Pose-based Body Language Recognition for Emotion and Psychiatric Symptom Interpretation

Zhengyuan Yang[1], Amanda Kay[2], Yuncheng Li[3], Wendi Cross[2], and Jiebo Luo[1]
[1]Department of Computer Science, University of Rochester
[2]Department of Psychiatry, University of Rochester Medical Center
[3]Google Inc.

***Abstract*—Inspired by the human ability to infer emotions from body language, we propose an automated framework for body language based emotion recognition starting from regular RGB videos. In collaboration with psychologists, we further extend the framework for psychiatric symptom prediction. Because a specific application domain of the proposed framework may only supply a limited amount of data, the framework is designed to work on a small training set and possess a good transferability. The proposed system in the first stage generates sequences of body language predictions based on human poses estimated from input videos. In the second stage, the predicted sequences are fed into a temporal network for emotion interpretation and psychiatric symptom prediction. We first validate the accuracy and transferability of the proposed body language recognition method on several public action recognition datasets. We then evaluate the framework on a proposed URMC dataset, which consists of conversations between a standardized patient and a behavioral health professional, along with expert annotations of body language, emotions, and potential psychiatric symptoms. The proposed framework outperforms other methods on the URMC dataset.**

I. Introduction

Humans have shown a remarkable ability to infer emotions in face-to-face conversations, and much of the inference is made through body language. For example, "touching one's nose" implies disbelief, and "holding one's head in the hands" expresses upset. It seems a natural ability for humans to understand the "meaning" of body language. To help machines acquire a similar ability, we propose a two-stage framework that predicts emotions based on body language with regular RGB video inputs. In the first stage, the model predicts body language from input videos based on estimated human poses. The predicted body language are then fed into the second stage for emotion interpretation. We define a body language as a certain maintained posture or a period with repeated short actions. It is similar to but different from human actions defined in previous studies [1], [2], which are usually shorter in time and contain dynamic motions. For example, "holding one's head in the hands" is a typical body language and is informative for emotion recognition, but it is not an action according to the definition in action recognition.

Automated body language based emotion recognition is useful in various application domains, such as health care, online chat, and computer-mediated communication [3]. Despite the shared automated body language and emotion recognition techniques in different application domains, the body language and emotion of interest vary. For example, online chatting systems are concerned with detecting people's moods, *i.e.*, whether they are happy or not, while applications in healthcare scenarios focus on identifying potential signs of mental disorders such as depression or panic attacks. Since a specific emotion can only be reflected by the corresponding body language[1], different applications require the annotation of different body language and emotions. Annotating videos for each application potentially lead to a high annotation cost. To alleviate the data annotation problem, we design our framework to learn from a small training set together with the expert knowledge, instead of being purely data-driven.

Specifically, we improve the method's transferability and reduce the required amount of annotations by 1). using the abstracted human poses as the framework input, 2). proposing a KNN based approach for body language recognition, and 3). conducting emotion recognition purely based on the predicted body language sequences. In the first-stage body language recognition, instead of directly predicting body language from RGB videos [4], [5], we use human poses as the input for body language recognition. Human poses are groups of human joint coordinates that provide abstracted high-level human structural information. Because of the abstractness, pose-based methods require less training data and have better transferrabilities, as proved empirically in our study. The recently avabile robust pose estimation methods [6], [7], [8] make our pose-based attempt more feasible. We adopt Openpose [6] for pose estimation, and propose a Spatio-Temporal Convolutional Pose Feature (ST-ConvPose) to encode the spatial-temporal relationships among the joints. With the learned pose representation, we propose a K-Nearest-Neighbors-based classifier for body language recognition. In the second-stage emotion recognition, the model predicts the emotion based on the predicted body language sequences.

Furthermore, as a concrete example of application, we adopt the proposed framework to help psychiatrists understand psychiatric symptoms from patients' body language and emotions. In collaboration with psychologists, we collected the URMC dataset under the mental healthcare scenario. The recorded videos are the conversations between a standardized patient and a psychiatrist or psychologist. Health experts define the

[1]Examples of body language and their meanings can be found in this website. https://www.enkiverywell.com/body-language-examples.html

body language, emotion, and psychiatric symptom classes adopted in this study. Multiple psychologists annotate the recorded videos in the URMC dataset with the defined classes. Experiments on the URMC dataset prove the effectiveness of the proposed framework.

Our main contributions are two-fold:

- We propose a framework to infer emotions from the body language. We design the framework to be interpretable and transferrable.
- We adopt the proposed framework to help mental health professionals predict psychiatric symptoms. A new URMC dataset is built for the study.

