Early prediction for physical human robot collaboration in the operating room

Tian Zhou · Juan Pablo Wachs

Received: 31 December 2016 / Accepted: 17 October 2017 / Published online: 6 November 2017
© Springer Science+Business Media, LLC 2017

Abstract To enable a natural and fluent human robot collaboration flow, it is critical for a robot to comprehend their human peers’ on-going actions, predict their behaviors in the near future, and plan its actions correspondingly. Specifically, the capability of making early predictions is important, so that the robot can foresee the precise timing of a turn-taking event and start motion planning and execution early enough to smooth the turn-taking transition. Such proactive behavior would reduce human’s waiting time, increase efficiency and enhance naturalness in collaborative task. To that end, this paper presents the design and implementation of an early turn-taking prediction algorithm, catered for physical human robot collaboration scenarios. Specifically, a robotic scrub nurse system which can comprehend surgeon’s multimodal communication cues and perform turn-taking prediction is presented. The developed algorithm was tested on a collected data set of simulated surgical procedures in a surgeon–nurse tandem. The proposed turn-taking prediction algorithm is found to be significantly superior to its algorithmic counterparts, and is more accurate than human baseline when little partial input is given (less than 30% of full action). After observing more information, the algorithm can achieve comparable performances as humans with a $F_1$ score of 0.90.

Keywords Turn-taking prediction · Recurrent neural network · Human–robot interaction · Robot nurse · Operating room · Sensor fusion · Multimodal · Long short-term memory · Dempster–Shafer theory

1 Introduction

Turn-taking prediction is about the capability to comprehend the on-going task progress and predict where, when and how to seize the next turn during multi-agent collaborations (Sacks et al. 1974). Fluent and natural turn-taking regulations would greatly increase team performances and lead to better social connections among team members (Sebanz et al. 2006; Marsh et al. 2009). Robots, designed to work in close proximity with humans, need to have the capability to understand turn-taking events and plan their actions accordingly. A critical requirement for fluent turn-taking regulation is the capability to make decisions and preparations for turn-taking events beforehand (Sacks et al. 1974). Specifically, robots which perform physical interactions with humans need significant amount of time to plan and execute their motions (Canny 1988), thus being able to make a prediction as early as possible would reduce human partners’ waiting time. During collaboration, such early-prediction behavior would minimize mutual silence (both parties relinquish the turn) and mutual conflict (both parties attempt to seize the turn simultaneously), leading to a more synchronized turn-taking regulation (Cutler and Pearson 1986).

The requirement for early turn-taking prediction stands out more clearly in high-risk and high-paced tasks like surgery. In the Operating Room (OR), the scrub nurse and...
the surgeon perform fast, accurate and highly coordinated turn-taking actions when exchanging surgical instruments. A nurse delivers surgical instruments to a surgeon based on explicit requests (e.g., uttering the words “scissors”) and implicit requests expressed by body language (e.g., leaning forward, evoking a hand gesture or looking at an instrument). All these forms of expressions are used to inform the nurse ahead of time that it is his/her turn to continue. Additionally, the nurses maintain a knowledge-base of common surgery procedures. Based on observing the current work-flow, the nurses can predict the most-likely next surgical operation and prepare instruments correspondingly. All such context knowledge about the surgical task helps nurses to better predict the timing (i.e., when) and the objects (i.e., what) of the next turn. Such complex and coordinated turn-taking behaviors between surgeons and nurses are learned, acquired and executed precisely in the OR through experience and teams’ practice. As robots are being introduced gradually to the OR to deal with nurses’ shortage problem (Buerhaus et al. 2009; Needleman et al. 2011; Zhou and Wachs 2016), they are expected to have the capability of performing fluent turn-taking actions as human nurses.

Such robotic scrub nurse (RSN) will need to understand subtle verbal and non-verbal cues and infer the surgeon’s intentions timely and accurately, in order to perform turn-taking actions correspondly. An illustration of the proposed RSN system is shown in Fig. 1. To design a fully functional RSN, it is critical to develop a computational framework for early turn-taking prediction and that is the focus of this paper.

2 Related work

This section presents the related work about turn-taking analysis, in the perspective of (1) human–human turn-taking; (2) human–robot turn-taking; (3) automatic turn-taking recognition and (4) predictive turn-taking. To conclude this section, the innovation points are presented.

2.1 Human–human turn-taking

The analysis of turn-taking in human–human interactions has drawn attention from researchers with psychology, linguistics and engineering background. The research has focused on conversational tasks, where linguistic turn-ending cues such as pause duration (Schlangen 2006), pitch levels (Ward et al. 2010) and intonation (Gravano and Hirschberg 2011) have been identified. Nonverbal behaviors such as gaze and posture shifts (Padilha and Carletta 2003) have also been studied during conversational exchanges, however, these are more difficult to spot during interaction.

2.2 Human–robot turn-taking

In human robot interaction area, the CHARM project [13] studied nonverbal cues as key contributors to timing coordination among human collaborators. Its goal was to develop a turn-taking aware robot assistant to work alongside human workers in a manufacturing environment. Calisgan et al. (2012) studied the different types and occurrences of implicit communication cues as turn-taking regulators in assembly tasks. Also, timing in multimodal turn-taking interaction (i.e., speech, gaze, gesture) was investigated between humans and robots through a collaborative Towers of Hanoi challenge by Chao and Thomaz (2012). The naturalness of human robot turn-taking has also been studied in robot-to-human handover tasks, from the perspective of distinct handover poses (Cakmak et al. 2011), object affordances (Chan et al. 2014) and unambiguous approach angles (Unhelkar et al. 2014). All the above-mentioned work focuses on observing and modelling turn-taking, without automatically recognizing them.

2.3 Automatic turn-taking recognition

Machine learning techniques have been applied to recognize turn-taking events automatically, mainly for spoken dialog systems. The speaker’s end-of-turn is detected by an AI agent using Support Vector Machines (Arsikere et al. 2015). Decision tree and its variants have also been used to detect turn-taking (Schlangen 2006; Saito et al. 2015; Raux and Eskenazi 2008). Such approaches have been extended to multimodal end-of-turn detection in multi-party meetings using Conditional Random Field (De Kok and Heylen 2009).

2.4 Predictive turn-taking

Turn-taking modelling has to enable early prediction so that turn-taking decisions are made much earlier before the
transition event occurs (Sacks et al. 1974). Given an observation of human actions, its meaning needs to be interpreted before the action is completely finished. For example, if an algorithm can only recognize the action meaning after it is fully conducted (100%), there is no prediction involved and the algorithm degenerates to classification. While on the other hand, predicting the action type given 0% of data would mean the earliest possible prediction (before the action even takes place). Heeman and Lunsford (2015) evaluated human’s performance when predicting who will speak next, achieving an accuracy of 61%. The work presented by Hart et al., (Gravano and Hirschberg 2011) achieved similar results using utterance-related cues and machine learning techniques. Conversely to utterance, gestures have also been studied as indicators for turn-taking. Early gesture recognition has been studied with dynamic time warping (Mori et al. 2006) and Naive Bayes (Escalante et al. 2016) previously.

A comprehensive comparison of the proposed turn-taking prediction algorithm with the relevant literature is presented in Table 1. The main scope of this table is to compare how early the turn-taking is analyzed (measured in percentage of full event), the involved modalities and the collaboration agents, as guiding criteria for the design of innovative turn-taking algorithms. This paper fits within the described criteria, and its innovation points are summarized as:

- Proposed a statistics-based feature selection process for raw sensor input.
- Leveraged a Long Short-Term Memory network (Guntakandla and Nielsen 2015) for early turn-taking prediction.
- Applied the Dempster–Shafer Theory of evidence for sensor fusion to increase robustness.
- Developed a system which enables multimodal sensing for human robot collaboration in the operating room.

This paper contributes to both the theory and implementation aspects of turn-taking. To the best of our knowledge, it is the first time that the concept of “early turn-taking prediction” is introduced, with a corresponding framework and implementation. The comprehensive experiments further validate the performance of the proposed framework.

### 3 Turn-taking in homogeneous humans team

This section describes the procedure used to collect observations during turn-taking activities in pure human’s team during a surgical task. The recorded observations were then
used to guide the design of the early turn-taking algorithm. In the following, the surgical task setup (Sect. 3.1), multimodal signal collection process (Sect. 3.2) and human-state annotations (Sect. 3.3) are discussed.

3.1 Surgical task setup

A simulation platform for surgical operations was used to capture turn-taking actions between surgeons and nurses, as shown in Fig. 1. The platform includes a patient simulator and a set of instruments required to complete a mock surgical task of abdominal incision and closure (Martyak and Curtis 1976). The detailed steps of the surgical task, together with the surgical instruments needed for each step are shown in Fig. 2. In this task, the surgeon and the nurse collaborate by delivering and retrieving surgical instruments to complete the surgery successfully. This simulation setup and the surgical task procedure have been used in our previous work [(gesture robot nurse (Jacob et al. 2012) and tele-mentoring in the OR (Andersen et al. 2016)].

