Learning Unseen Emotions from Gestures via Semantically-Conditioned Zero-Shot Perception with Adversarial Autoencoders

Abhishek Banerjee, Uttaran Bhattacharya, Aniket Bera

Department of Computer Science, University of Maryland
College Park, Maryland 20742, USA
{abanerj8, uttaranb, bera}@umd.edu

Abstract

We present a novel generalized zero-shot algorithm to recognize perceived emotions from gestures. Our task is to map gestures to novel emotion categories not encountered in training. We introduce an adversarial autoencoder-based representation learning that correlates 3D motion-captured gesture sequences with the vectorized representation of the natural-language perceived emotion terms using *word2vec* embeddings. The language-semantic embedding provides a representation of the emotion label space, and we leverage this underlying distribution to map the gesture sequences to the appropriate categorical emotion labels. We train our method using a combination of gestures annotated with known emotion terms and gestures not annotated with any emotions. We evaluate our method on the MPI Emotional Body Expressions Database (EBEDB) and obtain an accuracy of 58.43%. We see an improvement in performance compared to current state-of-the-art algorithms for generalized zero-shot learning by an absolute 25–27%. We also demonstrate our approach on publicly available online videos and movie scenes, where the actors' pose has been extracted and mapped to their respective emotive states.

Introduction

Emotion learning as an area of research is integral to a variety of domains, including human-computer interaction, robotics (Liu et al. 2017) and affective computing (Yates et al. 2017). Existing research in emotion recognition has leveraged aspects such as facial expressions (Liu et al. 2017), speech (Jacob and Mythili 2015), gestures and gaits (Bhattacharya et al. 2020a) to gauge an individual's emotional state. Studies in psychology indicate that humans perceive emotions by observing affective features such as arm swing rate, posture, and frequency of movements. Recent work such as that by Bhattacharya et al. (2020a) combine such affective features with pose dynamics extracted using spatial-temporal graph convolutional networks (ST-GCN) (Yan, Xiong, and Lin 2018) to map pose sequences to labeled emotions.

A major challenge in these machine learning-based emotion recognition algorithms is the requirement for significantly-sized, well-labeled datasets to build classification algorithms on previously labeled emotions. However, considering the wide spectrum of emotions for humans (Zhou et al. 2016) and different emotion representations, it is tedious and often prohibitively expensive to develop large-scale datasets with an adequate number of instances for every emotion. Zero-shot learning has recently drawn considerable attention to overcome such issues where labels of different classes are unavailable. It provides an alternative methodology that does not rely on existing labels. Instead, it relies on utilizing the relationships between various seen and unseen classes to automatically determine the appropriate labels.

In the generalized zero-shot learning (GZSL) paradigm,

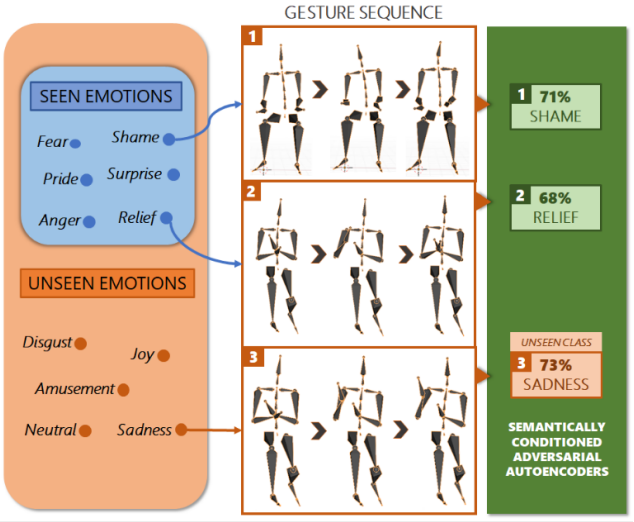

Figure 1: Generalized zero-shot Emotion Recognition from gestures[1]: We use gesture sequences from both seen and unseen classes of emotions as inputs to our AAE-based representation learning algorithm. We capture the spatial-temporal representation of 3D motion-captured gesture sequences in our network and correlate them with the semantic representation of the corresponding perceived emotion term. Our network can accurately recognize emotions not seen during training and has an overall accuracy of 58.43%.

