A Compact Embedding for Facial Expression Similarity

Raviteja Vemulapalli
Google AI
ravitejavemu@google.com

Aseem Agarwala
Google AI
aseemaa@google.com

Abstract

Most of the existing work on automatic facial expression analysis focuses on discrete emotion recognition, or facial action unit detection. However, facial expressions do not always fall neatly into pre-defined semantic categories. Also, the similarity between expressions measured in the action unit space need not correspond to how humans perceive expression similarity. Different from previous work, our goal is to describe facial expressions in a continuous fashion using a compact embedding space that mimics human visual preferences. To achieve this goal, we collect a large-scale faces-in-the-wild dataset with human annotations in the form: Expressions A and B are visually more similar when compared to expression C, and use this dataset to train a neural network that produces a compact (16-dimensional) expression embedding. We experimentally demonstrate that the learned embedding can be successfully used for various applications such as expression retrieval, photo album summarization, and emotion recognition. We also show that the embedding learned using the proposed dataset performs better than several other embeddings learned using existing emotion or action unit datasets.

1. Introduction

Automatic facial expression analysis has received significant attention from the computer vision community due to its numerous applications such as emotion prediction, expression retrieval (Figure 1), photo album summarization, candid portrait selection [14], etc. Most of the existing work [32, 37] focuses on recognizing discrete emotions or action units defined by the Facial Action Coding System (FACS) [13]. However, facial expressions do not always fit neatly into semantic boxes, and there could be significant variations in the expression within the same semantic category. For example, smiles can come in many subtle variations, from shy smiles, to nervous smiles, to laughter. Also, not every human-recognizable facial expression has a name. In general, the space of facial expressions can be viewed as a continuous, multi-dimensional space.

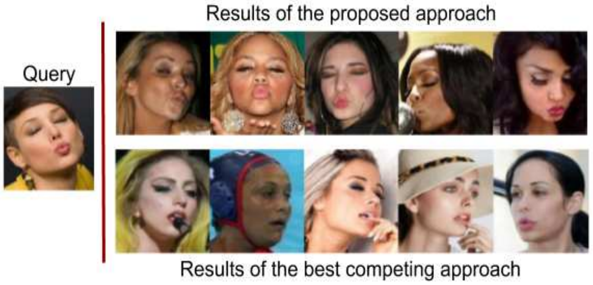

Results of the proposed approach
Query
Results of the best competing approach

Figure 1: Expression retrieval results for embeddings learned using the proposed dataset (top) and an existing emotion classification dataset (bottom).

In this work, we focus on learning a compact, language-free, subject-independent, and continuous expression embedding space that mimics human visual preferences. If humans consider two expressions to be visually more similar when compared to a third one, then the distance between these two expressions in the embedding space should be smaller than their distances from the third expression. To learn such an embedding we collect a new dataset, referred to as the Facial Expression Comparison (FEC) dataset, that consists of around 500K expression triplets generated using 156K face images, along with annotations that specify which two expressions in each triplet are most similar to each other. To the best of our knowledge, this is the first large-scale face dataset with expression comparison annotations. This dataset can be downloaded from https://ai.google/tools/datasets/google-facial-expression/.

We show that a compact (16-dimensional) expression embedding space can be learned by training a deep network with the proposed FEC dataset using triplet loss [48]. Based on the distances in the learned embedding space, we are able to predict the most similar pair in a triplet with an accuracy of 81.8% when evaluated on a held-out validation set. The accuracy of median human rater is 87.5% on this validation set, and the accuracy of random selection is 33.3%. We also show that the embedding learned using the FEC dataset performs better than several other embeddings learned using existing emotion or action unit datasets.

We experimentally demonstrate that the expression embedding learned using the FEC dataset can be successfully used for various applications such as expression retrieval, photo album summarization, and emotion recognition.

1.1. Our contributions

- We introduce the FEC dataset, which is the first large-scale face dataset with expression comparison annotations. This dataset is now publicly available.
- We experimentally demonstrate that a 16-dimensional expression embedding learned by training a deep neural network with the FEC dataset can be successfully used for several expression-based applications.
- We show that the embedding learned using the FEC dataset performs better than several other embeddings learned using existing emotion or action unit datasets.

