Enabling Early Gesture Recognition by Motion Augmentation

Rohit Agrawal
Department of Computer Science
Boston University
agrawroh@bu.edu

Ajjen Joshi
Department of Computer Science
Boston University
ajjendj@bu.edu

Margrit Betke
Department of Computer Science
Boston University
betke@bu.edu

ABSTRACT

In real-time gesture recognition algorithms, accurately classifying gestures early, when they are only partially observed, can be advantageous as it minimizes latency and improves user experience. This work investigates a novel approach for improving the results of an early gesture classification model. The method involves augmenting the input sequence of human poses of a partially observed gesture with a series of poses predicted by an auxiliary recurrent neural network sequence-to-sequence motion prediction model before being fed into a random forest gesture classifier. By concatenating the partially observed ground truth sequence with the forecasted motion sequence, we are able to significantly improve early gesture recognition accuracy. When forecasting 25 future frames of a partially observed input gesture sequence of 50 frames, recognition accuracy improves from 45% to 87% on average when evaluated on the MSRC-12 gesture dataset.

CCS CONCEPTS

• **Human-centered computing → Interaction paradigms**; • **Computing methodologies → Machine learning**;

KEYWORDS

gesture recognition, early prediction

ACM Reference Format:
Rohit Agrawal, Ajjen Joshi, and Margrit Betke. 2018. Enabling Early Gesture Recognition by Motion Augmentation. In *PETRA '18: The 11th PErvasive Technologies Related to Assistive Environments Conference, June 26–29, 2018, Corfu, Greece*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3197768.3197788

1 INTRODUCTION

The problem of recognizing human gestures is well-studied in the computer vision and machine learning research communities, with a broad scope of applications spanning human-computer interaction and motion synthesis for virtual and augmented reality. In applications that require real-time recognition of gestures, it is advantageous to accurately predict gestures as early as possible. Consider, for example, a rehabilitative gaming application where patients control the flow of the game with therapeutic gestures. Predicting a gesture when it is only partially observed instead of waiting for the gesture to be fully completed can decrease latency and therefore improve user comfort and experience.

Gesture recognition systems using a variety of machine learning based methods, such as nearest-neighbor based dynamic time warping (DTW) distances [9], hidden Markov models (HMM) [16], hidden conditional random fields (HCRF) [15], random forests [7] and recurrent neural networks (RNN) [5] have been proposed. A comprehensive survey of gesture recognition can be found elsewhere [2, 11]. Many modern gesture recognition systems rely on 3D skeletal data as input. Devices such as the Microsoft Kinect [14] have had a revolutionary impact on gesture-based HCI applications as they enabled the modeling of human motion based on accurate, low-dimensional skeletal pose features.

Related to the problem of recognition of human gestures and activities is that of the forecasting of human motion. Given a sequence of human skeleton poses, the task in motion forecasting is to predict a sequence of future poses conditioned on the input. This is a non-trivial problem as human motion is stochastic and exhibits non-linear dynamics. Following the impressive performance of deep learning methods in a wide spectra of problems in computer vision [13], recent work has focused on using deep recurrent neural networks (RNNs) to model human motion. Martinez et al. [10] trained a sequence-to-sequence Gated Recurrent Unit (GRU) network and achieved state-of-the-art short-term motion prediction results on the Human 3.6M dataset. Butepage et al. [4] employed an encoder-decoder architecture to predict future 3D poses, thereby learning a robust feature representation of human skeletal data.

Our focus is on building systems that can perform early gesture recognition. That is, we would like to build classifiers that can accurately predict labels of input gestures even when they are only partially observed. Solutions to the problem of early action and gesture recognition have been proposed previously [8, 12]. However, applying models trained on fully-observed sequences to partially observed gestures at test time does not always yield particularly encouraging results.

