Early Turn-taking Prediction in the Operating Room

Tian Zhou and Juan P. Wachs
School of Industrial Engineering, Purdue University. West Lafayette, IN
{zhou338, jpwachs}@purdue.edu

Abstract

This work presents the design and implementation of an early turn-taking prediction algorithm for a robotic scrub nurse system. The turn-taking prediction algorithm analyzes surgeon's implicit communication cues identifying among those surgical instrument requests before the request actually are explicitly evoked. Communication channels expressed through signals like EEG, EMG and physical signs were used to monitor surgeon's behaviors and automatically detect implicit instrument requests. Significant features were extracted from those signals, through an automatic feature selection process. Then recurrent neural networks were used for time-sensitive turn-taking prediction. Experimental results indicated that the proposed algorithm has higher prediction accuracies than human baseline when less than 70% of the entire action was observed. This is approximately 1.4 seconds after the action started, and 0.6 seconds before the action ends. At an extremely early stage (only 10% of data), the proposed turn-taking prediction algorithm achieves a F1 score of 82.8%.

Introduction

Surgery involves complex coordinated behaviors that are learnt, acquired and executed precisely in the operating room (OR) through experience, implicit and explicit communication. In order to include a robotic assistant into this context, key components involving communication, task-flow and turn-taking must be addressed accurately in such hybrid human-robot team. One especially challenging aspect is the ability to predict the partner's intention. An example of this challenge is surgical instruments handover task. A surgical nurse delivers surgical instruments to a surgeon based on explicit requests – the surgeon uttering the words "scissors" or/and implicit requests expressed by body language (e.g. changing posture, hand gestures, gaze, and head/neck movements). All these forms of expressions are used to inform the task partner that it is her turn to continue with a task. In such scenario, the main challenge is to integrate the communication channels meaningfully to distinguish signals meant to convey intention. To this end, optical, physiological and neurological information are integrated together to characterize the operator's intent. In this work, we focus on developing aspects related to the recognition of intent for timing and synchronization purposes rather than identifying the specific instrument requested which was subject to our previous work (Jacob, Li, & Wachs, 2012).

Collaborative work in hybrid teams of humans and robots is an area that is gaining major interest especially when it concerns time sensitive and cognitive demanding tasks. The concept of having robotic assistants to work along with human operators effectively is compelling and has the potential to reduce costs and time requirements. In this environment, how to coordinate the timing during a paired collaborative task is key for effective work. Haptic communication was used to predict intention and role of the participants in a joint object manipulation task by Groten (Groten, Feth, Klatzky, & Peer, 2013). In addition, Ehrlich (Ehrlich, Wykowska, Ramirez-Amaro, & Cheng, 2014) enabled a humanoid to determine the right timing and proper role to engage in interaction with its operators by means of gathering their EEG signals. In the CHARM project (Hart et al., n.d.), a robot assistant was developed to work alongside human workers in a manufacturing environment, where nonverbal cues were used for timing coordination. Timing in multimodal (speech, gaze, gesture) turn-taking interactions in human-robot interaction were considered in a collaborative Towers of Hanoi game by Chao (Chao & Thomaz, 2012). All the described work consider tasks that are well structured, demand collaboration, but are not time-sensitive. That is, the outcome of the task does not depend on the specific timing as long as the turns are respected. To the best of our knowledge, this is the first time that timing in multimodal human robot interaction is studied in the context of a high-sensitive and high-risk task such as surgery. An illustration of the system is shown in Figure 1.