2. Related work

Most of the existing research in the area of automatic facial expression analysis focuses on the following three topics: *(i) Categorical model:* Assigning discrete emotion category labels, *(ii) FACS model*: Detecting the presence/absence (and the strength) of various action units defined by FACS [13], and *(iii) Dimensional model*: Describing emotions using two or three dimensional models such as valence-arousal [47], pleasure-arousal-dominance [40], etc. Summarizing the vast amount of existing research on these topics is beyond the scope of this paper and we refer the readers to [29, 32, 37] for recent surveys on these topics.

Expression datasets: Several facial expression datasets have been created in the past that consist of face images labeled with discrete emotion categories [4, 10, 11, 12, 17, 18, 33, 36, 42, 43, 45, 56, 57], facial action units [4, 36, 38, 39, 45], and strengths of valence and arousal [27, 29, 30, 42, 46]. While these datasets played a significant role in the advancement of automatic facial expression analysis in terms of emotion recognition, action unit detection and valence-arousal estimation, they are not the best fit for learning a compact expression embedding space that mimics human visual preferences.

Expression embedding: A neural network was trained in [41] using an emotion classification dataset and category label-based triplet loss [48] to produce a 128-dimensional embedding. Emotion labels do not provide information about within-class variations and hence a network trained with label-based triplets may not encode fine-grained expression information. The proposed FEC dataset addresses this issue by including expression comparison annotations for within-class triplets.

A self-supervised approach was proposed in [28] to learn a 256-dimensional facial attribute embedding by watching videos, and the learned embedding was used for multiple tasks such as head pose estimation, facial landmarks prediction, and emotion recognition by training an additional classification or regression layer using labeled training data. However, as reported in [28], its performance is worse than existing approaches on these tasks. Different from [28], we follow a fully-supervised approach for learning a compact (16-dimensional) expression embedding. A modified Lipschitz embedding [24] was used in [5] to embed faces in a low-dimensional space for expression recognition.

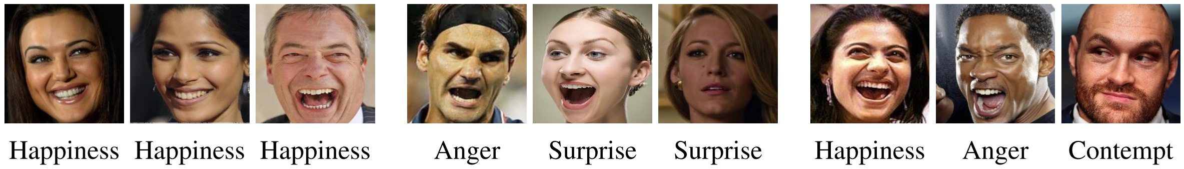
Triplet loss-based representation learning: Several existing works have used triplet-based loss functions for learning image representations. While majority of them use category label-based triplets [15, 20, 21, 34, 48, 50, 53, 58], some existing works [6, 52] have focused on learning fine-grained representations. While [52] used a similarity measure computed using several existing feature representations to generate groundtruth annotations for the triplets, [6] used text-image relevance based on Google image search to annotate the triplets. Different from these approaches, we use human raters to annotate the triplets. Also, none of these works focus on facial expressions.

3. Facial expression comparison dataset

In this section, we introduce the FEC dataset, which is a large-scale faces-in-the-wild dataset with expression comparison annotations provided by human raters. To the best of our knowledge, there is no such publicly-available expression comparison dataset. Most of the existing expression datasets are either annotated with emotion labels, or facial action units, or strengths of valence and arousal.

One may think that we could generate comparison annotations for the existing datasets using the available emotion or action unit labels. However, there are several issues with such an approach:

- Emotion labels do not provide information about within-class variations and hence we cannot generate comparison annotations within a class. For example, while all the expressions in Figure 2(a) fall into the *Happiness* category, the left and middle expressions are visually more similar when compared to the right expression. Such within-class comparisons are important to learn a fine-grained expression representation.
- Due to within-class variations and between-class similarities, two expressions from the same category need not be visually more similar when compared to an expression from a different category. For example, while the middle and right expressions in Figure 2(b) belong to the *Surprise* category, the middle expression is visually more similar to the left expression which belongs to the *Anger* category.



(a) One-class triplet (b) Two-class triplet (c) Three-class triplet

Figure 2: Different types of triplets based on the emotion labels used in the AffectNet [42] dataset.