II. Related Work

Human action recognition from videos is an important area in computer vision. Most of the RGB video based action recognition studies [4], [9], [5] start from low-level features. Intuitively though, high-level features such as human joints should be informative and beneficial for boosting the recognition accuracy. Joint-based action recognition [10], [11], [12], [13], [14] acquires reliable 3D joints with Kinect or similar RGB-D sensors and generates action predictions based on the joint sequences. However, one inherent problem in applications is the cost and inconvenience of depth sensors. Furthermore, depth sensors are difficult to integrate into the application system.

To better capture human-related features, Jhuang et al. [15] propose the scheme of using both RGB videos and poses for action recognition. Pose-CNN [16] proposes to use joint information to generate body parts sub-images, which are fed into a two-stream network for action recognition. Studies in JHMDB [15] and Pose-CNN [16] suggest that high-level information is rewarding for action recognition, while subject to the limited performance of pose estimation. The recent improvements in human pose estimation [6], [17] make pose-based recognition feasible and attractive. Several recent studies [18], [19], [20], [21] propose various further improvements on the RGB+Pose task. We conduct body language prediction from RGB videos with poses calculated by OpenPose [6], [7].

This study is also related to emotion recognition [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Several previous studies [32], [33] propose to detect psychological stresses with multi-modal contents. Xu et al. [34] recognize effects with body movements. Furthermore, this paper shares a similar goal with several previous studies on predicting human emotion from body motions and facial expressions. Traditional studies [35], [36] focus on analyzing facial expressions for emotion interpretation. De et al. [37] first propose to adopt gestures as a channel for emotion interpretation, and a motion capture system is built to record children's behaviors in the scenario of playing network games. Gunes et al. and Shan et al. [38], [39] propose to extract multi-modality information from videos for emotion recognition, which is more closely related to this paper. Shan et al. [39] propose to fuse body gestures and facial expressions with Canonical Correlation Analysis (CCA). Furthermore, Gunes et al. [38] combine features from facial expressions, hand poses, body parts location, and orientation with a late fusion method. Although previous studies have achieved good performances on a bi-modal face and body gesture database (FABO) [40], there are problems in real applications since the dataset is collected under ideal conditions. FABO contains face and body images in a front view with no body part occlusion, which simplifies the problem.

III. Methodology

In this study, we propose a framework that recognizes body language and human emotions with a small amount of data. The input of the framework is a long sequence of estimated human poses $p = \{p_t\}$ where $t \in 0, \ldots, T-1$ is the length of an untrimmed video. The framework contains two stages: the pose-based body language recognition stage and the body language based emotion interpretation stage. In the first stage, the model takes the pose sequence p as the input and outputs two body language sequences that represent the upper- and lower-body body language, respectively. The classifier is built with an easily-transferrable pose feature and an example based classifier, instead of being purely data-driven. The second stage takes the two predicted body language sequences as inputs and learns the emotions expressed by the person of interest. The framework is shown in Figure 1.

In this section, we first discuss the pose estimation and pre-processing methods, together with several previous high-level pose feature representations, *i.e.*, the *NTraj* and *NTraj+* [15]. Then we introduce the proposed Spatial-Temporal Convolutional Pose Feature (ST-ConvPose). We also discuss the methods for processing incorrect or missing pose estimations. After getting the pose feature representation, we predict the body language with an example-based classifier. Finally, we propose a temporal network for emotion interpretation based on the detected body language sequences.

A. Pose Feature Generation

The body language prediction stage uses only the poses as inputs to achieve the desired transferability and to reduce the required amount of data. We adopt a reliable pose estimator called OpenPose [6], [7] to generate 2D human poses from RGB videos. We then encode the estimated poses as the high-level pose features for body language recognition. In this section, we first introduce the pose pre-processing techniques and two previous pose features [15]. The proposed pose feature is introduced in the next section.

Pre-processing. For each frame, we calculate 18-point human poses with OpenPose. In the pre-processing step, we first normalize the predicted poses with the torso size. In previous works, a puppet mask is used to estimate the torso size, but it is not available in our task. Therefore, we propose to normalize the poses by the distance between the neck joint and the center of two hip joints, such that the length is fixed to 240 pixels. After the size normalization, all joints are "centered" with a reference point. The reference point is defined as the averaged neck joints in several adjacent frames.

