Anticipation in Human-Robot Cooperation: A recurrent neural network approach for multiple action sequences prediction

Paul Schydlo[1], Lorenzo Jamone[1,2] and José Santos-Victor[1]

***Abstract*— Close Human-robot cooperation is a key enabler for new developments in advanced manufacturing and assistive applications. Close cooperation require robots that can predict human actions and intent, and understand human non-verbal cues.**

Recent approaches based on neural networks have led to encouraging results in the human action prediction problem both in continuous and discrete spaces. Our approach extends the research in this direction.

Our contributions are three-fold. First, we validate the use of gaze and body pose cues as a means of predicting human action through a feature selection method.

Next, we address two shortcomings of existing literature: predicting multiple and variable-length action sequences. This is achieved by introducing the encoder-decoder recurrent neural network topology in the discrete action prediction problem.

In addition, we theoretically demonstrate the importance of predicting multiple action sequences as a means of estimating the stochastic reward in a human robot cooperation scenario.

Finally, we show the ability to effectively train the prediction model on a action prediction dataset, involving human motion data, and explore the influence of the model's parameters on its performance.

I. INTRODUCTION AND RELATED WORK

In a world with a growing number of autonomous systems and moving towards the coexistence and cooperation between humans and sophisticated robots, it is crucial to enable artificial systems to understand and predict human behaviour. The ability to predict how the human will behave in the near future finds applications in areas such as cooperative robotics [1], [2], auto-mobile safety [3], elderly care [4], among many others [5].

In addition to the use of speech for communicating and coordinating their next actions, humans rely extensively on non-verbal cues for action and movement prediction [6]. Situations where fast cooperation is essential, for example cooperative assembly, require the understanding of subtle non-verbal cues [2] about the human intention and future action. In these scenarios it is not enough to merely recognize the current action. Instead, it is fundamental to predict actions and anticipate the intent in order to guarantee seamless cooperation [7].

A. Non-verbal cues

There are several non-verbal cues that enable human action prediction [8], [9]. This paper takes into account two of them: gaze and body posture. Gaze is important, as it has both a role in social communication in conveying turn taking behaviour [10] or attention in conversation, but at the same time it is deeply related to the agent's Theory of Mind [11] and codifies the action goals through both visuo-motor coupling [12] and attention [9]. Body posture, similarly to gaze, can serve both a social and intention conveying signal while also indicating possible action targets.

Past works have focused on either gaze [1]–[3] or body pose [13] cues and their relation to action recognition and prediction. Both are important in understanding human behaviour and can give relevant information about the human's action goal.

Research on non-verbal cues in human robot-cooperation has a long history, including the bulk of work on mirror neurons [14] and its computational and robotic models and implementations [15]. Other examples include Admoni et al. [5] use of human eye-gaze as an addition to a joystick input as a means of estimating the human intent, modelling the relation between the eye-gaze and the action goal by their relative distance. Huang [1] quantified the importance of gaze features, successfully demonstrating the importance of gaze by proactively planning actions according to the human intent.

B. Prediction models

Human behaviour prediction can be solved on different levels of abstraction and is concerned with estimating a probability distribution over the set of next possible actions.

On a higher level of abstraction, models can predict actions in a discrete space [3], [16] where the actions are symbolic in nature and can represent underlying movement patterns, e.g. "press-button" or "grab-object". On a lower level of abstraction, movement can be directly anticipated in a continuous space [17], e.g. human walking trajectories.

Past works explored different architectures like probabilistic models with Markovian assumptions [18] and discriminative methods such as conditional random fields [19]. Recently, recurrent neural networks which do not assume limiting Markovian assumptions have shown excellent results [16], [17], [20].

Predicting continuous actions has been addressed in the context of body pose and human trajectory prediction. Example of relevant work include the use of recurrent neural networks by Martinez [17] as a means of predicting contextually coherent future joint trajectories.

The dual problem is action prediction in discrete outcome space. Examples of related work include a Conditional Ran-

*

[1]Vislab, Institute for Systems and Robotics, Instituto Superior Técnico, University Lisbon, Portugal

[2]Advanced Robotics, Queen Mary University of London, London, United Kingdom

dom Field based approach by Koppula [19] to capture temporal dependencies and Saponaro's Hidden Markov Model based approach [18]. More recently, Jain [16] introduced the structural RNN as a means of encoding past contextual information and predicting a fixed number of steps in the future. While the field has had a rapid evolution in the last couple of years, there are two shortcomings in the literature that this paper tries to address.

