Using Language-Aligned Gesture Embeddings for Understanding Gestures Accompanying Math Terms

Tristan Maidment[1], Purav J Patel[4], Erin Walker[1],[2],[3], Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Computer Science, [3]Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA, USA
tdm51@pitt.edu eawalker@pitt.edu kovashka@cs.pitt.edu

[4]University of Maryland, College Park, MD, USA
ppatel45@umd.edu

Abstract

In this paper, we introduce an approach for recognizing and classifying gestures that accompany mathematical terms, in a new collection we name the "GAMT" dataset. Our method uses language as a means of providing context to classify gestures. Specifically, we use a CLIP-style framework to construct a shared embedding space for gestures and language, experimenting with various methods for encoding gestures within this space. We evaluate our method on our new dataset containing a wide array of gestures associated with mathematical terms. The shared embedding space leads to a substantial improvement in gesture classification. Furthermore, we identify an efficient model that excelled at classifying gestures from our unique dataset, thus contributing to the further development of gesture recognition in diverse interaction scenarios.

1. Introduction

In recent years, there has been a growing interest in the development of advanced human-computer interaction methods. One critical area is understanding and interpreting human gestures, which are frequently integrated with spoken communication to convey complex ideas. In response to this need, we propose a novel dataset: Gestures Accompanying Math Terms (GAMTs). These gestures provide an additional layer of communicative context that enhances the expressive power of speech.

Gestures Accompanying Math Terms (GAMTs) refer to physical movements that were made by individuals while using math terms in context. These gestures typically involve hand movements or hand signs that correlate with the spoken math terms, as a form of non-verbal communication. Our intention in proposing the GAMT dataset is to assist in the development of more nuanced and effective human-computer interaction methods. Having the ability to recognize or distinguish between gestures may provide additional information in settings where gestures and speech work together to convey complex ideas.

The challenge in classifying GAMTs is due to their spontaneous nature and the inherent variability among them. Unlike most gesture-recognition datasets, where multiple participants are directed to execute the same gesture or hand signal, in our dataset, the only direction participants receive on what GAMTs to produce is a sentence containing a math term. Participants are free to make gestures in the manner that they feel matches the term and sentence the best, resulting in a diverse assortment of gestures for each math term. This diversity was a deliberate aspect of our collection methodology for the GAMT dataset, which resulted in a dataset with substantial within-class variance, reflecting the individual interpretative differences between the participants in conveying math terms through gestures.

Although these gestures may look visually different, they seek to express the same semantic concept. Thus, to give our classifier a glimpse into the meaning of these gestures and a bridge among visually diverse gestures, we seek help from language representations. Specifically, we leverage language-aligned gesture embeddings to improve the classification performance of GAMT. We compare our language-aligned gesture embeddings with action-recognition-based representations, further demonstrating their efficacy. By incorporating semantic information into the gesture embeddings, our approach can also exploit the inherent structure and relationships between the associated math terms, leading to a more robust feature space than action recognition-based representations that can better capture the nuances of GAMTs.

In this paper, we first describe the GAMT dataset and our data collection process. We then detail the various methods we attempted for classifying GAMT, including sequence classification models applied over the pose representations, various convolutional neural networks (CNNs) and transformers. We address the following questions:

- Can a CLIP-style training framework effectively create a shared embedding space for gestures and language that thereby enhances gesture classification performance in the context of mathematical terms?
- How does the proposed language-aligned gesture embedding approach compare with traditional action-recognition-style embeddings in recognizing gestures associated with mathematical terms?
- Can convolutional networks such as Temporal Convolutional Neural Networks (TCNN) [32] provide a more efficient and effective alternative for gesture classification compared to transformers?

2. Related Work

Gesture classification is an essential component of human-computer interaction. It is particularly significant in contexts where gestures synergize with spoken explanations to communicate complex ideas. Existing approaches for gesture classification have focused on techniques such as dynamic time warping [5, 14], hidden Markov models [5, 15, 39], recurrent neural networks [15], and more recently transformers [2, 12, 17].

A prominent deep learning technique in the literature on gesture classification is the use of 3D convolutional neural networks (3D CNNs) [3, 27, 58]. These networks have achieved state-of-the-art performance in various gesture recognition tasks by exploiting spatial and temporal information in video data [21, 52, 55]. However, most of these studies are conducted on datasets that capture the mechanics and physical execution of gestures, rather than trying to capture the meaning behind these gestures [4, 19, 29, 30, 56].

Typically, gesture classification is performed in pixel space, analyzing the RGB video stream directly. This approach is generally preferred for scenarios where large-scale data collection can ensure that there are sufficient data. In scenarios where data is more limited, analysis of the more compact human-skeleton representation is preferred [10, 11]. Traditionally, graph convolutional neural networks (GCNNs) were used for skeleton-based action recognition [16, 18, 22, 40, 41, 50, 51].

Some studies identified TCNNs as a good candidate for skeleton-based gesture recognition [45, 46]. TCNNs were once compared favorably against GCNNs for skeleton-based action recognition [24]. However, the success of the attention mechanism inspired various forms of applying attention to GCNNs [23, 36, 41, 42]. This culminated in the direct use of transformers [7, 34, 35, 37, 44, 44, 49, 54, 57].

The connection between speech and gestures is a well-known phenomenon in human communication and is often used to enhance interactions [1, 9, 20, 28, 47]. However, this connection has received very little attention from the machine learning community, with most focus on co-speech gesture generation [25, 26, 31, 53]. Co-speech gesture generation involves generating gestures that enhance the ability for a humanoid robot to have a conversation with a human counterpart. However, datasets commonly used for this task, e.g. the TED talks dataset [53], features many gestures that have no direct relation to the words spoken, and serve primarily as “beat” gestures that enhance the rhythm of speech, rather than “iconic” ones that exemplify the meaning of the spoken words.