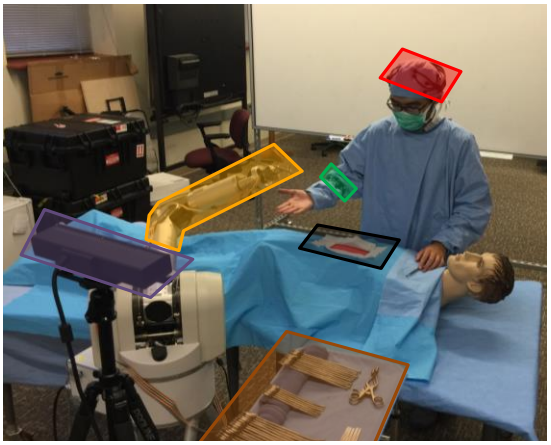

Figure 1. System setup for the robotic scrub nurse. The surgeon is conducting a surgery (black) when robotic nurse (orange) picks up the instrument from mayo stand (brown) and delivers to surgeon after requesting. The surgeon is monitored by Myo armband (green), Epoc headband (red) and Kinect (purple).

Turn-taking in Homogenous Human's Team

In this section we will describe the observation and recording of human teams' turn-taking activities during a surgical task. The associated experiment setup and data collection process will also be discussed below.

Task setup

To better understand how turn-taking activities were regulated between *surgeons* and *nurses*, we setup a simulation platform for surgical operations and recorded communication cues between *nurses* and *surgeons*. The team needs to collaborate in order to conduct the mock abdominal incision and closure task successfully. The detailed process of the surgical task is shown in Figure 2.

The human participant acting as *surgeon* was given instructions about the surgical procedure, and then was required to conduct the task on the simulation platform. The *surgeon* was explicitly asked to use verbal commands to request each instrument. In addition to the verbal commands, the *surgeon*'s body, gaze and arm motions were all used together as implicit communication cues to trigger the *nurse*'s actions (no explicit request was given to the subjects about this). The subject acting as the *nurse* must understand the *surgeon's* turn-taking communication cues (both implicit and explicit) in order to react according to the surgeon's expectations. One experiment assistant performed the role of the scrub *nurse*.

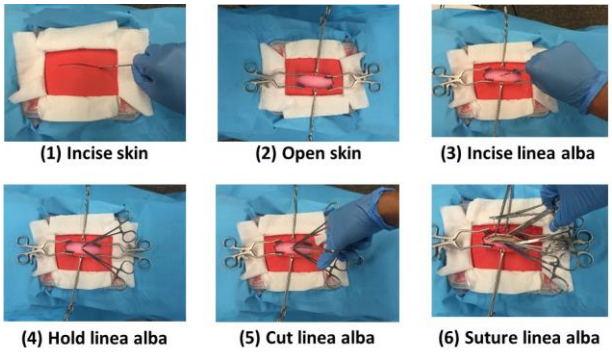

Figure 2. Mock Abdominal Incision and Closure Procedure

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 the task, resulting in around 14 turn-taking instances. These instruments are scalpel, hemostat, forceps, retractor, scissors and needle. Each participant acting as a *surgeon* repeats the surgical task 5 times. In total 5 participants were recruited, with ages in range 20-31 (mean = 24.8, std = 4.1). In total 348 turn-taking instances were created and served as the basic experiment dataset.

Data collection

The communication cues emitted by the *surgeon* were recorded for further analysis. Three different sensors were used together to record different aspects of communication cues, namely Myo armband, Epoc headband and Kinect sensor. The details of the recorded raw data for each sensor were given below, as well as the dimension of each signal.

Myo armband

Myo armband was used to capture the motion and EMG signals on the surgeon's dominant arm. The following information were recorded together with the data dimensionality:

- Orientation, 3D
- Acceleration, 3D
- Gyroscope, 3D
- EMG signals, 8D

Epoc headband

Epoc headband was used to capture surgeon's head motions and EEG signals. The following data was recorded:

- EEG signals, 14D
- Head gyro (pitch and yaw), 2D
- Emotion prediction (engagement, frustration, meditation, excitement and valence), 5D

Kinect

Kinect was used to capture head poses and body postures of the surgeon. The following information were recorded:

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

Signal Fusion

The real-time data from all three modalities were synchronized at a frame rate of 20 Hz. The raw values were concatenated together, forming a data level fusion. Unlike decision-level fusion in which each modality is classified individually, and then combined to convey a single outcome, fusion at the data level integrates all the information from the get-go and a single classification procedure is conducted. Such approach retains the low-level interactions between modalities and leads to models of higher expressiveness (Martínez & Yannakakis, 2014). For each time frame t, the fused sensor measurement r_t consists of 50 values (addition of all the dimension detailed in the previous subsection).