The first is concerned with predicting a fixed versus a variable number of steps into the future. While models like the one introduced in [16] have a remarkable ability to condense contextual past information their scope is limited to fixed step ahead prediction length. This paper aims at extending discriminative recurrent models in a classification setting with variable length action sequence prediction.

The second shortcoming is related to the single future action sequence versus multiple future action sequences. While models like the one introduced in [17] are able to effectively use recurrent models to predict a variable number of steps into the future their scope is limited to a regression setting where sampling multiple future action sequences is a non-trivial problem. This paper explores a multiple future action sequence predictor in the classification setting.

C. Contributions

The main contributions of this paper are the following:

- Quantifying the **relative importance of pose and gaze** features in an intention recognition scenario.
- Extending recurrent neural network fixed step action prediction with **variable length action prediction**.
- Introducing the simultaneous prediction of **multiple future action sequences**.

II. APPROACH

Our work aims to look at the action prediction problem from an end-to-end perspective, starting with the problem of non-verbal cues selection and moving on to develop an action sequence prediction model. Keeping in mind the final goal, predicting future human action given a sequence of non-verbal cues such as gaze and pose, this section is organized in a sequential bottom-up order.

First, we address the issue of establishing a qualitative metric for assessing the relative importance of pose and gaze features. Then, in Section II.B, we introduce the multiple action sequence prediction model that is one of the key contributions of this paper. Predicting action sequences introduces complexity issues which are handled in Section II.C. Finally, in Section III we use the distribution over future action sequences sampled from the model introduced in Section II to estimate the expected future reward in a human-robot cooperation scenario.

A. Feature importance

This section seeks to introduce a qualitative metric for the relative gaze and body pose cues importance, two commonly used features in non-verbal communication [8]. Selecting the right features is an important step to reduce the complexity and increase the robustness of our models.

There are different feature selection methods which can be categorized into *filter*, *wrapper* and *embedded* classes [21]. Since the relation between the features is unknown, it is assumed to be non-linear in nature. Following the non-linearity assumption the focus of this section will be on the *wrapper* class of feature selection methods.

This class of methods captures non-linear relation between the variables through a black-box model. It starts by training the model on subsets of the feature space and then ranks the features according to the models' accuracy [21].

In the case of this paper, the black-box model is the intention recognition model, a Recurrent Neural Network (RNN) sequence to sequence model. The structure of the model is defined by an embedding layer which at every step transforms the feature vector into an intermediary representation acting as an input to the model's RNN. For every input this RNN returns a discrete distribution over intentions. This distribution is obtained by projecting the recurrent neural network's internal state and normalizing it through a softmax layer.

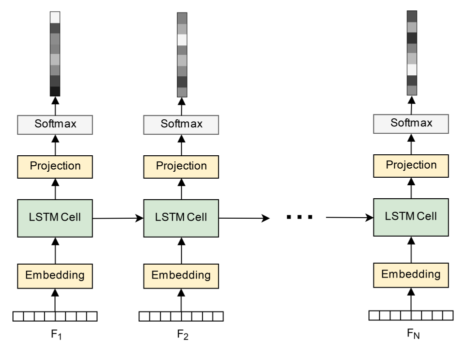


Fig. 1: **Intention recognition model.** This model maps a sequence of input features to a sequence of discrete distributions over the action vocabulary.

The prediction accuracy of the model with and without a given feature can be considered a proxy for the feature's added information.

Having established a qualitative measure of the gaze and pose features' importance, the next section introduces the prediction model.

B. Prediction model

This section introduces the discrete encoder-decoder recurrent neural network topology which seeks to solve the shortcomings enumerated in section II. The first part of the model is a contextual information encoder. The encoder condenses past information into a fixed length context vector through a Long Short Term Memory (LSTM) cell. The embedding is a fully connected layer (FeatureVectorDim x 50), where FeatureVectorDim is the size of the feature vector. The embedding layer includes dropouts which act as a regularization to the model [22]. The encoder LSTM's hidden state dimension is 20. This context vector is the initial state of the second part of the model, the decoder.

