Zero Shot Learning (ZSL) for Isolated Hand Gestures

Kaushik Manchella, Naveen Madapana, Minsoo Choi, Varad Satam

1 Abstract

Zero-shot learning (ZSL) is a new paradigm in machine learning that aims to recognize unseen object categories by just having a description of them. In contrast to conventional classification, ZSL powered systems can adapt to new gesture classes that were absent in the training phase without requiring to re-train the network with those new classes. Previous approaches that are concerned with this problem utilized a large number of gesture descriptors, linear classifiers to recognize new gesture categories and ignored the rich visual information present in the form of RGB-D data. To this end, we propose a Deep Learning based solution with a small set of descriptors to tackle the problem of ZSL for gesture recognition tasks. First, we relied on the literature to obtain a condensed list of 19 gesture attributes and annotated the 48 gesture classes from ChaLearn IsoGD gesture dataset w.r.t the attributes. Next, a pretrained VGG module was used to obtain feature representations for the RGB-D gestural data. These features were used to train 19 LSTM modules (one per each descriptor) with the goal of accurately recognizing the binary labels of the descriptors. Results show that our methodology yields unseen class accuracy of 25% which parallels performances obtained through state-of-the-art approaches.

2 Introduction

Zero-Shot Learning (ZSL) is a new paradigm in learning theory that aims to recognize unseen object categories or classes by just having a high-level description of them [4, 9]. ZSL is inspired by the way humans identify new species or children recognize unseen animals or new objects just by knowing their high-level properties such as color, shape, texture, etc [5, 17]. For instance, given the phrase, 'largest land animal with a long trunk and brown in color', humans can recognize this animal as elephant without requiring any visual samples of elephants. In a similar manner, ZSL relies on such attributes to transfer the knowledge gained from a finite set of seen classes to the categories that were never seen before [10].

ZSL approaches are particularly beneficial in the context of touchless interaction with machines as it allows the system to recognize unfamiliar gestures (never seen before) and characteristic of new users [13, 27]. Recently, there has been an increasing interest towards deploying gesture powered interfaces in gaming consoles [6, 26], smart devices and medical imaging systems in a touchless manner [12, 22, 23]. However, the current gestural systems are constrained by the pre-determined set of training gestures and such systems can not adapt to the new gestures that user might prefer. Furthermore, humans tend to create gestures on-the-fly and it is practically infeasible to train a system to recognize all gesture categories [7].

Previous works that addressed this issue of ZSL for gestures utilized skeleton features of the hand gestures and failed to incorporate visual information available in the form of RGB-D videos. Gesture videos contain rich information related to the posture, hand configuration and relative position of hands w.r.t each other and the whole body which determines the overall meaning of the gesture.

Moreover, it is a norm in the area of ZSL to develop end-to-end architectures that predict the labels of all descriptors instead of training a model for each descriptor individually. Given the complexity of the RGB-D data, we show that it is crucial to train and tune a model specifically for each descriptor and integrate the trained models to form an ensemble which acts as an overall ZSL model. In this regard, this work focuses on developing methodologies for ZSL in order to recognize and comprehend new gestures from RGB-D data.

The main contributions of this work is to: 1. Propose a new set of gesture descriptors that can represent a range of gesture categories; 2. Propose a methodology to incorporate RGB-D visual information into Deep architectures and thereby recognize unfamiliar hand gestures, and 3. Perform rigorous experiments to provide benchmarks on CGD 2016 gesture dataset [24].

3 Related Work

Deep learning has been a promising tool in achieving superior classification accuracies in objection recognition tasks when compared to traditional classification techniques, and gesture recognition is not an exception [2, 18]. Zhang et al. [28] showed that a temporal deep learning model can be a very effective tool in solving time-dependent problems such as gesture/speech recognition. Deep Learning was known to require massive amounts of training data which is infeasible for real-time applications. In such scenarios, transfer learning techniques are often used to utilize deep learning methods when there is limited data available in the target domain. While Cote-Allard et al. utilized transfer learning methods in the context of gesture recognition [1], none of these works have proposed to utilize transfer learning to effectively solve the problem of ZSL for gestures.