- It is difficult to predict the visual similarity relationships between expressions from three different emotion categories by using labels. For example, while the three expressions in Figure 2(c) belong to three different categories, the left and middle expressions are visually more similar when compared to the right expression. Such comparisons are useful for learning an embedding that can model long-range visual similarity relationships between different categories.
- It is unclear how the difference in the strengths of action units between two expressions could be converted into a distance function that mimics visual preferences.

3.1. Dataset

Each sample in the FEC dataset consists of a face image triplet (I_1, I_2, I_3) along with a label $L \in \{1, 2, 3\}$ that indicates which two images in the triplet form the most similar pair in terms of facial expression. For example, $L = 1$ means I_2 and I_3 are visually more similar when compared to I_1. Note that, different from the commonly-used triplet annotation [48, 54], these triplets do not have a notion of anchor, and each triplet provides two annotations: I_2 is closer to I_3 than I_1, and I_3 is closer to I_2 than I_1. Also, in this dataset, an image A can be (relatively) closer to another image B in one triplet and (relatively) farther from the same image B in another triplet. This is different from the triplets generated using category labels [48], in which any two images will either form a similar pair or a dissimilar pair in all the triplets they appear in.

The triplets in the FEC dataset were generated by sampling images from a partially-labeled [1] internal face dataset in which each face image has one or more of the following emotion labels [8, 9]: *Amusement, Anger, Awe, Boredom, Concentration, Confusion, Contemplation, Contempt, Contentment, Desire, Disappointment, Disgust, Distress, Doubt, Ecstasy, Elation, Embarrassment, Fear, Interest, Love, Neutral, Pain, Pride, Realization, Relief, Sadness, Shame, Surprise, Sympathy, and Triumph*. To reduce the effect of category-bias, we sampled the images such that all these categories are (roughly) equally represented in the triplet dataset. Each triplet was annotated by six human raters, and the raters were instructed to focus only on expressions ignoring other factors such as identity, gender, ethnicity, pose and age. A total of 40 raters participated in the process, each annotating a subset of the entire dataset.

Based on the existing emotion labels, each triplet in this dataset can be categorized into one of the following types [2]:

- *One-class triplets*: All the three images share a category label, see Figure 2(a). These triplets are useful for learning a fine-grained expression representation.
- *Two-class triplets*: Only two images share a category label and the third image belongs to a different category, see Figure 2(b). As explained in Section 3, images sharing a category label need not form the (visually) most similar pair in these triplets.
- *Three-class triplets*: None of the images share a common category label, see Figure 2(c). These triplets are useful for learning long-range visual similarity relationships between different categories.

One-class triplets are relatively the most difficult ones since the expressions could be very close to each other, and two-class triplets are relatively the easiest ones since the images sharing a label could potentially be different from the remaining image (though not always). While there are other possible types of triplets based on other label combinations (for example, I_1, I_2 sharing a label, and I_2, I_3 sharing another label), we prioritized the above three types while collecting the dataset as the other types could be confusing for the raters. Table 1 shows the number of triplets in this dataset along with the number of faces used to generate the triplets. The dataset is further divided into training (90%) and test (10%) sets, and we encourage the users of this dataset to use the training set for training their algorithms and the test set to validate them.

Annotation agreement: Each triplet in this dataset was annotated by six raters. For a triplet, we say that the raters *agree strongly* if at least two-thirds of them voted for the maximum-voted label, and *agree weakly* if there is a unique maximum-voted label and half of the raters voted for it. The number of such triplets for each type are shown in Table 1.

[1] The images in this dataset are not exhaustively labeled, i.e., an image may not have all the labels that are applicable to it.

[2] The images in the dataset (from which we sampled the faces) were not exhaustively labeled, and hence, a triplet classified as a two/three-class triplet based on the existing labels may not be be a two/three-class triplet if the images had been exhaustively labeled.

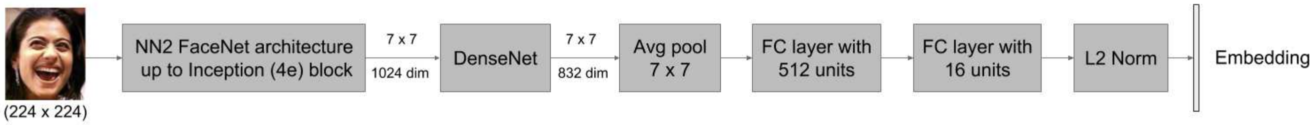


Figure 3: Proposed embedding network based on the NN2 FaceNet architecture [48]. Here, FC stands for fully-connected, the L2-Norm layer performs ℓ_2 normalization, and DenseNet consists of a 1×1 convolution layer (512 filters) followed by a Dense block [22] with 5 layers and growth rate of 64.