[1]https://gamma.umd.edu/unseen_gesture_emotions

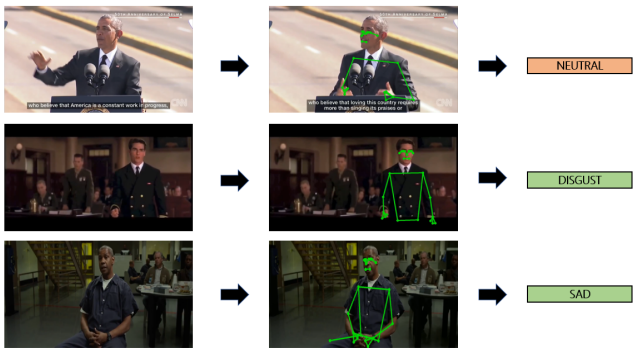

Figure 2: Public Video Results: In order to validate our results, we use publicly available videos wherein the individuals are unambiguous in their emotive state. Using the network introduced in (Pavllo et al. 2019), we extract the 3D pose of the person and then feed that into our network to recognize the emotion.

a network learns to recognize all classes, seen and unseen, while being trained with data annotations available only for seen classes. The model learns to generalize on the unseen classes by leveraging information from other modalities, such as language semantics, to create class embeddings corresponding to each label. Recent approaches to the zero-shot problem have used generative models (Mishra et al. 2018) to synthesize features for the unseen classes, which are then used for the classification task. GANs and VAEs have been the most prominent methods to synthesize these features. However, Shi et al. (Shi et al. 2019) have shown that the representation of multi-modal distributions by VAEs can result in sub-optimally learned representations. While GANs can create higher quality features than VAEs, the latent distribution spaces they learn can be susceptible to mode collapse(Goodfellow et al. 2014).

On the other hand, adversarial autoencoders (AAEs) create more closely aligned latent distributions than VAEs or GANs (Makhzani et al. 2015). Therefore, we build on the network by Makhazani et al. (Makhzani et al. 2015) to develop our network architecture.

Main Results. We present a generalized zero-shot algorithm to recognize perceived emotions from 3D motion-captured gesture sequences represented as upper-body poses. Zhan et al. (2019) have previously shown emotion perception from images in a zero-shot paradigm. To capture the semantic relationships between the emotion classes, we leverage the rich word embeddings of the pre-trained *word2vec* model (Mikolov et al. 2013). A fully supervised emotion recognition network generates a feature vector corresponding to a sequence of gesture inputs. We use an autoencoder architecture coupled with an adversarial loss to generate latent representations for the gesture-based feature vectors learned from the fully supervised network corresponding and another adversarial loss to align these latent representations with the semantically-conditioned distribution space of the emotion classes. Our main contributions include:

1. A generalized zero-shot learning (GZSL) algorithm, `SC-AAE`, based on a semantically-conditioned adversarial autoencoder architecture. We train it to learn a mapping between the gesture-feature vectors corresponding to 3D motion-captured gesture sequences and the seen and unseen perceived emotion classes expressed in natural language. To the best of our knowledge, our method is the first to classify unseen perceptual affective labels in a zero-shot learning fashion.
2. A fully supervised emotion recognition algorithm, `FS-GER` that classifies 3D motion-captured gesture sequences to seen emotion classes. We use this architecture to generate the feature vectors for input to our `SC-AAE` for generalized zero-shot learning.

Our fully supervised network achieves a validation accuracy of 77.61% with the seen emotion classes in the MPI Emotional Body Expressions Database (EBEDB) (Volkova et al. 2014), which outperforms state-of-the-art methods for fully supervised action and emotion recognition by 7–18% on the absolute. More importantly, we achieve an accuracy of 58.43% on EBEDB over the collective set of 11 seen and unseen emotion classes, outperforming state-of-the-art ZSL methods by 25–27% on the absolute.

Related Work

We provide an overview of emotion representation, emotion recognition from non-verbal body expressions, and relevant developments in zero-shot learning.

Emotion Recognition

Recent works in emotion recognition showcase the correlation between gaits and inherent psychological stress (Sanders et al. 2016). Sapiński et al. (2019) use deep learning methods to identify emotion states from gestures extracted from videos. Studies by Wegrzyn et al. (2017) identified peoples' emotional states through psychological studies of human facial expressions. With the advent of deep learning, various works have emerged that use vision-based methods (Akputu, Seng, and Lee 2013) to determine emotional state from facial expressions or audio signals using speech (Deng et al. 2017). Recently, a number of works have used multiple modalities, including speech and facial expressions, in determining emotions (Albanie et al. 2018). One distinction that needs to be made is between emotion recognition from gestures and action recognition. Action recognition methods learn a latent space fine-tuned for actions while we learn a latent space fine-tuned for emotions. Emotion recognition relies more on the relative movements of adjacent groups of nodes. On the other hand, as stated by Bhattacharya et al. (2020b), the action recognition methods STGCN, DGNN, and MS-G3D focus more on the movements of the leaf nodes, *i.e.*, hand indices, toes, and head. These nodes are useful for distinguishing between actions such as running and jumping but do not contain sufficient information to distinguish between perceived emotions.

