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.

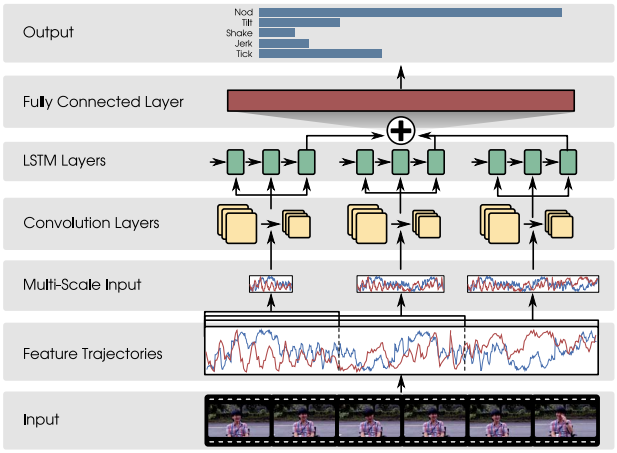


Figure 1. Proposed Multi-Scale ConvLSTM. At each step, take multiple temporal slices from the input, each being passed through Conv-LSTM layers. Concatenate representations from each temporal stream.

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-

978-1-5386-4886-5/18/$31.00 ©2018 IEEE
DOI 10.1109/WACV.2018.00050

Figure 2. Example frames in FIPCO. The users are free to act in any way.

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-

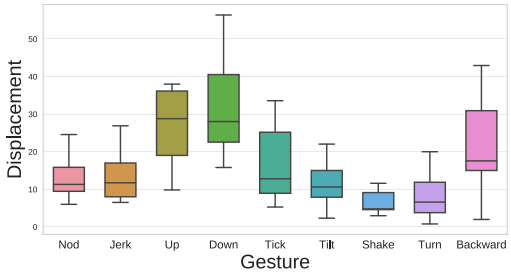


Figure 3. Intra-Person Motion Variance: Gesture motion for a user in FIPCO.

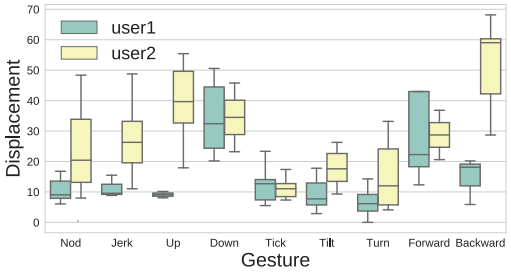


Figure 4. Inter-Person Motion Variance: Gesture motion for two different users for same set of gestures.

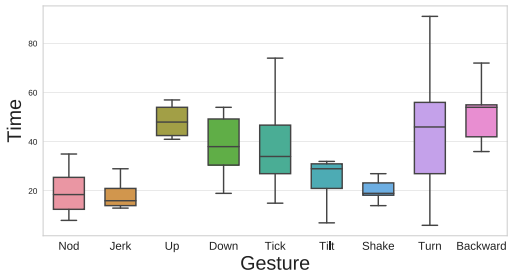


Figure 5. Intra-Person Time Variance: Time variance for different gestures for a user in FIPCO.

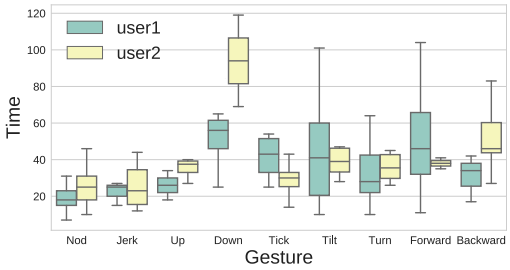


Figure 6. Inter-Person Time Variance: Gesture duration for two different users for same set of gestures.

tion for different gestures for a user. We can notice that there is a large motion difference between different gestures. For example, nods and jerks are much shorter than up and down gestures. Additionally, there is large variance while performing the same gesture. For example, backward and tick gestures can have small motion as well as large motion.

Inter-Person Motion Variance: Figure 4 compares the gesture motion between two different users for the same set of gestures. The above plot shows large amount of variance exhibited by different users while performing similar gesture. This is expected since some people can have larger nods as compared to others.

Intra-Person Time Variance: Figure 5 shows the large temporal variance between different gestures for a given user. It is worth noting that gestures such as backward (mean duration $\approx$ 60 frames) usually last much longer as compared to nod (mean duration $\approx$ 20 frames). We also notice that the same gesture (*e.g.*, turn) can be performed very quickly as well as slowly.