Here, we propose a framework for performing early gesture recognition. Instead of simply classifying partially observed gestures, we leverage recent work utilizing deep recurrent neural networks to forecast short-term human motion. We input a partially-observed gesture represented by a series of poses to a sequence-to-sequence motion forecasting model, which predicts a sequence of short-term future poses. This sequence of forecasted poses is concatenated with the partially-observed ground truth gesture and fed to a random forest gesture classification model. We show that augmenting the partially-observed gesture with the output from

PETRA '18, June 26–29, 2018, Corfu, Greece

ACM ISBN 978-1-4503-6390-7/18/06…$15.00
https://doi.org/10.1145/3197768.3197788

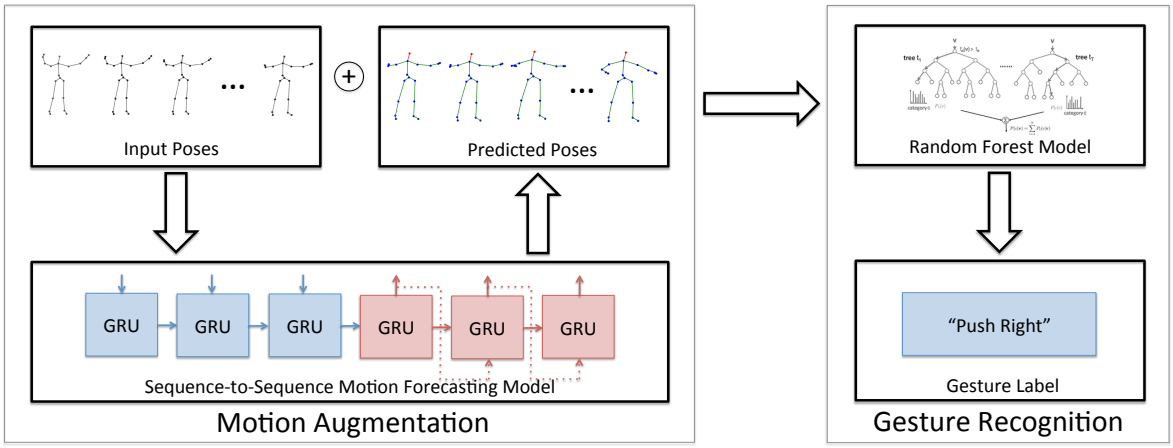

Figure 1: Pipeline view of our gesture recognition by motion augmentation model. The input sequence is first fed to a sequence-to-sequence model, which outputs a series of predicted poses. The combination of the ground truth input sequence and the predicted poses is then classified by a random forest classification model.

our sequence-to-sequence model significantly improves recognition accuracy during test time.

2 SYSTEM OVERVIEW

We here explain the components of our gesture recognition by motion augmentation model in detail. A pipeline view of the framework is illustrated in Figure 1.

2.1 Input

The input to our system consists of concatenated 3D pose descriptors. Human motion recorded with devices such as the Kinect is decoded into a concise skeletal descriptor. Each pose, which represents the configuration of the human skeleton in a single frame, is defined by a 60-dimensional vector constituting the x,y,z coordinates of 20 different body joints. These per-frame pose descriptors are concatenated to form a gesture descriptor. During training, gesture descriptors are computed from the entire gesture and are used to train the classification model. During testing, we assume the gesture is only partially observed. Therefore, the remaining poses are first inferred by the motion prediction model before being fed into the trained classifier.

2.2 Motion Prediction

For motion prediction, we utilize the sequence-to-sequence model developed by Martinez et al. [10]. The authors demonstrated the advantages of using a generic (as opposed to action-specific) GRU network based on a sampling-based loss. As in [10], a sequence of GRUs with 1024 units is used to map an input sequence of poses to an output sequence. GRUs are preferred as a computationally less-expensive alternative to LSTMs. The sequence-to-sequence model can be viewed as consisting of: (1) an encoder which receives the inputs and generates a robust internal representation, and (2) a decoder which receives the internal representation and produces a maximum likelihood estimate of subsequent poses for prediction. The network is trained to minimize the Root Mean Squared Error (RMSE) prediction error over the forecasted frames.