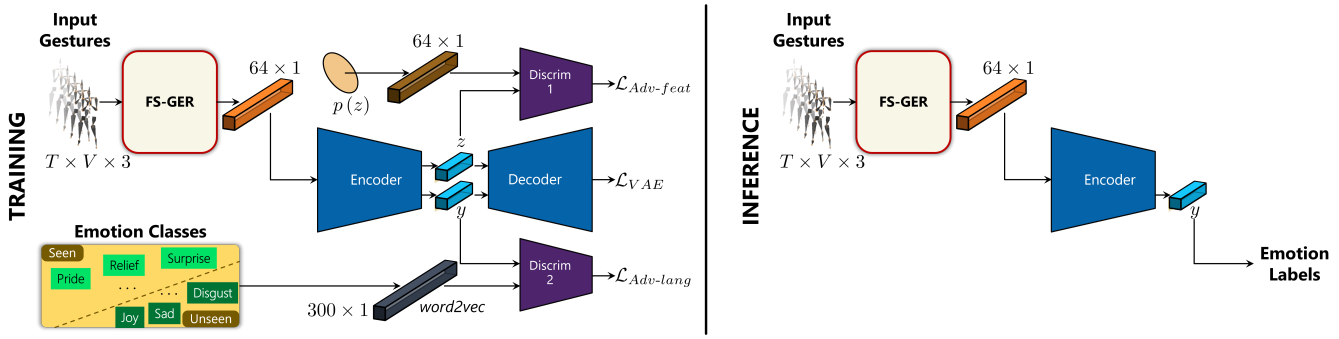

Figure 3: Network overview: Our network consists of a feature extraction pipeline that takes the sequence of gestures (T: time steps, V: joints or nodes) and extracts the relevant high-level features. We feed these features into a semantically-conditioned adversarial autoencoder, which projects them onto a latent space by aligning it with the word-level semantic features provided by *word2vec*. Our network encoder generates two latent vectors corresponding to the embeddings for the gestures and the word embeddings. We use two discriminators to adversarially train the latent distribution spaces of both the gestures ($\mathcal{L}_{Adv\text{-}feat}$) and the semantic embeddings ($\mathcal{L}_{Adv\text{-}lang}$). During inference, we only require the learned latent word embeddings to predict the emotion labels.

Generalized Zero-Shot Learning

In zero-shot learning (ZSL), both seen and unseen data are used for training, but label prediction is only attempted and evaluated on the unseen classes. By contrast, in generalized zero-shot learning (GSZL), the prediction task is executed for both seen and unseen classes. GZSL is more challenging than nominal ZSL because of the hubness problem (Dinu, Lazaridou, and Baroni 2014), which occurs when the model overfits to the trained classes. Recently, generative methods have become popular in GZSL, which uses either generative adversarial networks (GANs) (Mishra et al. 2018) or variational autoencoders (VAEs) (Schonfeld et al. 2019) to generate features for unseen classes. Traditional GZSL generative models rely on a data augmentation method, which generates features of interest that have been hitherto unseen by the model during training.

Hubert et al. (Hubert Tsai, Huang, and Salakhutdinov 2017) have shown that mapping the joint visual-language features to a joint latent space instead of the language space gives higher accuracy. Schonfeld et al. (2019) use unconditional VAEs and achieve multi-modal alignment via cross-reconstruction and distribution alignment. In our algorithm, we build on the network used by them to perform our latent space embedding and classification tasks. Considering the multiple modalities used to learn the distribution in our approach, *e.g.*, language semantics for emotions and gestures, we rely on methods that correlate the learned distributions of these modalities to estimate the semantic relation between the classes accurately.

Method

In this section, we define the problem statement and describe our approach in detail. We present an overview of our proposed algorithm in Figure 3. It consists of two main components, the fully supervised gesture emotion recognition (`FS-GER`) network and the semantically-conditioned adversarial autoencoder (`SC-AAE`). We train `FS-GER` to transform sequences of 3D motion-captured poses at the input to emotion-aware feature vectors. We also use the *word2vec* representation to obtain the semantic word-level embedding for the specific emotion label names. We use the feature vectors and the corresponding semantic embedding as inputs to `SC-AAE`. The encoder part of `SC-AAE` outputs a class semantic label as well as a latent vector. We pass the generated label and the latent vector through two corresponding discriminators that use the adversarial loss to discriminate between the generated and the ground-truth values for both the labels and the latent vectors. For classification, we use the encoder to output the corresponding semantic labels, which we then match with the relevant class labels.

Problem Definition

We formally define our problem statement in this section. Let

$$S = \{(x, y, c(y)) \mid x \in X, y \in Y^S, c(y) \in C\} \quad (1)$$

be a set of input data. Here, x denotes an input vector embedding representing a sequence of gestures, y is the corresponding class label, which in our approach is the associated emotion, and $c(y)$ is the semantic embedding corresponding to the class label. In our work, we use the *word2vec* representation for the semantic description (described later in Section). We also have the auxiliary training set

$$U = \{(u, c(u)) \mid u \in Y^U, c(u) \in C\} \quad (2)$$

for all the unseen classes. Here, u denotes an unseen class from the set Y^U, which is disjoint from Y^S. Our task at hand is the GZSL task, which evaluates the network on both seen and unseen classes, denoted by $f_{GZSL} : X \rightarrow Y^S \cup Y^U$.