Surgeon's state annotation

To establish ground truth, we annotated the recorded video based on different states of the *surgeon*. The two defined states of *surgeon* are (1) *operating*: engaged in the on-going surgical task; (2) *requesting*: expecting a new instrument. All the turn-taking activities happened during the *requesting* state, thus the goal is to predict as early as possible the transition from *operating* state to *requesting* state.

The *requesting* state is defined to begin at the earliest of the following events, and to end at the latest of the same events, as illustrated by Figure 3:

- Torso movement (t_{torso}). Body inclination was identified as one of the key communication cues in the OR (Moore, Butt, Ellis-Clarke, & Cartmill, 2010)
- Gaze shift (t_{gaze}). Gaze patterns were found to have high correlation with instrument handovers in the OR (MacKenzie, Ibbotson, Cao, & Lomax, 2001)
- Arm movement (t_{arm}). (Strabala et al., 2013) analyzed human-human handovers and observed a preparatory arm movement for triggering the timing of turn-taking.
- Speech command (t_{speech}). Though considered as the major source of communication errors in the OR, verbal command in still one of the most popular communication channels in the OR (Rabøl et al., 2011).
- Hand gestures (t_{hand}). Often hand gestures are used extensively in the OR to request certain type of instrument (Gulášová, Görnerová, Breza jr, & Breza,).

The key moments of the *requesting* state were annotated, and the remaining data between two consecutive *requesting* states was considered as *operating* state. A primary researcher segmented all the videos and labeled the *requesting* segments. An additional assistant labeled 20% of randomly picked segments (from both *requesting* and *operating* states) from all 5 subjects. Inter-rater reliability showed almost perfect agreement between the two sets of annotations with regard to the segmented states (Cohen's $\kappa = 0.95$) (Cohen, 1960).

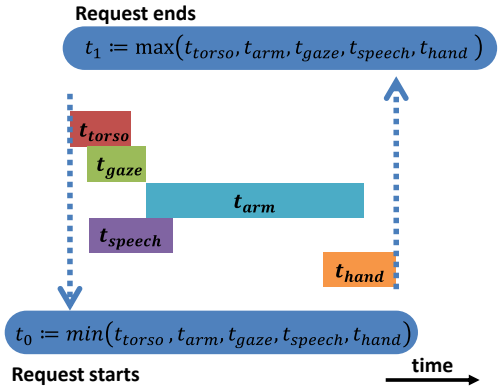

Figure 3. Illustration of the definition of request state. t_* indicates the time period when modality $*$ is active

Early Turn-taking Prediction

The turn-taking prediction framework is shown in Figure 4. The surgeon was monitored through three sensing devices. After data-level fusion and sampling, the recorded data was synchronized and concatenated. The most relevant features were retained through an automatic feature selection process. Then the features were temporally modelled for turn-taking prediction. The prediction results was transferred to a robotic arm, which picked up and delivered the surgical instrument to the surgeon accordingly.

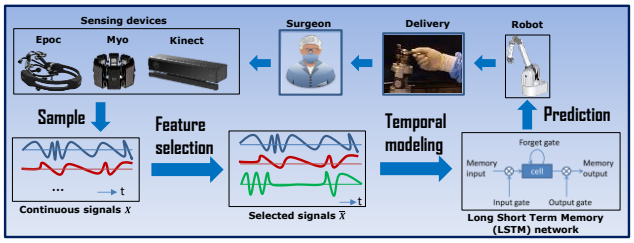

Figure 4. Turn-taking prediction algorithm framework

Channel preprocessing

The Exponentially Weighted Moving Average (EWMA) technique was applied on the raw signal for smoothing purposes. It is a common noise reduction technique for time-series data (Lucas & Saccucci, 1990), and is given by:

$$s_t = \alpha r_t + (1-\alpha)s_{t-1}, \quad t \in [1, L]; \quad s_0 = r_0$$

where r_t is the raw sensor measurement at time t, L is the length of raw signal and s_t is the filtered measurement at time t. The weighting parameter α controls the weight of raw measurement data, which was determined to be 0.2 for best performance in our environment.