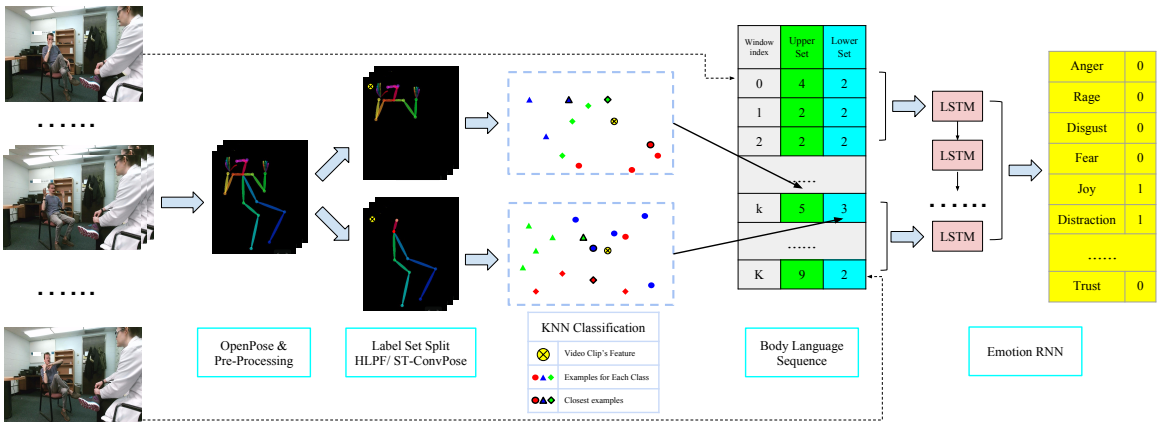

Fig. 1: Overview of the proposed framework. This is a two-stage framework for body language prediction and emotion interpretation. The first stage generates sequences of body language predictions with encoded pose features and a KNN-based classifier. The second stage predicts emotions from the predicted body language sequences.

NTraj. The *NTraj* feature [15] is one of the most basic high-level pose features, which only represents the joint position information. Five features are encoded for each joint. The centered x- and y- coordinates are selected as the first two dimensions that represent spatial structures. The differences in pixel locations between two frames are encoded to represent the temporal motion. The direction of motion $arctan(\frac{dy}{dx})$ is also included. Furthermore, the motion is calculated with a temporal gap s, i.e. $dx = x_{t+s} - x_t, dy = y_{t+s} - y_t$, which helps eliminate jitters and provide reliable motion. The gap lengths of $s = 1, 2, 3$ are computed, and the trajectory length T is set to $T = 5$ based on experiments. Finally, each dimension in the calculated *NTraj* features is normalized by the absolute sum value on a training set, i.e. $\frac{(F_i^t, \cdots, F_i^{t+T-1})}{\sum_{j=t}^{t+T-1} \|F_i^j\|}$. The normalization is calculated 5 times for each feature $F_i^j = \{x^j, y^j, dx^j, dy^j, arctan(\frac{dy}{dx})^j\}$, where j represents the timestamps.

NTraj+. Besides the 5 kinds of position-based features in *NTraj* calculated at each joint independently, the *NTraj+* feature further includes the relationship information among different joints. The orientations between every two joints and the inner angles of all permutations in a three-joint group are encoded to represent the relationship information.

The bag-of-features is calculated to encode the pose feature representations. For each feature type, a codebook of size N is formed by running k-means 10 times on all features available in a training set and pick the one with the lowest error. With a small feature dimensionality, the codebook size N is tested among $N = 10, 20, 50, 100, 200, 500$.

B. ST-ConvPose Feature

In previous high-level pose features, the spatial relations among joints are encoded with pre-defined features such as body part lengths, orientations, inner angles, and so on. Many RGB+Pose action recognition studies [16], [18] also only use the pre-defined spatial relation information among joints that can not be adaptively learned. Inspired by the recent studies on 3D skeleton representations [10], [41], [42], we propose a spatial-temporal convolutional pose feature (ST-ConvPose) that learns the spatial and temporal relations among joints simultaneously with 2D CNNs. In the proposed feature, the coordinates of poses are arranged as 2D matrices where each row is the chaining of joint coordinates at time-stamp t, and the column lists the chains in all timestamps. To arrange the joints in each row, we experiment with three different orders proposed by J-HMDB [15], PennAction [43], and NTU RGB+D [44]. The order defined in NTU RGB+D works the best on action recognition datasets. Therefore, we arrange the joints in each row following this order that joints in each body part are grouped first and then chained from upper left to lower right. To be specific, the left shoulder to the left wrist are placed in column 2 to 4, the right shoulder to the right wrist are in column 5 to 7, and so on. The encoded 2D matrices are then scaled into 0 to 255 and resized to a fixed width and height. The formatted multi-channel 2D matrices are referred to as pose images. Examples of the generated pose images are shown in Figure 2 (a). The pose images are then fed into CNNs to generate human activity representation in an end-to-end manner. The ResNet-50 [45] is used as the CNN structure for high-level pose feature encoding. The feature encoding structure is shown in Figure 2 (b).