We approach our problem of GZSL in the transductive setting (Wan et al. 2019). In the transductive setting, the pretrained network has access to the unseen classes, but the data points in these classes do not have any associated labels. We create a single dummy label for all the gestures belonging to all the unseen classes during feature extraction using the fully trained `FS-GER` network.

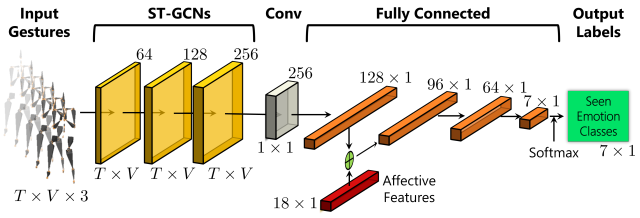

Figure 4: Fully Supervised Network for Emotion Recognition from Gestures (FS-GER): Our network comprises of three ST-GCN layers, followed by a single 1×1 convolution layer. The input data is of the form T: time Steps (510 at 30 fps) $\times$ V: nodes (10 joints) $\times$ 3 (dimension of nodes). The convolution output is appended with the affective features, A, and then passed through subsequent Fully Connected (FC) layers to generate a 64-dimensional feature description vector. This layer is passed through an FC of size 7 (total number of classes on which it is trained). The softmax layer uses this for classification. The 64-dimensional embedding is extracted from the network after the fully supervised step for the GZSL task.

Fully Supervised Gesture Emotion Recognition (FS-GER)

We show an overview of `FS-GER` in Figure 4. The input to the network is a sequence of poses of size T (time steps) $\times$ V (nodes) $\times$ 3 (position coordinates). Because gestures are a periodic sequence of poses, we use ST-GCN (Yan, Xiong, and Lin 2018) to capture the localized spatial and temporal relationships between the pose joints for the input gaits. The first ST-GCN layer has 64-layers while the second and third have 128 and 256-layers, respectively. We pass the output of each ST-GCN layer through a ReLU activation function and a BatchNorm layer. We feed the output of last ST-GCN layer through a 1×1 convolution layer, giving a 128-dimensional feature. We append this feature with an affective feature computed directly from the gestures (described next). We subsequently pass the appended vector through three fully connected (FC) layers to give a 7×1 feature vector for the seven emotion classes in our transductive setting (six seen classes Y^S and one dummy class for all the unseen emotions Y^U). We use softmax to produce the label probabilities for classification.

Affective features. Affective features are physiological measures that humans are known to observe when perceiving others' emotions. Following prior work (Randhavane et al. 2019), we consider two kinds of affective features, posture and motion features.

- **Posture features.** These consist of distances between pairs of joints, as well as angles and areas formed by three joints.
- **Motion features.** These consist of velocity and acceleration of joints of interest in the gesture.

However, different from prior work, we consider only the upper-body joints for computing affective features since gestures are predominantly expressed in the upper part of the body. Based on visual perception and psychology literature (Crenn et al. 2016), we use a total of 18 affective features as summarized in Table 1.

Features	Description
Volume	Bounding Box
Angle	With shoulders at neck With neck and left shoulder at right shoulder With neck and right shoulder at left shoulder With vertical-direction and back at neck With head and back at neck
Distance	Between right wrist and root joint Between left wrist and root joint
Area	Triangle between neck and wrists
Speed	Of left wrist Of right wrist Of head
Acceleration	Of left wrist Of right wrist Of head
Jerk	Of left wrist Of right wrist Of head

Table 1: Affective Features. We extract the posture and the motion features from an input gait using emotion characterization in visual perception and psychology literatures.

Language Embedding

The key idea in our zero-shot learning is to utilize the semantic relationship between multiple classes of emotions to determine the association between various gesture sequences and the seen and unseen emotion classes. The *word2vec* (Mikolov et al. 2013) representation gives a 300-dimensional embedding vector based on the semantics of the word. Using the vector representations for all emotions, we can ascertain the level of "closeness" or "disparity" between them. For the unseen classes, these representations give us the underlying relationship between instances of that class and other classes in the seen and unseen domains, allowing us to classify them into the appropriate categories.

We represent the set of emotions as

$$\mathcal{E} = \{e_1, e_2, e_3, \ldots, e_n\}, \tag{3}$$

where $\{e_i\} \in \mathbb{R}^{300}$ is the *word2vec* representation of the emotion-word. This way, we can relate two specific emotions by the Euclidean ℓ_2-norm distance to ascertain their adjacency.

Semantically-Conditioned Adversarial Autoencoder (SC-AAE)

In our current method, we build on the work of Makhzani et al. (2015) to create an adversarial autoencoder, which learns from the semantic distributions of data in the language space as well as the gesture space. We regularize the VAE in

such a setting by matching the posterior $q(z \mid x)$ to a prior $p(z)$ distribution. The training of the network takes place in two phases:

- the *reconstruction phase*, where the autoencoder updates the encoder and the decoder to minimize the reconstruction error of the inputs, and
- the *regularization phase*, where the adversarial network first updates its discriminative network to separate the true samples from the generated samples. The generator we use to compute the adversarial loss in our case comes from the encoder network of the VAE.

