Recognizing Visual Signatures of Spontaneous Head Gestures

Mohit Sharma   Dragan Ahmetovic   László A. Jeni   Kris M. Kitani
Robotics Institute, Carnegie Mellon University
{mohits1, dahmetov, laszlojeni, kkitani}@cs.cmu.edu

Abstract

Head movements are an integral part of human nonverbal communication. As such, the ability to detect various types of head gestures from video is important for robotic systems that need to interact with people or for assistive technologies that may need to detect conversational gestures to aid communication. To this end, we propose a novel Multi-Scale Deep Convolution-LSTM architecture, capable of recognizing short and long term motion patterns found in head gestures, from video data of natural and unconstrained conversations. In particular, our models use Convolutional Neural Networks (CNNs) to learn meaningful representations from short time windows over head motion data. To capture longer term dependencies, we use Recurrent Neural Networks (RNNs) that extract temporal patterns across the output of the CNNs. We compare against classical approaches using discriminative and generative graphical models and show that our model is able to significantly outperform baseline models.

1. Introduction

We address the problem of detecting spontaneous head gestures in everyday interactions to enable intelligent systems to better understand human interaction. Indeed, head movements coordinate speech production [14], regulate turn-taking [6], serve as back-channeling functions [12] and can convey attitudinal and emotional information [8]. Variation in head movement frequency and amplitude can also be an indicator of giving feedback, turn-taking or audiovisual prosody. Thus, head gestures play an invaluable role as a concurrent interaction channel during verbal communication [8]. They are also tied directly to the linguistic structure and perform semantic and communicative functions during speech production [14]. The ability to recognize human head gestures is therefore critical for intelligent systems to understand their human counterparts.

One of the key challenges with vision-based head gesture recognition is the large variation observed in spontaneous head gestures across different users and gesture categories. For example, a person’s nod during natural conversations can be very quick, where a single up-down motion can last less than 100 milliseconds. A nod can also extend over several seconds as a consecutive sequence of repeated nods. Across head gesture classes there can be an even larger difference in temporal scale e.g., a nod might last a second but other gestures like turning or tilting can take several seconds [14]. This implies that any successful detection algorithm must be able to extract the visual signatures of head gestures at multiple temporal scales.

Taking the above challenges into account, we propose a novel multi-scale deep convolution LSTM architecture for the task of spontaneous head gesture recognition. We use the representational power of convolutional neural networks (CNNs) to learn the features of head gesture primitives over multiple scales of the temporal data. Then the multi-scale features generated by each of the CNNs are passed to a two layer Recurrent Neural Network (RNN) with LSTM mod-
ules to model the temporal dependencies over a sequence of CNN features. The outputs of multiple RNN streams are fused and passed to a fully connected softmax layer to output gesture class probabilities. The network architecture is illustrated in Figure 1.

To the best of our knowledge, this is the first work to present state-of-the-art results on the problem of multi-class head gesture recognition, e.g., 5 to 11 categories beyond the basic nod or shake gesture categories [10, 20, 17]. The second contribution is our proposed Multi-Scale Convolution-LSTM architecture for the problem of spontaneous head gesture recognition in natural conversations. We show empirically that our proposed model outperforms the state-of-the-art CRF based models by a large margin on the NAIST Natural Conversation Dataset [21] and Cardiff Conversation Database [2].

2. Related Work

Despite the large role head gestures play in our communication, the problem of head gesture recognition in natural unconstrained environments has received limited attention in the vision community. Although there exists prior work on head gesture recognition [10, 20, 17, 16], most of these approaches are validated in constrained environments and with a small set of head gestures. Also, the size of existing public datasets are limited and thus difficult to use as meaningful benchmarks. Only recently have researchers [21] collected large datasets for head gesture recognition.

Past work in the field of head gesture recognition has been dominated by approaches that use Graphical models such as Conditional Random Fields (CRF) [13] or Hidden Markov Models (HMM). These approaches often involve extraction of manually engineered spatio-temporal descriptors which are then used directly by the classification model.

In [10] the authors use a combination of a pupil tracking system with a HMM based algorithm for real-time head nod and shake detection. In [5] the authors segment the eye using thresholding and then use eye tracking with a HMM classifier. However, one problem with using eye tracking as a feature descriptor is the inability to generalize to environments where the users face or eyes are occluded from the cameras view, a situation that arises often in conversation. Additionally, problems with eye tracking can occur if someone is wearing eyeglasses e.g., people with visual impairments.