The smoothed signal s_t in each modality was normalized using mean and variance values, following:

$$x_t = \frac{1}{\sigma^2}(s_t - \mu), t \in [0, L]$$

where μ and σ^2 are the mean and variance of signal s_t, and x_t is the corresponding normalized signal. Notice that μ and σ^2 were calculated based on the data from both states (*requesting* and *operating*). This step cancels any offset between data of different states. Thus, for each time stamp t, x_t consists of 50 values which is the total dimension of the raw information. The smoothed and normalized data x_t within a time window is cumulatively denoted as a segment. Each segment i, $\{x_t \mid t \in [0, L]\}$ was stacked together to form a raw feature representation $X_i \in \mathbb{R}^{L_i \times M}$, where L_i is the length of segment i and M is the dimension of raw data (i.e., $M = 50$). For each segment X_i, there is a corresponding label $y_i \in \{0,1\}$ to indicate whether segment X_i belongs to *requesting* state ($y_i = 1$) or *operating* state $y_i = 0$).

Feature selection

An automatic feature selection process was conducted initially to extract the relevant information of all the channels. More specifically, for each segment $X_i \in \mathbb{R}^{L_i \times M}$, the m most salient features were selected out of the total M features ($m \ll M$). This step results in a more succinct and salient representation. To that end, a statistics-based individual feature selection process was applied, similar to that of (Morency, de Kok, & Gratch, 2008). The major difference is that in (Morency et al., 2008), all the features were already binarized, but here the features have continuous values. Therefore, we need to encode the continuous signal into binary and then apply the feature selection procedure.

Each information channel was binarized using K-means (using 2 clusters, each representing a binary level). Then, a χ^2 test was conducted between the binarized data and the ground truth label. The M features were sorted based on the significance value of the corresponding χ^2 test, and the m most significant features were retained as the optimal feature set, represented by $\hat{X}_i \in \mathbb{R}^{L_i \times m}$.

Early prediction

To conduct early turn-taking prediction, we applied two techniques, Multi-Dimensional Dynamic Time Warping (MD-DTW) (ten Holt, Reinders, & Hendriks, 2007) and a recurrent neural network model named Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997). The DTW serves as a baseline and a representative for traditional temporal modeling algorithms.

Multi-Dimensional Dynamic Time Warping (MD-DTW)

Dynamic Time Warping is one of the most traditional and successful temporal modelling algorithms, and has been applied to speech processing (Abdulla, Chow, & Sin, 2003), gesture recognition (ten Holt et al., 2007) and trajectory navigation in robotics (Vakanski, Mantegh, Irish, & Janabi-Sharifi, 2012). (ten Holt et al., 2007) extended 1D-DTW to a multi-dimensional case and showed the superiority of MD-DTW over any 1D-DTW systems. DTW has also been shown as a good technique for early prediction (Mori et al., 2006).

In our scenario, we applied the MD-DTW algorithm proposed by (ten Holt et al., 2007). We used the 1-norm as the distance measure for two multi-dimensional points, i.e. the sum of the absolute differences in all dimensions. A Nearest Neighbor classification scheme was used to predict the *surgeon*'s current state based on the features described above.