Partition	Rater agreement	Triplet type				Faces
		One-class	Two-class	Three-class	All	
Training set	Strong	115,544	124,665	117,540	357,749	130,516
	Strong + Weak	137,266	138,034	132,435	407,735	
	All	152,674	150,234	146,235	449,143	
Test set	Strong	13,046	14,607	13,941	41,594	25,427
	Strong + Weak	15,411	15,908	15,404	46,723	
	All	17,059	17,107	16,894	51,060	
Full dataset	Strong	128,590	139,272	131,481	399,343	155,943
	Strong + Weak	152,677	153,942	147,839	454,458	
	All	169,733	167,341	163,129	500,203	

Table 1: Number of triplets and faces in the proposed FEC dataset.

Raters agree strongly for about 80% of the triplets suggesting that humans have a well-defined notion of visual expression similarity.

4. Facial expression embedding network

In the recent past, the performance of face recognition systems has improved significantly [1, 2, 25, 44] in part due to the availability of large-scale (several million data samples) training datasets such as MS-Celeb-1M [19], MegaFace [44], SFC [51] and Google-Face [48]. Neural networks trained on these large-scale datasets see images with significant variations along different dimensions such as lighting, pose, age, gender, ethnicity, etc. during training.

Compared to these large-scale face datasets, our facial expression comparison dataset is significantly smaller (just 130K training faces). Hence, in order to leverage the power of a large training set, we build our facial expression embedding network using the pre-trained FaceNet proposed in [48], see Figure 3. We use the NN2 version of pre-trained FaceNet [48] up to the inception (4e) block [3] whose output is a 7×7 feature map with 1024 channels. This feature map is processed by a DenseNet which consists of a 1×1 convolution layer (512 filters) followed by a Dense block [22] with 5 layers [4] and growth rate of 64. The output of DenseNet is passed to a 7×7 average pooling layer followed by a fully connected (FC) layer with 512 hidden units and an embedding layer (a linear FC layer + ℓ_2 normalization layer). Batch normalization [23] and ReLu6 [31] activation function are used in the DenseNet and the first FC layer. We also use dropout for regularization.

The input to our network is an aligned (rotated to undo roll transformation and scaled to maintain an inter-ocular distance of 55 pixels) 224×224 face image I, and the output is a d-dimensional embedding e_I of unit ℓ_2 norm.

4.1. Triplet loss function

For training the embedding network using the proposed FEC dataset, we use a triplet loss function that encourages the distance between the two images that form the most similar pair to be smaller than the distances of these two images from the third image. For a triplet (I_1, I_2, I_3) with the most similar pair (I_1, I_2), the loss function is given by

$$\begin{aligned} l(I_1, I_2, I_3) = max(0, \|e_{I_1} - e_{I_2}\|_2^2 - \|e_{I_1} - e_{I_3}\|_2^2 + \delta) \\ + max(0, \|e_{I_1} - e_{I_2}\|_2^2 - \|e_{I_2} - e_{I_3}\|_2^2 + \delta), \end{aligned} \tag{1}$$

where δ is a small margin.

5. Experiments

In this section, we demonstrate the usefulness of the expression embedding learned from the proposed FEC dataset for various applications such as expression retrieval, photo album summarization, and emotion classification. In all our experiments, we only use the triplets with strong rater agreement for both training and evaluation. We also tried using the triplets with weak rater agreement for training, but the results did not improve (see Section 5.4). In the rest of the paper, we refer to the proposed expression embedding network trained on the proposed FEC dataset as *FECNet*.

[3] We also experimented with features from inception 4d, 5a and 5b blocks, and features from 4e block performed the best.

[4] Adding more layers did not improve the results.