One recent gesture-language alignment work by Abzaliev et al. [2] trains a model to align gestures and language using a CLIP-style training technique and create a joint embedding space, which is then used to classify the LIWC tags for the words accompanying the gesture. LIWC tags [33] denote that a word is a preposition or pronoun, that it refers to the future, or it talks about emotions, etc. Rather than classifying the *type of word*, we directly classify the gesture itself. We leverage the CLIP-style [38] training for learning language-aligned gesture embeddings, similar to Abzaliev et al. [2]. However, our use of this training differs in three key ways. First, unlike Abzaliev et al. [2] who focused primarily on LIWC tags, our focus is on the mathematical terminologies tied to the gestures. No prior work explores classifying such gestures, nor leveraging semantics and language to improve classification. Second, we compare different representations (using transfer learning from different tasks and using different architectures) for classifying the gestures. Third, we present modifications to the models used in the CLIP-style training method proposed by Abzaliev et. al. [2] to reduce the complexity of the model.

3. Dataset and Data Collection

We introduce the Gestures Accompanying Math Terms (GAMT) dataset, a collection of video clips featuring paid volunteers saying sentences with math terms to the viewer. The dataset is designed to capture the diverse range of gestures that individuals may use to accompany an explanation of math concepts. The data set comprises N = 176 samples, each of which belongs to one of the eight classes. The classes represent different math terms and the associated set of gestures. The terms (classes) and associated language are shown in Table 1. During the data collection process, participants received scripted lines containing math-related terms. Each participant was instructed to recite their scripted line three times, while producing a different gesture each time. The emphasis was placed on associating the gesture with the specific math-related word that was emphasized in the scripted line.

For example, if the scripted line was “You should *CONVERT* time to an easier-to-understand variable,” the partic-

Math Term	# of Clips	Example Scripted Line
Amount	20	So, you're multiplying by three to get the new AMOUNT.
Convert	8	You should CONVERT time to an easier-to-understand variable.
Multiply	26	You would MULTIPLY two and three.
Plus	24	One times two PLUS one equals nine.
Ratio	26	The given RATIO was six to two.
Three	22	So I have THREE turns with many people.
Zero	28	They accelerated from ZERO to 60 in four seconds.
Slope	15	I can solve for the SLOPE of the line.
Units	7	The UNITS cancels out and the result is ten.

Table 1. GAMT dataset classes and example lines

ipant would produce three different gestures, each emphasizing the word "CONVERT" in a distinct manner to convey its meaning. This process allowed for the creation of a diverse dataset of gestures associated with math terms.

We assembled the GAMT dataset using an automated data collection process on the Prolific platform, which was followed by a data cleaning phase. Given the nature of this automated collection, there was no real-time opportunity to correct volunteers, leading to instances where misunderstood directions resulted in unusable gesture data. These unusable iterations were identified and filtered out during the cleaning phase to ensure the quality of the dataset. However, since the amount of data removed varied between classes, it led to a disparity in representation among classes.

The primary challenges posed by the GAMT dataset are its modest size, multi-class nature, and gesture variability. The small size limits the capacity to train a robust gesture recognition model from scratch. The multi-class nature of the dataset adds to this challenge, as models must distinguish between a wide variety of gestures, which may exhibit subtle differences or overlap in their characteristics. Lastly, the data imbalance issue means that certain classes are underrepresented in the dataset, which can lead to biased model predictions and reduced classification performance. To address these challenges, we tested a range of methods for feature extraction, including transfer learning using models trained for other gesture classification tasks.

According to research ethics guidelines, the dataset used in this study has been anonymized to protect the privacy and confidentiality of the participants. An anonymized version of the dataset will be made publicly available at https://tmaidment.github.io/gamt. By making the dataset available, other researchers and interested parties can access the data and further investigate gesture recognition and its association with math terms.

To preprocess the video data in the GAMT dataset, we first extracted the skeleton-based pose representation for each frame within the gesture. These pose representations serve as input features for our subsequent classification models. Pose representations provide a rich low-dimensional representation of the volunteer's body position and how they move throughout each clip. To obtain these pose representations, we used AlphaPose [13] to predict poses in the representation defined by OpenPose [6]. The OpenPose representation provides a set of 17 keypoints that represent the positions of various body parts in each frame. Each keypoint also has a confidence score, which we exclude. The resulting pose representation is a one-dimensional vector that encodes the spatial configuration of the keypoints.

Figure 1 showcases three gestures from the collected dataset. Each dot represents a keypoint location in a specific frame. The keypoints are color-coded according to their type, corresponding to different body part locations as defined by OpenPose. To illustrate the time component and motion, keypoints are plotted for each frame in the gesture.

In the gesture for the math term "Amount", the top two participants emphasize the concept by using large circular hand and arm movements, likely to demonstrate a mass, captured by the round shapes formed by the keypoints in the image. For the gesture related to "Fraction", the top most participant used their right hand to make diagonal movements and waved their left hand in an oval shape, while the bottom participant opted to make diagonal motions with both hands. As for the gestures for "Zero", we see that the participants made gestures that involve bringing their hands close together. These keypoint representations demonstrate the complexity and richness of non-verbal communication in conveying math concepts.

