Self-supervised Contrastive Learning of Multi-view Facial Expressions

Shuvendu Roy, Ali Etemad
Department of Electrical and Computer Engineering & Ingenuity Labs Research Institute
Queen’s University
Kingston, Canada
{shuvendu.roy,ali.etemad}@queensu.ca

ABSTRACT
Facial expression recognition (FER) has emerged as an important component of human-computer interaction systems. Despite recent advancements in FER, performance often drops significantly for non-frontal facial images. We propose Contrastive Learning of Multi-view Facial Expressions (CL-MEx) to exploit facial images captured simultaneously from different angles towards FER. CL-MEx is a two-step training framework. In the first step, an encoder network is pre-trained with the proposed self-supervised contrastive loss, where it learns to generate view-invariant embeddings for different views of a subject. The model is then fine-tuned with labeled data in a supervised setting. We demonstrate the performance of the proposed method on two multi-view FER datasets, KDEF and DDCF, where state-of-the-art performances are achieved. Further experiments show the robustness of our method in dealing with challenging angles and reduced amounts of labeled data.

CCS CONCEPTS
• Computer systems organization → Embedded systems; Redundancy; Robotics; • Networks → Network reliability.

KEYWORDS
Contrastive Learning, Self-Supervised Learning, Multi-View Learning, Affective Computing

ACM Reference Format:
Shuvendu Roy, Ali Etemad. 2021. Self-supervised Contrastive Learning of Multi-view Facial Expressions. In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI ’21), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3462244.3479955

1 INTRODUCTION
Facial expression recognition (FER) is an important attribute of non-verbal communication. FER systems play an important role in human-machine systems by enabling customization and adaptation of such systems to user reactions. Examples of such devices include emotion-aware multimedia and smart devices [9], personal mood management systems [35, 40], driving assistants [20], health-care assistants [41], and others. However, despite the potential for a wide variety of applications, FER remains challenging due to various factors like illuminations, scene backgrounds, occlusions, and challenging viewing angles among others.

Despite the recent progress in FER systems with deep learning [21, 22, 30], most prior work in the area focus on only utilizing frontal facial views [16, 30]. As a result, developed models may not be able to recognize expressions in faces captured from sharp side angles. Another problem in deep learning-based FER systems is the requirement for large amounts of training data given the millions of trainable parameters.

In this work, to deal with the problems of view-point sensitivity and labeled training set size, we propose Contrastive Learning for Multi-view facial Expressions (CL-MEx). Our proposed framework learns facial expression representations through a two-step training process. The first step is self-supervised pre-training, whereby our novel loss function forces the latent representations of ‘different’ views with the same expression from the same subject to be grouped together. This will allow for challenging viewpoints (e.g. side view) to be augmented by and benefit from the more effective representations of frontal views, allowing better performance (see Figure 1). After pre-training, we fine-tune the model with the labeled data in a supervised setting. Our experiments show that the CL-MEx training protocol obtains better results than fully supervised training and outperforms the state-of-the-art solutions.
on the two datasets used (KDEF [23] and DDCF [10]). Moreover, we demonstrate that our model does not require all the available labels to reach the performance of a fully supervised model. Lastly, we demonstrate that as a result of our proposed loss, our model is better equipped to deal with challenging views.

In summary, our contributions are as follows.

- We propose a novel contrastive learning framework, CL-MEx, for multi-view data. Our model utilizes a novel loss function designed specifically for multi-view images.
- Experiments demonstrate the strong performance of our approach by setting a new state-of-the-art, and reduced sensitivity to challenging views. Our model also requires less labeled data to obtain competitive results with respect to fully supervised training.

2 RELATED WORKS

In this section, we review recent works in two key areas related to this paper: self-supervised and contrastive learning, and multi-view facial expression recognition.

2.1 Self-Supervised and Contrastive Learning

To reduce the influence of output labels and human-annotations, self-supervised learning has been proposed to use specific augmentations [5] or transformations [12, 27] applied to input data, and produce pseudo-labels to train or pre-train the model. Successive to training or pre-training the model with the augmented/transformed data and associated pseudo-labels (‘ pretext’ tasks), the main building blocks of the model, for instance the convolutional blocks in a CNN, can then be used for ‘downstream’ tasks, e.g. classification [12]. In this phase, the convolutional blocks are typically frozen and new fully connected layers are added and trained from scratch using the original output labels [5]. The pre-trained components can also be fine-tuned to help achieve better performance [26].