Participants were recruited to serve as surgeons. After signing the consent form, they were instructed about the steps of the mock abdominal incision and closure task through a video tutorial. Then the participants performed a “warm-up” trial on the surgical simulator to increase familiarity with the task. Afterwards, each participant repeated the surgical task five times in order to reach the expertise level required. During the course of the surgical operations, the surgeons requested surgical instruments from the nurses, and handed back the used instruments. The surgeons would reach out for the instrument when it is presented by the nurse, and then continue to work on the task. The surgeons were explicitly required to use verbal commands to request each instrument, in order to diminish the effects of potential false nurse predictions. The noise resulting from medical equipment and reverberations was not simulated. In addition to the verbal commands, the surgeons’ body, gaze and arm motions were all used together as implicit communication cues to trigger the nurses’ actions, but no explicit request was given to the subjects about this. The nurse had to understand the surgeon’s turn-taking communication cues (both implicit and explicit) in order to react according to the surgeon’s expectations. Those implicit and explicit communication cues were collected through a set of sensors for the following turn-taking analysis, but the participants were not notified of any information of the specific sensing channels. The surgeon-nurse tandem forms a type of asymmetric collaboration, where the surgeon leads the task (i.e., a dominant agent) while the nurse mainly follows the task (i.e., a submissive agent). In this scenario, the focus is on enabling the follower to predict the leader’s turn-taking intention in order to collaborate efficiently. Thus, this paper only focuses on developing algorithms to enable robotic nurses to predict surgeons’ turn-taking intentions, while ignoring the aspect of humans predicting robot motions, or co-prediction scenarios.

In this surgical task, the instrument request event was treated as the main turn-taking activity. The surgeon needs around 14 surgical instruments to finish one trial of the task, resulting in around 14 turn-taking instances. The instruments used are scalpel, hemostat, forceps, retractor, scissors and needle. Each participant repeated the surgical task 5 times. Each trial takes 3–7 min to finish, with subjective variations. The study was approved by IRB (protocol number 1305013664), and the participants were recruited through emails and personal inquiries. In total 12 participants were recruited with ages in range 20–31 (mean = 25.7, std = 2.93). The participants are all graduate students from the College of Engineering at Purdue University.

3.2 Multimodal signal collection

The communication cues expressed by the surgeon during the surgical task were recorded for further analysis. Three sensors were used to record those communication cues, namely Myo armband, Epoc headset and Kinect. An illustration of the captured multimodal signals is presented in Fig. 3. The details of each sensor channel is given below.

3.2.1 Myo armband

a gesture capturing device worn on the forearm, capturing the motion and Electromyography (EMG) signals on the surgeon’s dominant arm. The following information was recorded:

- Arm orientation (roll, pitch and yaw), 3D
- Arm acceleration (xyz), 3D
- Arm gyroscope (xyz), 3D
- Arm muscle EMG signals, 8D

![Fig. 2 Steps of the mock abdominal incision and closure surgical task with the surgical instruments used in each step](image-url)
3.2.2 EpoC headset

A brain–computer interface based on electroencephalography (EEG) technology. It is used to capture surgeon’s head motions and EEG signals. The following data was recorded:

- Head EEG signals (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF42), 14D
- Head gyro (pitch and yaw motion), 2D
- Emotion classification (engagement, frustration, meditation, excitement and valence), 5D

3.2.3 Kinect

A motion sensing device. Joint, body and face tracking algorithms were used to extract participants’ head poses, body postures and utterances. The following information was recorded:

- Face orientation (roll, pitch and yaw motion), 3D
- Body postures (left-right leaning and forward-backward leaning), 2D
- Left hand extension (vector from joint SpineMid to joint LeftHand), 3D
- Right hand extension (vector from joint SpineMid to joint RightHand), 3D
- Acoustic amplitude, 1D

3.2.4 Synchronization

The real-time data from all three modalities was synchronized at a frame rate of 20Hz and then concatenated together, forming a basic data level fusion. More advanced sensor fusion techniques will be introduced in Sect. 5, for now the data level fusion only serves for data preparation and formatting purposes. After the data level fusion, for each time frame \( t \), the fused sensor measurement (denoted as \( \vec{r}_t \)) consists of 50 values (addition of all the dimensions detailed above, denoted as \( M \)). Besides, the color images from Kinect were also recorded for annotation purposes (not included in \( \vec{r}_t \)).

3.3 Surgeon state annotation

The recorded communication signal (\( \vec{r}_t \)) is obtained during periods when the surgeon is either focusing on the surgical operation or intending to request an instrument. These periods were further segmented into different surgeon states. It is assumed that the surgeon is always in one of the two states. The two defined states of a surgeon are (1) operating: when the surgeon is engaged in current surgical operation and has no intention to relinquish the turn. This corresponds to the period when surgeon holds the current turn; or (2) requesting: the surgeon is requesting an instrument from the nurse. This corresponds to the period when surgeon is approaching the end of his/her turn, and wants to pass the turn to the nurse who will fetch and deliver the requested instrument. A finite state machine to define the surgeon states is shown in Fig. 4. In this scenario, instrument request events all happen during the switch from operating state to requesting state. It is assumed by this paper that the human won’t change his mind after switching to requesting state, and won’t return back to operating state until receiving the requested surgical instrument. Thus, the goal of the early turn-taking prediction algorithm is to predict the transition from operating state to requesting state as early as possible. Such binary end-of-turn detection approach (i.e., whether the human wants to keep or relinquish the turn) is a common practice in turn-taking analysis (De Kok and Heylen 2009; Guntakandla and Nielsen 2015; Heeman and Lunsford 2015; Arsikere et al. 2015), and is also sufficient for the subsequent robot action decisions (i.e., engage interaction or not).

The recorded video was reviewed by humans to generate time indexes for each surgeon state. From the experiment recordings, it is observed that the participants use a combination of communication modalities to express their intent to request an instrument. Some modality starts early (e.g., changes in body stance) while others start late in the process (e.g., explicit hand gestures and specific verbal commands verbal requests). Nurses can recognize the request intent as early as the earliest clue starts, or as late as the latest clue ends.

![Finite state machine to represent surgeon’s states. The black dot indicates initial state](image-url)
4 Turn-taking prediction

The main contribution of this paper is a computational framework for early turn-taking prediction, which can predict surgeon’s turn-taking intentions before they are fully conceived. The proposed early turn-taking prediction framework is shown in Fig. 6. The surgeon was monitored through three sensors, which are Epoc headset, Myo armband and Kinect sensor. The raw data was sampled and encoded, and the most relevant features were retained through a feature selection process. Then, the selected features were used for temporal modelling for turn-taking prediction. The prediction results from different resources were fused using the Demspiter-Shafer method to achieve a final result. This result triggered the robot motion planning algorithm, which aims to pick up and deliver the surgical instrument to the surgeon at the right time. This paper mainly focuses on the turn-taking prediction and fusion part, while neglecting the robot motion planning aspect. However, the entire closed-loop process flow was presented here for integrity and illustration purposes. In the reminder of this section, detailed descriptions of channel preprocessing (Sect. 4.1), feature construction and selection (Sect. 4.2) and temporal modelling (Sect. 4.3) are presented.

4.1 Channel preprocessing

The first step in the signal processing pipeline was noise reduction. The Exponentially Weighted Moving Average (EWMA) method was used, which is a common noise reduction technique for time-series data (Lucas and Saccucci 1990). This filter was applied to smooth the raw signal \( \tilde{r}_t \):

\[
\tilde{s}_t = \alpha \tilde{r}_t + (1 - \alpha) \tilde{s}_{t-1}, \quad t \in [1, L]; \quad \tilde{s}_0 = \tilde{r}_0
\]

where \( \tilde{r}_t \) is the raw sensor measurement at time \( t \), \( L \) is the discrete length of signal and \( \tilde{s}_t \) is the filtered measurement at time \( t \). The smoothing was applied to each of the \( M \) channels of \( \tilde{r}_t \). The weighting parameter \( \alpha \) controls the relative importance of raw measurement, which was empirically set to 0.2 for best performance in our environment.

The smoothed signal \( \tilde{s}_t \) was then normalized to have zero mean and unit variance in each of the \( M \) channels, following: \( \tilde{s}_t \{ c \} = (\tilde{s}_t \{ c \} - \mu_c) / \sigma^2_c \), where \( \mu_c \) and \( \sigma^2_c \) are the global mean and variance of channel \( c \) in signal \( \tilde{s}_t \). Here the statistics \( \mu_c \) and \( \sigma^2_c \) were calculated based on \( \tilde{s}_t \{ c \} \) from both states (requesting and operating), hence the name global. Such approach preserves any offset between data of different states, while at the same time enforces the multimodal signals to be in similar magnitudes for further comparison and aggregation purposes.