Network Architecture

FS−GER (Figure 3) outputs a 64-dimension feature vector for the respective gesture input sequence. Correspondingly, we get the 300-dimensional language embedding using *word2vec*. The encoder for SC−AAE predicts the latent vector corresponding to the gesture z and the class semantic label, $\hat{y}$. We pass the generated labels and vectors through two separate discriminators that help discriminate between the desired samples from the prior and those generated by the encoder. After the training, we use the encoder to generate the relevant semantic labels, which identifies the predicted emotion labels corresponding to that gesture inputs.

Loss Functions

We aim to minimize the cross-alignment loss between gestures and the word-labels. Since we have two separate modalities, we utilize two separate VAEs akin to those in (Schonfeld et al. 2019) to map the inputs to a common latent space. We write our VAE loss as

$$\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta\left(x^{(i)} \mid z\right)\right] - \beta D_{KL}\left(q_\phi\left(z \mid x^{(i)}\right) \| p_\theta(z)\right). \tag{4}$$

Here, D_{KL} denoted the Kullback–Leibler (KL) divergence that aligns the desired distributions. In our algorithm, we further use the adversarial losses to align the prior distributions with the encoder output.

Adversarial Loss. Following standard formulation, we write the adversarial loss for a discriminator as

$$\mathcal{L}_{Adv} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{\overline{\mathbf{x}}\sim p_\theta(\tilde{\mathbf{x}}|\mathbf{z},\mathbf{a})}\left[\log\left(1 - D(\tilde{\mathbf{x}})\right)\right]. \tag{5}$$

We use two adversarial losses in our network, corresponding to the two discriminators. For the label discriminator, a corresponds to $c(y)$, which is an element of $\mathcal{E}$, in Equation 3. We denote this by $\mathcal{L}_{Adv\text{-}lang}$. For the feature discriminator, a corresponds to an element from the generated features from a prior distribution $p(z)$ and we denote the adversarial loss for this by $\mathcal{L}_{Adv\text{-}feat}$.

Collectively, we can write our net loss as

$$\mathcal{L}_{net} = \mathcal{L}_{VAE} + \gamma\mathcal{L}_{Adv\text{-}lang} + \delta\mathcal{L}_{Adv\text{-}feat}, \tag{6}$$

where γ and δ are weighing functions.

Experiments and Results

We present experiments and results for our zero-shot classification task in this section, including the details of our network and the hardware configuration.

Dataset

We train and evaluate our network on the MPI Emotional Body Expressions Database (EBEDB) (Volkova et al. 2014). It consists of $1,447$ 3D motion-captured sequences of natural-emotion body gestures from actors as they narrated specific lines. All body movements were captured at 120 fps. The original dataset consists of information regarding 23 joints in the body. However, because we are interested in gestures made by the upper body, we select $V = 10$ joints: the head, neck, right-shoulder, left-shoulder, right-elbow, left-elbow, right-wrist, left-wrist, backbone, and pelvis. We ignore the lower-body joints as there is no significant motion in those joints. Each sequence is annotated with one of 11 categorical emotion classes.

To evaluate our model, we split the 11 available emotion classes in MPI EBEDB into a roughly equal split of six seen classes and five unseen classes. To ensure an unbiased evaluation, we split the 11 emotion labels randomly into two sets of seen and unseen labels five times. The results we present here are the mean values of all five experiments. During the training phase, the model learns only from the six seen classes. Since there are multiple possible combinations for choosing these five unseen classes and there are no fixed criteria in particular for this dataset for zero-shot learning, we conduct five experiments in which we successively select five random classes from the available 11 classes. Our results are averaged over these five experiments. We use a train-test split of $80\% - 20\%$.

Currently, MPI EBEDB is the only publicly dataset that maps human gestures to their emotional states. As a result, we evaluate and discuss our results only on this particular dataset. However, to validate our results, we use popular online videos and movie scenes that are largely unambiguous with regard to the emotional states of the people or the characters. For validation, we further labeled these videos with expert annotators. We show some of these video snapshots in Figure 2.

Training Details

All our encoders and decoders are multi-layer perceptrons with two hidden layers. More hidden layers reduce the performance because the gesture-features and language embeddings are very high-level representations and generally sparse; hence more layers would result in loss of crucial features for classification.

We use 100 hidden units each for the encoder and the decoder. The discriminators consist of two hidden layers with 100 hidden layers each for the language-embedding model, while the discriminator for the gesture-feature vector has two hidden layers of size 100 and 32, respectively. In our work, we use our FS−GER to generate a 64-dimension feature vector corresponding to the gestures and a 300-dimension *word2vec* feature encoding the emotions.