With the ability to jointly learn the spatio-temporal joint relations, the proposed ST-ConvPose features show a better performance in modeling human activities compared to previous high-level pose features. Furthermore, the ST-ConvPose feature can be trained with a limited amount of training data and possesses a good transferability. With these desired properties provided by pose features, we do not include RGB frames as inputs to the framework and learn purely based on estimated poses. More details are discussed in the experiment section.

C. Body Language Sequence Prediction

The proposed body language recognition stage is designed to work on untrimmed videos, and a fixed-length sliding

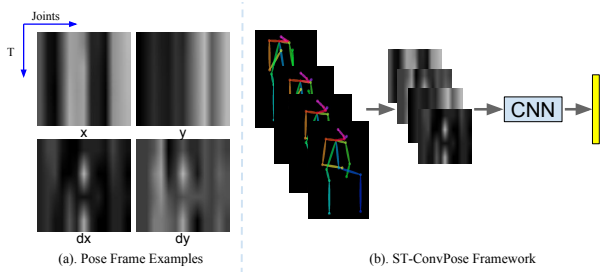

Fig. 2: (a). Pose image examples. (b). The framework for the spatio-temporal convolutional pose feature (ST-ConvPose).

window is adopted to convert the localization and recognition problem into a regular classification task.

In practice, different body parts could perform different body language simultaneously, and the required joints for recognition are different. For example, people might keep a catapult posture while having their legs crossed. Because of this, we divide body language into two label sets and predict them separately. The lower body pose set contains leg postures, including locking ankles, leg crossed, walking around, and so on. The upper body pose set includes both arm motions and hand movements. A background class is added to both pose sets. Full label name list is shown in Figure 3. The split of the labeling set not only solves the multi-label problem but also provides classifiers the prior knowledge on which groups of joints should be focused on to recognize a specific body language. Joints on the torso and both legs are used for lower-body body language recognition, and the upper body recognition task takes the joints on the arms and head as input.

After obtaining the feature representations with either *NTraj+* or ST-ConvPose, we adopt a k-nearest neighbor classifier with manually selected label examples. For each body language, we manually select five to seven video clips and use their pose features as the KNN data points. Classification is conducted on each video clip under the sliding window. The output of the model is the two sequences of upper and lower body language predictions. The body language prediction framework is shown in the left part of Figure 1.

D. Emotion Interpretation

After getting the body language sequence predictions, we propose an emotion RNN network for emotion interpretation as the second stage of the framework. We predict the body language sequences for all video clips in a dataset. Each output contains the lower- and upper-body body language prediction sequences with length K, where K is the number of sliding windows in an input video. We apply another sliding window on the predicted body language sequences and calculate the body language histogram under each window. The histogram sequences are fed into an LSTM network in the temporal order. The LSTM outputs at all timestamps are fed into another dense layer for the final video-level emotion prediction. The output is an N-hot vector representing the predicted emotions.

We have also experimented with the end-to-end emotion recognition model. We replace the KNN classifier with dense layers and teach the model to simultaneously predict body language and emotions. Unsurprisingly, due to the limited size of the available emotion labels, the end-to-end framework fails to learn effectively.

IV. The URMC Dataset

Among the various applications of the proposed body language and emotion recognition framework, psychotherapy is an area especially suitable for model evaluation. There exist solid and complete theoretical proofs in the medical field for the emotions of interest and the corresponding body language. Therefore, the proposed framework is employed to predict psychiatric symptoms. We construct a new URMC dataset under realistic simulated mental health care scenarios with expert annotations. A typical scenario involves the conversation between a standardized patient and a psychiatrist, with an initial intention to infer the standardized patient's potential symptoms by analyzing their body language. There are 144 30-second long video segments cropped from 12 20-minute videos.

A. Dataset Collection

Scene Recording. We collaborated with psychologists to collect 12 videos recording the diagnostic sessions between standardized patients (highly trained actors) and mental health professionals, with the view focused on the standardized patients (SPs). The mental health professionals complete a survey after each diagnostic session to control the quality of the dataset. Positive feedbacks on questions, including the confidence level of diagnosis and other session-related questions, prove the quality of the recordings. All participants in the recordings provided permission for the data collection for research. Both RGB videos and conversation audios are recorded, and poses are later estimated using RGB videos. Example frames and estimated poses are shown in Figure 1. Four standardized patients and eight mental health professionals (psychologists, psychiatrists) are involved. For the 12 videos, either the standardized patient or the psychiatrist is different. Also, the subjects' clothing is different in all videos.