For each time stamp \( t \), the smoothed and normalized signal \( \tilde{X}_i \) consists of \( M \) values. The collection of \( \tilde{X}_i \) within a given time window \( i \) (e.g., \( t \in [t_i', t_i''] \)) is cumulatively denoted as...
segment $\tilde{X}_i$ in time window $i$ was stacked together to form a matrix representation $\tilde{X}_i \in \mathbb{R}^{L_i \times M}$, where $L_i$ is the length of segment $i$ (i.e., $L_i = t^i_f - t^i_i$) and $M$ is the dimension of raw data (i.e., $M = 50$). For each stacked segment $\tilde{X}_i$, a label $y_i \in \{0, 1\}$ was assigned to indicate whether segment $\tilde{X}_i$ belongs to requesting state ($y_i = 1$) or operating state ($y_i = 0$). The combination $\{\tilde{X}_i, y_i\}$ forms the initial data set for turn-taking analysis, which aims to infer $y_i$ from $\tilde{X}_i$.

4.2 Feature construction and selection

To construct a compact yet relevant feature representation, a feature construction and selection process was carried out. The basic principle is that the raw signal was first convolved with a set of filters to characterize different patterns in the original signal (i.e., feature construction stage). Then through correlation analysis, the most relevant features were selected for the subsequent analysis (i.e., feature selection stage). These two stages are described in more detail in the following.

Feature construction the signal after channel preprocessing (i.e., $\tilde{X}_i$) was temporally encoded by convolving with a set of filter banks. The set of filter banks $\{F_q|q = 1, \ldots, Q\}$ includes $Q$ different filters to encode different temporal characteristics of the original signal. The filter bank includes the identity transformation, Sobel operator (window size 3), Canny edge detector (approximated with derivative of Gaussian), Laplacian of Gaussian (LoG) and Gabor filter (gabor1 with central frequency 100Hz and phase 0 and gabor2 with central frequency of 100Hz and phase $\pi/2$). For each segment $\tilde{X}_i \in \mathbb{R}^{L_i \times M}$, each of its $M$ channels was convolved with all $Q$ filters from the filter bank, resulting in $\bar{X}_i \in \mathbb{R}^{L_i \times MQ}$.

Feature selection the features constructed from the $MQ$ channels (i.e. $\bar{X}_i$) showed high redundancy and hence only those channels highly correlated with the turn-taking labels were kept. Specifically, the $m$ most salient features were selected out of $MQ$ encoded features ($m \ll MQ$), based on a statistics-based approach. This method is inspired by Morency et al. (2008), with the major difference that the previous work requires manually annotated discrete inputs, while in this case the input data can have continuous values and is automatically encoded. This extension from discretely valued inputs to continuously valued input can increase the applicability of the feature selection algorithm, and also saves time and effort in manual annotation.

The feature selection process works as the following. First, each of the $MQ$ continuous signals was converted into binary representations using clustering ($K$-means with 2 clusters and each cluster represents a binary level). Then, a $\chi^2$ test was carried out between each binarized signal and the ground truth label. The $MQ$ features were then ranked based on the test statistics of the corresponding $\chi^2$ test, and the $m$ features of the largest test statistics were retained as the optimal feature set, represented as $X_i \in \mathbb{R}^{L_i \times m}$. This representa-
4.3 Early prediction

We propose the usage of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), a type of recurrent neural network, for early turn-taking prediction. For comparison purposes, the Multi-Dimensional Dynamic Time Warping (MD-DTW) (ten Holt et al. 2007) is also tested. The DTW serves as a baseline and a representative of traditional temporal modelling algorithms.

4.3.1 LSTM

It is a recurrent neural network architecture which has been successfully applied to handwriting recognition (Graves et al. 2009) and emotion recognition (Willmer et al. 2013), among other applications. This network structure has the intrinsic capability to automatically extract spatial-temporal patterns. Even though there are many different recurrent neural network structures [(e.g. Gated Recurrent Unit (Cho et al. 2014)], a recent comparison finds out that most popular variants perform similarly (Greff et al. 2015). Thus, we decided to choose LSTM as the sequence modelling network for our scenario. The basic structure of a cell of LSTM is shown in Fig. 7.

The LSTM structure can be described by a set of formulas:

\[
f_t = \sigma(W_f[x_t, h_{t-1}] + b_f) \\
i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \\
g_t = \tanh(W_g[x_t, h_{t-1}] + b_g) \\
o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \\
c_t = f_t \odot c_{t-1} + i_t \odot g_t \\
h_t = o_t \odot \tanh(c_t) \\
y_t = \text{softmax}(W_y h_t)
\]

where \(x_t\) is the input temporal sequence at time \(t\) (corresponding to a row in \(X_i \in \mathbb{R}^{L_i \times m}\) as defined above), \(h_{t-1}\) is the output of the memory cell at time \(t-1\). \(W_f, W_i, W_g, W_o\) are weight matrices, and \(b_f, b_i, b_g, b_o\) are bias terms respectively. \(\sigma\) denotes a sigmoid function, \(\tanh\) indicates the hyperbolic tangent function and \(\odot\) denotes an element-wise multiplication. \([x_t, h_{t-1}]\) denotes the concatenation of vector \(x_t\) and \(h_{t-1}\). \(f_t, i_t, g_t, o_t, c_t, h_t\) are forget gate, input gate, candidate gate, output gate, cell unit and hidden state respectively. The memory unit \(c_t\) is generated by the coupling between the input and forget gates. LSTM can model long-term temporal dependencies because the memory unit can selectively “remember” or “forget” past information. The strategy to open or close gates is embedded in the learned weights and biases.

During the training stage, the segment-label pairs \(\{(X_i, y_i) | X_i \in \mathbb{R}^{L_i \times m}, y_i \in \{0, 1\}, i \in [1, N_0 + N_1]\}\) from both requesting (in total \(N_1\) instances) and operating states (in total \(N_0\) instances) were introduced into the network. The learning algorithm then updates the weights and biases for each gate, by minimizing the softmax cross entropy between the ground truth and the predicted labels. The Adam optimizer (Kingma and Ba 2014) was used to iteratively estimate the parameters of the network, based on stochastic gradient descent algorithm (SGD).

During the testing stage, a given unknown sequence \(X_k\) is presented to the network. The memory cell output of the last time step \(h_t\) is multiplied by \(W_y\) and then applied the softmax function to determine the model output \(\hat{y}_k \in \{0, 1\}\), which is the predicted label for segment \(X_k\).

4.3.2 Multi-dimensional dynamic time warping (MD-DTW)

DTW is one of the most traditional and successful temporal modelling algorithms, and has been successfully applied to speech recognition (Abdulla et al. 2003), early gesture recognition (Mori et al. 2006) and robot trajectory navigation (Vakanski et al. 2012). In ten Holt et al. (2007), 1D-DTW was extended to a multi-dimensional case (MD-DTW) and the superiority of MD-DTW over any 1D-DTW systems was shown.

In this scenario, the MD-DTW algorithm (ten Holt et al. 2007) was applied with L1-norm as the distance measurement for two multi-dimensional signals. A nearest-neighbor classification scheme is used, which assigns the label of the closest example to the unknown sequence.

The training stage consists of finding the most representative instance (known as the template) for each surgeon state. Take the requesting state as an example (\(y_i = 1\)). The selection of the template for this state (denoted as \(X^1_i\)) is based on within-group consensus. The instance \(X_i\) which has the least cumulative DTW distances with the remaining instances of the same state was chosen as the template, i.e:
\[ X^*_k = \arg \min_{i=1}^{N_1} \sum_{j,j \neq i}^{N_1} DTW(X_i, X_j) \]  

The same procedure is repeated for the operating state and the resultant template is denoted as \( X^*_k \). This pair of templates \( \{X^*_1, X^*_k\} \) was selected based on the training instances from all the operation steps, and can generalize to novel operations/surgeries.

During testing stage, for a given unknown sequence \( X_k \), its DTW distance with the template pairs \( \{X^*_1, X^*_k\} \) is calculated. Then the label associated with the smaller distance is chosen as the prediction \( \hat{y}_k \) for sequence \( X_k \), i.e.:

\[ \hat{y}_k = \arg \min_{y \in [0,1]} DTW(X_k, X^*_k) \]  

5 Decision fusion with Dempster–Shafer theory

Turn-taking intent is expressed through several modality channels, together with additional context knowledge. This section discusses how to fuse the information from multiple channels together in order to achieve a robust and comprehensive turn-taking reasoning.

Fusion can happen at data level, feature level and decision level. A decision-level fusion technique was utilized due to its low computational cost and robustness. First, independent classifiers were trained using information from different resources, and then the outputs of these classifiers were combined together using weighted average. Common approaches to find the weights include grid search (Hinton 2010), random search (Bergstra and Bengio 2012) and supervised retraining on confidence outputs (Mittal et al. 2011). However, the searching approaches are time-consuming and suffer from local minimum. Supervised retraining approach cannot generalize to scenarios where modalities are missing or added. Instead, the Dempster–Shafer Theory (DST) was resorted for decision fusion.