We train the model for 200 epochs by stochastic gradient descent using the Adam optimizer (Kingma and Ba 2014) and a batch size of 6 for features. Each batch consists of pairs of extracted gesture features and matching attributes from different seen classes. Pairs of data always belong to

Method	*Accuracy*
ST-GCN (Xian et al. 2018)	59.12%
STEP (Bhattacharya et al. 2020a)	70.38%
FS-GER (Ours)	77.61%

Table 2: Classification accuracies for fully supervised emotion recognition methods on the seen emotion classes (bold is best).

the same class. We keep the values of γ and δ constant and discuss how we choose their values in Section . Our network takes around 6 minutes to train on an Nvidia RTX 2080 GPU.

Performance of FS-GER

We compare the performance of FS-GER with previous methods on emotion recognition (Bhattacharya et al. 2020a), as well as action recognition (Yan, Xiong, and Lin 2018). In (Yan, Xiong, and Lin 2018), the authors introduce ST-GCNs to perform action recognition. The network takes a sequence of gaits as input and uses the spatial relation between the various joints and their temporal locations to create a mapping between the motion sequences and their actions. In (Bhattacharya et al. 2020a), the authors develop an emotion-specific embedding method to augment the graph convolution network's ability to map motion patterns to perceived emotions. In addition to capturing the spatial and temporal variance of the joints, they extract certain affective features that capture semantics more specific to emotions.

We show the overall network architecture for our network, FS-GER in Figure 4. We train all networks from scratch using all the upper-body joints as per their input requirements. We classify for the same set of six seen classes and one dummy class corresponding to the five unseen classes. Based on the MPI EBEDB (Volkova et al. 2014), we have $T = 510$ time steps and $V = 10$ joints in the upper body.

We report the performance of all the methods in Table 2. We observe that our method outperforms the other methods by 7–18% on the absolute, as a result of using the relevant set of joints and affective features. We use our proposed emotion classifier network to generate features for the subsequent GZSL framework.

Related Methods for GZSL

We compare with GZSL methods for image classification, which are the closest existing alternatives to our GZSL approach. Similar to our method, these methods also attempt to learn mappings from visual as well as spatial-temporal feature vectors to semantic descriptions.

We compare with state-of-the-art image classification problems in the GZSL paradigm, such as CADA-VAE (Schonfeld et al. 2019), f-CLSWGAN (Xian et al. 2018), and CVAE-ZSL (Mishra et al. 2018). Schonfeld et al. (2019) implement two separate VAEs and use cross-reconstruction losses to align them. Xian et al. (2018) use a GAN-based reconstruction to generate unseen features and leverage the Wasserstein distance to align the multiple-distributions. Mishra et al. (2018), the authors implement a standard VAE architecture and add semantic labels to the inputs for calculating the reconstruction loss.

For a fair comparison, we trained all these methods from scratch on MPI EBEDB (Volkova et al. 2014).

Evaluation Metric for GZSL

Following prior methods in the GZSL paradigm (Schonfeld et al. 2019; Xian et al. 2018), we evaluate our performance using the harmonic mean of the accuracies on the seen and the unseen classes. The harmonic mean is given by

$$H = \frac{2 \cdot acc_{seen} \cdot acc_{unseen}}{acc_{seen} + acc_{unseen}}, \tag{7}$$

where acc_{seen} and acc_{unseen} represent the accuracy of gestures from the seen and the unseen classes, respectively. The harmonic mean is preferred over the more conventional arithmetic mean in this paradigm because the arithmetic mean gives a large value if the seen class accuracy is much greater than the unseen class accuracy. By contrast, the harmonic mean only gives a large value both the seen, and the unseen class accuracies are large, providing a more accurate reflection of performance.

Performance of SC-AAE

We evaluate our proposed GZSL approach (SC-AAE) with the other approaches for the GZSL task in Table 3. We report the harmonic mean of the accuracies for the seen and the unseen classes, as achieved by each method. We observe that SC-AAE outperforms the other approaches by 25–27% on the absolute. f-CLSWGAN (Xian et al. 2018), which conditioned GANs on image classification, suffers from mode collapse. CADA-VAE (Schonfeld et al. 2019), while aligning the language-semantic and gesture-feature spaces effectively, fails to create representative features for the unseen classes, which can help in recognition. CVAE-ZSL (Mishra et al. 2018), which was built for the action recognition task, does not generate robust features for emotion recognition. We show some of the visual results for our method in Figure 5.

Analysis of the Generalized Zero-Shot Model

In this section, we present an analysis of our zero-shot learning architecture, including the choice of hyperparameters and the size of the latent space. For additional analysis and details, please refer to the technical appendix.

