
9	1,2,5	0.2769	0.2959
5	1,5	0.3015	0.2959
3	5	0.3692	0.4184

Dressing strategy and dissimilar end effector heights (EE height bias) were not captured in this data and so can't be used for comparison to the predictive models. However, head orientation (X_5) based on voice localisation was a key attribute along with arm angle which wasn't considered. Models for the end effector selection may be improved by including dressing strategy, dissimilar end effector height categories and also knowing the dressing history (in this case, if one arm is in the jacket already).

D. *Generalisation*

The proposed scenario is very specific to one particular aspect of robot-assisted dressing namely the phrase "up" whilst receiving assistance with putting on a jacket. However, the proposed disambiguation method is not limited to this phrase. As highlighted in section III-B, the experiments have produced various examples of ambiguous statements, so this method can be adapted to resolve all other instances of ambiguous utterances. In a possible future implementation of a robotic dressing assistant that has multimodal capability, redundant inputs coming from different modalities can have independently assigned probabilities with which they contribute to the disambiguation of user intentions. In other words, in such a system the interpretation of a potentially ambiguous input in one modality may be resolved by observing a redundant input in a different modality.

V. CONCLUSIONS

We attempted to show the possibility of providing dressing assistance based on inputs excluding vision. The data used for the assessment was based on 18 participants in a Human-Human Interaction study. The phrase "up" was used very frequently and found to have ambiguous meaning without contextual information, referring to direction and end effector selection. Determining these contextual factors from the other modalities was the focus of a categorical decision tree analysis. A model for determining the direction indicated an 87.8% accuracy based on 2 contextual factors, user arm angle and dressing task segment. A model for determining the correct end effector had 70.41% accuracy based on the head orientation of the user and the angle of the arm. We present ideas on how these inputs may be attained through non-visual means, possibly through haptic perception of end effector position, proximity sensors and acoustic source localisation. We have built predictive models from the HHI data and the accuracy of these models indicate consistency of the decisions based on measurable inputs. We propose that additional inputs such as dressing strategy, end effector height bias and dressing history may improve the predictive model for end effector selection. The next phase of this work is to implement these models on a robotic platform and repeat the tests reported here.

Acknowledgment

The authors gratefully acknowledge funding from EU CHIST-ERA, UK EPSRC and the assistance of the staff at the Bristol Robotics Laboratory and Institut de Robòtica. Also Dr. C. L. Wright for advice regarding regression analysis.

References

[1] WHO, “Facts about ageing,” 2016. [Online]. Available: http://www.who.int/ageing/about/facts/ en/. [Accessed: 09-Sep-2016].

[2] A. Davies and A. James, *Geographies of ageing : social processes and the spatial unevenness of population ageing*. Ashgate Pub, 2011.

[3] Christie & Co, “The UK Nursing Workforce,” 2015.

[4] R. Foster, V. Fender, and Office for National Statistics, “Valuing Informal Adultcare in the UK,” no. June, pp. 1–23, 2013.

[5] A. Tinker, L. Kellaher, J. Ginn, and E. Ribe, “Assisted Living Platform - The Long Term Care Revolution,” *Housing Learning & Improvement Network*, no. September, King’s College London, 2013.

[6] LTCR, “Long Term Care Revolution,” 2016. [Online]. Available: http://www.ltcr.org/index.html. [Accessed: 09-Sep-2016].

[7] B. J. Dudgeon, J. M. Hoffman, M. A. Ciol, A. Shumway-Cook, K. M. Yorkston, and L. Chan, “Managing Activity Difficulties at Home: A Survey of Medicare Beneficiaries,” *Arch. Phys. Med. Rehabil.*, vol. 89, no. 7, pp. 1256–1261, 2008.

[8] A. Tinker and P. Lansley, “Introducing assistive technology into the existing homes of older people: feasibility, acceptability, costs and outcomes,” *J. Telemed. Telecare*, 2005.

[9] E. Broadbent, R. Stafford, and B. MacDonald, “Acceptance of healthcare robots for the older population: review and future directions,” *Int. J. Soc.*, 2009.

[10] T. Tamei, T. Matsubara, A. Rai, and T. Shibata, “Reinforcement learning of clothing assistance with a dual-arm robot,” *Int. Conf. Humanoid Robot.*, pp. 733–738, 2011.

[11] T. Matsubara, D. Shinohara, and M. Kidode, “Reinforcement learning of a motor skill for wearing a T-shirt using topology coordinates,” *Adv. Robot.*, vol. 27, no. 7, pp. 513–524, 2013.

[12] N. Koganti, T. Tamei, T. Matsubara, and T. Shibata, “Estimation of Human Cloth Topological Relationship Using Depth Sensor for Robotic Clothing Assistance,” in *Proceedings of Conference on Advances In Robotics*, 2013, p. 36:1--36:6.

[13] S. D. Klee, B. Q. Ferreira, R. Silva, P. Costeira, F. S. Melo, and M. Veloso, “Personalized Assistance for Dressing Users,” in *International Conference on Social Robotics (ICSR 2015)*, 2015, vol. 9388, pp. 359–369.