The training stage aims to find a most representative instance out of all the trials for each *surgeon* state, known as the *template*. Assume that there are a total of N_1 instances of state *requesting* ($y_i = 1$), the *template* $\hat{X}_*^1$ was selected based on within-group consensus. The instance $\hat{X}_i$ which has the least cumulative DTW distances with the rest of the group was chosen as the *template* $\hat{X}_*^1$, i.e.:

$$\hat{X}_*^1 = \underset{i}{argmin} \sum_{j,j \neq i}^{N_1} DTW(\hat{X}_i, \hat{X}_j)$$

During testing stage, for a given unknown sequence $\hat{X}_k$, its DTW distance with the *templates* of each state was calculated. Then the label associated with the minimum distance was chosen as the prediction y_k for sequence $\hat{X}_k$, i.e.:

$$y_k = arg \underset{y \in \{0,1\}}{min} DTW(\hat{X}_k, \hat{X}_*^y)$$

Long Short-Term Memory (LSTM)

LSTM is a recurrent neural network architecture which has been successfully applied to handwriting recognition (Graves et al., 2009) and emotion recognition (Wöllmer, Kaiser, Eyben, Schuller, & Rigoll, 2013) among other applications. This network structure has the intrinsic temporal capabilities to automatically extract spatial-temporal features and predict an outcomes based on these values. The basic structure of a cell of LSTM is shown in Figure 5.

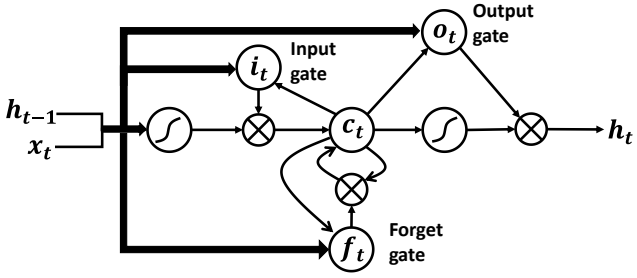

Figure 5. Basic cell of LSTM

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

$$g_t = \phi(\mathbf{W}_{xg} * X_t + \mathbf{W}_{hg} * h_{t-1} + b_g),$$
$$i_t = \sigma(\mathbf{W}_{xi} * X_t + \mathbf{W}_{hi} * h_{t-1} + b_i),$$
$$f_t = \sigma(\mathbf{W}_{xf} * X_t + \mathbf{W}_{hf} * h_{t-1} + b_f),$$
$$o_t = \sigma(\mathbf{W}_{xo} * X_t + \mathbf{W}_{ho} * h_{t-1} + b_o),$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t,$$
$$y_t = softmax(\mathbf{W}_y * h_t),$$

where X_t is the input temporal sequence at time t (corresponding to $\hat{X}_i$ defined above), h_{t-1} is the output of the memory cell at time $t-1$. $\mathbf{W}_{xg}, \mathbf{W}_{xf}, \mathbf{W}_{hg}, \mathbf{W}_{hf}$ are weight matrices, and b_g, b_i, b_f, b_0 are bias terms respectively. ϕ denotes a $tanh$ function, σ denotes a $sigmoid$ function and $\odot$ denotes an element-wise multiplication. i_t, f_t, o_t, c_t are input gate, forget gate, output gate and cell unit respectively. The memory unit is generated by the coupe of input gate and the forget gate. Generally speaking, LSTM can model long-term dependencies in temporal dimension because the cell unit can selectively "remember" or "forget" past information. The strategy to open or close each gates is data driven and is embedded in learned weights and biases.

During training stage, the segment-label pairs $\{(\hat{X}_i, y_i) | \hat{X}_i \in \mathbb{R}^{L_i \times m}, y_i \in \{0,1\}, i \in [1, N_0 + N_1]\}$ from both *requesting* state and *operating* state were supplied together into the network. The learning algorithm calculates the weights and biases, as the output of the training stage.

During testing stage, the given unknown sequence $\hat{X}_k$ was delivered to the network and the memory cell output h_t of the last time step was multiplied by $\mathbf{W}_y$ and then transformed by the *softmax* function to compute the model output $y_k \in \{0,1\}$, which is the predicted label for segment $\hat{X}_k$.