Dataset Labeling. Three mental health professionals are involved in the dataset labeling process. In the first step, three mental health professionals go through each of the 12 recorded long videos, remove less informative segments, select 12 30-second video segments, and then make the symptom decisions. Removing the less informative segments can reduce the labeling cost, and also reduce the noises in the dataset. After getting the 144 video segments, three mental health experts label the video clips with discernible body language and symptoms. Each video clip is labeled by all three health professionals, and conflicts are resolved in consensus discussions. Both the video recordings and the conversation audios are used for labeling, as certain symptoms can only be reflected from the content of the conversation. It is worth mentioning that instead of generating body language and emotion labels based on previous studies in the computer science field, the labels for annotation in our dataset are designed by the mental health

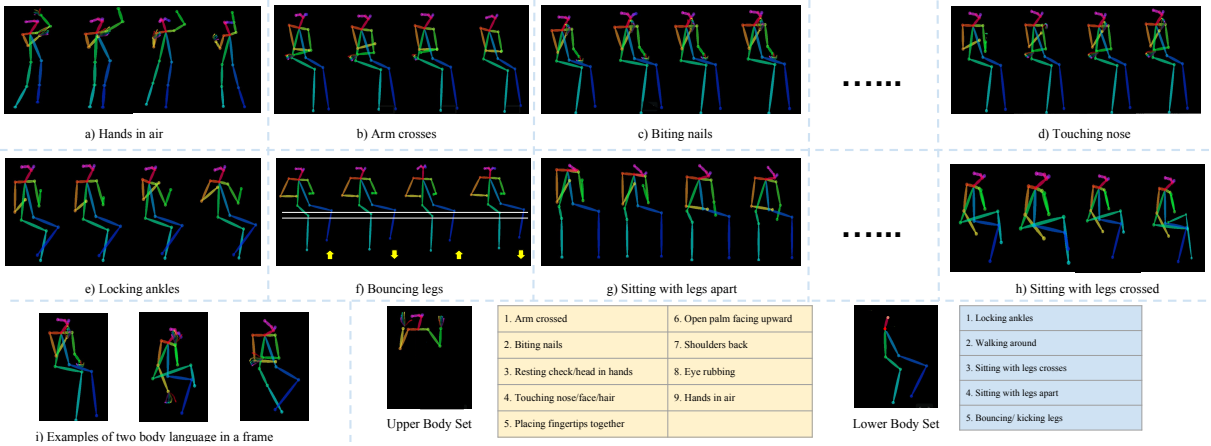

Fig. 3: Examples from the URMC dataset. First row: examples from the upper body set. Second row: examples from the lower body set. Third row: examples of two body language appearing simultaneously. The complete label sets are also included.

experts with evidence in the medical field. Based on the psychiatry theories, 32 body language, 24 emotions, and 24 symptoms are selected for labeling. A final psychiatric verdict of whether the standardized patient has the major depressive disorder (MDD) or manic episode (ME) is also included. This dataset is built to help inferring emotions and psychiatric symptoms from body language.

The 32 body language includes *arm crossed, finger tapping, ear pulling and etc.* Once a body language appears in a 30-second video, the video clip is marked with that label. Because certain body language rarely appear and several of them are extremely similar, we merge a number of body language classes. The final 14 body language classes are shown in Figure 3, which are *arm crossed, biting nails, and etc*. Furthermore, a subset of the videos clips has frame-level body language labels that are annotated under sliding windows with a length of 6 frames.

B. Dataset Summary

There are 144 30-second video clips cropped from 12 20-minute videos. Each clip has multiple labels from the 14 body language, 24+1 emotion labels, and 24+1 symptom labels. Among the 144 videos, 48 videos are selected for training and have frame-level body language labels. Additionally, 48 videos are for validation, and 48 videos are for testing. The split follows a cross-subject setting. We will release the collected and processed poses.

V. Experiments

In this section, we first experiment with the proposed ST-ConvPose feature. We then evaluate the proposed body language and emotion recognition framework on the URMC dataset. Section V-C presents the results of body language prediction, and Section V-D shows the performance of emotion recognition and psychiatric symptom prediction.