The decoder is responsible for generating a coherent future sequence of actions. At each step the decoder, an LSTM cell, returns a discrete distribution over possible future actions. This distribution is obtained by projecting the decoder's internal state and normalizing it using a softmax layer. The decoding process samples an action from the distribution and feeds it back as an input to the next decoding iteration. The projection is a fully connected layer (HiddenStateDim x VocabDim), where HiddenStateDim is the size of the hidden state, 20, and VocabDim the dimension of the action discrete possible actions vocabulary, 11. The decoder LSTM's hidden state dimension is 20.

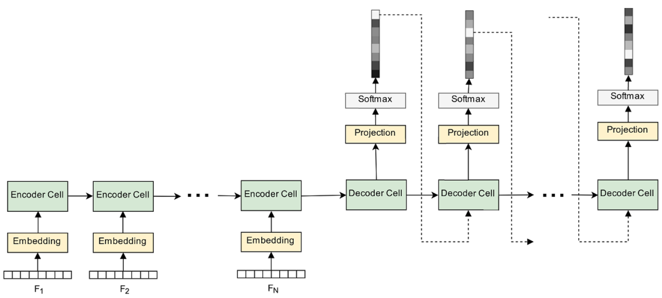


Fig. 2: **Encoder-decoder model.** Left part summarises past information into a fixed length context vector. Right part expands this context vector into future action sequences.

The model is trained on a sequential cross entropy loss function approximating the discrete next action distribution. The cross entropy cost (1) is a measure of difference between two distributions. The sequential cross entropy is obtained by summing the cross entropy cost over the prediction steps.

$$H(p,q) = -\sum_{x} p(x) \log\left(q(x)\right) \tag{1}$$

After training, the decoding process allows for variable length action sequence prediction. Expanding every possible future action sequences becomes NP hard and computationally intractable. The next section looks more closely at this issue and introduces one possible solution to the problem.

C. Complexity issues

The previous section hints at the complexity underlying the decoding process. At every decoding step the decoder samples one or more actions from the output distribution as possible actions at a given time step, it then expands these actions by branching and feeding them individually as input to the next prediction step decoder iteration. There are two strategies that could be applied to this decoding process.

Naively expanding the space of all possible action sequences and selecting the most probable action sequence in the end seems like a reasonable idea. Nevertheless, expanding the actions at each step results in a vocabulary sized multiplier in the number of possible action sequences at every prediction step. In terms of complexity this means the number of action sequences increases exponentially with

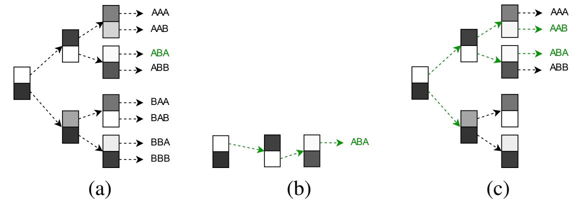


Fig. 3: **Search methods comparison.** a) Exhaustive search expands all possible action sequences. b) Greedy search picks the most probable action at every step. c) Beam search keeps a set of the best K action sequences, expanding and pruning the set at every step.

the number of prediction steps. Considering a 10 actions vocabulary size, the first decoding step results in 10 action sequences, expanding the 10 action sequences results in 100 possible action sequences for a two step ahead prediction, a three step ahead prediction would result in 1000 action sequence alternatives.

Greedily expanding only the best option, could be a solution to the exponentially expanding trajectory space, nevertheless it has the shortcoming that this method only returns one action sequence prediction.

A common solution to these two problems is the implementation of a *beam search* based decoder [23]. This method keeps a set of the top K best future action sequences at every decoding step, expanding by the action vocabulary size and pruning the action sequence set back to the top K best future action sequences. The end result is a sample of the top K most probable future action sequences ordered by likelihood. These trajectories are called beams and K is called the beam width parameter.

III. APPLICATION SCENARIO

Anticipating a set of possible future actions is crucial in cooperative assembly scenarios were two agents work together in a fast paced joint action setting. This scenario aims to clarify the importance and some caveats inherent to the action prediction problem in human robot cooperation scenarios.

Fig. 4: **Joint action scenario.** The human and robot act independently towards maximizing a joint reward function.