4. Methodology

In this section, we describe our testing methodology used to find a reliable method to classify gestures in the GAMT dataset. At a high level, our approach consists of training a pose encoder, which is used to extract important features of gestures. The extracted features are then used by a second model to perform classification (Sec. 4.4). This allows us to test the efficacy of aligning gestures with language (Sec. 4.2) compared to representing them without in-

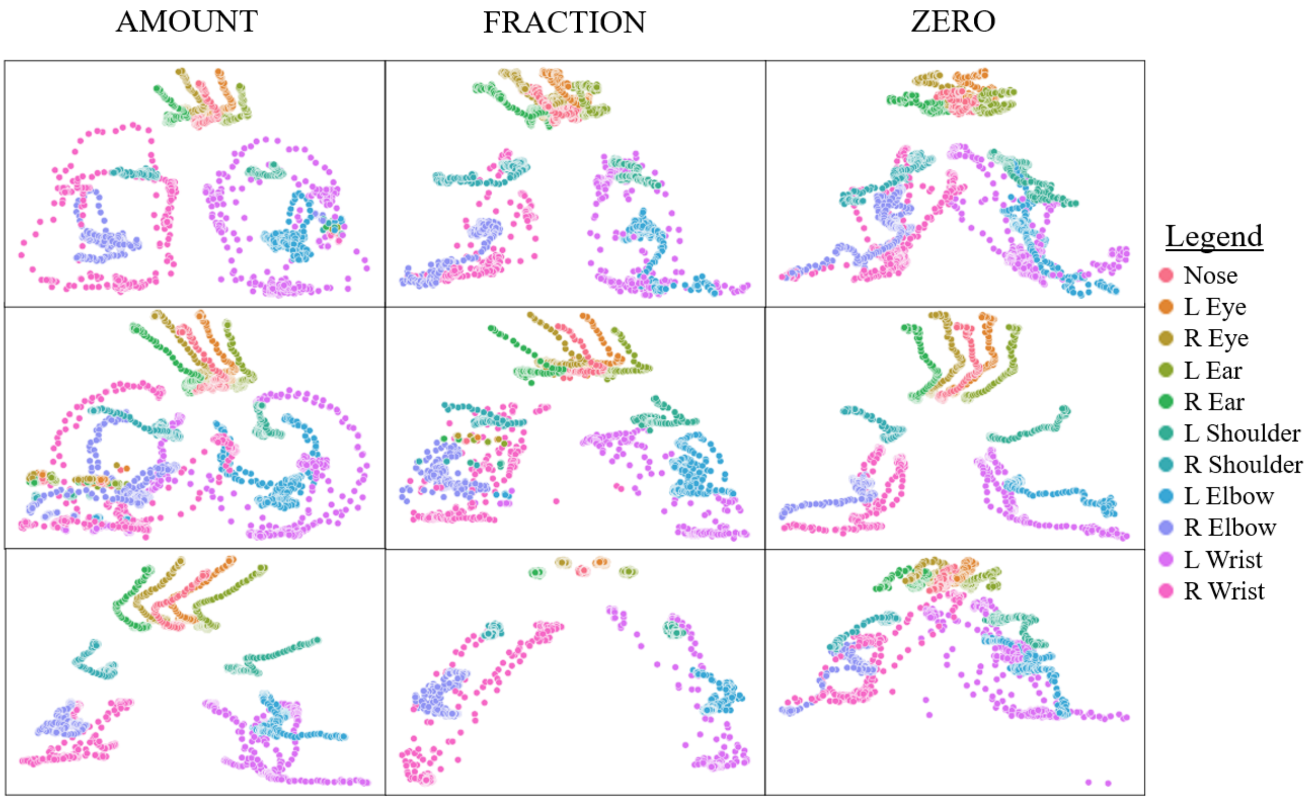

Figure 1. Examples of gestures for "Amount", "Fraction", and "ZERO" math terms from the GAMT dataset, represented as scatter plots. Points represent the locations of joints from the pose, plotted as they move over time.

corporating any information about the underlying language (Sec. 4.1). We tested a variety of pose encoders (Sec. 4.3).

4.1. Gesture Embeddings without Lang. Context

We use a multi-class classification objective to train pose encoders that do not consider the language with which the gesture was performed. To test the efficacy of using transfer learning to improve the performance of classifying GAMT, we first train the pose encoder to perform multi-class classification on a gesture-classification dataset (Sec. 5). The final classification layer of the trained pose encoder was excluded, and the model used for gesture feature extraction.

4.2. Language-Aligned Gesture Embeddings

To build a pose encoder that aligns gesture with language, we used a CLIP-style training technique. Inspired by Abzaliev et al. [2], we used XLM-RoBERTa [8] as our text encoder. For the pose/gesture encoder, we tested transformer [48] and convolution architectures. Figure 2 illustrates an overview of the CLIP-style. During the training phase, paired video clips consisting of a pose representation and its associated language are fed into the respective language and pose encoder. A joint embedding space is then learned, aligning gestures with their corresponding textual descriptions. To take advantage of this joint embedding space, a pose representation (without an associated word) is used as input to the pose encoder. The pose encoder then maps the pose representation to the language-aligned space. This representation is then used for classification. We experiment with CLIP training on two different datasets.

The CLIP-style training uses contrastive learning to align gestures and language. We seek to maximize the similarity between the corresponding text and gesture representations, while minimizing the similarity between nonmatching pairs. Given a batch of N text-gesture pairs, we use a text encoder f and gesture encoder g to obtain the text and gesture representations for the text-gesture pair i, denoted v_i and u_i, respectively.

$$\begin{aligned} v_i &= f(\text{text}_i), \qquad && i = 1, \ldots, N \\ u_i &= g(\text{gesture}_i), \qquad && i = 1, \ldots, N \end{aligned}$$

We then compute the similarity between all text and gesture pairs using the Multi-class N-Pair Loss [43], defined as:

$$L = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \sum_{j \neq i}^{N} \exp(\text{sim}_{ij} - \text{sim}_{ii}) \right)$$