Inter-Person Time Variance: Figure 6 compares the time duration for two different users for a set of gestures. Notice that the same gesture can be performed at varying speeds by different users. As expected some users can have faster gestures (*e.g.*, nod) compared to others.

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

Inter annotator agreement				
Nod	Tilt	Shake	Forward	Mean
0.397	0.383	0.348	0.561	0.430
0.433	0.306	0.337	0.484	0.401

Table 1. The inter annotator agreement between three annotators on FIPCO [21]. First row contains raw values, second row contains mean normalized values. [21]

(chin tip, cheek contours, *etc.*) [4]. Both of these are useful in gesture and facial expression understanding [4], and therefore we select a few landmarks from both of these sets.

Since primary landmarks are easy to track and can be estimated accurately over time we select a small set of these landmarks (nose tip and eye corner). Among the secondary landmarks we select a small set of cheek contour landmarks and the chin tip. Notice that, we choose these landmarks such that we are able to span the entire frontal face *i.e.*, left and right side (cheek contours), bottom (chin tip) and top (nose tip and eye corner). Later, we empirically show that choosing landmarks heuristically is better than selecting all the landmarks as well as randomly selecting a subset of landmarks from one face side.

4.2. Multi-Scale Convolution-LSTM (ConvLSTM)

We now propose our model for head gesture recognition. Since head gestures display large spatio-temporal variance using a local representation for gesture classification is insufficient. Hence we propose our multi-scale deep neural architecture (Figure 1) to extract the gesture context at different temporal scales and use them together for final classification.

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 process each of them independently. Precisely, to classify the frame at time t we extract features from multiple temporal windows that extend from $[t-k, t+k]$ with $k \in \{16, 32, 64\}$.

Each of the above temporal windows is then processed by a separate ConvLSTM layer. The convolution layers in the network are used to learn hierarchical representations for short temporal data streams, while the LSTM [9] layers are used to extract meaningful information from this temporal representation. We then use late fusion to combine the representations from each different scale. Finally, we use a fully connected layer before classification layer.

Architecture Details: Our model is a 3-stream network architecture with inputs at scales 16, 32, and 64. Each stream consists of 3 temporal convolution layers (1-D convolution). For each convolution layer we use, stride 1 and 128 channels. For the 16 and 32 temporal streams we use convolution kernels of size 3, while for 64 temporal stream kernels of size 5. For the LSTM model we use the basic LSTM architecture [7] with 256 nodes. The LSTM outputs from multiple temporal scales are concatenated and passed to a fully-connected layer with 128 nodes. Finally, we have the classification layer with the number of outputs equal to the number of head gestures to classify. In experiments below we look at how each part of the network is useful for classification.

5. Datasets

To evaluate the performance of our proposed architecture on different datasets we use two publicly available datasets. Namely, we use NAIST Natural Conversation Dataset [21] and Cardiff Conversation Database [2] respectively.

5.1. Natural Conversation Dataset (FIPCO)

Recently, a large corpus of data for human face to face conversations has been made available in [21]. The dataset contains annotations for every frame based on 11 head gesture classes. The classes include *None* (background class), *Nod* (head up and down), *Jerk* (head down and up), *Up* (pitch up for a while), *Down* (pitch down for a while), *Tick* (repeated nods), *Tilt*, *Shake*, *Turn*, *Forward* (lean forward), *Backward* (lean backward).

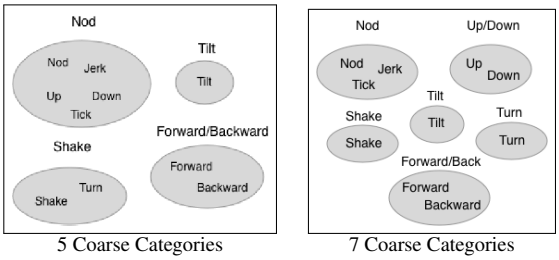
Overall the dataset consists of 30 conversations with close to 15 hours of recorded video data from a wearable and a static tripod camera. We use the video recorded with the tripod camera. Figure 2 shows some of the video frames in the dataset.

Despite the abundance of data there exists a large skew in the data distribution between different classes. In particular, close to 70% of the data belongs to one class, *None*. Also, of the total 5000 annotated gestures, there exist 6 classes which have less than 100 samples in the dataset. This distribution reflects a common feature of data collected from natural conversations – certain target gestures occur very infrequently.