A. ST-ConvPose Feature Evaluation

Because there lack public benchmarks on body language recognition, we evaluate the proposed ST-ConvPose feature on a similar task of action recognition and compare our method to state-of-the-art methods. It is important to note that although body language, as defined earlier, *are different from actions*, it is the best comparison we could perform due to the lack of existing body language datasets. Experiments prove that the proposed ST-ConvPose feature outperforms both RGB-based and pose-based state-of-the-art with only the pose information. Furthermore, experiments show that the ST-ConvPose feature achieves the following two desired properties:

1) The pose feature generates good results when the amount of training data is limited.
2) The pose feature captures more dataset-invariant human representations instead of learning the dataset bias, thus has a better transferability.

We detail the analyses as follows.

Action Recognition Results. We evaluate the proposed ST-ConvPose feature on the PennAction [43] and UCF-Motion [14] datasets. PennAction contains 1,212 videos for training and 1,020 videos for testing, with 2D full body joints manually labeled on each frame. Half of the training videos are used for training, and the rest is the validation set. UCF-Motion dataset extends UCF-101 [2] by including estimated poses [46] on all video frames. UCF-Motion contains 23 classes from UCF-101 with 3172 videos in total. Table I shows the action recognition accuracy on the PennAction dataset. Our ST-ConvPose feature outperforms both the RGB-based and the pose-based state-of-the-art on the PennAction dataset. It is gratifying even when compared with the methods using both RGB and pose information, the proposed pose feature can achieve comparable results. The experiments show the effectiveness of the proposed ST-ConvPose feature and prove that poses can adequately represent the information needed to distinguish human actions. A similar improvement is also observed on UCF-Motion, as shown in Table II.

Training with Less Data. Our method also performs well with a limited amount of training data. We design the experiments by sampling different amounts of training data from

TABLE I: Recognition accuracy compared to the state-of-the-art on PennAction. The input data format is also shown.

State-of-the-art	Acc.	Pose	RGB
C3D [5]	86.0	-	✓
idt-fv [20]	92.0	-	✓
NTraj+ [20]	79.0	✓	-
JDD [18]	87.4	✓	✓
Pose+idt-fv [20]	92.9	✓	✓
ST-ConvPose	**94.4**	✓	-

TABLE II: The action recognition accuracy compared to the state-of-the-art methods on the UCF-Motion dataset.

State-of-the-art	Acc.	Pose	RGB	Flow
HLPF [15]	71.4	✓	-	-
C3D [5]	75.2	-	✓	-
Flow CNN [47]	85.1	-	-	✓
ST-ConvPose	**88.1**	✓	-	-

PennAction. Starting from using all the training data, we train the model with 50% and 20% of the training set. All 1,020 testing videos are used for testing. The training and validation sets are still split. As shown in Table III, when 20% of the training data is used, the performance of RGB-based methods drop $(86.0\% - 65.0\%)/86.0\% = 24.4\%$, while ST-Convpose only drops 15.7%. The performance drop is even smaller when ST-Convpose is pretrained on NTU RGB+D [44] of 5.2%. This experiment shows that the ST-ConvPose works well with a limited amount of training data.

Domain Transfer. Furthermore, pose feature captures dataset-invariant action representations and thus has a better ability to transfer across datasets, compared to RGB-based methods. We adopt the ST-ConvPose feature trained on NTU RGB+D with projected x and y coordinates for action recognition on the PennAction dataset with or without fine-tuning.

As shown in Table III, the ST-ConvPose feature achieves an 82.4% accuracy using the pose features pretrained on NTU RGB+D without fine-tuning, which is already better than the 79.0% accuracy generated by previous $NTraj+$ pose feature. The high recognition accuracy proves that the pose feature captures invariant information to represent actions and has an excellent ability for domain transfer even without fine-tuning. The results can be further improved with fine-tuning.

In this experiment, we show that the proposed ST-ConvPose feature outperforms the state-of-the-art action recognition methods on PennAction and UCF-Motion. Furthermore, it requires less training data and has good transferability.

B. Evaluation Settings on the URMC Dataset

We evaluate the proposed body language and emotion recognition framework on the URMC dataset. For the first stage of the framework, we treat the body language sequence prediction as a video-level multi-label classification task. In the experiment, an N-hot vector is generated from the predicted body language sequence and compared with the video-level ground truth body language labels. Metrics for multi-label classification are used for evaluation, *i.e.*, multi-label accuracy, precision, recall, and F1-score.

TABLE III: Experiments on training with less data. $100\%, 50\%, 20\%$ are the percentage of used training data. *"NTU Pretrain"* indicates that the ST-ConvPose feature is pretrained on NTU RGB+D. *"Drop"* is the accuracy decrease percentage when using 20% training data compared to using all data.