2.3 Motion Augmentation

The sequence of poses forecasted by the sequence-to-sequence model is concatenated with the partially observed input gesture before being fed to the gesture recognition model. The intuition behind the idea of motion augmentation is simple. Because the gesture recognition model is trained on fully observed gestures during the training phase, it cannot be expected to produce accurate predictions when given only a fraction of the full gestures. This is because the model might have learned to discriminate between different gestures in the training vocabulary by utilizing dynamics during the latter part of the gesture. We posit that this can be alleviated if a separate model is trained to peek into the future and provide a reasonable representation of the future motion conditioned on the partially observable input during the testing phase.

2.4 Gesture Recognition using Random Forests

The training set is defined as $D = ((X_1, y_1), ..., (X_n, y_n))$ of pairs of gesture feature descriptors and their corresponding gesture labels. A random forest model [3] consists of an ensemble of decision trees, each of which is trained using a separate subset of the training data in order to maximize generalizability. In a decision tree, split points are chosen at internal nodes by finding the attribute and the value of that attribute that results in the lowest cost. At each internal node of the tree, m features are randomly selected from the available d, where d is the dimensionality of the feature vector of the inputs, such that $m < d$. From the random selection of features, the feature that most reduces the cost function, such as Information Gain or the Gini index, is chosen to split the tree. We train our random forest model with feature descriptors of the fully observed gestures in the training set. During testing, we compare the performance of this model when provided with feature descriptors of partially observed gestures with and without motion augmentation.

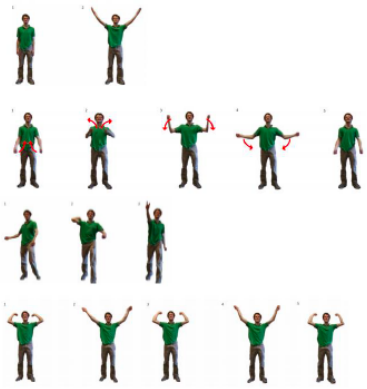

Figure 2: Example of gestures from the MSRC-12 dataset, which consists of gestures for playing a shooting game.

3 DATASET

We used the Microsoft Research Cambridge-12 (MSRC-12) [6] Kinect gesture dataset to evaluate our system. The dataset consists of sequences of 12 different human gestures (Figure 2) that could be used while playing a shooting game along with the corresponding gesture labels. The gestures are performed by 30 different subjects for a total of ∼ 6000 gesture instances, encompassing more than ∼ 700000 frames, approximately six hours and forty minutes. The gestures are represented in motion files that contain tracks of the world positions of twenty body joints estimated using the Kinect's pose estimation algorithm. The body poses are captured at a sample rate of 30Hz with an error of up to two centimeters in joint positions.

4 EXPERIMENTAL RESULTS

In all our experiments, we used the same network configuration using a single gated recurrent unit (GRU) with 1024 units as presented by Martinez et al. [10]. We changed the learning rate from the initial value of 0.005 to 0.01 by continuing to use the same batch size of sixteen samples. During training, we fed the network a total of fifty frames, which is equivalent to 1.67 seconds of motion, to the encoder and predicted the next frame in the sequence, which is equivalent to 33.33 milliseconds of motion. A generalized network was trained for all different gestures using the above configuration. The whole architecture was implemented in Tensorflow [1] on a grid of three NVIDIA Titan GPUs.

4.1 Motion Prediction

We evaluated the performance of our motion prediction model using the RMSE metric which measures the Euclidean distance between the ground truth and prediction in pose space. In order to compute the RMSE for a particular gesture, we computed the sum of Euclidean distances between each of the consecutively predicted frames and their corresponding ground truth frames and normalized it over the total number of body joints and the total number of frames.

Table 1: RMSE For different gestures at different temporal horizons of motion prediction. (fr. stands for frames.)