DST provides a framework to combine degree of beliefs derived from independent evidence channels. The DST framework is easily extensible and allows for incremental modality addition or subtraction. Also, it provides a measurement of uncertainty levels and modality disagreement. Compared to other uncertainty reasoning frameworks such as fuzzy logic (Klir and Yuan 1995), DST has the advantage of preserving uncertainty levels from each evidence channel and providing a boundary for the certainty levels of the final belief (Harmanec and Klir 1994). Due to DST’s wide applicability, it has been used for decision fusion in visual tracking (Li et al. 2013), human activity recognition (McKeever et al. 2010) and robot localization (Hughes and Murphy 1992), to mention a few examples.

In this scenario, the DST was used to accumulate confidences from both spatial domain (across different modalities and context knowledge) and temporal domain (across consecutive time frames), similar to the spatial-temporal weighted Dempster–Shafer scheme (Li et al. 2013). The major difference is that our approach in addition considers contextual knowledge about task progress.

5.1 DST formulation

DST is considered as a generalization of Bayesian methods, with the difference that Bayesian method only assigns weights (i.e., probabilities) to each individual state of the system, while DST assigns weights (i.e., belief) to each combinatory propositions of the system states. Such a generalization grants DST more flexibility in modelling uncertainties and fusing beliefs from multiple modalities (Wu et al. 2002).

Let \( \epsilon = \{\epsilon_1, \epsilon_2, ..., \epsilon_n\} \) represents the \( n \) possible states of the system under consideration. The power set \( 2^\epsilon \) contains all subsets of \( \epsilon \) and thus represents all propositions about the actual state of the system. The DST theory then assigns a Basic Belief Assignment (BBA) \( b \), to each element of the power set \( 2^\epsilon \), i.e., \( b : 2^\epsilon \rightarrow [0,1] \) such that \( b(\emptyset) = 0 \) and \( \sum_{A \in 2^\epsilon} b(A) = 1 \). For state \( A \in 2^\epsilon \), its BBA value \( b(A) \) expresses the evidence to support the claim that the actual system state belongs to \( A \). The DST theory also provides a framework to combine confidences from different sources, known as the Dempster’s Rule of Combination (DRC). It calculates a joint belief \( b_{1,2}(\cdot) \) from two independent beliefs \( b_1(\cdot) \) and \( b_2(\cdot) \), according to:

\[ b_{1,2}(A) = \frac{1}{1 - K} \sum_{B \cap C = A} b_1(B) b_2(C) \]  

where \( K = \sum_{B \cap C = \emptyset} b_1(B) b_2(C) \) is a measure of disagreement between the two beliefs \( b_1(\cdot) \) and \( b_2(\cdot) \). A large \( K \) value implies strong disagreement between the two beliefs regarding the actual state of the system.

5.2 DST in turn-taking prediction

We created a BBA function from each modality available, and then used DRC to combine the evidence from all the BBA dimensions to reach a single final decision. The construction of the BBA for each modality is based on the output of each classifier. For LSTM, the output of memory cell of the last time step \( h_t \) is multiplied by \( W_y \) and then transformed by the softmax function, i.e. \( softmax(W_y h_t) \). The result is a vector of values in the range \([0,1]\) that add up to 1. This vector is used to initialize the BBA values, by using the corresponding \( i_{th} \) entry in the vector for \( b(A = \{y_i\}) \).

In our system, three sensors were used to capture human communication cues, namely Epoc headset (\( b_e \)), Kinect (\( b_k \))
and Myo armband \((b_m)\). A LSTM network is trained individually on features corresponding to each sensor. During testing, the output from each sensor is used to construct each BBA function, and finally a joint belief function \(\tilde{b}\) is calculated by combining individual BBA functions using DRC. The ultimate decision is made solely based on the joint belief function \(\tilde{b}\).

The DRC fusion rules provide a robust and flexible framework for spatial-temporal confidence fusion. For example, in situations where one of the sensors is not available due to potential hardware failures, the DRC combination rules can still accumulate confidences from the working sensors and make a final decision.

To make a final decision \(\tilde{b}^t\) for time frame \(t\), we used 1) the individual BBA based on the output of the LSTM network for each modality, denoted as \(b_c^t, b_r^t\) and \(b_m^t\); 2) the final BBA from the previous time frame \(\tilde{b}^{t-1}\) and 3) a context BBA \(b_c^t\) which characterizes the current context cues. For now, the context cue \(b_c^t\) only characterizes the task progress. The task progress is described by the amount of time past since the beginning of current operation, normalized by the total duration of this operation. The task progress is described by the amount of time past since the beginning of current operation (in percent of the total duration). We assume that the surgeon’s intentions of requesting an instrument are expressed in a stronger manner as time goes on. More specifically, \(b_c^t(A = \text{Requesting}) \triangleq a + \text{step} \cdot t/L\) where \(t (0 \leq t \leq L)\) is current time, \(L\) is the window length for each segment, \(a\) and \(\text{step}\) are offset and normalization constants, respectively (set to \(a = 0.4\) and \(\text{step} = 0.2\)). The complement value then defines the BBA for the other state: \(b_c^t(A = \text{Operating}) \triangleq 1 - b_c^t(A = \text{Requesting})\). As time moves on, the probability of instrument request increases linearly over time.

6 Experiments

To validate the performance of the proposed turn-taking prediction algorithm, a number of experiments were conducted. In the following, the experiment setup, associated evaluation schemes and metrics are first described (Sect. 6.1). These metrics are used to evaluate the performance of: (1) automatic feature selection (Sect. 6.2); (2) LSTM vs DTW for early prediction (Sect. 6.3); (3) DST fusion methods (Sect. 6.4); (4) algorithms vs human’s performance (Sect. 6.5).

6.1 Experiment setup and evaluation scheme

This section describes the general experiment setup and the evaluation metrics for all the following experiments. The aim of all following experiments is to evaluate the performance of the turn-taking prediction algorithm. The general evaluation scheme is machine learning driven, where the data set collected from Sect. 3 is used to train and test the proposed turn-taking prediction algorithm. The metrics used to evaluate the performances of the turn-taking prediction algorithm are also described below.

6.1.1 Experiment setup

To evaluate the generalization capability of the trained system, the following machine learning experiment setup follows a leave-one-subject-out (loso) cross validation (cv) scheme was used (Esterman et al. 2010). Under loso, in each of the \(R\) folds (here \(R = 12\) since we have 12 subjects’ data), the data from a single subject was left out as testing while the rest \(R - 1\) subjects’ data used as training. Under such scheme, the training and testing split never contains data from the same subject, so that the generalization capability of the proposed algorithm can be tested. Under such cross validation scheme, the performance of the trained algorithm when working with novel subjects’ data is evaluated. This scenario is very common when deploying trained systems into actual usage, thus loso has been commonly used when evaluating machine learning algorithms related to humans (Bartlett et al. 2003; Zheng et al. 2006; Lucey et al. 2010; Ohn-Bar and Trivedi 2014). There were a total of 846 segments of requesting state and 1305 segments of operating state. The operating states were segmented to have a duration equal to the median length of all requesting states (which is approximately 40 frames, or 2 s).

The number of participants \((R = 12)\) was determined based on power analysis. To reach a power level of 0.8 (i.e., \(\beta = 0.2\)) for differentiating DTW and LSTM under \(T E_1\) feature set, a two-sample t-test power analysis requires 56 observations for each algorithm (pooled variance, sample mean difference as the detected difference, alternative hypothesis is inequality and \(\alpha = 0.05\)). 12 subjects with 5 trials for each subject leads to 60 observations for each algorithm to reach this power requirement.

The hyper-parameters for the LSTM algorithm consisted of training iterations (100, 000), learning rate (0.001), batch size (128) and number of hidden units (32). The LSTM training and testing was obtained through the Tensorflow library (Abadi et al. 2015).

6.1.2 Evaluation scheme and metrics

To test the algorithm’s performance in early prediction instead of recognition, only the beginning fraction of the unknown action segment is used to infer its class. This can evaluate the performance of predicting the action type before the action is completely finished. The fraction percentage is characterized by the parameter \(\tau\), which is referred as
“percentage of the complete action/event” in the following content. Given an unknown segment \( X_i \in \mathbb{R}^{L_i \times m} \) of an event with window length \( L_i \), its class \( \hat{y}_i^f \) (requesting or operating) is calculated for each fraction value \( \tau \in T \) where \( T = \{10\%, 20\%, \ldots, 100\%\} \). At fraction \( \tau \), the data during \([0, \tau L_i]\) was used for testing and the resultant data set is \( \{X_i^f \in \mathbb{R}^{\tau L_i \times m}\} \).