State-of-the-art	100%	50%	20%	$Drop\%$
C3D [5]	86.0	74.1	65.0	24.4
ST-ConvPose	100%	50%	20%	$Drop\%$
ST-ConvPose from Scratch	94.4	84.3	79.6	15.7
NTU Pretrain. w/o Fine-tune	82.4	77.6	70.0	15.0
NTU Pretrain. with Fine-tune	95.4	93.2	90.4	**5.2**

The second stage of the framework is emotion interpretation and symptom prediction. For emotion interpretation, we compare the proposed temporal network with other approaches as a multi-label classification task. Instead of predicting all 24 labeled psychiatric symptoms, the model only learns to distinguish major symptoms of Major Depressive Disorder (MDD) with Manic Episode (ME). As mentioned in Section IV-A, certain symptoms can only be reflected with other modalities, such as the audio track. Although predicting detailed psychiatric symptoms is a very interesting problem of great importance, we focus on a more feasible task of inferring emotions and major psychiatric symptoms in this study.

C. Body Language Recognition

Table IV shows the body language recognition accuracy. Other than the accuracy metrics, we also indicate the models' interpretability, transferability, and the required amount of training data. We first compare the model with two end-to-end methods: the two-stream action recognition and the proposed ST-ConvPose feature trained end-to-end with dense layers. The two-stream action recognition is pretrained on UCF101 [2] and HMDB51 [1]. Since both methods are trained end-to-end, they lack the interpretability. The ST-ConvPose feature requires less data since it is trained from pose information instead of RGB frames. Furthermore, the ST-ConvPose feature provides a good ability for domain transfer. For SVM, we adopt the one-vs-one approach for multi-class classification with linear kernels. Overall, the KNN classifier provides interpretable results and requires less data for training. The proposed ST-ConvPose feature provides the desired transferability and further reduces the required amount of training data.

In Table IV, $NTraj^+$ + SVM and ST-ConvPose + Dense layers have the lowest recognition accuracy because of the limited training data size of 48 video clips. Although the two-stream network requires even more training data, the model is transferred from the one pretrained on UCF101 and HMDB51. In contrast, our KNN based methods have the best performance with a small training set. Overall, our proposed ST-ConvPose and the KNN classifier has the best recognition performance.

D. Emotion and Symptom Prediction

The ultimate goal of the framework is to understand emotions and infer psychiatric symptoms from human body language. We conduct emotion recognition purely based on

TABLE IV: Multi-label body language recognition performance compared with different approaches.

Lower Body Set	Acc.	Prec.	Recall	F1	Interpretability	transferability	Required Data Size
Two Stream	0.445	**0.581**	0.497	0.526			Large
$NTraj^+$+SVM	0.327	0.336	**0.690**	0.424			Medium
$NTraj^+$+KNN	0.483	0.516	0.616	0.538	✓		Small
ST-ConvPose+Dense	0.384	0.397	0.606	0.460		✓	Medium
ST-ConvPose+KNN	**0.488**	0.520	0.658	**0.554**	✓	✓	Small
Upper Body Set	Acc.	Prec.	Recall	F1	Interpretability	transferability	Required Data Size
Two Stream	0.341	0.472	0.567	0.492			Large
$NTraj^+$+SVM	0.346	0.388	**0.766**	0.485			Medium
$NTraj^+$+KNN	0.398	0.504	0.578	0.502	✓		Small
ST-ConvPose+Dense	0.374	**0.522**	0.486	0.473		✓	Medium
ST-ConvPose+KNN	**0.400**	0.497	0.641	**0.519**	✓	✓	Small

TABLE V: The performance of emotion interpretation by different learning methods and body language sequences. L is the length of the sliding window and S is the stride of the window.

LSTM+ST-ConvPose	Acc.	Prec.	Recall	F1
$L = 1$, $S = 1$	0.468	0.644	0.630	0.637
$L = 7$, $S = 3$	**0.564**	0.775	**0.674**	**0.721**
$L = 48$	0.145	0.162	0.587	0.254
Other Methods	Acc.	Prec.	Recall	F1
LSTM+$NTraj^+$	0.510	**0.839**	0.565	0.675
Conv 1D+$NTraj^+$	0.490	0.788	0.565	0.658
Conv 1D+ST-ConvPose	0.556	0.789	0.652	0.714

the predicted body language sequence to achieve the desired transferability. A LSTM-based temporal network is proposed for the task. The output of the network is an N-hot vector with a length of 25, representing the 24 labeled emotions and a background class. For baseline comparisons, we implement a network with a 1D convolution for temporal information learning. The idea of adopting a convolutional layer for sequence learning is inspired by works in natural language processing, where CNN are used for sentence analysis [48], [49]. The body language sequences predicted by the proposed ST-ConvPose feature is also compared with the one predicted with *NTraj+*. We also consider an end-to-end design of the framework. However, the result is not promising due to the limited amount of body language labels and emotion labels.