Gesture Name	1 fr.	5 fr.	25 fr.	50 fr.	100 fr.
Beat Both Arms	0.0194	0.0201	0.0516	0.0790	0.1208
Bow	0.0141	0.0150	0.0456	0.0952	0.0800
Change Weapon	0.0220	0.0244	0.0557	0.0695	0.1104
Duck	0.0198	0.0218	0.0363	0.0531	0.1077
Goggles	0.0353	0.0764	0.0635	0.0456	0.0459
Had Enough	0.0161	0.0183	0.0298	0.0491	0.0905
Kick	0.0257	0.0239	0.0452	0.0863	0.2545
Push Right	0.0141	0.0169	0.0418	0.0902	0.1247
Shoot	0.0277	0.0308	0.0677	0.0983	0.0795
Start System	0.0475	0.0676	0.0799	0.0783	0.0698
Throw	0.0239	0.0288	0.0858	0.0960	0.1893
Wind It Up	0.0176	0.0190	0.0411	0.0660	0.1004

In Table 1, we list the RMSE errors separately for each gesture in the dataset when forecasted at different temporal horizons. On average, given an initial 50 frames as input, the network is most accurate in forecasting the subsequent pose for the gestures "Bow" and "Push Right." The RMSE error expectedly increases as the network is asked to forecast motions over a longer temporal horizon. While producing an additional 100 future poses when given a ground truth of a sequence of 50 frames, the network is most accurate in predicting future poses for the gestures "Goggles" and "Start System."

4.2 Gesture Recognition

Our gesture recognition model consists of a Random Forest classifier with fifty trees trained to a max depth of five hundred, which gives an optimal classification rate on this dataset. The Gini Index is used as our cost function for node-splitting.

We experimented with two paradigms of motion prediction. First, we assumed that the gesture prediction model has access to only the first fifty frames of the test gesture. The motion predictor model takes fifty frames as input and predicts the next frame in the sequence. In order to predict subsequent future frames, successive early frames were dropped from the sequence after each iteration while adding the predicted frame from the previous iteration. This accounted for the new incoming frames predicted by the motion predictor by sliding this fixed sized window of fifty frames.

In the second setting, we used the same setup with the only exception of iteratively feeding the ground truth poses itself instead of the obtained forecasted poses for predicting successive frames while sliding the fixed sized window. The motivation behind this online setting was the fact that the ground truth pose itself becomes available as we progress through the gesture and capture more data frames from the user performing the gesture.

When the classifier is fed only the first 50 frames of the test gesture, the average classification accuracy is a mere 45%. However, when the partially observed gesture is augmented with an additional 25 frames, classification accuracy improves to 87% (as depicted by the orange curve in Figure 3 Left). The classification accuracy continues to increase, when the input gesture is augmented with more frames from the sequence-to-sequence model, peaking at

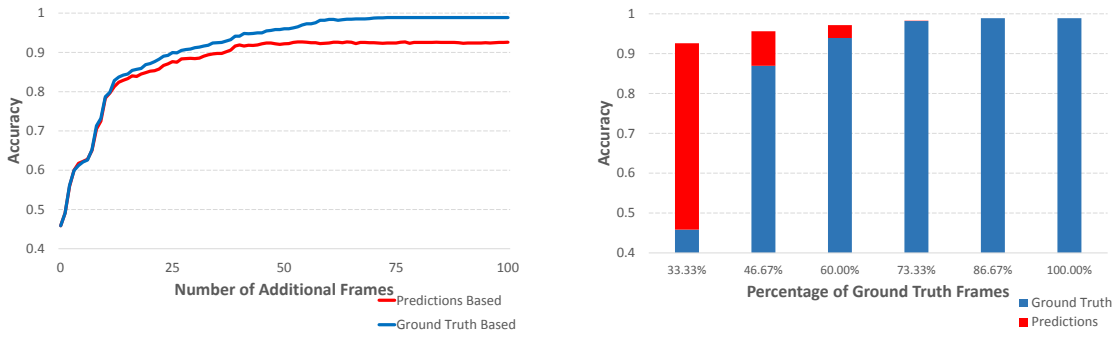

Figure 3: Left - Classifier accuracy plotted against additional frames received as input comparing performance for ground-truth-based predictions vs forecasting-based predictions. Right - Classifier accuracy comparisons based on different percentages of ground truth & augmented predictions (#Total Frames = 150). The orange columns demonstrate the effect of augmenting the input with the forecasted frames.