Since generative models such as HMM above assume observations to be conditionally independent, they often fail to learn a shared common structure between different classes [20].

To alleviate these problems, conditional models such as CRF [13] have been used extensively for gesture recognition. For example, [20] used Hidden Conditional Random Fields (HCRF) for head gesture recognition. However, [20] uses pre-segmented gestures for training which prevents it from capturing the dynamics between gesture labels.

To capture these inter-gesture dynamics [17] proposed Latent-Dynamic Conditional Random Field (LDCRF). LDCRF uses latent state variables to model the sub-structure of a gesture class which allows it to learn inter-label dynamics. However, to keep inference in LDCRF tractable, the authors assume a disjoint set of hidden states per class label.

Unlike the above approaches, in this work we focus on large datasets of natural unconstrained conversations with a much larger set of gesture categories (5 to 11) [21, 2]. We do not assume input videos to be pre-segmented for our model. Our proposed architecture can also better learn the substructure for each gesture and the inter-gesture dynamics since in our model all the gesture labels are associated with the same set of network parameters.

3. Head Gesture Properties

As noted above, one of the main challenges of head gesture recognition is the large variances observed in spontaneous head gestures. Our aim is to quantitatively observe these variances to identify the main challenges of head gesture recognition. Since in this work we focus on head gesture recognition in everyday interactions, for our analysis we use the recently released FIPCO dataset [21]. The FIPCO dataset is a natural conversation dataset containing 15 hours of recorded video of about 20 participants, with each video frame annotated using a dense 11 gesture categorization. Thus, the large amount of data from different people along with multiple categories make FIPCO a good source to identify the above challenges.

The two main properties of all head gestures include (1) motion: how large is the head movement and (2) duration: how long is the head movement. We show here that these properties can display large variances across different users when comparing the same gesture (inter-person) as well as when comparing different gestures for the same user (intra-person).

**Intra-Person Motion Variance:** Figure 3 shows the mo-
Figure 3. Intra-Person Motion Variance: Gesture motion for a user in FIPCO.

Figure 4. Inter-Person Motion Variance: Gesture motion for two different users for the same set of gestures.

Figure 5. Intra-Person Time Variance: Time variance for different gestures for a user in FIPCO.

Figure 6. Inter-Person Time Variance: Gesture duration for two different users for the same set of gestures.

Figure 3. Intra-Person Motion Variance: Gesture motion for a user in FIPCO.

Figure 4. Inter-Person Motion Variance: Gesture motion for two different users for the same set of gestures.

Figure 5. Intra-Person Time Variance: Time variance for different gestures for a user in FIPCO.

Figure 6. Inter-Person Time Variance: Gesture duration for two different users for the same set of gestures.

Due to this large variance in head gestures in natural conversations, an effective recognition scheme would require appropriate gesture representation at multiple spatio-temporal scales. Taking these challenges into account, we now propose our approach for head gesture recognition.

4. Proposed Approach

Different feature inputs (e.g., head pose, eye gaze etc.) can be used for head gesture recognition. Hence we first discuss our gesture representation scheme. We then propose our Multi-Scale ConvLSTM architecture which uses the above representation from multiple spatio-temporal scales for effective head gesture recognition.

4.1. Gesture Representation

What are the motion features that are most informative for head gesture recognition? Head pose, i.e., the 6 degrees of motion from 3D translation and 3D rotation (roll, pitch, yaw), is a feature that has been used extensively for head gesture recognition [17, 18].

While estimating head pose over time does provide important visual signal for recognizing head gestures, relying on head pose alone might be insufficient for very subtle gestures. For instance, detecting a subtle nod requires very accurate head pose estimation which might not be possible in natural environments with frequent pose changes.