The experiment results on emotion interpretation are shown in Table V. Two special cases are worth noting. First, when the sliding window length and stride both equal to one, the predicted vector sequences are directly fed into the temporal network without forming a histogram. Second, when the window length is the entire video, a video-level body language histogram is calculated with no temporal information used, similar to directly adopting a dense layer at the top. The limited performance of the two special cases proves the effectiveness of the proposed temporal structure and histogram features. The experiment results in Table V also indicate that the LSTM structure has a better ability in representing the temporal information in body language sequences, compared to 1D convolutional neural networks. Additionally, using the proposed ST-ConvPose feature generates a better overall performance compared to previous high-level pose features.

Furthermore, we try to infer the psychiatric symptoms based on the body language predictions. The result of the binary classification between Major Depressive Disorder (MDD) or Manic Episode (ME) is promising. We achieve an accuracy of 90.3% with the ground truth body language sequences and 79.9% with the predicted sequences.

VI. Conclusion and Future Work

Teaching machines to recognize human emotions from body language is a challenging but useful task that has various potential applications. In this paper, we propose a framework starting from regular RGB videos for body language prediction and emotion interpretation. Aiming at building a system capable of easy transfer between different application scenarios and producing interpretable results, we design a body language recognition system with the ST-ConvPose feature and the KNN classifier. The proposed pose feature shows good performances on both public datasets and the new URMC dataset, while requiring less training data and showing good transferability. Finally, we show that emotions or psychiatric symptoms can be predicted reliably from recognized human body language. Next on our agenda is to apply the framework to more domains and continue exploring detailed psychiatric symptom detection based on body language.

VII. Acknowledgement

This work is partially supported by NSF award #17228477, and the Morris K.Udall Center of Excellence in Parkinson's Disease Research by NIH.

References

[1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2556–2563.

[2] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[3] C. Beyan, V.-M. Katsageorgiou, and V. Murino, "Moving as a leader: Detecting emergent leadership in small groups using body pose," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1425–1433.

[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[7] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[8] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[10] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 4570–4579.

[11] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018, pp. 7444–7452.

[13] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with visual attention on skeleton images," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3309–3314.

[14] ——, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[15] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[16] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3218–3226.

[17] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *CVPR*, vol. 3, no. 4, 2017, p. 6.

[18] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3d deep convolutional descriptors." in *IJCAI*, 2016, pp. 3324–3330.

[19] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2923–2932.

[20] U. Iqbal, M. Garbade, and J. Gall, "Pose for action-action for pose," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 438–445.

[21] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3725–3734.

[22] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[24] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.

[25] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark." in *AAAI*, 2016, pp. 308–314.

[26] T. Rao, M. Xu, and D. Xu, "Learning multi-level deep representations for image emotion classification," *CoRR*, vol. abs/1611.07145, 2016.

[27] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 890–897.

[28] S. Zhao, G. Ding, Y. Gao, and J. Han, "Learning visual emotion distributions via multi-modal features fusion," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 369–377.

[29] C. Chen, Z. Wu, and Y.-G. Jiang, "Emotion in context: Deep semantic feature fusion for video emotion recognition," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 127–131.

[30] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 83–92.

[31] Z. Yang, Y. Zhang, and J. Luo, "Human-centered emotion recognition in animated gifs," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1090–1095.

[32] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 507–516.

[33] M. Sidorov and W. Minker, "Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 81–86.

[34] J. Xu and S. Sakazawa, "Temporal fusion approach using segment weight for affect recognition from body movements," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 833–836.

[35] S. Gong, C. Shan, and T. Xiang, "Visual inference of human emotion and behaviour," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 22–29.

[36] Y. Tang, "Deep learning using linear support vector machines," *CoRR*, vol. abs/1306.0239, 2013.

[37] P. R. De Silva, M. Osano, A. Marasinghe, and A. P. Madurapperuma, "Towards recognizing emotion with affective dimensions through body gestures," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 269–274.

[38] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[39] C. Shan, S. Gong, and P. W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures." in *BMVC*, 2007, pp. 1–10.

[40] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 1148–1153.

[41] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[42] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *e Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[43] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2248–2255.

[44] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[47] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[48] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.

[49] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *CoRR*, vol. abs/1404.2188, 2014.