For each fraction point \( \tau \), the resultant \( F1 \) score was calculated to evaluate the prediction performance, based on the following equations. \( N \) indicates the number of examples in the test split, \( y_i \in \{0,1\} \) is ground truth and \( \hat{y}_i^f \in \{0,1\} \) is the estimated class type at fraction value \( \tau \):

\[
\text{True Positive (TP)} = \sum_{i=1}^{N} \hat{y}_i^f y_i \\
\text{False Positive (FP)} = \sum_{i=1}^{N} \hat{y}_i^f (1 - y_i) \\
\text{True Negative (TN)} = \sum_{i=1}^{N} (1 - \hat{y}_i^f)(1 - y_i) \\
\text{False Negative (FN)} = \sum_{i=1}^{N} (1 - \hat{y}_i^f)y_i
\]

\[
\text{Precision (P)} = \frac{TP}{TP + FP} \\
\text{Recall (R)} = \frac{TP}{TP + FN} \\
\text{F1 score (F1)} = \frac{2PR}{P + R}
\]

We also defined the metrics to study how early and confident a prediction is made, based on metrics introduced in Murphy (2012). Point of First detection (PoF) measures the time (in percentage of total duration) that it takes to detect the class correctly for the first time. Point of Confident detection (PoC) indicates the first time (in percentage of total duration) that it takes to detect the class correctly for the first time.

Assume that the number of positive and negative examples are the same and always predicting \( \hat{y}_i^f = 1 \) or \( \hat{y}_i^f = 0 \), the average PoF and PoC value would be \((0.1 + 1.0)/2 = 0.55\). This forms an upper boundary (worst performance) for the evaluation of the algorithms. Notice that flipping a coin randomly would lead to PoF of 0.198, but a poor performance of PoC and \( F1 \) score.

6.2 Feature construction and selection

The aim of this experiment is to evaluate the effect of feature construction and selection on prediction performances. To that end, we compare the prediction performance and confidences with different feature sets, named below:

- \textbf{Raw:} using all \( M \) channels of raw values obtained from the sensors (i.e. using \( \hat{X}_i \in \mathbb{R}^{L_i \times M} \), in this case \( M = 50 \))
- \textbf{TE}_m: using the \( m \) most significant features selected from the statistics-based analysis on Temporally-Encoded (TE) features, i.e., using \( \hat{X}_i \in \mathbb{R}^{L_i \times m} \), and \( m = 10, 30 \) and 50.
- \textbf{BTE}_m: same as \textbf{TE}_m but using the binarized signal instead of the continuous signal as final features.

10% of action segments were randomly drawn out of the entire data set and were used for feature selection process. The remaining 90% action segments were left for training and testing purposes. An alternative approach would sample 10% data from the training split of each fold for feature selection process. But the selected top features between the two approaches were found to be almost identical (with only a few features ranked differently). The reason is that the contribution of one subject’s data out of 12 subjects’ in \( \chi^2 \) analysis is small and negligible, thus the resultant features were almost identical. For simplicity and faster training purposes, feature selection was conducted only once based on the samples drawn from the entire dataset. To evaluate the effect of feature selection on performance, the DST fusion method was not used here. The data from all the modalities was concatenated together and a single network was trained on all the data. We used LSTM as the only temporal prediction algorithm for a more controlled comparison.

The cv-averaged \( F1 \) scores are shown in Fig. 8, with arrows indicating plus one standard error. The different feature sets are represented by different curves in the figure. The solid lines represent the \textbf{TE}_m groups while the dashed lines represent the \textbf{BTE}_m groups, with the same color indicating same \( m \) value. The block-design ANOVA analysis (Kirk 1982) was conducted to find out the significance of different feature selection methods. The response is the \( F1 \) score and the two factor levels are with feature selection (i.e., \( F1 \) scores with all \textbf{TE}_m and \textbf{BTE}_m methods) and without feature selection (i.e., \( F1 \) scores using raw features). The cross-validation
To examine the performance of early prediction for each feature set, we present the average PoF and PoC for each feature set, as shown in Fig. 9. As we can see, all the TE_m and BTE_m groups have smaller PoF and PoC values than Raw, indicating a better early prediction performances. The TE_10 group achieves the best PoF and PoC scores among all the groups. As shown by Fig. 8, TE_10 can achieve the highest F1 score when little input is given (i.e., when $\tau < 0.5$), leading to an accurate early and confident detection, therefore higher PoF and PoC results are achieved.

Notice that adding more features lead to better classification scores when looking at the complete event (i.e., corresponding to the right-most points at $\tau = 100\%$ in Fig. 8). But it does not guarantee the optimal performance in early prediction (low $\tau$). One potential reason is that a simpler model (i.e., a smaller value of $m$) can generalize better to partial input, thus achieving better early prediction performances. Such behavior of achieving optimal performance when the complete action is finished, but inferior performance when little action is observed can be explained: adding extra features in the model generates a more complex model, which is more easily to over-fit and does not generalize well to partial inputs. On the contrary, the TE_10 feature set is very accurate in early prediction, and can achieve comparable performance with the whole feature set in later stages. Table 2 shows the 10 selected features in the TE_10 feature set along with their $\chi^2$ statistics. The $\chi^2$ statistics is the Pearson’s cumulative test statistics, which is calculated based on expected frequency and observed frequency. The larger the $\chi^2$ statistics is (for a fixed degree of freedom), the more correlated the feature is with the action labels. For this degree of freedom ($df = 1$), the upper-tail critical values of $\chi^2$ distribution to achieve 99.9% significance level is 10.828, and all the shown statistics greatly surpass that number, indicating a high correlation of selected features with ground truth labels.

### Table 2

| Rank | Feature name + Filter name               | $\chi^2$ statistics |
|------|------------------------------------------|---------------------|
| 1    | Epoc.gyro_y + identity                   | 1479.2              |
| 2    | Epoc.gyro_y + gabor1                     | 1456.6              |
| 3    | Epoc.gyro_y + gabor2                     | 1430.9              |
| 4    | kinect.audioConfidence + gabor1          | 1424.7              |
| 5    | kinect.audioConfidence + identity        | 1408.5              |
| 6    | kinect.audioConfidence + gabor2          | 1388.0              |
| 7    | myo.orientation_x + gabor1               | 990.3               |
| 8    | myo.orientation_x + gabor2               | 975.9               |
| 9    | myo.acceleration_y + gabor1              | 975.1               |
| 10   | myo.acceleration_y + gabor2              | 971.1               |

Notice that adding more features lead to better classification scores when looking at the complete event (i.e., corresponding to the right-most points at $\tau = 100\%$ in Fig. 8). But it does not guarantee the optimal performance in early prediction (low $\tau$). One potential reason is that a simpler model (i.e., a smaller value of $m$) can generalize better to partial input, thus achieving better early prediction performances. Such behavior of achieving optimal performance when the complete action is finished, but inferior performance when little action is observed can be explained: adding extra features in the model generates a more complex model, which is more easily to over-fit and does not generalize well to partial inputs. On the contrary, the TE_10 feature set is very accurate in early prediction, and can achieve comparable performance with the whole feature set in later stages. Table 2 shows the 10 selected features in the TE_10 feature set along with their $\chi^2$ statistics. The $\chi^2$ statistics is the Pearson’s cumulative test statistics, which is calculated based on expected frequency and observed frequency. The larger the $\chi^2$ statistics is (for a fixed degree of freedom), the more correlated the feature is with the action labels. For this degree of freedom ($df = 1$), the upper-tail critical values of $\chi^2$ distribution to achieve 99.9% significance level is 10.828, and all the shown statistics greatly surpass that number, indicating a high correlation of selected features with ground truth labels.

### 6.3 Early prediction

The goal of this experiment is to compare the performance of traditional temporal modelling algorithms (MD-DTW) with recurrent neural network algorithms (LSTM network). For comparison purposes, the DST fusion algorithm is not used. The raw feature, together with the TE_m feature sets for $m = 10, 30, 50$ are used to compare LSTM against MD-DTW. TE_10 is used.

In order to achieve early prediction, the MD-DTW and LSTM algorithm have different decision schemes. Given an
unknown segment $X_i \in \mathbb{R}^{L_i \times m}$ and an evaluation fraction $\tau$ ($0 < \tau < 1$), MD-DTW makes a decision by comparing the distances between the beginning $\tau$ fraction of the unknown segment and the template $(X^1_i, X^2_i)$, following:

$$\hat{y}_i = \arg\min_{y \in [0,1]} DTW(X_i([0, \tau L_i], X^y_i)$$

(9)

The LSTM network, alternatively, predicts the class label based on the output of memory cell at time $\tau L_i$ (instead of from the last timestamp $L_i$). The cv-averaged $F1$ scores for each algorithm is shown in Fig. 10. The solid lines represent LSTM performances and the dashed lines represent DTW performances, with the same color indicating same feature set. As shown by the image, each LSTM curve outperforms the corresponding DTW curve at each percentage point. The block-design ANOVA analysis was conducted. The response is $F1$ score and the two factor levels are using LSTM and using DTW. The feature sets (i.e., Feature Set={$\text{Raw}, \text{TE}_10, \text{TE}_30, \text{TE}_50$}), cross-validation fold index and fraction number are all treated as random factors. The resultant ANOVA test shows $p < 0.001$ and the Tukey’s post-hoc test shows a significant difference between LSTM and DTW ($\alpha = 0.05$).