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

Figure 7. Example frames in CCDb. As can be seen CCDb does not capture subtle body motion cues.



5 Coarse Categories 7 Coarse Categories

Figure 8. Coarse gesture category proposals, as described in section 5.1. Both of the above categorization also include *None*.

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) merge 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.

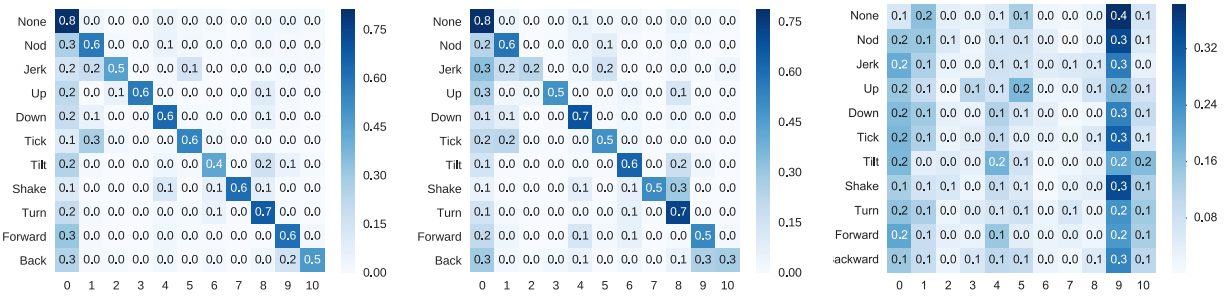
Figure 7 shows some of the video frames in CCDb. Notice the difference from FIPCO videos (Figure 2), in CCDb the camera is very close to the user while in FIPCO the camera is much further away, this causes subtle motion to be exaggerated in CCDb as compared to FIPCO. Given these differences we use both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



MS-ConvLSTM SS-ConvLSTM LDCRF [17]

Figure 9. Confusion matrix for the original 11-class classification on FIPCO [21].

	11 class		5 class	
Model	F1	Wt-F1	F1	Wt-F1
MS-ConvLSTM	**0.542**	**0.751**	**0.629**	**0.779**
SS-ConvLSTM	0.493	0.731	0.612	0.757
LSTM	0.426	0.667	0.571	0.732
LDCRF [17]	0.076	0.081	0.180	0.207
CRF [13]	0.071	0.084	013	0.16

Table 2. Performance on FIPCO [21] with original labels (11 categories) and coarse gesture categories (5 categories).

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 HCRF 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 (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

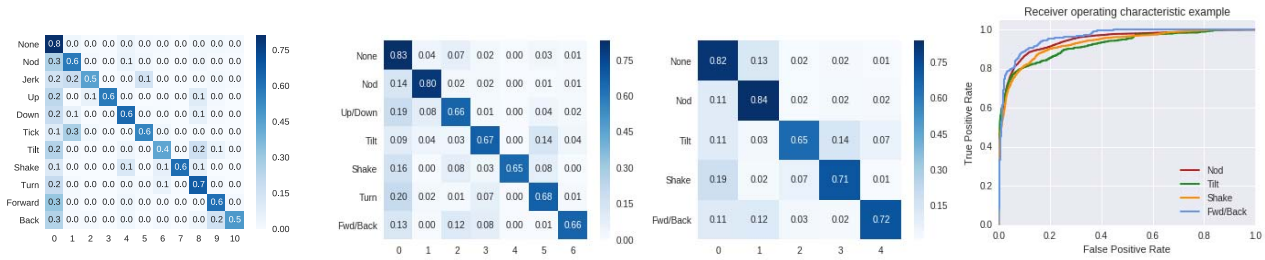


Figure 10. MS-ConvLSTM results on fine to coarse gesture categories on FIPCO. *Right*: ROC curve for 5-class classification using MS-ConvLSTM.

	F1-score	Weighted F1
MS-ConvLSTM	**0.522**	**0.840**
SS-ConvLSTM	0.507	0.820
MS-ConvLSTM - pretrain	**0.528**	**0.866**
SS-ConvLSTM - pretrain	0.493	0.838
LDCRF [17]	0.341	0.370
CRF [20]	0.216	0.338

Table 3. Results for training and evaluation on CCDb.

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.