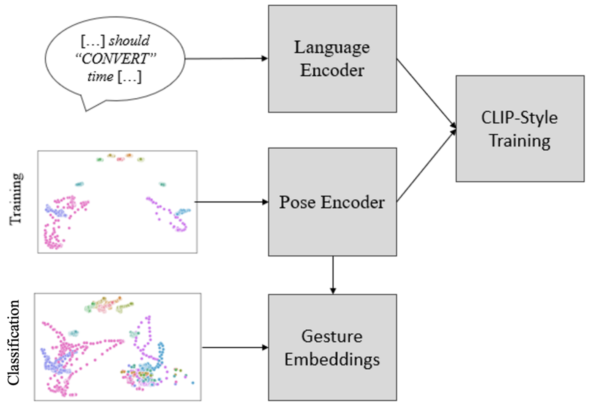

Figure 2. The general architecture for generating the gesture and language joint embedding. During training, the Pose Encoder aligns the gestures and underlying language via CLIP-style training. For GAMT classification, the Pose Encoder is used to produce a set of language-aligned gesture embeddings.

Here, sim_{ij} denotes the similarity between the text representation v_i and the gesture representation u_j, and sim_{ii} represents the similarity within the matching text and gesture pair, calculated using dot product: $\text{sim}_{ij} = v_i \cdot u_j$.

4.3. Pose Encoders

First, we tested a transformer architecture that has previously been effective for gesture-language alignment. Second, we tested a TCNN, due to its success prior to the large use of attention for skeleton-based action recognition. Lastly, we tested a 2D CNN to capture both the temporal dependencies, and those between the body's joints.

The **Pose Transformer** follows the form of the CLIP image encoder, as described in [2]. The model uses the same set of modifications made to the width of the model width to match the size of the text encoder, XLM-RoBERTa [8].

The **Temporal Convolutional Neural Network** (TCNN) [32] was tested as a pose encoder, due to its brief success for skeleton-based action recognition. TCNNs employ dilated causal convolutions, which introduce spacing between input elements, effectively increasing the receptive field without introducing additional parameters. This feature is particularly valuable for capturing long-term dependencies within sequential data and handling input sequences of varying lengths. Moreover, the incorporation of causal convolutions ensures that each output at a given time step is solely influenced by past and present input elements, preserving temporal order. The structure of a TCNN closely resembles a 1-Dimensional (1D) CNN, with the 1D kernel applied over the time dimension.

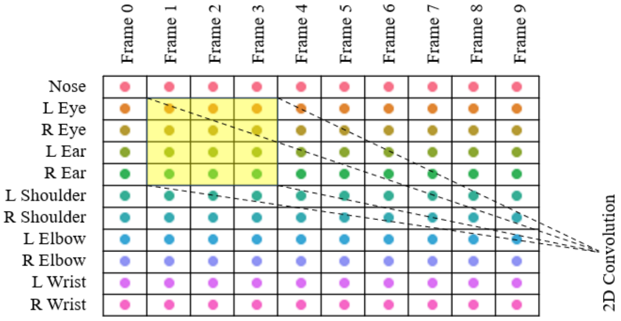

Figure 3. An example of the 2D CNN being applied on the pose representations

We further explored the potential of using a **2D Convolutional Neural Network** (CNN) as an alternative modification to the pose encoder. The inspiration came from the remarkable success of 3D CNNs, which are widely recognized as state-of-the-art algorithms for gesture classification tasks. 3D CNNs excel at capturing both spatial and temporal relationships in video data by applying 3D convolutions over the spatial dimensions (height and width of the frame) and the time dimension (across frames). The success of 3D CNNs in this space is therefore partly due to the relationship between the spatial dimensions and the time dimension, which is captured by the third dimension of the CNN. Training a 3D CNN requires a significantly large dataset and are therefore too complex for classifying GAMT. Given that the skeleton-based pose representation converts the spatial information about the pose into a 1D vector, it can be hypothesized that applying a 2D CNN over the pose representation and time should retain the benefits of the Temporal Convolutional Neural Network, and capture dependencies between joints in the pose.

Figure 3 illustrates the application of the 2D CNN on the pose representations and time in our gesture classification approach. The chart shows the keypoints on the y-axis, representing different body part locations as defined by OpenPose, while the x-axis represents the frames (time) of the video. The 2D CNN operates solely on the pose representation and does not consider any pixel-level information. The 2D CNN could provide an efficient approach to capture the interactions between joints and time, without relying upon attention.

4.4. Classifying GAMT Gestures

In order to validate the quality of the language-aligned joint embeddings we adopted a binary classification approach for the evaluation process. We treated the multi-class classification problem as a series of binary tasks, one for each class. For each class i ($i = 1, \ldots, C$), we create a binary dataset D_i consisting of an equal number of positive and negative examples, P_i and N_i, respectively. Let x represent a gesture clip and y be the binary label, with $y = 1$

Model	No Training	GAMT	Jester	GAMT (LA)	TED (LA)
Pose Transformer	0.5309	0.5290	0.5148	0.5660	**0.5808**
TCNN	0.5263	0.5433	0.5669	0.5771	**0.5876**
ResNet 18	0.5247	0.5213	0.5136	0.4828	**0.5301**
ResNet 152	0.5238	0.5236	0.4897	0.5473	**0.5476**

Table 2. The LR binary classification performance (accuracy) of the embeddings from different pose encoder models and configurations. The title of the columns describes the dataset that was used for training, and whether or not the model was Language-Aligned, indicated by the inclusion of (LA). The best performer per row is bolded.

for positive examples and $y = 0$ for negative examples: $D_i = \{(x, y)|x \in P_i \cup N_i, y \in \{0, 1\}\}$ For each binary dataset D_i, we train a logistic regression (LR) classifier. We use 30-fold cross-validation on the validation dataset to estimate the performance of the LR classifier, and thus the information contained in the embeddings, then average the performance across the folds.