The PoF and PoC are also better when using LSTM compared to using the corresponding DTW algorithms. Taking TE_10 as an example, the average PoF is 22.24 for DTW and 18.84 for LSTM, and the average PoC is 32.83 for DTW and 28.03 for LSTM. TE_10 is used_10 is used_10 is used_10. Therefore, it can be concluded that LSTM outperforms DTW in all metrics, including $F1$ score, PoF and PoC. These findings support the superiority of the LSTM algorithm against a more traditional temporal modelling algorithm, namely DTW.

6.4 DST fusion performance

In this experiment we evaluated the effect of adding context cues and previous decisions using the DRC fusion algorithm. The baseline condition only uses multimodal signals without any contextual cues or historic information. As comparison, the contextual cues and previous decisions are added to the multimodal signals using the DRC fusion technique, and their performances are evaluated.

Next we describe several fusion configurations which were evaluated in this experiment. The first configuration ($\text{TE}_10$) only fuses decisions from current multimodal signals ($b_e^i, b_t^i$ and $b_m^i$). The second configuration ($\text{TE}_10$ + context) adds the contextual information $b_c^i$ to all the multimodal inputs. The third configuration ($\text{TE}_10$ + prev) adds the previous decision $\tilde{y}_i^{-1}$ to all the multimodal inputs. And last, the fourth configuration ($\text{TE}_10$ + prev + context) includes multimodal signal, contextual cues and previous decisions. For all configurations in the DRC fusion process, the same weights were assigned to all the input decisions.

Figure 11 shows the cv-averaged $F1$ score for all 4 different configurations. The block-design ANOVA analysis was conducted, where response is $F1$ score and four factor levels are the four different DRC configurations. Cross-validation fold index and fraction number are treated as blocks same as before. The resultant ANOVA test shows $p = 0.005 < 0.01$ and the Tukey’s post-hoc test shows that $\text{TE}_10$ + context + prev is significantly different from $\text{TE}_10$ and $\text{TE}_10$ + prev group ($\alpha = 0.05$). This indicates that the complete fusion set ($\text{TE}_10$ + prev + context) achieves the best prediction performance. It is slightly outperformed by ($\text{TE}_10$ + context) with a small margin in $\tau \in [0.4, 0.7]$ ranges, but it is still the best combination due to its superior detection performance.
in the beginning fraction (τ = 0.1) and full fraction of the event (τ = 1.0). When comparing the best performed set (TE_10 + prev + context) against using only TE_10, there is an average F1 score increment of 2.5%.

6.5 Human baseline comparison

We compared the performance of the proposed early prediction algorithm against human baseline. To that end, we recruited the same participants who took part in the data collection process (Sect. 3). These participants all repeated the surgical procedure five times in order to reach a plateau in the learning curve. The participants observed recorded surgery videos. Based on those observations they had to answer whether the surgeon wanted an instrument or not at different time periods. The illustration of the process is shown in Fig. 12. These answers were used to compare human predictions against machine predictions. The video recording and replay approach is resorted to avoid interruption to the original surgery flow. It also allows evaluation of human reactions at different event fractions, since the exact duration of each turn-taking event was already spotted beforehand. The recorded videos from the human turn-taking experiment (Sect. 3) were used instead of videos from real operations so that a more controlled experiment can be conducted and a fairer comparison can be made.

Recorded videos of surgery were played to each participant under a cross-participant setting, i.e. every participant watched other’s videos. The video was paused at random times, and participants were asked to mark whether the surgeon intended to request an instrument or not. Those answers were recorded and then compared with the ground truth. The evaluation metric (F1 score) was then calculated as human baseline performance. To determine how the video would be paused, the following procedure was followed. The video was paused within each requesting and operating state. Assume that a video clip contains a turn-taking event starting at \( t_s \) and ending at \( t_e \), this video would be paused at times \( t_a (t_s < t_a \leq t_e) \), according to a discrete uniform distribution \( t_a \sim \mathcal{U}(t_s, t_e) \).

We selected the best LSTM algorithm configuration determined from the previous experiments and compared it against the human baseline. The chosen LSTM used \( TE_10 \) feature set and the full DST fusion scheme (TE_10 + prev + context). The performance of the human baseline, compared with the best LSTM algorithms is shown in Fig. 13. The LSTM algorithm outperforms the human baseline for \( \tau \in [0, 0.3] \), and is worse than human baseline afterwards with a margin of around 6%. The median length of the entire action is about 2 s. Therefore the proposed algorithm can deliver better prediction performance in early stages of the action (about 0.6 s after the action starts, and 1.4 s before the action is completed), and then is outperformed by human baseline with a margin of about 6% achieve comparable performance to the human baseline. Such behavior grants this type of predictor the name of “early”.

7 Discussion

Regarding the results of feature selection (Sect. 6.2), it has been found that higher F1 scores are achieved when using the feature selection method compared to using the raw features. This indicates the effectiveness of the proposed feature selection process.

There are some observations about different encoding methods (\( TE_m \) vs \( BTE_m \)). Binary encoding (i.e., \( BTE_m \)) degrades the performance compared to using multi-level encoding (i.e., \( TE_m \)). This is due to the potential loss of information during binarization. Another observation is regarding different feature dimensions (i.e., value of \( m \)). When testing complete actions (i.e., \( \tau = 1.0 \)), the performance was found to increase as feature dimension increases.
However, when testing on partial inputs (i.e., small \( \tau \)), fewer features actually perform better. This is potentially because a simpler model (i.e., fewer interconnected nodes in the network) can generalize better to partial observations. Therefore, a trade-off exists here between having expressive models (a large \( m \)) to achieve high performances with complete actions, and good generalization capability to partial input (a small \( m \)). The best performance is achieved by balancing between these two factors, and among all the \( m \) values tested, an optimal value of 10 was found. More experiments can be conducted to evaluate the optimality of \( m \) values in a larger range with a more refined resolution (Mladenović and Hansen 1997).

Careful analysis of Table 2 can shed some light on the importance of different features when predicting turn-taking. According to the table, the most significant features are mainly motion-based (encoding head, face and arm movement) and audio cues. Similar findings have been reported by Saito et al. (2015); De Kok and Heylen (2009). Notice that surgeons in the OR have different preferences in the usage of head motions for instrument requests. Some prefer to turn their heads towards the nurse, while others prefer to focus on the operation area. Both cases are contemplated with the algorithm because of the cross-subject design.

Regarding the results about comparison between DTW and LSTM algorithm (Sect. 6.3), the LSTM network is found to outperform DTW in all metrics, including F1 score, PoF and PoC. The LSTM is one type of deep learning algorithm, and can deal with large amounts of data effectively, as has been seen in handwriting recognition (Graves et al. 2009), emotion recognition (Wllmer et al. 2013) and gesture recognition fields (How et al. 2014). LSTM can model timed decision-making processes as is conducted by humans, and therefore we expect to see an analogous performance in turn-taking events. DTW was selected as the benchmark algorithm since it is still actively used in early gesture recognition (Izuta et al. 2015) and multivariate sequence classifications (Mei et al. 2016). Moreover, empirical studies have shown that DTW-based classifiers can perform at least equal (and generally better) than other classification algorithms across dozens of data sets (Shokoohi-Yekta et al. 2016).

Regarding the results about DRC fusion (Sect. 6.4), it was found that adding more information channels and combining them in a rational manner (i.e., DRC) helps increasing prediction performances. Currently we have used task progress as the only context cue, including additional context information and more sophisticated sensor fusion approaches (e.g. weighted spatial-temporal DRC fusion (Li et al. 2013)) have the potential to further improve prediction performances.

When comparing performances of the proposed algorithm vs human baseline (Sect. 6.5), we have found that the proposed algorithm achieves higher F1 scores (12% in average) than human in the beginning stage of prediction (\( \tau \in [0, 0.3] \)). In later stages (\( \tau \in [0.4, 1] \)), we observed a higher F1 score (6% in average) in humans than the proposed algorithm. Thus, the algorithm’s relative strength when compared with humans occurs when only partial observation is given and early prediction results are preferred. The proposed algorithm can clearly predict the incoming turn-taking intention much earlier than humans do, thus providing early prediction capability instead of just classification.

Putting the findings of this paper in the context of the related work yields some observations about feature importance and turn-taking recognition performance. Regarding feature importance for turn-taking understanding, we have found that multimodal non-verbal signals (head, body and eye motions) are very salient in predicting turn-taking types. Similar results were found by Saito et al. (2015) (using gaze and head motion) and De Kok and Heylen (2009) (using body postures and eye movements) in the context of conversational turn-taking recognition. Regarding turn-taking recognition performance, we have reached a F1 score range of [0.7, 0.9], depending on the prediction ranges. Saito et al. 2015 reported a similar F1 score of 0.87 with linear SVM for speaker’s turn-taking recognition. Heeman and Lunsford (2015) reported human’s performance in predicting speaker’s turn switches and found a F1 score of 0.61. Gravano and Hirschberg 2011 reported results in similar ranges for the same type of task, but instead using machine learning algorithms and utterance features. De Kok and Heylen (2009) studied the multi-party turn-taking prediction problem with Conditional Random Fields, but only achieved a F1 score of 0.061. The F1 scores mentioned above are acquired under different tasks (e.g., conversational turn-taking, multi-party meetings etc) and thus cannot be directly compared to ours, but they can still show the general level of turn-taking recognition performances. In this scenario, since the number of negative instances is 1.5 times larger than that of the positive instances, the F1 score is actually deflated by a small amount (Jeni et al. 2013). Such data skewness issue won’t bias the comparison of different feature/algorithm/fusion methods, since they all used the same unbalanced dataset and thus the F1 score measurement is fair to all the configurations.