Alternatively, facial landmark motion features can also provide key cues for head movement and have been used for head gesture recognition [1]. Thus, in addition to head pose we explore the use of facial landmarks to provide an alternate modality for recognizing head gestures. Facial landmarks are divided into “fiducial” (primary) landmarks (nose tip, eye corner etc.) and “ancillary” (secondary) landmarks...
Our architecture consists of a Multi-Scale Convolution-LSTM (MS-ConvLSTM) model. At each video frame we extract the above gesture representation at different temporal scales and use them together for final classification. We approximate each feature of the gesture representation using multiple function approximators with radial basis functions. To create new gestures we sample from these approximations by varying the spatio-temporal scale. We cluster gestures together based on their motion and time values. Data augmentation. To address this data imbalance we use data augmentation. Since video data is very high dimensional we use the extracted gesture representations for these augmentations. To create these augmentations we cluster gestures together based on their motion and time values. Since gestures in each cluster will have similar spatio-temporal (motion-time) scale we can now create new gestures by manipulating these scales i.e., by slightly varying the speed of these gestures as well as by varying their motion. We approximate each feature of the gesture representation using multiple function approximators with radial basis functions. To create new gestures we sample from these approximations by varying the spatio-temporal scale. We only vary the scales by a small amount therefore the augmented gesture is similar in structure to the original gesture. **Coarse Gesture Categories:** We now look at the original 11 gesture categories (as defined above) used in FIPCO in more detail. The 11 gesture categories used in FIPCO can

<table>
<thead>
<tr>
<th>Inter annotator agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nod</td>
</tr>
<tr>
<td>0.397</td>
</tr>
<tr>
<td>0.433</td>
</tr>
</tbody>
</table>

Table 1. The inter annotator agreement between three annotators on FIPCO [21]. First row contains raw values, second row contains mean normalized values. [21]
result in ambiguous labelings i.e., although the categories are well defined the occurrences sometimes may be hard to label. As a result, it is hard to gather consistent ground truth labels. It has been reported [21] that the inter annotator agreement is quite low (mean for all gestures less than 0.5, Table 1). Hence in addition to the original 11 categories, we propose an alternative categorization of head gestures using 5 or 7 categories, as shown in Figure 8. Below, we discuss the rationale behind these alternative categorizations.

**Nod, Tick and Jerk:** The difference between these gestures is the initial direction of head movement and periodicity. Both of these are difficult visual cues for human annotators. For example, the difference between one small head nod versus two small nods is almost indiscernible. Similarly, distinguishing between small nod and small jerk requires identifying the onset of a small movement at the beginning of the gesture, which is very difficult to perceive. For these reasons we group Nod, Jerk and Tick into one group.

**Up and Down:** These gestures are defined as keeping pitch up and down for a long duration respectively. Thus the time period while the pitch is either up or down is crucial. However, notice in Figure 6 these gestures can also have a very short duration, this contradiction makes them almost indistinguishable from nods, since a very quick down gesture looks similar to a nod for most human observers. Hence we experiment with two settings (1) merge these two gestures together (7 categories) and (2) merging them with above defined nod (5 categories).

**Shake and Turn:** There exists a dichotomy in the actual labeling and definition of turn in FIPCO, by definition it involves significant amount of head rotation, however in FIPCO a small head rotation coupled with eye gaze changes in the same plane (similar to shake) are also annotated as turn. Hence we compare two settings (1) keeping both gestures separate (7 categories) and (2) merging them together (5 categories).

**Forward and Backward:** These gestures are defined as leaning forward and backward respectively. As seen in Table 1 there exists large confusion among human annotators for these gestures. We believe this confusion exists since human annotators find it hard to precisely annotate the boundaries of these gestures, as they usually occur in sequence. Hence we combine these gestures together in our coarse categories.

Thus in addition to the original 11 gesture categories we also evaluate coarse gesture categories in our experiments.

### 5.2. Cardiff Conversation Database (CCDb)

The Cardiff Conversation Database (CCDb) [2] is a 2D conversation database which contains natural conversations between people. It contains a total of 30 conversations of which only 8 are fully annotated for head motion. The dataset consists of labels for only three gestures Nods, Shakes and Tilts. As before, the dataset consists of less than 100 gestures each for both shakes and tilts. Also, the definition of these gestures are slightly different from FIPCO e.g. Nod in FIPCO consists of one down and up motion while in CCDb it is defined to be a rigid repetitive head motion.

Figure 7 shows some of the video frames in CCDb. Notice the difference from FIPCO videos (Figure 2), in CCDb the camera is very close to the user while in FIPCO the camera is much further away, this causes subtle motion to be exaggerated in CCDb as compared to FIPCO. Given these differences we use both these datasets together to evaluate our proposed architecture.

### 6. Evaluation

We now compare our proposed architecture against several baseline models on the above datasets. We first discuss the experimental setup including the different baseline models, metrics and the hyperparameters used during evaluation. We then observe the results of our proposed architecture and the baseline models on both the datasets.