This setting (Fig. 4) is defined by a set of possible world states, S, a set of joint human and robot actions, A_s, transition probabilities between states over joint actions, T, and a joint reward function $R(S, a_H, a_R)$. For the sake of this example, the world state is a set of pre-conditions, T a set of action-effect axioms defined in a Planning Domain Definition Language (PDDL) space and R a reward function on the sub-goal completion.

Both the human and the robot choose their actions, a_H and a_R respectively, independently maximizing a joint reward function $R(S, a_H, a_R)$ and influencing the environment Markov process. We assume a leader-follower paradigm where the robot chooses his action as a function of the possible next human actions and the humans actions are not influenced by the robots action choice.

Given an initial state, S_0, and an action sequence, A_s, which is a series of action pairs (a_H, a_R), the total reward after N steps is given by (2). Where $a_{H,i}$ and $a_{R,i}$ correspond respectively to the action performed by the human and robot at time step i and N the length of the sequence.

$$R(S_0, A_s) = \sum_{i=0}^{N} R(S_i, a_{H,i}, a_{R,i}) \tag{2}$$

Since the behaviour of the human is non-deterministic from the perspective of the robot the future reward associated to a chosen action sequence can only be determined stochastically as an expected reward. This expected reward of a chosen robot action sequence, A_R, is given by the marginalization of the stochastic reward along the space of possible human actions (3), where V is the dimension of the action vocabulary, $f_{H,i}$ represents the human action distribution at time step i, a_j an action in the human action vocabulary and N the action sequence length.

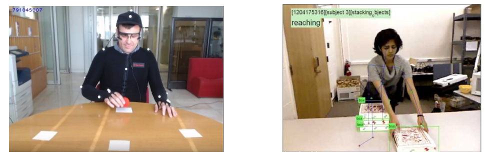
$$\mathbb{E}\left[R(S_0, a_s)\right] = \sum_{i=0}^{N} \sum_{j=0}^{V} R(S_i, H_j, a_{s,i}) f_{H,i}(H_j) \tag{3}$$

Estimating the complete future action distribution would require expanding all possible action sequences which is NP hard and computationally intractable. The beam search introduced in the previous section enables us to approximate the future action distribution.

Considering the set of most probable action sequences returned by the multiple action sequence prediction as representative of the future human behaviour, the marginalization and therefore the expected reward can be approximated by summing and weighting each human and robot action sequence's reward by the action sequence probability (4). Increasing the number of beams in the decoder, K, approximates the reward better but is computationally more demanding. Here $p(b_k)$ represents the beams probability, $a^k_{H,i}$ the action performed by the human at time step i in the beam k and $a_{s,i}$ the action performed at step i in the robot action sequence a_s.

$$\mathbb{E}\left[R(S_0, a_s)\right] = \frac{\sum_{k=0}^{K} \sum_{i=0}^{N} \left[R(S_i, a^k_{H,i}, a_{s,i})\right] p(b_k)}{\sum_{k=0}^{K} p(b_k)} \tag{4}$$

As the beam count tends to the total number of possible action sequence combinations this expression approximates the expected reward (3).



(a) (b)

Fig. 5: **Datasets.** a) ACTICIPATE motion and eye gaze dataset. b) CAD120 RGB-D motion dataset.

IV. EXPERIMENTS AND RESULTS

We start by describing the datasets used in the experimental part of our work. Then, we move on to compare the non-verbal cues importance and finish by evaluating the action sequence prediction model on a dataset that includes body pose information.

A. Datasets

The feature importance is evaluated on a combined gaze and skeleton dataset which was acquired and published in the ISR Vislab ACTICIPATE[1] project (Fig. 5a). This dataset consists of a human actor's gaze and skeleton movement while performing either one of six actions (Place Left, Place Center, Place Right, Give Left, Give Center, Give Right). This dataset was recorded using the Optitrack motion capture system, and Pupil Labs binocular eye gaze tracking system, synchronised at a 120Hz frequency. The total number of action sequences is 120. The sequences have an average length of 220 frames. Every sequence corresponds to one action and is labelled accordingly.