93% after augmenting hundred successive predictions. When the model is incrementally fed with the ground truth instead of its own predictions, we expectedly observe better maximal performance in the classification accuracy. In this setting, the average classification accuracy approaches 99%.

We also quantified when this method of motion augmentation is most beneficial to the classifier. When only a third of the gesture (50 frames) is observed, the benefit of augmenting the input with forecasted motion is significant as seen in the disparity in classification accuracy (first column in Figure 3 Right). The model achieves 93% accuracy when fed with an additional 100 forecasted frames. In the absence of the forecasted frames, the model only achieves a 45% accuracy. As more of the gesture is observed, the advantage of augmenting future motion wanes. After more than seventy percent of the input gesture is observed, the benefits of motion augmentation to classifier performance is negligible.

5 CONCLUSIONS

In this paper, we proposed a framework for performing early gesture recognition. A partially-observed gesture represented by a series of poses was inputted to a sequence-to-sequence motion forecasting model, which produced a sequence of predicted poses. This sequence of forecasted poses was concatenated with the partially-observed ground truth gesture and fed to a random forest gesture classification model. We showed that augmenting the partially-observed gesture with the output from our sequence-to-sequence model significantly improved recognition accuracy. This can be attributed to the additional signal provided by the pose estimates of the motion forecasting model, which improves the discriminative capacity of the classification model. In experiments with the MSRC-12 gesture recognition dataset, we observed that gesture classification accuracy increased from 45% to 87%, when a partially observed gesture of 50 frames was augmented with an additional 25 frames of predicted motion and to 93% when augmented with 100 frames of predicted motion.

6 ACKNOWLEDGMENTS

This work was supported in part by NSF grants 1337866 and 1551572.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).
[2] Jake K Aggarwal and Lu Xia. 2014. Human activity recognition from 3d data: A review. *Pattern Recognition Letters* 48 (2014), 70–80.
[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[4] Judith Bütepage, Michael Black, Danica Kragic, and Hedvig Kjellström. 2017. Deep representation learning for human motion prediction and classification. *arXiv preprint arXiv:1702.07486* (2017).
[5] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1110–1118.
[6] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. 2012. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1737–1746.
[7] Ajjen Joshi, Camille Monnier, Margrit Betke, and Stan Sclaroff. 2015. A random forest approach to segmenting and classifying gestures. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 1. IEEE, 1–7.
[8] Shugao Ma, Leonid Sigal, and Stan Sclaroff. 2016. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1942–1950.
[9] Sotiris Malassiotis, Niki Aifanti, and Michael G Strintzis. 2002. A gesture recognition system using 3D data. In *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*. IEEE, 190–193.
[10] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. *arXiv preprint arXiv:1705.02445* (2017).
[11] Sushmita Mitra and Tinku Acharya. 2007. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, 3 (2007), 311–324.
[12] Akihiro Mori, Seiichi Uchida, Ryo Kurazume, Rin-ichiro Taniguchi, Tsutomu Hasegawa, and Hiroaki Sakoe. 2006. Early recognition and prediction of gestures. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, Vol. 3. IEEE, 560–563.
[13] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks* 61 (2015), 85–117.
[14] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (2013), 116–124.
[15] Yale Song, David Demirdjian, and Randall Davis. 2011. Multi-signal gesture recognition using temporal smoothing hidden conditional random fields. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 388–393.
[16] Thad Starner and Alex Pentland. 1997. Real-time american sign language recognition from video using hidden markov models. In *Motion-Based Recognition*. Springer, 227–243.