The problem of ZSL has been predominantly studied in the domains such as scene understanding [19] and animal/bird recognition [4]. In such domains, the function that maps the image of an object to the name of an object is one to one in many cases. For instance, an image of dog is unarguably the animal, 'dog' and an image of mountain is unambiguously a 'mountain' in most cases. This distinct mapping coupled with several publicly available attribute based datasets such as AwA [9] and SUN [19] encouraged researchers to thoroughly investigate ZSL for object recognition. Hence we have seen a consistent increase in the unseen class accuracies on AwA dataset from 57% in 2013 [9] to 86% in 2017 [8]. The need for Zero Shot learning for gestures is becoming increasingly important with the advancement of human-computer interfaces [12, 13, 16]. To this end, our goal is to propose techniques to address this issue of ZSL for gestures in order to recognize the natural and unseen gestures of humans.

Previous works concerned with gesture representations proposed to use a finite set of descriptors or attributes to represent a wide range of gesture categories [11, 16]. Madapana et al. proposed to recognize unseen gesture classes (the categories for which there is no training data) by learning to identify key semantic descriptors [14]. Additionally, a large set of descriptors were used to describe each gesture which leads to issues such as extremely sparse distributions for some descriptors and increased cost of annotation. In this work, we propose a new set of semantic descriptors which are fewer in number but are able to describe a large number of gesture classes.

Furthermore, existing benchmarks for ZSL in the area of gesture recognition were obtained using skeletal features while ignoring the rich visual information that is present in RGB-D videos [14, 27]. Given the abundant RGB-D data, especially for gestures, we hypothesize that the ZSL accuracies can be greatly improved by incorporating the RGB-D data into the ZSL architectures.

4 Technical Approach

Our approach to recognizing the unseen gesture categories can be divided into two major parts: 1. Creation of gesture descriptors that are integral to the ZSL framework, and 2. Develop a deep architecture that takes RGB-D data as an input and recognizing those descriptors. The overall ZSL framework consisted of two models: Semantic Descriptor Detector (SDD) and a Gesture Mapper. The SDD provides a binary output regarding the presence of a particular descriptor in the gesture. Once the descriptor is predicted, the Gesture

Mapper applies KNN algorithm to assign the class label to the gesture.

4.1 Gesture Attributes

We have relied on the previous work in semantics, pragmatics of language, computational linguistics and zero shot learning [11, 15] to develop a set of gesture properties or attributes that are relevant to a range of gesture categories. The main goal was to represent a large number of categories with a small subset of attributes. We considered ChaLearn Isolated gesture dataset (CGD 2016) for our experiments. Hence, we selected the descriptors that can explain the diversity in these gesture categories.

Overall, we represented each gesture using a set of 19 semantic descriptors ($\mathbb{R}^{19} \rightarrow \mathbb{R}$). The descriptors are a semantic representation of human gestures. Rather than having a label per video, our gesture descriptors map a video to a binary vector $[0, 1]^{19}$, where each element represents a semantic meaning of human gestures such as upward motion, iterative motion, etc. The proposed descriptors are divided into 4 high level categories: Hand Usage (notated by $Both\_Hands$), Motion (notated by $M\_$), Orientation of Palm (notated by $O\_$), and Finger State (notated by $F\_$). Each of the high level categories are further classified into subcategories as shown by table 1.