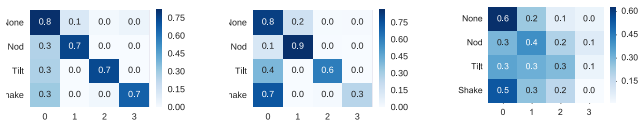
We also look at the qualitative results for the above models in Figure 11. As seen above, the CRF based models fail to generalize across different gesture categories as they predict all gestures belonging to a small set of categories. As before, notice that our proposed MS-ConvLSTM performs well across 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

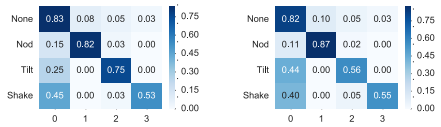
In this section, we look at the effect of using different temporal scales and feature representations in our proposed architecture. We report all results using the FIPCO [21] dataset, with the same train and test data as used above.

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 rep-



MS-ConvLSTM SS-ConvLSTM LDCRF [17]

Figure 11. Confusion Matrix for CCDb dataset [2]



MS-ConvLSTM SS-ConvLSTM

Figure 12. Transfer Learning results with pretraining on FIPCO [21].

Model	Scales	F1-score	Weighted F1
Multi Scale-ConvLSTM	16, 32, 64	0.542	0.751
Single Scale-ConvLSTM	32	0.493	0.731
Single Scale-ConvLSTM	16	0.457	0.679
Single Scale-ConvLSTM	64	0.501	0.734
Two Scale ConvLSTM	16, 32	0.509	0.732
Two Scale ConvLSTM	16, 64	0.515	0.745
Two Scale ConvLSTM	32, 64	0.514	0.732
LSTM	32	0.426	0.667

Table 4. Performance with different models and scales using the original gesture categories in FIPCO.

Features	F1-score	Wt-F1
6DOF Head Pose + Selected Landmarks	0.629	0.779
6DOF Head Pose + All Face Landmarks	0.545	0.756
6DOF Head Pose + Fiducial landmark	0.614	0.760
6DOF Head Pose + Ancillary landmark	0.616	0.772
6DOF Head Pose + 1 face side landmarks	0.622	0.766
6DOF Head Pose Only	0.597	0.750
All Face Landmarks Only	0.547	0.760

Table 5. Performance with different input features for 5 class classification using MS-ConvLSTM architecture.

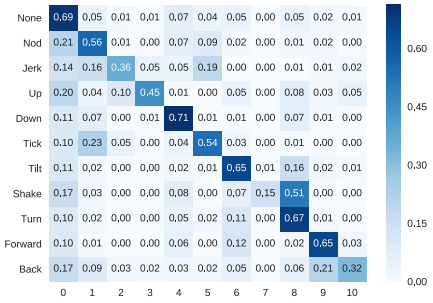


Figure 13. Classification results on FIPCO using LSTM with input temporal scale 32.

resentations from subtle gestures. Also, 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 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, effective 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

[1] H. Ç. Akakın and B. Sankur. Robust classification of face and head gestures in video. *Image and Vision Computing*, 29(7):470–483, 2011. 3

[2] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendeventer, D. W. Cunningham, and C. Wallraven. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282, 2013. 2, 4, 5, 8

[3] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016. 6

[4] O. Çeliktutan, S. Ulukaya, and B. Sankur. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13, Mar 2013. 4

[5] H.-I. Choi and P.-K. Rhee. Head gesture recognition using hmms. *Expert Systems with Applications*, 17(3):213–221, 1999. 2

[6] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972. 1

[7] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 2016. 4

[8] U. Hadar, T. J. Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985. 1

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[10] A. Kapoor and R. Picard. A real-time head nod and shake detector, 2001. In Proceedings from the Workshop on Perspective User Interfaces, November 2001. 2

[11] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[12] M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013. 1

[13] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 2, 6

[14] E. Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of pragmatics*, 32(7):855–878, 2000. 1

[15] L.-P. Morency. HCRF Library, 2007. Available at http://sourceforge.net/projects/hcrf/. 6

[16] L.-P. Morency and T. Darrell. Head gesture recognition in intelligent interfaces: the role of context in improving recognition. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 32–38. ACM, 2006. 2

[17] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 2, 3, 5, 6, 7, 8

[18] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24. ACM, 2005. 3

[19] G. V. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on pattern analysis and machine intelligence*, pages 306–307, 1979. 8

[20] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1521–1527. IEEE, 2006. 2, 7

[21] Y. Wu, K. Akiyama, K. Kitani, L. Jeni, and Y. Mukaigawa. Head gesture recognition in spontaneous human conversations: A benchmark. In *Workshop on Egocentric (First-Person) Vision (CVPR), 2016*, 2016. 2, 4, 5, 6, 7, 8