Self-supervised learning has been utilized in a variety of different problem domains including image analysis [4, 26], wearable-based activity recognition [31, 34], and affective computing with biosignals [36–38]. Specifically for image-based recognition, a variety of different self-supervised solutions have been proposed in recent years [11, 26, 43]. A key differentiator in many of these methods is the design of novel and interesting pretext tasks. Examples include jigsaw puzzle solving (where the images are divided into pieces and shuffled with the pretext aiming to solve the puzzle) [11, 26], inpainting (where patches are removed and the pretext aims to complete the image) [29], transformation classification (where the pretext aims to classify the type of transform applied to the image, e.g., rotation, mirroring, etc.) [14], and colorization (where the pretext aims to color the image converted to gray-scale) [43].

Contrastive learning is a variation of self-supervised learning that has recently shown great success, achieving state-of-the-art performance in a variety of tasks such as object detection [42], medical image analysis [3], image classification [2, 13, 15, 28], and video-based affect classification [33]. A recent method called SimCLR [6] proposed the use of transformations, referred to as ‘ augmentations’, to create different variations of input images and maximize the agreement between positive pairs (augmented versions of the same image) via contrastive loss. A number of different variations of SimCLR have been proposed to exploit this framework and provide robust results [7, 8, 19]. Later, a variation of contrastive learning was proposed to also make use of the output labels, effectively creating a ‘supervised’ contrastive learning framework [19]. In this approach, the positive examples were generated not only from different augmentations of the same input, but also the augmented versions of other examples in the ‘same class’.

In the context of FER, which is the focus of this paper, self-supervised or contrastive learning have rarely been explored. In a recent work, FER was performed with a combination of contrastive learning and rotation prediction as pretext tasks [30].

2.2 Multi-view FER

Multi-view FER exploits facial images obtained from different angles to learn a mapping between images of the face performing specific types of expressions and the expression class. This approach allows for the methods to capture information that might otherwise be hidden from the camera in certain view-points, resulting in more robust performance. Deep learning solutions have recently been proposed to jointly learn the different view-points and perform FER in a variety of different settings [21, 22, 24].

A multi-channel pose-aware CNN (MPCNN) was proposed in [22], and showed strong performance for multi-view FER. In PhaNet [21], a pose-adaptive hierarchical attention network was proposed that could jointly recognize the facial expressions and poses. At any angle of view, the attention module helped the model identify the most relevant regions for detecting the expression. In [39], a new LSTM cell architecture was proposed to jointly learn different perspectives, which was then used for a number of different tasks, including multi-view FER.

3 METHOD

We propose a framework called CL-MEx for exploiting images acquired from multiple views to perform FER with self-supervised contrastive learning. Like SimCLR [6], our proposed method uses positive and negative examples to carry out contrastive learning. However, we do not limit the definition of positive pairs as augmentations of a particular input, and expand this notion to include the augmentations of all the views of the input image. CL-MEx then uses our proposed loss function to bring the representations of different views of a subject with the same expression closer in the embedding space (as shown earlier in Figure 1). Successive to the pre-training step, we fine-tune the model by supervised learning using the labels from the dataset. The proposed method uses the multiple views of an image in the self-supervised pre-training step only, while at inference time, it predicts the expression using only a single-view image. Following we present the details of our proposed method including the encoder architecture, proposed loss, and implementation details.

3.1 CL-MEx Components

The main components of our proposed framework are as follows.

Data Augmentation Module, \( \text{Aug}() \). This module randomly applies combinations of the following augmentations: random cropping, random resizing, random horizontal flip, random color distortion, and random gray-scaling, as \( x' = \text{Aug}(x) \), where \( x \) is the
We use the embedding dimension of 512 for all our experiments. In which gives an output dimension of 128. We normalize the output of this layer to enable using a linear product to measure distances in the projection space. This projection has proven to be important for learning effective representations by self-supervised pre-training [6]. After the pre-training, we discard the projection head and only fine-tune the encoder for the FER task.

3.2 Multi-view Contrastive Loss

For the purpose of pre-training the model in a self-supervised setting, we propose a contrastive loss function which is inspired by SupCon [19]. However, unlike SupCon, we only need the multi-view information to train the model and do not use the output labels in the pre-training stage.

In a batch of size $2N$, the images are obtained by applying two random augmentations on $N$ random input samples. These images are passed through the encoder and the projection head to generate the embedding vector $Z$, on which the contrastive loss is calculated. The multi-view contrastive loss is defined as:

$$
\mathcal{L}_{\text{CL-MEx}} = \frac{2N}{i=1} \left( \frac{1}{2Nv_i} \sum_{j=1}^{2N} \eta_{ij} \cdot \mathbb{1}_{v_i=v_j} \cdot \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2N} \exp(z_i \cdot z_k / \tau)} \right),
$$

where $v_i$ is the view-invariant id of image $i$, which indicates that for $v_i = v_j$, $i$ and $j$ are different views of the same subject. $\mathbb{1}_{[a,b]} \in \{0, 1\}$ is a function of $a$ and $b$, which outputs 1 when $a = b$, and 0 otherwise. $N_v$ is the number of instances with id $v_i$.