Table 1: Semantic Descriptor

Characteristic	Binary Attributes
Hand Usage	(i) Both Hands
Motion	(i) Upward (ii) Downward (iii) Inward (iv) Outward (v) Forward (vi) Backward (vii) Iterative
Orientation of Palm	(i) Upward (ii) Downward (iii) Inward (iv) Outward (v) Forward (vi) Backward
Finger State	(i) Thumb Visible (ii) Index Finger Visible (iii) Middle Finger Visible (iv) Ring Finger Visible (v) Pinky Finger Visible

4.2 ZSL Architecture

Let us start by defining the notations. Let $\mathcal{S}$ be a set of training (seen) classes, $\mathcal{U}$ be a set of unseen classes, z_s be the number of seen classes, z_u be the number of unseen classes, and a be the number of attributes. Note that $\mathcal{S}$ and $\mathcal{U}$ share the attribute space i.e. they have equal number of attributes. Let m_s be the number of instances in seen data. For ZSL, note that $\mathcal{S} \cap \mathcal{U} = \phi$. Let $X \in \mathbb{R}^{d \times m_s}$ be the input feature matrix, $Y \in \{-1, 1\}^{m_s \times z_s}$ be the ground truth labels, $S \in [0, 1]^{a \times z_s}$ be the per-class semantic descriptions of the seen classes. Without the loss of generality, it is assumed that each frame of a gesture sample is represented by a fixed one dimensional vector of dimension d. A gesture instance would be of size ($M \times T$), where T is the number of frames in the instance. Note that each gesture instance is re-sampled to a fixed number of frames (T).

4.2.1 Feature Extraction Using Transfer Learning Each of the T frames sampled from the video data were passed through a VGG16 model [21] that is pretrained on the ImageNet dataset [3]. Feature maps were extracted per frame per video for both RGB and Depth videos separately and concatenated before passing into a Temporal model and a Static model. Specifically, the last two fully connected layers were extracted from the VGG16 model which act as a features for ZSL. The output of each feature map is of the dimension $7 \times 7 \times 512$ i.e each frame of the video is represented as a $7 \times 7 \times 512$ vector. This approach was used to extract the feature maps for both RGB and Depth frames, and the resulting vectors were concatenated. Next, global average pooling is applied on per-frame basis to represent each frame as a 512 dimensional vector. Overall, each video was represented as an array of (T frames x 512 features).

Table 2: Distribution of Descriptor Values

Descriptor	Both_Hands	M_Iterative	M_Out	M_Front	O_Down	O_In	F_Thumb	F_Index	F_Middle
0's	0.86	0.83	0.76	0.94	0.73	0.86	0.65	0.66	0.55
1's	0.14	0.17	0.24	0.06	0.27	0.14	0.35	0.34	0.45

Table 3: Static vs. Temporal model (Validation AUC Scores)

Descriptor	M_In	M_Front	O_Down	O_In	O_Out	O_Front	O_Back	F_Thuml
Static AUC	0.86 ± .06	0.81 ± .10	0.86 ± .04	0.80 ± .06	0.39 ± .54	0.92 ± .04	0.85 ± .04	0.79 ± .0
Temporal AUC	0.92 ± .03	0.84 ± .06	0.87 ± .03	0.85 ± .04	0.79 ± 0.45	0.91 ± .04	0.76 ± .06	0.71 ± .0

4.2.2 Semantic Descriptor Detector (SDD)

The extracted features were used to train our zero shot learning model (SDD). The idea is to train a LSTM (Long Short Term Memory) model followed by a fully connected network for each descriptor with the objective of minimizing the descriptor prediction loss.

Our temporal network resembles a sequence to one (Seq2One) architecture i.e. a recurrent neural network where each cell takes an input value at time t, X_t, and returns a cell state h_t which has processed the information from an entire sequence, $X_0, X_1, ..., X_t$. The cell state will be passed to the next cell for time $t + 1$. The cell state at the last cell X_T is used to train a fully connected network that outputs a binary prediction, zero if a descriptor is absent and one otherwise. Fig 1 depicts the flowchart of our architecture. Overall, our LSTM model consisted of 25 cells (one cell per each frame), 128 hidden units and dropout probability was tuned on per-descriptor basis.