5.1. Comparative approaches

Most of the existing large-scale expression datasets focus on the task of classification. One can train a classification network with such a dataset, and use the output of the final or penultimate layer as an expression embedding. Here, we train two networks: *AFFNet-CL* for emotion recognition using the AffectNet dataset [42], and *FACSNet-CL* for facial action unit detection using the DISFA dataset [38]. AffectNet is a large-scale faces-in-the-wild dataset manually labeled with eight emotion categories. This dataset has around 288K training and 4K validation images. DISFA is a widely-used spontaneous facial actions dataset manually labeled with the presence/absence of 12 action units [5]. This dataset has around 260K images, out of which 212K images are used for training and 48K images are used for validation. We create four expression embeddings using these two classification networks:

- *AFFNet-CL-P* and *AFFNet-CL-F*: Penultimate and final layer outputs of AFFNet-CL.
- *FACSNet-CL-P* and *FACSNet-CL-F*: Penultimate and final layer outputs of FACSNet-CL.

Another way to learn an embedding using a classification dataset is to train an embedding network with a category label-based triplet loss similar to [48]. So, we also train an embedding network (referred to as *AFFNet-TL*) on AffectNet dataset using triplet loss.

For a fair comparison, the input and architecture of all the networks are chosen to be same as FECNet (Figure 3) except that the embedding layer is replaced by a softmax classifier for AFFNet-CL and separate binary classifiers for FACSNet-CL.

5.2. Training and validation

We define *triplet prediction accuracy* as the percentage of triplets for which the distance (in the embedding space) between the visually most similar pair is less than the distances of these two images from the third. Without loss of generality, let (A_i, B_i) be the most similar pair in the test triplet (A_i, B_i, C_i). Then, the triplet prediction accuracy is given by

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left[\begin{array}{l}\|e_{A_i}-e_{B_i}\|_2 < \|e_{A_i}-e_{C_i}\|_2 \text{ and} \\ \|e_{A_i}-e_{B_i}\|_2 < \|e_{B_i}-e_{C_i}\|_2\end{array}\right],$$

where N is the number of test samples.

As for the validation measure during training, we use triplet prediction accuracy on the FEC test set for FECNet, (following [55]) average area under ROC curve (AUC-ROC) on the AffectNet validation set for AFFNet-CL and AFFNet-TL [6], and (following [7, 49]) average F1-score on the DISFA validation set for FACSNet-CL.

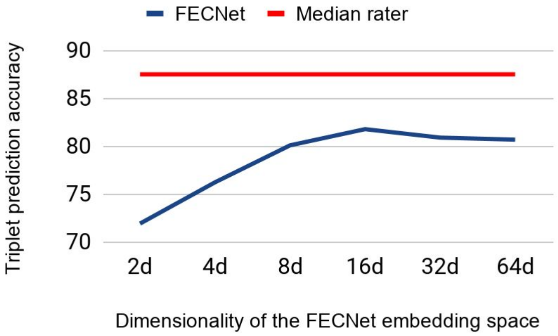


Figure 4: Triplet prediction accuracy on the FEC test set.

For all the networks, the parameters of the FaceNet layers were kept fixed and the newly-added DenseNet and FC layers were trained starting from Xavier initialization [16] using Adam optimizer [26] with a learning rate of 5×10^{-4} and dropout of 0.5. FECNet was trained on the FEC training set with mini-batches of 90 samples (30 triplets from each of the triplet types) for 50K iterations, AFFNet-CL and AFFNet-TL were trained on the AffectNet training set with mini-batches of 128 samples (16 samples from each of the eight emotion categories) for 10K iterations, and FACSNet-CL was trained on the DISFA training set with mini-batches of 130 samples (at least 10 positive samples for each action unit and 10 samples with no action units) for 20K iterations. For training FECNet, the value of margin δ was set to 0.1 for one-class triplets, and 0.2 for two-class and three-class triplets. For training AFFNet-TL, the loss margin was set to 0.2 and the embedding dimensionality was set to 16. All the hyper-parameter values were chosen based on the corresponding validation measures.

5.3. Dimensionality of the FECNet embedding

While we want to represent facial expressions in a continuous fashion using an embedding, it is unclear how many dimensions should be used for the embedding space. To answer this question, we trained FECNet for different values of the output dimensionality. Figure 4 shows how the triplet prediction accuracy on the FEC test set varies with the dimensionality of the embedding space. The accuracy increases till 16 dimensions and drops slightly after that. Based on these results, we choose 16-dimensions to represent the expression embedding space (referred to as FECNet-16d).

Figure 4 also shows the median rater accuracy. Accuracy for a human rater is computed based on how often they agree with the maximum-voted label. Please see the supplementary material for accuracy values of individual raters. Using 16 dimensions, the proposed FECNet is able

[5] The frames with action unit intensities greater than 2 are treated as positives and the remaining are treated as negatives.