The multiple action sequence prediction model is evaluated on the CAD120 dataset (Fig. 5b, [24]). This dataset consists of a human actor's skeleton movement while performing a sequence of actions like "pouring" and "eating". This dataset is of especial interest since it covers the scope of action sequences and it is not limited to one action per video segment. It is one of the few datasets which has varying order of action sequences. This dataset consists of body pose features in a binary feature format together with the respective action labels at a sample frequency of 5Hz. The total number of action sequences is 120 and the sequences have an average length of 25 time steps.

B. Feature Importance

In our first experiment, we train the model on the combined body pose and gaze features to confirm that it yields the expected behaviour. As the movement progresses, the model receives more information and identifies the intention, correctly converging to the true label, 5. The movement takes 220 frames (about 2 seconds), the model is able to predict the intention target after seeing less than half of the total trajectory, about 100 frames.

The second experiment is concerned with quantifying the relative importance of the different non-verbal cues in predicting human intent.

[1]The ACTICIPATE dataset can be downloaded from the following web page: http://vislab.isr.tecnico.ulisboa.pt/datasets/

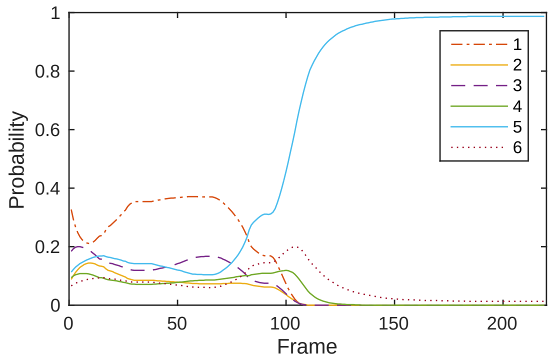


Fig. 6: **Action distribution temporal dynamic.** Action probability temporal evolution. The model starts with uniform probability and after about 100 frames converges to the correct label.

The model is trained on two sets of features: (i) combined gaze and pose cues, and (ii) body pose only. Figure 7) shows the model performance under these two conditions and the importance of the gaze information for the model to anticipate human actions.

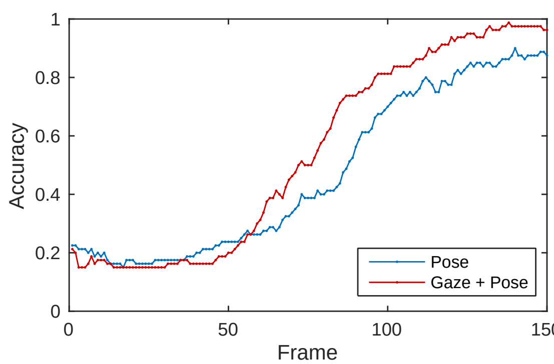


Fig. 7: **Gaze and pose accuracy.** Accuracy of a model trained on (i) gaze only features, and (ii) trained on combined gaze and body pose features.

The difference in accuracy between the two sets of cues hints at the importance of gaze. Despite the model performing similarly with and without gaze, the results show that gaze has an important role in early prediction of human activity. The model trained on both gaze and body pose cues predicts the correct action 92 ms before the model with only body cues. An interesting result is that this delay in predictive ability coincides with the range of delays between eye and hand movement observed in research on eye-arm movement coupling [25].

Having established the relative importance of both gaze and body pose features in action prediction, in the next section we will evaluate the multiple action sequence prediction model on a multi-action pose feature dataset.

C. Prediction Model

The model takes the pose features, observed over three time steps, as input in order to predict future actions as accurately as possible. We will investigate how the prediction model's parameters affect the performance. The model is evaluated on the CAD120 dataset, already introduced before.

Performance will be assessed with the F1 score [26]. The F1-score is evaluated on a four fold cross validation scheme, with the final score being an average over the folds results. As there are folds without instances of some label, the F1 score is calculated directly on the true positive, false negative and false positive rate (5).

$$F1 = \frac{2 \cdot TruePos}{2 \cdot TruePos + FalseNeg + FalsePos} \tag{5}$$

While the model is dynamic in its ability to predict variable length action sequences, the accuracy of action sequence prediction is influenced by the prediction length the model is trained on (Fig. 8). This correlation is related to the ability of the decoder to manage its internal state. When the network is trained on a long future action sequence, it learns to keep and manage the decoder's internal state predicting longer sequences with more accuracy.