5. Experimental Design

Models. We test each of the four pose encoders, specifically the Pose Transformer (with 60 million parameters), Temporal Convolutional Neural Network (20 thousand, the lowest overall), and two 2D convolutional networks, ResNet 18 (11.6 mil) and ResNet 152 (86 mil).

Datasets. The pose encoders' performance is evaluated by testing the resulting embeddings' performance when classifying the GAMT dataset. Each of the four pose encoders are trained in two ways: with language-alignment (Sec. 4.2) or without language-alignment (Sec. 4.1). For training the pose encoders with language-alignment, we opt to use the TED dataset [53]. For training without language-alignment, we use the Jester dataset [29]. The GAMT gestures have both an associated set of language and have assignment to a specific class, and therefore can be used to train pose encoders that are either language-aligned or are without language-alignment.

Without Language Alignment: Qualcomm Jester Dataset. The Qualcomm Jester dataset is a large-scale video dataset containing 148,092 video clips capturing human gestures performed by various individuals. The dataset contains 27 gesture classes, such as swiping left, swiping right, zooming in, and zooming out, among others. These gestures are primarily focused on human-computer interaction scenarios, making the dataset suitable for training and evaluating models that aim to identify specific motions. Notably, the dataset's classes are not designed to decode the semantic significance of gestures but rather focused on classifying the precise physical movement enacted.

With Language Alignment: TED Dataset. The TED dataset is a collection of 52 hours of video clips from TED Talks, featuring speakers delivering presentations on various topics. Each clip contains both the spoken transcripts and the associated videos, from which we can extract pose and gesture information. The dataset encompasses a vast array of gestures, including iconic gestures that mimic real-world actions, metaphoric gestures that symbolize abstract ideas, deictic gestures that point to or indicate a location or object, and beat gestures that emphasize the rhythm of speech.

6. Results

6.1. Logistic Regression Binary Classification

Table 2 displays the LR binary classification performance (average accuracy) of GAMT using the embeddings from different models and configurations of pose encoders. The entries are separated by pose encoder model and type of training. We also indicate whether or not a specific form of training was Language-Aligned, designated by the inclusion of (LA). In this table, we include a case where pose encoders do not receive training. This is to provide a baseline for the performance of the LR when performing binary classification and no form of transfer learning occurs.

It is clear that the language-aligned (LA) style of training provides a larger lift in performance than the pose encoders that are trained without language-alignment. Pose encoders trained through classification (on GAMT and Jester) did not see the same boost in performance overall.

Training the pose encoder to perform classification on the GAMT dataset provides a slight improvement, but only for the TCNN. This provides some evidence that the model structure of the TCNN is capable of capturing information about GAMT that is elusive to the ResNet and Pose Transformer models.

However, when training the pose encoders on the Jester dataset, only the embeddings produced by the TCNN provide an improvement in classifying GAMT. This finding suggests that the gesture types and contexts in the Jester dataset differ substantially from those in the GAMT dataset, limiting the transferability of the learned features. Interestingly, the embeddings produced by the other models resulted in a decrease in the performance for classifying GAMT. The TCNN differs from the ResNet and Pose Transformer models in one important way; the TCNN only models the temporal dependencies of the gesture but does not

Model	**AMOUNT**	**CONVERT**	**MULTIPLY**	**PLUS**	**RATIO**	**THREE**	**ZERO**	**SLOPE**	**UNITS**
Pose Transformer	0.5663	**0.7466**	0.4081	0.6941	0.4577	0.5193	**0.5103**	0.3901	0.6064
±	0.0580	**0.0323**	0.0125	0.0268	0.0132	0.0212	**0.0431**	0.0260	0.0362
TCNN	**0.6769**	0.6252	0.4346	0.6195	**0.5166**	**0.5404**	0.5073	0.4439	**0.6780**
±	**0.0235**	0.0588	0.0177	0.0294	**0.0220**	**0.0398**	0.0124	0.0061	**0.0439**
ResNet 18	0.5506	0.6566	0.4384	0.6605	0.5120	0.4548	0.4048	**0.4607**	0.4922
±	0.0462	0.0547	0.0064	0.0164	0.0424	0.0143	0.0172	**0.0256**	0.0323
ResNet 152	0.4773	0.6549	**0.4435**	**0.7148**	0.4941	0.4995	0.4571	0.4449	0.5514
±	0.0308	0.0282	**0.0074**	**0.0167**	0.0147	0.0199	0.0269	0.0165	0.0419

Table 3. The average accuracy and standard error for GAMT classification per model. The values in this table average the performance from the four methods of training the pose encoders; on GAMT without Language Alignment, Qualcomm Jester without Language Alignment, GAMT with Language alignment, and TED Talks with Language Alignment. The best performer per column is bolded.

directly model the relationship between joints.

The performance of the pose encoders was higher when using CLIP-style training on GAMT than when training using classification, with an exception for the ResNet 18 pose encoder. This improved performance extends to CLIP-style training on the TED Talks dataset. The joint embedding space learned from the semantics of the language from the GAMT dataset and the TED Talks dataset appears to offer a broader context for the pose embeddings.

It may be surprising, however, that the pose encoders that utilize CLIP-style training on the TED dataset outperform those that use the GAMT dataset. The TED Talks dataset has some advantages: dataset size, high diversity of gesture and high diversity of language. Furthermore, the topics of the TED Talks vary; a subset of the gestures contain math terms, but we qualitatively observed those have higher within-class variance than GAMT.

6.2. Per-Class Performance across Models

Table 3 presents the average accuracy for each class of gestures for the different forms of pose encoder, along with the corresponding standard error (SE) values.