To summarize our findings in the context of related work, we have reached similar conclusions regarding the importance of multimodal signals for turn-taking comprehension, and achieved high performance for turn-taking prediction. In addition, we addressed the early prediction problem and presented results related to early predictability.

There are some limitations of the proposed turn-taking algorithm. Currently, the proposed early turn-taking algorithm generates a decision based only on the information included in an isolated and segmented event, independently from the other nearby events. Conversely, humans continuously gain context information about the surgical task, such as task progress, timing of previous instrument requests.
and emergency levels. Based on the continuous information acquisition and building, a decision is made. All such context cues contribute to the identification of turn-taking events. Future work will focus on incorporating more contextual cues in a continuous manner, mimicking more of the human behavior.

8 Conclusion

Early turn-taking prediction is critical for natural human–robot collaborations, since it can provide more time for robot to plan and execute its actions. Such prediction capability would induce a more fluent and smooth turn-taking transition. This paper discusses an approach consisting of an early turn-taking prediction algorithm for human robot collaboration using multimodal signals. Specifically, the design and implementation of a turn-taking prediction algorithm for a robotic scrub nurse system is presented. The proposed algorithm is evaluated on a simulated surgical procedure data set, and it is found to be superior to its algorithmic counterparts, and is better than the human baseline when little partial input is given (less than 30% of full action). When given only 30% of full action (0.6 s after the action starts and 1.4 s before the action finishes), our algorithm achieves a F1 score of 0.80. After the beginning stage, our algorithm can achieve comparable performances with humans as the action progresses, reaching a F1 score of 0.90 when observing the entire action. Such performances justify that the proposed early turn-taking prediction algorithm can reach high prediction performances early in the turn-taking process, granting robots more time to plan and execute their motions. Specifically, when the robot serves as an assistant to the surgeon, planning time is required to figure out where the surgical instruments should be placed for prompt delivery. When extending these concepts to other scenarios, early turn-taking prediction can also shed light on the design of other proactive and anticipatory robotic systems.

Future work includes proposing a more comprehensive human state definition, giving a richer spatial-temporal feature construction method and including more contextual information to improve the early prediction algorithm. The runtime dynamic feature selection capability is also going to be investigated. Methods on how to integrate the turn-taking prediction results with the proactive robot motion planning will also be explored. We also plan to validate the proposed turn-taking prediction algorithm with clinical faculty in the operating room.

Acknowledgements The authors would like to thank Dr. Rashid Mazhar and Dr. Carlos Velasquez from Hamad Medical Corporation (Qatar) for their collaboration and discussion of the project. That authors would also like to thank all the members of ISAT lab for their inspiring discussions.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org (Vol. 1).

Abdulla, W. H., Chow, D., Sin, G. (2003). Cross-words reference template for DTW-based speech recognition systems. In Conference on convergent technologies for the Asia-Pacific region TENCON 2003 (Vol. 4, pp. 1576–1579). IEEE. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1273186.

Andersen, D., Popescu, V., Cabrera, M. E., Shahghavi, A., Gomez, G., Marley, S., et al. (2016). Virtual annotations of the surgical field through an augmented reality transparent display. The Visual Computer, 32(11), 1481–1498.

Arsikere, H., Shirberg, E., Ozertem, U. (2015). Enhanced end-of-turn detection for speech to a personal assistant. In 2015 AAAI spring symposium series, March 2015. Available: https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10256.

Arsikere, H., Shirberg, E., Ozertem, U. (2015). Enhanced end-of-turn detection for speech to a personal assistant. In 2015 AAAI Spring symposium series.

Bartlett, M. S., Littlewort, G., Fasel, I., Movellan, J. R. (2003). Real time face detection and facial expression recognition: Development and applications to human computer interaction. In Conference on computer vision and pattern recognition workshop, 2003. CVPRW’03 (Vol. 5, pp. 53–53). IEEE.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

Buerhaus, P. I., Auerbach, D. I., & Staiger, D. O. (2009). The recent surge in nurse employment: Causes and implications. Health Affairs, 28(4), w657–w668.

Cakmak, M., Srinivasa, S. S., Lee, M. K., Kiesler, S., Forlizzi, J. (2011). Using spatial and temporal contrast for fluent robot–human hand-overs. In Proceedings of the 6th international conference on human–robot interaction (pp. 489–496). ACM.

Calisgan, E., Haddadi, A., Van der Loos, H. M., Alcazar, J. A., Croft, E. A. (2012). Identifying nonverbal cues for automated human-robot turn-taking. In Robot and human interactive communication (ROMAN 2012). 21st IEEE International Symposium (pp. 418–423). IEEE.

Canny, J. (1988). The complexity of robot motion planning. MIT press. Available: https://books.google.com/books?hl=en&lr=&id=VRM_sczrKgC&oi=fnd&pg=PR11&dq=robot+motion+planning+&ots=ZGk1puFOs&sig=3rJFFS7FYe9ROsnV4mBAX2bPFs.

Chan, W. P., Kakiuchi, Y., Okada, K., Inaba, M. (2014). Determining proper grasp configurations for handovers through observation of object movement patterns and inter-object interactions during usage. In 2014 IEEE/RSJ international conference on intelligent robots and systems (IROS 2014) (pp. 1355–1360). IEEE.

Chao, C., Thomaz, A. (2012). Timed petri nets for multimodal interaction modeling. In ICMI 2012 workshop on speech and gesture production in virtually and physically embodied conversational agents. Available: https://robotics.usc.edu/~icmi/2012/docs/2012ChaoThomaz_ICMI-WS1.pdf.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. arXiv:1406.1078 [cs, stat], June 2014. arXiv:1406.1078. Available: https://arxiv.org/abs/1406.1078.

Cutler, A., Pearson, M. (1986). On the analysis of prosodic turn-taking cues. Intonation in Discourse (pp. 139–156). Available: https://pubman.mpdl.mpg.de/pubman/item/escidoc:76883:7/component/escidoc:506929/Cutler_1985_On. Springs
Auton Robot (2018) 42:977–995

De Kok, J., Heylen, D. (2009). Multimodal end-of-turn prediction in multi-party meetings. In Proceedings of the 2009 international conference on Multimodal interfaces (pp. 91–98). ACM. Available: https://dl.acm.org/citation.cfm?id=1647332.

Dumas, B., Ingold, R., Lalanne, D. (2009). Benchmarking fusion engines of multimodal interactive systems. In Proceedings of the 2009 international conference on multimodal interfaces (pp. 169–176). ACM. Available: https://dl.acm.org/citation.cfm?id=1647345.

Ehrlich, S., Wykowska, A., Ramirez-Amaro, K., Cheng, G. (2014). When to engage in interaction #x2014; And how? EEG-based enhancement of robot’s ability to sense social signals in HRI. In 2014 14th IEEE-RAS international conference on humanoid robots (humanoids), November 2014, pp. 1104–1109.

Escalante, H. J., Morales, E. F., Sucar, L. E. (2016). A naive Bayes baseline for early gesture recognition. Pattern Recognition Letters, 73, 91–99. Available: https://www.sciencedirect.com/science/article/pii/S0167865516000258.

Esterman, M., Tamber-Rosenau, B. J., Chiu, Y.-C., & Yantis, S. (2010). Avoiding non-independence in fmri data analysis: Leave one subject out. Neuroimage, 50(2), 572–576.

Gravano, A., Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. Computer Speech & Language, 25(3), 601–634. Available: https://www.sciencedirect.com/science/article/pii/S0885230810000690.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwritten recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(5), 855–868. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4851750.

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., Schmidhuber, J. (2015). LSTM: A search space odyssey. arXiv:1503.04069 [cs], March 2015, arXiv:1503.04069. Available: http://arxiv.org/abs/1503.04069.

Gulov, I., Gnerov, L., Breza, J. Communication in the operating room. Available: https://www.szhorizont.eu/home/archiv/roc122014/GUL%C3%81%5%AOOV%C3%81%5%20%20%20%20%20%20al.%20Communication%20in%20the%20operating%20room.pdf.

Guntakandla, N., Nielsen, R. D. (2015). Modelling turn-taking in human conversations. In 2015 AAAI spring symposium series. Available: https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10313.