#### 6.1. Experimental Setup

**Baseline Models:** Following [17] we use the Latent Dynamic Conditional Random Field (LDCRF) and Conditional Random Field (CRF) as our baseline models. We
vary the window size for CRF models between 8 and 32. We also vary the number of parameters between 5 and 30. Similar to previous work [17] we use the BFGS optimizer during training. We use the HCPRF library [15].

Additionally, to show the efficacy of using multiple temporal scales, we compare our approach against a Single-Scale ConvLSTM model (SS-ConvLSTM). SS-ConvLSTM is similar to MS-ConvLSTM but with input limited to one temporal scale i.e., 32. To examine the role of convolution layers we also compare against a pure LSTM model.

**Feature vector:** We use OpenFace [3], a state-of-the-art facial analysis algorithm. It outputs 6DOF head pose and 68 facial landmark positions. Based on the 6DOF signal, we use both position and velocity profiles, which yield a 12 dimension vector. Among facial landmarks we select 5 landmarks from cheek contours, chin tip, nose tip and 1 eye corner. From these 8 landmarks we get 16 features (since both \((x, y)\)). Thus our final feature representation is of size 28.

**Training Hyperparameters:** We use the architecture as described in Section 4.2. We use the Adam optimizer [11] for training and set an initial learning rate of \(10^{-4}\). We use Categorical Cross-Entropy as our loss function. The input to the model is taken by centering at each frame of the input video and taking temporal slices of size 16, 32, and 64.

**Evaluation metrics:** For our evaluation metric we perform dense classification i.e., detect gesture at every video frame and report both F1-score and the weighted F1-scores. Although, both of the metrics are influenced by the skew in the data we believe that collectively they give an appropriate assessment of our approach.

6.2. Comparative Performance Analysis on FIPCO

We now analyze the performance of our proposed architecture against the baseline models on FIPCO. For comparative analysis we use both the original gesture categories as well as the coarse gesture categories.

**Original Category Recognition Performance:** Quantitative results for both baseline models and our proposed architecture are given in Table 2. As seen above, our proposed Multi-Scale ConvLSTM architecture performs the best with a weighted-F1 of 0.751. The LSTM based models (LSTM, SS-ConvLSTM and MS-ConvLSTM) perform better than the CRF models presumably due to the large number of parameters and non-linearities modeled by the underlying LSTM. Also, Conv-LSTM (SS-ConvLSTM and MS-ConvLSTM) models perform better than LSTM which shows that the convolution layers are helpful in extracting a better representation than the raw gesture signals.

Next, we analyze the qualitative performance of our architecture. Figure 9 shows the gesture confusion matrix for the above models. The CRF based models are not able to discover the underlying structure of each gesture as they predict all gestures belonging to one class. Conversely, both Conv-LSTM architectures perform well on most classes and hence are able to discover the general substructure of the gestures. However, our proposed architecture (MS-ConvLSTM) has better performance uniformly across all gesture classes while SS-ConvLSTM is skewed towards certain gestures. In particular, our architecture performs better for *Forward* and *Backward* gestures, both of which are long duration gestures. Also notice that SS-ConvLSTM displays large confusion between *Shake* and *Turn*, both of which involve visually similar head motion but with different duration. Thus we observe that encoding information at multiple scales helps our model perform better.

Additionally, we look at the failure cases for the Conv-LSTM model. Notice the major confusing gesture in Figure 9 is *None*. This confusion is least for our proposed MS-ConvLSTM, however it is seen for all models and across all gestures. We believe this occurs partly because of (1) mislabeling by human annotators and (2) very subtle head motion.
which confuses our models. Another source of confusion is between a small subset of valid gestures, for example between nod and jerk gestures as well as between forward and backward gestures. As discussed above, this confusion also manifests itself among human annotators which highlights the inherent problems of classifying subtle motion.

**Coarse Category Recognition Performance:** We also show the performance for the reduced set of aggregated head gesture categories in Table 2. Notice that our proposed MS-ConvLSTM still performs better than all other models. Also, as expected, the weighted F1 score increases to 0.84 as frequently confused fine-grained categories have been merged together. Notice in Figure 10 that, as we aggregate more gestures the confusion between fine-grained gesture categories disappears.