The results highlight the varying performance of each model across different gesture classes. The Pose Transformer achieved relatively high accuracy for certain gesture classes, such as "CONVERT" and "PLUS", both words related to mathematical operations. However, it exhibited relatively lower accuracy for some classes, performing particularly poorly for classes such as "RATIO" and "SLOPE". The TCNN model excels at recognizing gestures related to "RATIO" and "AMOUNT."

Both the Pose Transformer and the TCNN models perform well in recognizing gestures related to "THREE" and "ZERO." However, the Pose Transformer achieves higher accuracy in "THREE," while TCNN excels in "ZERO." Both of these gestures involve very little movement and typically involve the participant simply holding up their hand to present the corresponding number.

The 2D CNNs ResNet 18 and ResNet 152 demonstrate moderate performance across the various gesture classes. Although they do not outperform the Pose Transformer or TCNN overall, Resnet 152 excels on "MULTIPLY" and "PLUS". A qualitative analysis indicates that these two terms typically involve iconic gestures that mimic the associated symbols. The 2D CNNs show more *consistent* performance across gestures: Resnet 18 has the lowest standard deviation across classes (0.091 vs 0.122 for the Pose Transformer).

6.3. CLIP-style Training on TED Talks Dataset

The embeddings from the CLIP-style training on the TED Talks dataset resulted in the highest performance for all models. We unpack the performance of CLIP-style training on the TED Talks Dataset in Table 4, which displays the average accuracy of each model for individual gesture classes. The results provide insight into how the models perform when exposed to a more diverse set of gestures and language from real-world TED Talks.

Overall, results on TED are consistent with those on GAMT, in terms of relative performance of methods, and in several cases absolute performance numbers. Across all models, recognizing gestures related to numbers poses a significant challenge, as is evident by the comparatively lower average accuracy scores. This difficulty may stem from the inherent ambiguity of representing numbers in gestures - these gestures often involve subtle hand configurations that are not discernible using the OpenPose representation. As such, the gestures for "ZERO" and "THREE" are difficult to distinguish. Improved performance in this class would require specialized pose representations that capture the structure of the hand or incorporate additional data to disambiguate number-related gestures.

Similarly, the "RATIO" and "SLOPE" gestures exhibit lower accuracy across the models. Recognizing gestures associated to proportional quantities can be challenging due to the abstract nature of this concept. Ratios involve a top-to-bottom comparison between quantities, often represented by the participant using two hands to demonstrate the spa-

Model	AMOUNT	CONVERT	MULTIPLY	PLUS	RATIO	THREE	ZERO	SLOPE	UNITS
Pose Transformer	0.6264	**0.7475**	0.4065	**0.7558**	0.5016	0.5187	**0.6000**	0.5097	0.5613
TCNN	**0.7401**	0.5893	0.4182	0.6361	**0.5325**	**0.5612**	0.5719	**0.5417**	**0.6976**
ResNet 18	0.6075	0.6253	0.4347	0.6188	0.5012	0.5067	0.4031	0.4697	0.6042
ResNet 152	0.5821	0.6717	**0.4644**	0.7000	0.4989	0.5559	0.4188	0.4771	0.5598

Table 4. The average accuracy and performance per class of the pose encoders, specifically for the language-aligned embeddings that were trained using the TED Talks Dataset. The best performer per column is bolded.

tial relationships. Gestures associated with slopes typically involve hand movements representing an incline or decline. These motions may be more obvious than the more physically abstract concept of ratios, but often take a long time to execute. This may be the reason why the TCNN significantly outperformed other models on “SLOPE”.

Furthermore, the manifestation of the “RATIO” and “SLOPE” in the TED Talks differs greatly from the case in GAMT. In TED Talks, “SLOPE” does not necessarily refer to the mathematical use of the word. This is also true about “RATIO”, but to a lesser extent. While “RATIO” is used to compare two quantities, the speaker is not necessarily explaining the mathematical concept of ratios. A common example is speakers talking about the ”aspect *ratio*” of a display.

7. Discussion and Conclusions

In this work, we set out to find the following answers:

- The CLIP-style training framework effectively created a shared embedding space for gestures and language, significantly enhancing gesture classification performance in the context of mathematical terms, through the semantic information language provides.
- The proposed language-aligned gesture embedding approach outperformed traditional action-recognition-style embeddings in recognizing gestures associated with mathematical terms. Transfer learning from CLIP-style training on the TED Talks dataset demonstrated superior performance compared to other methods, showcasing the versatility of the diverse joint embedding space.
- The Temporal Convolutional Neural Network showed promising results, proving to be a more efficient and effective alternative to other methods for classification within GAMT. Its ability to model long-range dependencies while maintaining a significantly smaller footprint makes it a practical choice, particularly in real-time inference scenarios.

Our experiments showed that transfer learning from the Jester dataset did not lead to substantial improvements in classification performance compared to the GAMT dataset. We note that TCNN and ResNet-18 showed slight improvement via transfer learning, both of which are smaller models. This may highlight the differences in gestures between the Jester and GAMT datasets, which limits the applicability of the learned features. This difference in types of gestures may be able to be bridged by using language-infused gesture representations.

The CLIP-style training on the TED Talks dataset demonstrated superior performance compared to training on the GAMT dataset or attempting to leverage transfer learning from an action-recognition-style method. This suggests that the joint embedding space between language and gesture acquired from the TED Talks dataset offers a more versatile feature space that is advantageous for classifying GAMT.

The Pose Transformer and TCNN model both worked well. There are commonalities between these models that may drive their superior performance compared to the 2D CNNs. The Pose Transformer and TCNNs can both model long-range dependencies in their input, either through the self-attention mechanism or dilated causal convolutions. In particular, this may highlight the fact that the motion of the gestures is important.

However, the TCNN has the added benefit of having a very small footprint. In the case of gesture understanding, the models will likely have to be run in real-time. The TCNN requires significantly less resources than the Pose Transformer and is easily run in real-time.