Hyperparameters. We use two hyperparameters, γ and δ, for regularizing the loss of our network (Equation 6). These weigh the effect of the adversarial losses of our two discriminators, for the language embedding and for the extracted gait features, on the training process. Fixing γ at 1, we varied δ between 0.1 and 2 during training. On account of the heavier usage of the *word2vec* embedding in the determination of classification accuracy, we found $\delta = 1.5$ to give us the highest harmonic mean of accuracies, and therefore we have used this value to report our results. Changing γ while keeping δ fixed at 1.5 did not result in any significant changes, as

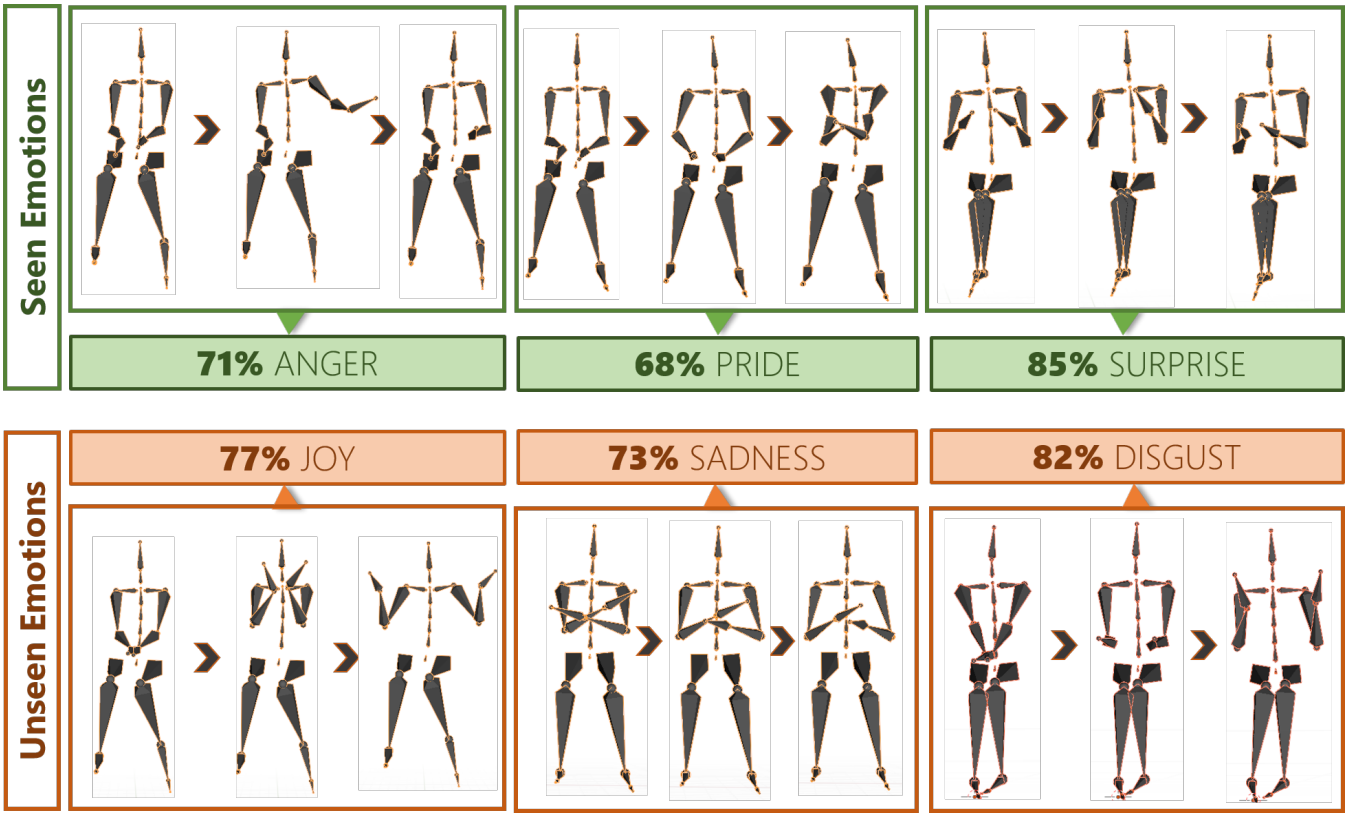

Figure 5: Visual Results: The top row shows three sets of gestures in temporal order from left to right, which map to the correct seen emotions during classification. The bottom row consists of three gestures mapped to the correct unseen emotions during training.

Method	*Harmonic Mean*
CADA-VAE (Schonfeld et al. 2019)	33.27%
f-CLSWGAN (Xian et al. 2018)	30.18%
CVAE-ZSL (Mishra et al. 2018)	31.74%
SC-AAE (Ours)	58.43%

Table 3: Harmonic mean of classification accuracies on seen and unseen classes by different methods on our GZSL task (bold is best).

these changes were largely overshadowed by the gains from changing δ. Hence, we set $\gamma = 1$ for our experiments.

Size of Latent Embedding. The latent embedding refers to the size of the gesture feature vector used in our latent space. We changed the sizes of the latent embeddings, d, from $d = 2$ to $d = 32$ in steps of one. We obtained the best results for $d = 16$ and used this in our final network.