### 6.3. Comparative Performance Analysis on CCDb

We analyze the performance of our proposed architecture on CCDb in two different settings. First we use CCDb for both training and testing, *i.e.*, we randomly split the dataset into a train and test set and report performance on it. However, since CCDb has a small amount of data compared to FIPCO, we also experiment with transfer learning by training a model on FIPCO and finetuning it on CCDb. Since the definitions of gestures between the two datasets are not identical, we pre-train the model on coarse gesture categories in FIPCO (Figure 8).

**Comparative Performance Analysis:** We compare the quantitative performance of our proposed model against the baseline models in Table 3. As observed before, the ConvLSTM based non-linear models are able to outperform the linear CRF based models. Within the ConvLSTM models we see that extracting multiple temporal scales from the input helps the classifier, as our weighted-F1 increases to 0.840. This indicates that our proposed MS-ConvLSTM model is able to outperform baseline models even for a small set of gesture categories.

We also look at the qualitative results for the above models in Figure 11. As seen above, the CRF based models fail to generalize across different gesture categories as they predict all gestures belonging to a small set of categories. As before, notice that our proposed MS-ConvLSTM performs well across all gesture categories. SS-ConvLSTM is skewed towards certain gestures (nod) while it performs poorly on shakes and tilts. This shows that our proposed MS-ConvLSTM is able to use the multiple temporal scales to better generalize across different gestures.

**Transfer Learning Performance:** We now look at the transfer learning performance when using a pretrained model from FIPCO in Table 3. Notice that our proposed model performs better than SS-ConvLSTM with a weighted-F1 of 0.866. Also notice, in Figure 12, that our finetuned model has almost similar performance when training a model from scratch. Thus, even though the recorded videos in FIPCO and CCDb are quite different, the network is able to learn the appropriate gesture representations and generalize across datasets.

### 6.4. Ablation Study

In this section, we look at the effect of using different temporal scales and feature representations in our proposed architecture. We report all results using the FIPCO [21] dataset, with the same train and test data as used above.

To observe the effect of using convolution layers in our proposed architecture we compare the performance of a LSTM model against the proposed ConvLSTM architectures. Figure 13 shows the confusion matrix for a LSTM model. Notice that, compared to Figure 10, the performance of LSTM on the background class (None) is much worse. Since the background class includes subtle unintended head motion, this shows that the LSTM model fails to differentiate between subtle gestures and unintended head motion. Thus, the convolution layers help in extracting useful repre-
sentations from subtle gestures. Additionally, large confusion exists between gesture groups such as shake and turn as well as nod, jerk and tick. This further shows the inability of an LSTM only model to fully capture the inter-gesture dynamics.

Next, we observe the effect of different input feature representations on recognition performance in Table 5. Notice that just using the 6DOF head pose as input representation performs better than all the facial landmarks only. This shows that head pose provides a better signal for head gesture recognition compared to facial landmarks. However, interestingly, combining head pose and all facial landmarks together seems to perform worse than using head pose only. We believe that this happens because of the curse of dimensionality [19], since the facial landmark features (136 features) are far greater than the head pose features (12 features). This might cause our network to fit to irrelevant noise. Also, notice that, when we combine head pose with a subset of the selected landmarks, i.e., primary and secondary landmarks separately, we get improved performance. Moreover, using secondary landmarks from one face side also improve performance (0.622), but it is outperformed by heuristically chosen landmarks (0.629). This shows that rather than choosing landmarks randomly, selecting a small set of both primary and secondary landmarks leads to a better classification performance.

We also analyze the results using different temporal scales for head gesture recognition. Table 4 shows the results for multiple different architectures with different scales. Among single scale architectures, using a small scale of 16 performs worse than 32 and 64, which indicates that using a longer temporal context provides better gesture recognition. Also, notice that two scale Conv-LSTM architectures perform better than single scale architectures. This is expected since using two scales allows the model to better generalize across gestures as compared to single scale models. However, using a multi-scale architecture which combines information from short, medium and large temporal scales performs better than all the other models.

### 7. Conclusion

In this paper, we address the problem of head gesture recognition in natural conversations. We show that, since head gestures display large inter-person and inter-gesture spatio-temporal variance, head gesture recognition requires appropriate temporal context around each gesture. To extract this temporal context we design a simple and intuitive multi-scale architecture for continuous gesture detection. We test the efficacy of our method by evaluating it on two very different head gesture datasets and observe that our model has significantly better performance than existing methods.
Acknowledgements: This work was sponsored in part by JST CREST grant (JP-MJCR14E1) and NSF NRI grants (1637927, 1208598).

References