Experiment

To validate the performance of the proposed early turn-taking prediction algorithm, we conducted three experiments using the recorded multimodal data and videos as described in section *Turn-taking in Homogenous Human's Team*. The three experiments aim to: 1) evaluate the performance of automatic feature selection; 2) evaluate the performance of two early prediction algorithms; 3) evaluate the algorithm performance when compared with a human (acting as a nurse). The experiment setting follows a k-fold validation scheme, with the number k= 3. Under such scheme, the training and testing split contains data from different trials of the same subject. The F1 score for the *requesting* event prediction was used as the single metric to evaluate performance. The average F1 score, as well as the standard deviation, are shown in the Figures 6 and 7. There were a total of 348 segments of state *requesting* and 536 segments of state *operating* (roughly 1.5 times the size of *requesting*). All the *operating* events were segmented to have a window length equal to the median of all *requesting* events.

Feature selection experiment

We conducted an experiment to compare the performance of using all M raw features (i.e., using $X_i \in \mathbb{R}^{L_i \times M}$) with using the selected m optimal subset of features (i.e., using $\hat{X}_i \in \mathbb{R}^{L_i \times m}$). There were in total 50 (i.e., $M = 50$) raw channels of information obtained from the sensors used. For the comparison, we randomly draw 10% of the entire dataset, and conducted the feature selection process on this portion. The remaining 90% was left for training and testing purposes.

After running the statistics-based feature selection, we selected 20% features (i.e., $m = 10$) who are the most significant as indicated by the χ^2 tests. They were described in Table 1.

TABLE 1. SELECTED TOP FEATURES

Rank	χ^2 stat	Feature Name	Description
1	1299	myo.orientation.x	Dominant hand orientation X
2	963	myo.orientation.y	Dominant hand orientation Y
3	825	myo.acceleration.y	Dominant hand acceleration Y
4	805	myo.gyroscope.y	Dominant hand gyroscope Y
5	626	kinect.faceOrient.y	Face orientation Y
6	606	kinect.faceOrient.z	Face orientation Z
7	508	kinect.lean_forward	Forward-backward leaning
8	478	kinect.audioConfi	Acoustic amplitude
9	428	epoc.gyro.x	Head gyroscope (pitch)
10	420	epoc.gyro.y	Head gyroscope (yaw)

It is shown that features automatically selected here match those ones which are manually selected based on literature for surgeon state annotation (Figure 3). We also include the binary encoded version of the selected subset features as a comparison feature set. Here the performance of different feature sets were tested with both DTW and LSTM. The performance is shown in Figure 6. The best performance is achieved by just using the automatically selected feature subset, for both DTW and LSTM cases. The binary encoded version performed poorly, which was potentially due to the loss of information during binarization.

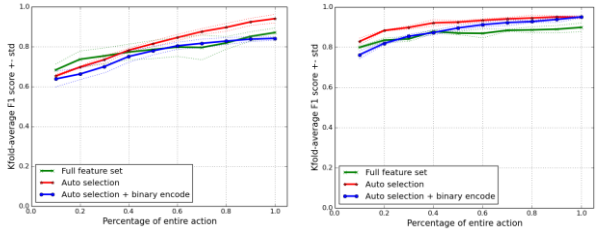

Figure 6. Comparison of different feature sets. (left) using DTW; (right) using LSTM

Early prediction experiment

In this experiment we compared MD-DTW with the LSTM network. To evaluate early prediction performances, only the beginning fraction of data was used. The specific fraction value follows a 10-grid uniform distribution from 10% of the full data to 100% of the full data. Assume that a segment $\hat{X}_i \in \mathbb{R}^{L_i \times m}$ has window length L_i , at fraction $\tau \in [0.1, 0.2, \dots, 1]$, only the data in range $[0, l_i]$ was used for training and testing, where $l_i = \tau * L_i$. In such a scenario, at data fraction point τ, we work with the sub dataset of $\{\tilde{X}_i \in \mathbb{R}^{l_i \times m}\}$. All the following experiments follows such setup.