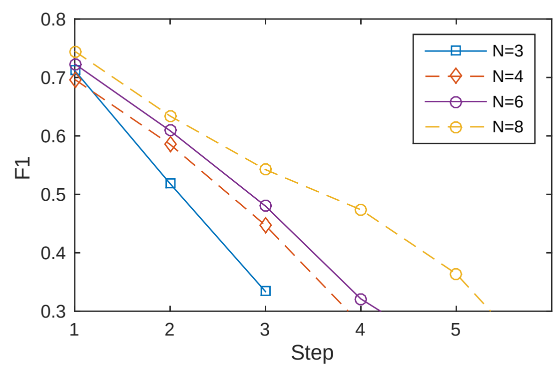


Fig. 8: **Accuracy as a function of prediction length.** Prediction accuracy across time steps is positively correlated with the prediction length the model is trained on. ("N" corresponds to the prediction length used for training the model, and "Step" the position in the predicted sequence.)

The second parameter to analyse, is the number of beams (action sequences) which determine the space of action sequences the model is able to capture (Fig. 9). The cumulative sum of the beams' probabilities is a measure of the solution space we are able to cover with a given number of beams.

The space of possible solutions grows exponentially with the number of prediction steps. While a beam width of 11 beams is able to capture 100% of the outcome probability space in a one-step ahead prediction scenario, the same number of beams only captures around 75% of the outcome probability space in the two-step ahead prediction scenario. As the solution space grows, a fixed number of beams captures a cumulative probability outcome space that decays with the number of prediction steps.

It is well known that the generalization error is related to the model's capacity [27], thus representing the model's

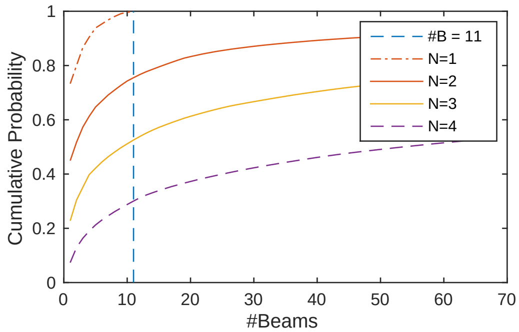


Fig. 9: **Beam cumulative probability.** Cumulative probability of the outcome space the model is able to capture. "N" represents the length of the predicted trajectory, and "#Beams" the length of the predicted action sequences.

ability to learn complex patterns. The dimensionality of the context vectors can be used as a parameter to define the model capacity. Increasing this dimension reduces the informational bottleneck, increasing the model capacity and the generalization error. Increasing the generalization error makes the model prone to over fitting to the training set and not generalizing to new samples (Fig. 10). Hence, the context vector dimensionality acts as a regularizer of the model.

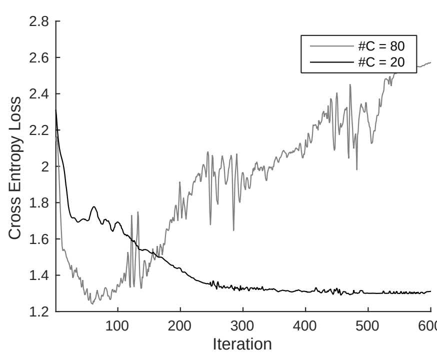


Fig. 10: **Validation loss as a function of the context dimensionality.** The iteration represents the number of training steps while #C represents the dimensionality of the context vector parameter. As the dimensionality parameter is increased, the network starts to overfit to the training data.

V. CONCLUSIONS

We showed the importance of both body pose and gaze cues for the accurate prediction of human intent. More specifically, the experiments demonstrated that a model trained on both body and gaze cues predicts the correct action about 92ms before a model trained only on body pose cues.

We introduced a recurrent neural network topology designed to predict multiple and variable length action sequences. Predicting action sequences introduces combinatorial complexity issues which were successfully mitigated using a pruning method.

We demonstrated the theoretical value of predicting multiple and variable action sequences prediction for estimating the expected future reward in a human robot cooperation scenario.

We studied how different training procedure and parameter combinations affect the model performance. All tests were carried out on real publicly available datasets.

Our approach extends the state of the art in directions that are key to enable more efficient human-robot cooperation, particularly involving non-verbal communication.