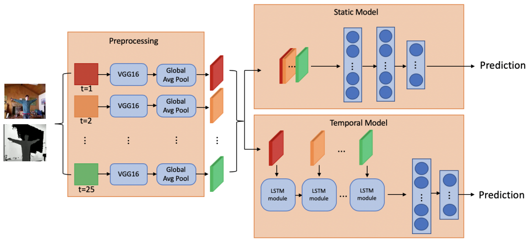

Figure 1: Flowchart of our proposed architecture.

4.2.3 Inference: Gesture Mapping

Once the training is completed, there are 19 trained LSTM models (one per each descriptor) and each model will output a binary value (s^i) corresponding to the descriptor i. Let $\hat{s}$ denote the predicted description vector, i.e. $\hat{s} = [s^1, s^2, ..., s^{19}]^T$. At this juncture, we need to assign a gesture class label to each predicted vector. To achieve this, a K-Nearest Neighbors classifier was used. Let $\bar{s}_i \in \mathbb{R}^{1 \times a}$ be the description vector of the i^{th} unseen gesture category.

$$y_{pred} =_{i \in [1, z_u]} < \hat{s}.\bar{s}_i >$$

Note that, in ZSL, we assume that the gesture examples witnessed during testing are unseen [4, 10, 20].

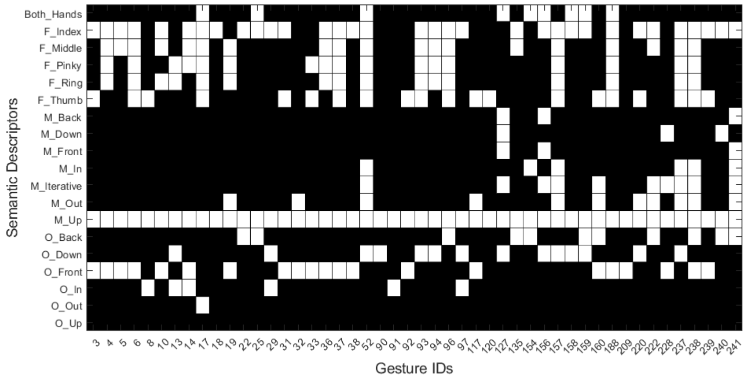

Figure 2: Representing description vectors as a binary image.

5 Evaluation

5.1 Dataset

Our proposed methodology was validated using the Chalearn 2016 IsoGD dataset. This dataset consists of 47933 gesture examples in the form of RGB-D videos. There are total of 249 gesture labels [25]. A subset of 48 gesture categories that occurred frequently were considered for our experiments. A seen-unseen split proposed by Lampert et al. was used to for seen and unseen classes [10]. Overall, seen and unseen set consisted of 38 and 10 gesture categories respectively, and there were 12,000 gesture examples approximately in total.

5.2 Annotation Procedure

The next step in the pipeline was to annotate these 48 gesture categories with respect to the 19 descriptors as mentioned in the section 4.1. Three subjects were asked to annotate ten sample gesture examples corresponding to each category. Annotations from each subject were processed by taking a *mode* at the descriptor level in order to obtain a final description vector for each category.

Once the annotation is finished, we observed that there is a considerable data imbalance not at the instance level but at the descriptor level. Table 2 shows the percentage of zeros and ones w.r.t each descriptor which indicates the extent of data imbalance. To combat this issue, we up-sampled the sparse descriptor labels during the training procedure.

Moreover, we observed that the gesture samples corresponding to the same class were not consistent with each other. In other words, there are significant differences in the way the gesture is performed by different individuals. This makes the problem more challenging and contributes towards reducing the zero shot learning accuracies. Hence we annotated ten gesture examples per category and aggregated the annotations using *mode* operator.

5.3 Zero Shot Learning

During preprocessing, we reshaped the gesture videos to 240 x 240 to be consistent with the pretrained VGG model and re-sampled the gesture examples to 25 frames ($T = 25$). Next, each frame is passed through a pre-trained VGG Network to represent each frame using a 512 dimensional vector. Hence each video is represented by 12,800 dimensional vector (25 frames x 512).