Our research underscores the potential to utilize language-aligned gesture embeddings to improve performance in classifying gestures, particularly those related to mathematical terminology. This is especially valuable when dealing with smaller, imbalanced datasets like GAMT.

In future work, further enhancements to the transformer architecture can be considered, as well as novel strategies to integrate language information into the gesture classification process more effectively.

Acknowledgements: This work was supported by Grant No. 2024645 from the National Science Foundation.

References

[1] Co-Speech Gesture in Communication and Cognition. 2

[2] Artem Abzaliev, Andrew Owens, and Rada Mihalcea. Towards Understanding the Relation between Gestures and Language. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5507–5520,

Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 2, 4, 5

[3] Norah Alnaim, Maysam Abbod, and Rafiq Swash. Recognition of Holoscopic 3D Video Hand Gesture Using Convolutional Neural Networks. *Technologies*, 8(2):19, 2020. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. 2

[4] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. IPN Hand: A Video Dataset and Benchmark for Real-Time Continuous Hand Gesture Recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4340–4347, Milan, Italy, 2021. IEEE. 2

[5] Alina Delia Calin. Gesture Recognition on Kinect Time Series Data Using Dynamic Time Warping and Hidden Markov Models. In *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 264–271, 2016. ISSN: 2470-881X. 2

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017. ISSN: 1063-6919. 3

[7] Yi-Bin Cheng, Xipeng Chen, Dongyu Zhang, and Liang Lin. Motion-transformer: self-supervised pre-training for skeleton-based action recognition. *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, pages 1–6, 2021. Conference Name: MMAsia '20: ACM Multimedia Asia ISBN: 9781450383080 Place: Virtual Event Singapore Publisher: ACM. 2

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020. Association for Computational Linguistics. 4, 5

[9] Anthony Steven Dick, Susan Goldin-Meadow, Uri Hasson, Jeremy I. Skipper, and Steven L. Small. Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping*, 30 (11):3509–3526, 2009. 2

[10] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 579–583, 2015. Conference Name: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) ISBN: 9781479961009 Place: Kuala Lumpur, Malaysia Publisher: IEEE. 2

[11] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting Skeleton-based Action Recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2959–2968, New Orleans, LA, USA, 2022. IEEE. 2

[12] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. A Transformer-Based Network for Dynamic Hand Gesture Recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632, 2020. ISSN: 2475-7888. 2

[13] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7157–7173, 2023. 3

[14] Nurfazlin Muhamad Feizal Franslin and Giap Weng Ng. Vision-Based Dynamic Hand Gesture Recognition Techniques and Applications: A Review. In *Proceedings of the 8th International Conference on Computational Science and Technology*, pages 125–138, Singapore, 2022. Springer. 2

[15] Volkmar Frinken, Tim Peter, Andreas Fischer, Horst Bunke, Trinh-Minh-Tri Do, and Thierry Artieres. Improved Handwriting Recognition by Combining Two Forms of Hidden Markov Models and a Recurrent Neural Network. In *Computer Analysis of Images and Patterns*, pages 189–196, Berlin, Heidelberg, 2009. Springer. 2

[16] Xiang Gao, Wei Hu, Jiaxiang Tang, Pan Pan, Jiaying Liu, and Zongming Guo. Generalized Graph Convolutional Networks for Skeleton-based Action Recognition. *ArXiv*, 2018. 2

[17] Basavaraj Hampiholi, Christian Jarvers, Wolfgang Mader, and Heiko Neumann. Convolutional Transformer Fusion Blocks for Multi-Modal Gesture Recognition. *IEEE Access*, 11:34094–34103, 2023. Conference Name: IEEE Access. 2

[18] Xiaoke Hao, Jie Li, Yingchun Guo, Tao Jiang, and Ming Yu. Hypergraph Neural Network for Skeleton-Based Action Recognition. *IEEE Transactions on Image Processing*, 30: 2263–2275, 2021. 2

[19] Alexander Kapitanov, Andrew Makhlyarchuk, and Karina Kvanchiani. HaGRID - HAnd Gesture Recognition Image Dataset, 2022. arXiv:2206.08219 [cs]. 2

[20] Zohreh Khosrobeigi, Maria Koutsombogera, and Carl Vogel. Gesture and Part-of-Speech Alignment in Dialogues. 2022. 2

[21] Zhiping Lai, Xiaoyang Kang, Hongbo Wang, Weiqi Zhang, Xueze Zhang, Peixian Gong, Lan Niu, and Huijie Huang. STCN-GR: Spatial-Temporal Convolutional Networks for Surface-Electromyography-Based Gesture Recognition. In *Neural Information Processing*, pages 27–39, Cham, 2021. Springer International Publishing. 2

[22] Chaolong Li, Zhen Cui, Wenming Zheng, Chunyan Xu, Rongrong Ji, and Jian Yang. Action-Attending Graphic Neural Network. *IEEE Transactions on Image Processing*, 27(7): 3657–3670, 2018. 2

[23] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3590–3598, 2019. Conference Name: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ISBN: 9781728132938 Place: Long Beach, CA, USA Publisher: IEEE. 2

[24] Yong Li, Zihang He, Xiang Ye, Zuguo He, and Kangrong Han. Spatial temporal graph convolutional networks for

skeleton-based dynamic hand gesture recognition. *EURASIP Journal on Image and Video Processing*, 2019(1):78, 2019. 2

[25] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. SEEG: Semantic Energized Co-speech Gesture Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10463–10472, New Orleans, LA, USA, 2022. IEEE. 2

[26] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-Driven Co-Speech Gesture Video Generation, 2022. arXiv:2212.02350 [cs]. 2