When making early predictions, MD-DTW only calculated the distance from the beginning τ fraction of the *template* $\{\hat{X}_*^1, \hat{X}_*^0\}$ to the same fraction of the unknown sequence. In the LSTM case, the neural network was retrained with each beginning segment of the full data, and tested on the same fraction.

The LSTM training and testing was based on Tensorflow library (Abadi et al., 2015). The number of iterations was set to be 10000, with a learning rate of 0.001 and a hidden layer number of 32. The performance of the two algorithms are shown in Figure 7 (the human baseline curve is explained below). The LSTM greatly outperforms DTW in all regions.

Human baseline comparison

We also compared the performance of the proposed early prediction algorithm with that of human performance. To that end, we recruited the same participants in the data collection process and get their early prediction performances when acting as *nurses*. Videos of recorded surgical tasks were played to each participant (ensuring a cross-participant setting, i.e. every one watched other's videos). The video was played from the beginning to the end, and paused at random time instances. When the video was paused, the participant was asked whether they think that the *surgeon* intends to request an instrument or not. Those answers were compared with the ground truth and, thus human baseline performance was calculated.

The specific time marks when the video was paused were determined as follows. First, the video was paused within each *requesting* state. Assume that a video clip starts at t_0 and ends at t_1. It was randomly paused at time $t_*(t_0 < t_* \leq t_1)$, according to a discrete uniform distribution of $\mathcal{U}\{t_0, t_1\}$. Second, one or two video clips (with equal probability) were selected as *operating* states at random locations between two consecutive *requesting* states. The length of each video clips equals to the median length of all the *requesting* events. For each video clip of *operating* state, the pausing time was determined in the same way as explained above for the *requesting* state.

The performance of the human baseline, compared with the DTW and LSTM algorithms is shown in Figure 7. The LSTM algorithm outperforms the human baseline in fraction range $\tau \in [0,0.7]$, performs as well as human baseline in fraction range $\tau \in [0.7, 0.9]$, and is slightly worse than human baseline when full action was observed ($\tau = 1$). The median length of the entire action is about 2 seconds. Therefore the proposed algorithm can deliver better early prediction performance in early stages of the action (about 1.4 seconds after the action starts and 0.6 seconds before the action ends).

It is worth noting that humans watch the video continuously and thus gain context information about the surgical task. Such context information can contribute in learning to identify potential turn-taking instances. This occurs due to the correlation between consecutive events. Currently, the proposed early turn-taking algorithms enable decision making based solely on a window of the segmented data and independently from the other time window, thus lacking the capability of utilizing contextual cues and task dynamics.

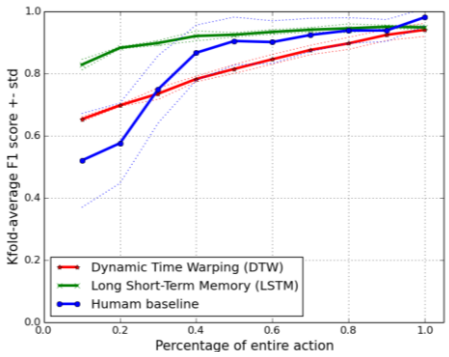

Figure 7. Comparison of the proposed early turn-taking prediction algorithm with DTW baseline and human baseline

Conclusions

This paper presented the design and implementation of an early turn-taking prediction algorithm for a robotic scrub nurse system. Sensors were utilized to capture surgeon's communication cues, automatic feature selection was carried to retain significant features, and lastly LSTM networks were used for early turn-taking prediction. The effectiveness of automatic feature selection process was verified through three experiments. It was found that the proposed early turn-taking prediction algorithm can outperform human performance before 70% of entire action finishes (about 0.6 seconds before the end of the event).

Future work includes proposing a more sophisticated surgeon state definition, giving a better spatial-temporal feature construction and including more contextual information to enrich the early prediction algorithm. We plan to validate the proposed turn-taking prediction algorithm and the robotic scrub nurse system with surgeons in the operating room.