The extracted features were used to train two deep learning models: 1. Static model and 2. Temporal model. Static model is a simple two layer neural network that takes 12,800 dimensional vector as an input and predicts if the descriptor is present in a gesture. The first and second hidden layers consisted of 2000 and 200 nodes respectively. Next, the temporal model is the LSTM module that is trained to predict the descriptor labels. The LSTM module consisted of 25 cells (one per frame) followed by a fully connected network at the end which predicts the label of the descriptor. In both cases, *binary cross-entropy* loss was used to train the algorithms.

Figure 2 depicts the semantic description matrix as a binary image where zero and one indicate the absence and presence of a descriptor respectively. While some descriptors are well distributed across the gesture classes, attributes such as *O_Up* and *M_Up* are highly imbalanced i.e. attribute is either present or absent all classes.

Given the imbalanced nature of semantic descriptor distributions, the ROC-AUC metric was used to assess the performance of the SDD models on a validation set of unseen gesture instances. Performance comparisons between the static and temporal model are depicted in the Table 3 which shows the mean and standard deviations of AUC scores from a 5-fold cross validation. An AUC score of 1 indicates perfect detection of semantic descriptors and hence good classification performance by the model.

Comparing performance between the static and temporal models across the different descriptor categories (Hand Usage, Motion, Orientation, Finger State), it is noteworthy that one model is generally better than the other for some categories. Looking at model performance on Finger State descriptors (F_Thumb, F_Index, F_Middle), it is evident from Table 3 that the static model outperforms the temporal model. Alternatively, for the Motion descriptor type, the temporal Model outperforms the static model as seen by the AUC scores of M_In, M_Front. For The Orientation descriptors, there is no significant evidence to suggest that one model outperforms the other as the static model yields higher scores on O_Front and O_Back while the temporal models yields higher scores on O_In and O_Down. For descriptors which were highly under represented by the gestures in the dataset (O_Out), the AUC scores exhibited a very large standard deviation.

Zero-shot experiments were conducted on our dataset using three approaches: 1. Static, 2. Temporal and 3. Semantic Auto Encoder (SAE) [8] method on five folds. Note that SAE is a linear model while other methods have non-linear activation functions. On an average, non-linear approaches obtain significantly higher ($> 10\%$) unseen class accuracies. Note that results vary considerably across the folds as the folds are created in a completely random manner. Given the data imbalance, few folds contain gesture categories for which a large number of attributes are completely present or absent. This makes it very hard for the learning algorithm to recognize such attributes. Nevertheless, we conclude that the non-linear temporal approaches are desirable for this problem of ZSL for gesture recognition.

6 Conclusions

Deep Learning has greatly pushed the limits of gesture classification accuracies. However, the problem of Zero Shot Learning for unseen gesture recognition (ZSGL) is yet a challenging problem due to distinct source and target data distributions. Previous works have proposed a large number of gesture descriptors and linear approaches to tackle this problem of ZSGL. However, the existing benchmarks were obtained using the skeletal features ignoring the rich visual information present in RGB-D data. In this regard, the main goal of this work is to leverage on the success of deep learning and develop novel approaches to recognize unfamiliar gestures of humans. First, a small subset of 19 gesture attributes were obtained and 48 gesture classes from the ChaLearn IsoGD gesture dataset were annotated w.r.t those attributes. Next, transfer learning methods were utilized by passing the RGB-D data through a pretrained VGG model in order to construct features for the gesture examples. The obtained features were used to train the temporal

Table 4: Comparison of unseen accuracy of ZSL approaches(%). Columns indicate ZSL approaches and rows depict the folds.

Method	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean**
Static	28.7	12.5	28.2	20.4	36.1	25.2
Temporal	25.2	12.7	28.3	17.9	35.0	23.8
SAE	30.12	9.25	7.17	9.44	8.67	12.9

models such as LSTMs to identify the presence/absence of a descriptor. Results show that non-linear models perform significantly better at ZSL tasks in comparison to linear models such as Semantic Auto-Encoder (SAE). Moreover, our dataset will be made publicly available to encourage researchers to further investigate this problem of ZSGL.