Harmanec, D., & Klir, G. J. (1994). Measuring total uncertainty in Dempster–Shafer theory: A novel approach. International Journal of General System, 22(4), 405–419.

Hart, J. W., Gleeson, B., Pan, M., Moon, A., MacLean, K., Croft, E. (2014). Gesture, gaze, touch, and hesitation: Timing cues for collaborative work. Available: https://milab.idc.ac.il/timingham/wp-content/uploads/2014/02/Hart-TimingHRI-2014.pdf.

Heeman, P., Lunsford, R. (2015). Can overhearers predict who will speak next? In 2015 AAAI spring symposium series, March 2015. Available: https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10269.

Heger, D., Putze, F., & Schultz, T. (2011). An EEG adaptive information system for an empathic robot. International Journal of Social Robotics, 3(4), 415–425. https://doi.org/10.1007/s12369-011-0107-x.

Hinton, G. (2010). A practical guide to training restricted Boltzmann machines. Momentum, 9(1), 926.

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6795963.

How, D. N. T., Sahari, K. S. M., Yuhuang, H., Kiong, L. C. (2014). Multiple sequence behavior recognition on humanoid robot using long short-term memory (LSTM). In 2014 IEEE international symposium on robotics and manufacturing automation (ROMA) (pp. 109–114). IEEE. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7295871.

Hughes, K. F., Murphy, R. R. (1992). Ultrasonic robot localization using Dempster–Shafer theory. In San Diego ’92 (pp. 2–11). International Society for Optics and Photonics.

Izuta, R., Murao, K., Terada, T., Tsuchamoto, M. (2015). Early gesture recognition method with an accelerometer. International Journal of Pervasive Computing and Communications, 11(3), 270–287. Available: https://www.emeraldinsight.com/doi/abs/10.1108/IJPCC-03-2015-0016.

Jacob, M. G., Li, Y.-T., Wachs, J. P. (2012). Gestonurse: A multimodal robotic scrub nurse. In Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction (pp. 153–154). ACM. Available: https://dl.acm.org/citation.cfm?id=2157731.

Jeni, L. A., Cohn, J. F., De La Torre, F. (2013). Facing imbalanced data recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction (ACCI) (pp. 245–251). IEEE.

Kingma, D., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980 [cs], December 2014, arXiv:1412.6980. Available: https://arxiv.org/abs/1412.6980.

Kirk, R. E. (1982). Experimental design. New York: Wiley.

Kilr, G., & Yuan, B. (1995). Fuzzy sets and fuzzy logic (Vol. 4). New Jersey: Prentice Hall.

Kose-Bacig, H., Dautenhahn, K., Nehaniv, C. L. (2008). Emergent dynamics of turn-taking interaction in drumming games with a humanoid robot. In RO-MAN 2008—The 17th IEEE international symposium on robot and human interactive communication, August 2008, pp. 346–353.

Li, X., Dick, A., Shen, C., Zhang, Z., van den Hengel, A., Wang, H. (2013). Visual tracking with spatio-temporal Dempster–Shafer information fusion. IEEE Transactions on Image Processing, 22(8), 3028–3040. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6482637.

Lucus, J. M., Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. Technometrics, 32(1), 1–12. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1990.10484583.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010). The extended Cohn–Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW) (pp. 94–101). IEEE.

MacKenzie, L., Ibbotson, J. A., Cao, C. G. L., & Lomax, A. J. (2001). Avoiding non-independence in fmri data analysis: Leave one subject out. Neuroimage, 109–114. IEEE. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7295871.

Matsuyama, Y., Kobayashi, T. (2015). Towards a computational model of multi-party meetings. In Proceedings of the 2009 international conference on multimodal interfaces (humanoids), November 2014, pp. 1104–1109.

McKeever, S., Ye, J., Coyle, L., Bleakley, C., Dobson, S. (2010). Activity recognition using temporal evidence theory. Journal of Ambient Intelligence and Smart Environments, 2(3), 253–269. Available: https://content.iopress.com/articles/journal-of-ambient-intelligence-and-smart-environments/aio701.
Mei, J., Liu, M., Wang, Y. F., & Gao, H. (2016). Learning a mahalanobis distance-based dynamic time warping measure for multivariate time series classification. *IEEE Transactions on Cybernetics*, 46(6), 1363–1374.

Mittal, A., Zisserman, A., Torr, P. H. (2011). Hand detection using multiple proposals. In *BMVC*. CiteSeer (pp. 1–11). Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.3602&rep=rep1&type=pdf.

Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, 24(11), 1097–1100.

Moore, A., Butt, D., Ellis-Clarke, J., Cartmill, J. (2010). Linguistic analysis of verbal and non-verbal communication in the operating room. *ANZ Journal of Surgery*, 80(12), 925–929. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1445-2197.2010.05531.x/full.

Morency, L.-P., de Kok, I., Gratch, J. (2008). Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *Proceedings of the 10th international conference on multimodal interfaces* (pp. 181–188). ACM. Available: https://dl.acm.org/citation.cfm?id=1452426.

Mori, A., Uchida, S., Kurazume, R., Taniguchi, R.-I., Hasegawa, T., Sakoe, H. (2006). Early recognition and prediction of gestures. In *18th International conference on pattern recognition (ICPR’06)* (Vol. 3, pp. 560–563). IEEE. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1699588.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press. Available: https://books.google.com/books?hl=en&lr=&id=RC43AgAAQBAJ&oi=fnd&pg=PR7&dq=machine+learning:+a+probabilistic+perspective&ots=ukmzEFTu-adj&sig=IE9AmDT0EVdltWhuq_ZKTX7LA.

Muñoz, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N. (2009). Footing in human–robot conversations: How robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on human robot interaction* (pp. 61–68). ACM. Available: https://dl.acm.org/citation.cfm?id=1514109.

Needelman, J., Buershan, P., Pankritz, V. S., Leibson, C. L., Stevens, S. R., & Harris, M. (2011). Nurse staffing and inpatient hospital mortality. *New England Journal of Medicine*, 364(11), 1037–1045.

Ohn-Bar, E., & Trivedi, M. M. (2014). Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 15(6), 2368–2377.

Padilha, E., Carletta, J. (2003). Nonverbal behaviours improving a simulation of small group discussion. In *Proceedings of the first international Nordic symposium of multi-modal communication* (pp. 93–105). CiteSeer. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.1326&rep=rep1&type=pdf.

Rabli, L. I., Andersen, M. L., Stgaard, D., Bjrn, B., Lilja, B., Mogensen, T. (2011). Republished error management: Descriptions of verbal communication errors between staff. *An analysis of 84 root cause analysis-reports from Danish hospitals. Postgraduate Medical Journal*, 87(1033), 783–789. Available: https://pmj.bmj.com/content/87/1033/783.

Raux, A., Eskenazi, M. (2008). Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of the 9th SIGdial workshop on discourse and dialogue* (pp. 1–10). Association for Computational Linguistics. Available: https://dl.acm.org/citation.cfm?id=1622066.

Raux, A., & Eskenazi, M. (2012). Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing*, 9(1), 1:1–1:23. https://doi.org/10.1145/2168748.2168749.

Sacks, H., Schegloff, E. A., Jefferson, G. (1974). A simplest systems for the organization of turn-taking for conversation. *Language* (pp. 696–735). Available: https://www.jstor.org/stable/412243.
Zhou, T., Wachs, J. (2016). Early turn-taking prediction in the operating room. Available: https://www.aaai.org/ocs/index.php/FSS/FSS16/paper/view/14074.

Tian Zhou is currently a Ph.D. student at the School of Industrial Engineering at Purdue University, working in Intelligent Systems and Assistive Technologies (ISAT) Lab. He received his M.S. in Electrical and Computer Engineering at Purdue University in 2016, and B.S. in Information Science and Engineering at Southeast University (Nanjing, China) in 2013. His research focuses on gesture recognition, machine vision, robot teleoperation and human–robot collaboration. He is the recipient of Robotics Fellowship of 2016 AAAI, and best presentation award of 2015 AAAI poster session.

Juan Pablo Wachs is an Associate Professor in the Industrial Engineering School at Purdue University. He is the director of the Intelligent Systems and Assistive Technologies (ISAT) Lab at Purdue, and he is affiliated with the Regenstrief Center for Healthcare Engineering. He completed postdoctoral training at the Naval Postgraduate School’s MOVES Institute under a National Research Council Fellowship from the National Academies of Sciences. Dr. Wachs received his B.Ed.Tech. in Electrical Education in ORT Academic College, at the Hebrew University of Jerusalem campus. His M.Sc. and Ph.D. in Industrial Engineering and Management from the Ben-Gurion University of the Negev, Israel. He is the recipient of the 2013 Air Force Young Investigator Award, and the 2015 Helsley Senior Scientist Fellow. Currently he is working at UBA as part of his 2016 Fulbright U.S. Scholar Program to Argentina. He is also the associate editor of IEEE Transactions in Human-Machine Systems, Frontiers in Robotics and AI, and the Journal of Real-Time Image Processing.