VI. FUTURE WORK

This work establishes a strong base for the implementation of a joint action scenario on a humanoid robotics platform such as the iCub.

More specifically, it would be interesting to extend the model by exploring the connection between non-verbal cues and semantic features related to the context, through composing classical graphical probabilistic methods with discriminative neural network based methods.

ACKNOWLEDGMENT

This work has been partially supported by the Portuguese Foundation for Science and Technology (FCT) project [UID/EEA/50009/2013] and the RBCog-Lab research infrastructure.

REFERENCES

[1] C. M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, vol. 2016-April, no. Section V, pp. 83–90, 2016.

[2] K. Sakita, K. Ogawara, S. Murakami, K. Kawamura, and K. Ikeuchi, "Flexible cooperation between human and robot by interpreting human intention from gaze information," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 1, pp. 846–851, 2004.

[3] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent Neural Networks for driver activity anticipation via sensory-fusion architecture," *Proceedings - IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2016-June, 2016.

[4] T. Yonezawa, H. Yamazoe, A. Utsumi, and S. Abe, "Attractive, Informative, and Communicative Robot System on Guide Plate as an Attendant with Awareness of User's Gaze," *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 113–122, 2013.

[5] H. Admoni and S. Srinivasa, "Predicting user intent through eye gaze for shared autonomy," *AAAI Fall Symposium - Technical Report*, vol. FS-16-01 -, pp. 298–303, 2016.

[6] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration.," *Frontiers in psychology*, vol. 6, no. July, p. 1049, 2015.

[7] N. Sebanz and G. Knoblich, "Prediction in Joint Action: What, When, and Where," *Topics in Cognitive Science*, vol. 1, no. 2, pp. 353–367, 2009.

[8] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of Nonverbal Communication on Efciency and Robustness of Human-Robot Teamwork," *International Conference on Intelligent Robots and Systems (IROS)*, 2005.

[9] M. Argyle, R. Ingham, F. Alkema, and M. McCallin, "The Different Functions of Gaze," 1973.

[10] S. Ho, T. Foulsham, and A. Kingstone, "Speaking and listening with the eyes: Gaze signaling during dyadic interactions," *PLoS ONE*, vol. 10, no. 8, pp. 1–18, 2015.

[11] S. Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, 1997.

[12] L. Lukic, J. Santos-Victor, and A. Billard, "Learning robotic eye-arm-hand coordination from human demonstration: A coupled dynamical systems approach," *Biological Cybernetics*, vol. 108, no. 2, pp. 223–248, 2014.

[13] C. Perez-D'Arpino and J. A. Shah, "Fast Target Prediction of Human Reaching Motion for Cooperative Human-Robot Manipulation Tasks using Time Series Classification," *International Conference on Robotics and Automation (ICRA)*, pp. 6175–6182, 2015.

[14] V. Gallese and A. Goldman, "Mirror neurons and the simulation theory of mind-reading," *Trends in Cognitive Sciences*, vol. 2, no. 12, pp. 493 – 501, 1998.

[15] M. Lopes and J. Santos-Victor, "Visual learning by imitation with motor representations," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 35, no. 3, pp. 438–449, 2005.

[16] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN : Deep Learning on Spatio-Temporal Graphs," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] G. Saponaro, G. Salvi, and A. Bernardino, "Robot anticipation of human intentions through continuous gesture recognition," *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, no. Cts, pp. 218–225, 2013.

[19] H. S. Koppula and A. Saxena, "Anticipating Human Activities Using Object Affordances for Reactive Robotic Response," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 1, pp. 14–29, 2016.

[20] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human Trajectory Prediction in Crowded Spaces," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, 2016.

[21] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[23] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," 2014.

[24] H. S. Koppula, R. Gupta, and A. Saxena, "Learning Human Activities and Object Affordances from RGB-D Videos," *International Journal of Robotics Research*, 2012.

[25] R. W. Angel, W. Alston, and H. Garland, "Functional relations between the manual and oculomotor control systems," *Experimental Neurology*, vol. 27, no. 2, pp. 248–257, 1970.

[26] C. J. Van Rijsbergen, *Information Retrieval*. Butterworth-Heinemann, 1979.

[27] Y. Bengio, I. J. Goodfellow, and A. C. Courville, *Deep Learning*. The MIT Press, 2016.