References

Abadi, Martin, Ashish Agarwal, Paul Barham, et al. 2015 TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software Available from Tensorflow. Org 1.

Abdulla, Waleed H., David Chow, and Gary Sin 2003 Cross-Words Reference Template for DTW-Based Speech Recognition Systems. In TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region Pp. 1576–1579. IEEE.

Chao, Crystal, and A. Thomaz 2012Timed Petri Nets for Multimodal Interaction Modeling. In ICMI 2012 Workshop on Speech

and Gesture Production in Virtually and Physically Embodied Conversational Agents.

Cohen, J. 1960 A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1).

Ehrlich, S., A. Wykowska, K. Ramirez-Amaro, and G. Cheng 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) Pp. 1104–1109.

Graves, Alex, Marcus Liwicki, Santiago Fernández, et al. 2009 A Novel Connectionist System for Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(5): 855–868.

Groten, R., D. Feth, R.L. Klatzky, and A. Peer 2013 The Role of Haptic Feedback for the Integration of Intentions in Shared Task Execution. IEEE Transactions on Haptics 6(1): 94–105.

Gulášová, Ivica, Lenka Görnerová, Ján Breza jr, and Ján Breza N.d. Communication in the Operating Room.

Hart, Justin W., Brian Gleeson, Matthew Pan, et al. N.d. Gesture, Gaze, Touch, and Hesitation: Timing Cues for Collaborative Work.

Hochreiter, Sepp, and Jürgen Schmidhuber 1997 Long Short-Term Memory. Neural Computation 9(8): 1735–1780.

ten Holt, Gineke A., Marcel JT Reinders, and E. A. Hendriks 2007 Multi-Dimensional Dynamic Time Warping for Gesture Recognition. In Thirteenth Annual Conference of the Advanced School for Computing and Imaging.

Jacob, Mithun George, Yu-Ting Li, and Juan P. Wachs 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.

Lucas, James M., and Michael S. Saccucci 1990 Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements. Technometrics 32(1): 1–12.

MacKenzie, L., J. A. Ibbotson, C. G. L. Cao, and A. J. Lomax 2001 Hierarchical Decomposition of Laparoscopic Surgery: A Human Factors Approach to Investigating the Operating Room Environment. Minimally Invasive Therapy & Allied Technologies 10(3): 121–127.

Martínez, Héctor P., and Georgios N. Yannakakis 2014Deep Multimodal Fusion: Combining Discrete Events and Continuous Signals. In Proceedings of the 16th International Conference on Multimodal Interaction Pp. 34–41. ACM.

Moore, Alison, David Butt, Jodie Ellis-Clarke, and John Cartmill 2010 Linguistic Analysis of Verbal and Non-Verbal Communication in the Operating Room. ANZ Journal of Surgery 80(12): 925–929.

Morency, Louis-Philippe, Iwan de Kok, and Jonathan Gratch 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.

Mori, Akihiro, Seiichi Uchida, Ryo Kurazume, et al. 2006Early Recognition and Prediction of Gestures. In 18th International Conference on Pattern Recognition (ICPR'06) Pp. 560–563. IEEE.

Rabøl, Louise Isager, Mette Lehmann Andersen, Doris Østergaard, et al. 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.

Strabala, Kyle Wayne, Min Kyung Lee, Anca Diana Dragan, et al. 2013 Towards Seamless Human-Robot Handovers. Journal of Human-Robot Interaction 2(1): 112–132.

Vakanski, Aleksandar, Iraj Mantegh, Andrew Irish, and Farrokh Janabi-Sharifi 2012 Trajectory Learning for Robot Programming by Demonstration Using Hidden Markov Model and Dynamic Time Warping. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42(4): 1039–1052.

Wöllmer, Martin, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll 2013LSTM-Modeling of Continuous Emotions in an Audiovisual Affect Recognition Framework. Image and Vision Computing 31(2): 153–163.