[14] K. Yamazaki, R. Oya, K. Nagahama, K. Okada, and M. Inaba, “Bottom dressing by a life-sized humanoid robot provided failure detection and recovery functions,” *2014 IEEE/SICE Int. Symp. Syst. Integr. SII 2014*, pp. 564–570, 2014.

[15] Y. Gao, H. J. Chang, and Y. Demiris, “User modelling for personalised dressing assistance by humanoid robots,” in *IEEE International Conference on Intelligent Robots and Systems*, 2015, vol. 2015–Decem, pp. 1840–1845.

[16] Y. Gao, H. J. Chang, and Y. Demiris, “Iterative path optimisation for personalised dressing assistance using vision and force information,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4398–4403.

[17] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and a. Waibel, “Natural human-robot interaction using speech, head pose and gestures,” *2004 IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IEEE Cat. No.04CH37566)*, vol. 3, pp. 2422–2427, 2004.

[18] A. Jevtic, G. Doisy, Y. Parmet, and Y. Edan, “Comparison of Interaction Modalities for Mobile Indoor Robot Guidance: Direct Physical Interaction, Person Following, and Pointing Control,” *IEEE Trans. Human-Machine Syst.*, vol. 45, no. 6, pp. 653–663, Dec. 2015.

[19] M. Khoramshahi, A. Shukla, S. Raffard, B. G. Bardy, and A. Billard, “Role of Gaze Cues in Interpersonal Motor Coordination: Towards Higher Affiliation in Human-Robot Interaction,” *PLoS One*, vol. 11, no. 6, p. e0156874, Jun. 2016.

[20] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, “Recognizing engagement in human-robot interaction,” in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 375–382.

[21] M. S. Bartlett, G. Littlewort, I. Fasel, and J. Movellan, “Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.,” *Comput. Vis. Pattern Recognit. Work.*, vol. 5, pp. 53–53, 2003.

[22] D. Kuli and E. a Croft, “Affective State Estimation for Human – Robot Interaction,” vol. 23, no. 5, pp. 991–1000, 2007.

[23] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, “Providing the basis for human-robot-interaction,” *Proc. 5th Int. Conf. Multimodal interfaces - ICMI ’03*, no. July 2004, p. 28, 2003.

[24] Z. Wang, K. Mulling, M. P. Deisenroth, H. Ben Amor, D. Vogt, B. Scholkopf, and J. Peters, “Probabilistic movement modeling for intention inference in human-robot interaction,” *Int. J. Rob. Res.*, vol. 32, no. 7, pp. 841–858, Jun. 2013.

[25] G. Chance, A. Camilleri, B. Winstone, P. Caleb-solly, and S. Dogramadzi, “An Assistive Robot to Support Dressing – Strategies for Planning and Error Handling,” in *6th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2016, pp. 774–780.

[26] G. Chance, A. Jevtić, P. Caleb-Solly, and S. Dogramadzi, “A Quantitative Analysis of Dressing Dynamics for Robotic Dressing Assistance,” *Front. Robot. AI*, vol. 4, p. 13, 2017.

[27] O. Lemon and A. Gruenstein, “Multithreaded Context for Robust Conversational Interfaces,” *ACM Trans. Comput. Interact.*, vol. 11, no. 3, pp. 241–267, 2004.

[28] R. Mead, A. Atrash, and M. J. Matarić, “Automated Proxemic Feature Extraction and Behavior Recognition: Applications in Human-Robot Interaction,” *Int. J. Soc. Robot.*, vol. 5, no. 3, pp. 367–378, 2013.

[29] P. Rota, N. Conci, and N. Sebe, “Real time detection of social interactions in surveillance video,” *Eur. Conf. Comput. Vis.*, 2012.

[30] H. Zhang and L. Parker, “4-dimensional local spatio-temporal features for human activity recognition,” *Robot. Syst. (IROS), 2011 IEEE/ …*, 2011.

[31] T. Lourens, R. Van Berkel, and E. Barakova, “Communicating emotions and mental states to robots in a real time parallel framework using Laban movement analysis,” *Rob. Auton. Syst.*, 2010.

[32] T. Kundu, “Acoustic source localization,” *Ultrasonics*, vol. 54, no. 1. pp. 25–38, Jan-2014.

[33] S. Paulose, E. Sebastian, and B. Paul, “Acoustic Source Localization,” *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, vol. 2, no. 2, pp. 933–939, 2013.

[34] A. Kapusta, W. Yu, T. Bhattacharjee, C. K. Liu, G. Turk, and C. C. Kemp, “Data-driven haptic perception for robot-assisted dressing,” *Robot Hum. Interact. Commun. (RO-MAN), 2016 25th IEEE Int. Symp.*, pp. 451–458, 2016.

[35] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees. Wadsworth & Brooks*, 1 edition. Boca Raton: Chapman & Hall/CRC, 1984.