[27] Hammad Mansoor, Nidhi Kalra, Piyush Goyal, Muskan Bansal, and Namit Wadhwa. Hand Gesture Recognition Using 3D CNN and Computer Interfacing. In *Inventive Systems and Control*, pages 727–736, Singapore, 2022. Springer Nature. 2

[28] Lars Marstaller and Hana Burianová. The multisensory perception of co-speech gestures – A review and meta-analysis of neuroimaging studies. *Journal of Neurolinguistics*, 30: 69–77, 2014. 2

[29] Joanna Materzynska, Guillaume Berger, I. Bax, and R. Memisevic. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019. 2, 6

[30] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4207–4215, Las Vegas, NV, USA, 2016. IEEE. 2

[31] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *Computer Graphics Forum*, 42(2):569–596, 2023. arXiv:2301.05339 [cs]. 2

[32] Ashutosh Pandey and DeLiang Wang. TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6875–6879, 2019. ISSN: 2379-190X. 2, 5

[33] James Pennebaker, Martha Francis, and Roger Booth. Linguistic inquiry and word count (LIWC). 1999. 2

[34] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial Temporal Transformer Network for Skeleton-Based Action Recognition. pages 694–701, Cham, 2021. Springer International Publishing. Book Title: Pattern Recognition. ICPR International Workshops and Challenges Series Title: Lecture Notes in Computer Science. 2

[35] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208-209:103219, 2021. 2

[36] Xiaofei Qin, Rui Cai, Jiabin Yu, Changxiang He, and Xuedian Zhang. An efficient self-attention network for skeleton-based action recognition. *Scientific Reports*, 12(1):4111, 2022. Number: 1 Publisher: Nature Publishing Group. 2

[37] Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-Temporal Tuples Transformer for Skeleton-Based Action Recognition, 2022. arXiv:2201.02849 [cs]. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. ISSN: 2640-3498. 2

[39] Gerhard Rigoll, Andreas Kosmala, and Stefan Eickeler. High performance real-time gesture recognition using Hidden Markov Models. In *Gesture and Sign Language in Human-Computer Interaction*, pages 69–80, Berlin, Heidelberg, 1998. Springer. 2

[40] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Adaptive Spectral Graph Convolutional Networks for Skeleton-Based Action Recognition. *ArXiv*, 2018. 2

[41] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-Based Action Recognition With Directed Graph Neural Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7904–7913, 2019. Conference Name: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ISBN: 9781728132938 Place: Long Beach, CA, USA Publisher: IEEE. 2

[42] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1227–1236, 2019. Conference Name: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ISBN: 9781728132938 Place: Long Beach, CA, USA Publisher: IEEE. 2

[43] Kihyuk Sohn. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 4

[44] Yan Sun, Yixin Shen, and Liyan Ma. MSST-RT: Multi-Stream Spatial-Temporal Relative Transformer for Skeleton-Based Action Recognition. *Sensors*, 21(16):5339, 2021. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute. 2

[45] Panagiotis Tsinganos, Bruno Cornelis, Jan Cornelis, Bart Jansen, and Athanassios Skodras. Improved Gesture Recognition Based on sEMG Signals and TCN. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1169–1173, 2019. ISSN: 2379-190X. 2

[46] Panagiotis Tsinganos, Bart Jansen, Jan Cornelis, and Athanassios Skodras. Real-Time Analysis of Hand Gesture Recognition with Temporal Convolutional Networks. *Sensors*, 22(5):1694, 2022. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute. 2

[47] Paweł Urbanik and Jan Svennevig. Action-Depicting Gestures and Morphosyntax: The Function of Gesture-Speech

Alignment in the Conversational Turn. *Frontiers in Psychology*, 12, 2021. 2

[48] Satish E. Viswanath, Prathyush V. Chirra, Michael C. Yim, Neil M. Rofsky, Andrei S. Purysko, Mark A. Rosen, B Nicolas Bloch, and Anant Madabhushi. Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on T2-weighted MRI: a multi-site study. *BMC Medical Imaging*, 19, 2019. 4

[49] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for Skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 537(C):164–186, 2023. 2

[50] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. 2

[51] Dun Yang, Qing Zhou, and Ju Wen. Interactive two-stream graph neural network for skeleton-based action recognition. *Journal of Electronic Imaging*, 30(03), 2021. 2

[52] Yang Yi, Feng Ni, Yuexin Ma, Xinge Zhu, Yuankai Qi, Riming Qiu, Shijie Zhao, Feng Li, and Yongtao Wang. High Performance Gesture Recognition via Effective and Efficient Temporal Modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1003–1009, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization. 2

[53] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. pages 4303–4309, 2019. 2, 6

[54] Qipeng Zhang, Tian Wang, Mengyi Zhang, Kexin Liu, Peng Shi, and Hichem Snoussi. Spatial-temporal Transformer For Skeleton-based Action Recognition. In *2021 China Automation Congress (CAC)*, pages 7029–7034, 2021. ISSN: 2688-0938. 2

[55] Wei Zhang, Zeyi Lin, Jian Cheng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. STA-GCN: two-stream graph convolutional network with spatial–temporal attention for hand gesture recognition. *The Visual Computer*, 36(10-12):2433–2444, 2020. 2

[56] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018. Conference Name: IEEE Transactions on Multimedia. 2

[57] Yuhan Zhang, Bo Wu, Wen Li, Lixin Duan, and Chuang Gan. STST: Spatial-Temporal Specialized Transformer for Skeleton-based Action Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3229–3237, Virtual Event China, 2021. ACM. 2

[58] Nan Zhou and Jun Du. Recognition of Social Touch Gestures Using 3D Convolutional Neural Networks. In *Pattern Recognition*, pages 164–173, Singapore, 2016. Springer. 2