3.3 Downstream FER

While the projection head is an important component for learning representations in the contrastive pre-training step, it is no longer required for the final downstream FER classification task. During the second step for FER, we discard the projection head and add a linear layer for predicting the output class probability. We then freeze the encoders and train the new linear layer with class labels using categorical cross-entropy loss. This is followed by fine-tuning the entire model including the encoders to maximize performance.

3.4 Implementation Details

During the pre-training step, the model is trained with the proposed multi-view contrastive loss function for 500 epochs with an Adam optimizer, a learning rate of 0.0001 with cosine learning rate decay, and weight decay of 1e-4. The downstream FER step consists of a total of 60 training epochs. First, we train the newly added classification layer for 10 iterations, and then fine-tune the full model for 50 additional iterations. During this step, an Adam optimizer was again used, this time with a plateau learning rate decay with an initial learning rate of 1e-4, decay factor of 0.5, and patience of 3. The method is implemented with PyTorch and trained on 4 NVIDIA V100 GPUs. All image resolutions are 224 $\times$ 224.

4 EXPERIMENTS AND RESULTS

This section describes the experimental setup and results for CL-MEx on two benchmark datasets that contain multi-view facial expressions, Karolinska Directed Emotional Faces (KDEF) [23] and Dartmouth Database of Children’s Faces (DDCF) [10].

4.1 Datasets

KDEF [23] is a multi-view emotion recognition dataset with 7 expressions (afraid, angry, happy, sad, surprised, disgust, and neutral), where each subject is imaged from 5 different views (-90$^\circ$: full left (FL), -45$^\circ$: half left (HL), 0$^\circ$: straight (S), +45$^\circ$: half right (HR), and +90$^\circ$: full right (FR)). The dataset was collected from 140 subjects. The DDCF [10] is also a multi-view dataset collected from 80 subjects. It contains 8 expression classes (afraid, angry, happy, sad, surprised, disgust, pleased, and neutral), from 5 different views (-60$^\circ$: full left (FL), -30$^\circ$: half left (HL), 0$^\circ$: straight (S), +30$^\circ$: half right (HR), and +60$^\circ$: full right (FR)).

4.2 Results

Accuracy is measured for the downstream network which predicts the expression labels. The input to this network is a single image (a single view of a subject). The accuracy is therefore measured as the rate of correct classifications over all the test images in 10-fold cross-validation. Table 1 compares the results of our proposed method with previous methods on KDEF and DDCF datasets. We
Table 1: Results and comparison with other methods.

(a) KDEF

| Method   | Acc. ± SD |
|----------|-----------|
| SVM [25] | 70.5±1.2  |
| SURF [32]| 74.05±0.9 |
| TLCNN [44]| 86.43±1.0 |
| PhaNet [21]| 86.5     |
| MPCNN [22]| 86.9±0.6 |
| RBFNN [24]| 88.87    |
| CL-MEx   | 94.64±0.92|

(b) DDCF

| Method   | Acc. ± SD |
|----------|-----------|
| LBP [18] | 82.3      |
| SVM [1]  | 91.27     |
| RBFNN [1]| 91.80     |
| CL-MEx   | 95.26±0.84|

Figure 4: Drop in performance for different viewing angles with respect to the frontal view, for self-supervised CL-MEx and its fully supervised counterpart.

Figure 5: Performance of self-supervised CL-MEx and its fully supervised counterpart when different amounts of labeled data are used.

Next, we perform a sensitivity study on the proposed method for the amount of labeled data by comparing the performance of CL-MEx to its fully supervised counterpart when 75%, 50%, 25%, 10%, and 5% of the labels are used. The results are presented in Figure 5, where when all the data are used, our method shows 4% and 3% improvements over the fully supervised model for KDEF and DDCF datasets respectively. It is then seen that as the amount of output training labels are reduced, our method experiences smaller drops in performance versus the fully supervised model, showing a clear advantage by reduced reliance on the output labels.

5 CONCLUSION AND FUTURE WORK

This paper presents a self-supervised contrastive framework specialized in learning facial expressions from multiple views. Our method uses the multi-view images of the subjects during the pre-training and learns view-invariant embedding representations. The model is then fine-tuned in a supervised setting. We test our solution on KDEF and DDCF dataset and perform rigorous experiments. The model outperforms existing methods and shows robustness towards challenging viewing angle. Experiments also show the robustness of our model towards reduced amounts of labeled data. For future work, the proposed method can be used for other multi-view or multi-modal image-based applications in self-supervised and contrastive learning settings, for example multi-view face/activity recognition or even multi-view object recognition.

ACKNOWLEDGEMENTS

We would like to thank BMO Bank of Montreal and Mitacs for funding this research.