References

[1] U. C. Allard, C. L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, K. Glette, F. Laviolette, and B. Gosselin. Deep learning for electromyographic hand gesture signal classification by leveraging transfer learning. *CoRR*, abs/1801.07756, 2018.

[2] H. Cheng, L. Yang, and Z. Liu. Survey on 3d hand gesture recognition. *IEEE Trans. Circuits Syst. Video Techn.*, 26(9):1659–1673, 2016.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, June 2009.

[5] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, Apr. 2006.

[6] H. Istance, A. Hyrskykari, L. Immonen, S. Mansikkamaa, and S. Vickers. Designing gaze gestures for gaming: An investigation of performance. In *Proceedings of the 2010 Symposium on Eye-Tracking Research &#38; Applications*, ETRA '10, pages 323–330, New York, NY, USA, 2010. ACM.

[7] A. Kendon. Do Gestures Communicate? A Review. *Research on Language and Social Interaction*, 27(3):175–200, 1994.

[8] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017.

[9] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, June 2009.

[10] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, Mar. 2014.

[11] A. Lascarides and M. Stone. A Formal Semantic Analysis of Gesture. *Journal of Semantics*, 26(4):393, 2009.

[12] N. Madapana, G. Gonzalez, R. Rodgers, L. Zhang, and J. P. Wachs. Gestures for picture archiving and communication systems (pacs) operation in the operating room: Is there any standard? *PLOS ONE*, 13(6):1–13, 06 2018.

[13] N. Madapana and J. Wachs. Zsgl: Zero shot gestural learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 331–335, New York, NY, USA, 2017. ACM.

[14] N. Madapana and J. Wachs. Database of gesture attributes: Zero shot learning for gesture recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[15] N. Madapana and J. Wachs. Database of gesture attributes: Zero shot learning for gesture recognition. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–8, May 2019.

[16] N. Madapana and J. P. Wachs. A semantical & analytical approach for zero shot gesture learning. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 796–801. IEEE, 2017.

[17] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot Learning with Semantic Output Codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418. Curran Associates, Inc., 2009.

[18] S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct. 2010.

[19] G. Patterson and J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, June 2012.

[20] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France, July 2015. PMLR.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[22] J. Wachs, H. Stern, Y. Edan, M. Gillam, C. Feied, M. Smith, and J. Handler. A Real-Time Hand Gesture Interface for Medical Visualization Applications. In *Applications of Soft Computing*, pages 153–162. Springer, Berlin, Heidelberg, 2006. DOI: 10.1007/978-3-540-36266-1_15.

[23] J. Wachs, H. Stern, Y. Edan, M. Gillam, C. Feied, M. Smith, and J. Handler. Gestix: A Doctor-Computer Sterile Gesture Interface for Dynamic Environments. In *Soft Computing in Industrial Applications*, pages 30–39. Springer, Berlin, Heidelberg, 2007. DOI: 10.1007/978-3-540-70706-6_3.

[24] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 761–769, June 2016.

[25] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016.

[26] Y. Wang, T. Yu, L. Shi, and Z. Li. Using human body gestures as inputs for gaming via depth analysis. In *2008 IEEE International Conference on Multimedia and Expo*, pages 993–996, June 2008.

[27] J. Wu, K. Li, X. Zhao, and M. Tan. Unfamiliar dynamic hand gestures recognition based on zero-shot learning. In L. Cheng, A. C. S. Leung, and S. Ozawa, editors, *Neural Information Processing*, pages 244–254, Cham, 2018. Springer International Publishing.

[28] L. Zhang, G. Zhu, L. Mei, P. Shen, S. A. A. Shah, and M. Bennamoun. Attention in convolutional lstm for gesture recognition. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1953–1962. Curran Associates, Inc., 2018.