Conclusion, Limitations and Future Work

In this work, we proposed a novel SC-AAE architecture for generalized zero-shot learning of perceived emotions from 3D motion-captured gesture sequences. We used an adversarial loss to learn mappings between the gestures and the semantically-conditioned space of emotion words to classify gestures into both seen and unseen emotions. We evaluated our approach on the MPI Emotional Body Expressions Database (EBEDB), using feature-embeddings extracted from gestures and language-embeddings from *word2vec*. Our proposed approach outperforms previous state-of-the-art algorithms for GZSL by 25–27% on MPI EBEDB.

Our work has some limitations. Since *word2vec* is a generic language-embedding model, not specific to emotions, it may not capture all aspects of psychological and emotional diversity. Therefore, we plan to leverage affective-based semantics from words in the future. We also plan to incorporate more affective modalities, including speech and eye movements, to ensure a more robust classification. Furthermore, we plan to use the dimensional emotional space spanned by VAD (Valence-Arousal-Dominance) to learn relationships between disparate categorical emotions.

Acknowledgements

This work was supported by a grant from the Brain and Behavior Institute at the University of Maryland at College Park, USA.

References

Akputu, K. O.; Seng, K. P.; and Lee, Y. L. 2013. Facial emotion recognition for intelligent tutoring environment. In *2nd International Conference on Machine Learning and Computer Science (IMLCS'2013)*, 9–13.

Albanie, S.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, 292–301.

Bhattacharya, U.; Mittal, T.; Chandra, R.; Randhavane, T.; Bera, A.; and Manocha, D. 2020a. STEP: Spatial Temporal Graph Convolutional Networks for Emotion Perception from Gaits. In *AAAI*, 1342–1350.

Bhattacharya, U.; Roncal, C.; Mittal, T.; Chandra, R.; Kapsaskis, K.; Gray, K.; Bera, A.; and Manocha, D. 2020b. Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping. In *European Conference on Computer Vision*, 145–163. Springer.

Crenn, A.; Khan, R. A.; Meyer, A.; and Bouakaz, S. 2016. Body expression recognition from animated 3D skeleton. In *2016 International Conference on 3D Imaging*, 1–7. IEEE.

Deng, J.; Xu, X.; Zhang, Z.; Frühholz, S.; and Schuller, B. 2017. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1): 31–43.

Dinu, G.; Lazaridou, A.; and Baroni, M. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hubert Tsai, Y.-H.; Huang, L.-K.; and Salakhutdinov, R. 2017. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, 3571–3580.

Jacob, A.; and Mythili, P. 2015. Prosodic feature based speech emotion recognition at segmental and supra segmental levels. In *SPICES*, 1–5. IEEE.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, Z.; Wu, M.; Cao, W.; Chen, L.; Xu, J.; Zhang, R.; Zhou, M.; and Mao, J. 2017. A facial expression emotion recognition based human-robot interaction system.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mishra, A.; Krishna Reddy, S.; Mittal, A.; and Murthy, H. A. 2018. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2188–2196.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7753–7762.

Randhavane, T.; Bhattacharya, U.; Kapsaskis, K.; Gray, K.; Bera, A.; and Manocha, D. 2019. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv:1906.11884*.

Sanders, J. B.; Bremmer, M. A.; Comijs, H. C.; Deeg, D. J.; and Beekman, A. T. 2016. Gait speed and the natural course of depressive symptoms in late life; an independent association with chronicity? *Journal of the American Medical Directors Association*, 17(4): 331–335.

Sapiński, T.; Kamińska, D.; Pelikant, A.; and Anbarjafari, G. 2019. Emotion recognition from skeletal movements. *Entropy*, 21(7): 646.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Shi, Y.; Siddharth, N.; Paige, B.; and Torr, P. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, 15718–15729.

Volkova, E. P.; Mohler, B. J.; Dodds, T. J.; Tesch, J.; and Bülthoff, H. H. 2014. Emotion categorization of body expressions in narrative scenarios. *Frontiers in psychology*, 5: 623.

Wan, Z.; Chen, D.; Li, Y.; Yan, X.; Zhang, J.; Yu, Y.; and Liao, J. 2019. Transductive zero-shot learning with visual structure constraint. In *Advances in Neural Information Processing Systems*, 9972–9982.

Wegrzyn, M.; Vogt, M.; Kireclioglu, B.; Schneider, J.; and Kissler, J. 2017. Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PloS one*, 12(5): e0177239.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5542–5551.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI conference on artificial intelligence*.

Yates, H.; Chamberlain, B.; Norman, G.; and Hsu, W. H. 2017. Arousal detection for biometric data in built environments using machine learning. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, 58–72.

Zhan, C.; She, D.; Zhao, S.; Cheng, M.-M.; and Yang, J. 2019. Zero-shot emotion recognition via affective structural embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1151–1160.

Zhou, D.; Zhang, X.; Zhou, Y.; Zhao, Q.; and Geng, X. 2016. Emotion distribution learning from texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 638–647.