[6] Nearest neighbor classifier with 800 neighbors is used.

Embedding	Distance		
	ℓ_1	ℓ_2	Cosine
FACSNet-CL-F	47.1	47.1	40.7
FACSNet-CL-P	45.3	44.2	48.3
AFFNet-CL-F	49.0	47.7	49.0
AFFNet-CL-P	52.4	51.6	53.3
AFFNet-TL	-	49.6	-
FECNet-16d	-	81.8	-

Table 2: Triplet prediction accuracy on the FEC test set.

Triplet type	AFFNet-CL-P	FECNet-16d	Median rater
One-class	49.2	77.1	85.3
Two-class	59.8	85.1	89.3
Three-class	50.4	82.6	87.2
All triplets	53.3	81.8	87.5

Table 3: Triplet prediction accuracy for different types of triplets in the FEC test set.

to achieve an accuracy of 81.8%, which is fairly close to the median rater accuracy (87.5%). Also, note that the triplet prediction accuracy of random choice is 33.3%.

5.4. Comparison of different embeddings

Table 2 shows the triplet prediction accuracy of various embeddings on the FEC test set using different distance functions. Among all the AFFNet and FACSNet embeddings, the combination of AFFNet-CL-P and cosine distance gives the best accuracy, and hence, we use this combination for comparison with FECNet-16d in the rest of the experiments. It is worth noting that the proposed FECNet-16d (81.8%) performs significantly better than the best competing approach (AFFNet-CL-P + Cosine distance; 53.3%).

We also trained FECNet-16d by adding the triplets with weak rater agreement to the training set, but the test accuracy dropped from 81.8% to 80.5%.

5.5. Performance for different triplet types

Table 3 shows the triplet prediction accuracy of median rater, FECNet-16d and AFFNet-CL-P for each triplet type in the FEC test set. As expected, the performance is best (85.1%) for two-class triplets, which are relatively the easiest ones, and is lowest (77.1%) for one-class triplets, which are relatively the most difficult ones.

5.6. Visualization of the FECNet embedding space

Figure 5 shows a 2D t-SNE [3] visualization of the learned FECNet-16d embedding space using the AffectNet validation set. The amount of overlap between two categories in this figure roughly tells us about the extent of visual similarity between them. For example, fear and surprise have a high overlap indicating that they could be confused easily, and both of them have a very low overlap with contempt indicating that they are visually very distinct from contempt. Also, the spread of a category in this figure tells us about the visual diversity within that category. For example, happiness category maps to three distinct regions indicating that there are three visually distinct modes within this category. Please refer to the supplementary material for a visualization of the face images that fall into different regions in Figure 5.

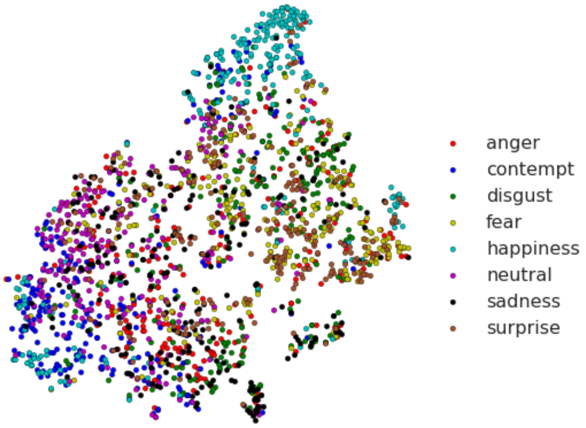


Figure 5: 2D visualization of the FECNet-16d embedding space using t-SNE [3].

5.7. Applications

Image retrieval:
We can perform expression-based image retrieval by using nearest neighbor search in the expression embedding space. To compare the retrieval performance of FECNet-16d and AFFNet-CL-P embeddings, we use a query set consisting of 25 face images and a database (CelebA dataset [35]) consisting of 200K face images. For each query, we retrieved N nearest neighbors (N varied from 1 to 10) using both the embeddings and ranked the $2N$ retrieved images based on how close they are to the query as judged by ten human raters. Since ranking all $2N$ images at once is difficult for human raters, we asked them to rank two images at a time. In each pairwise ranking, the winner and looser get a score of +1 and -1, respectively. If it is a tie, i.e., the two images get equal number of rater votes, then both of them get a score of zero. We obtained such pairwise ranking scores for all pairs and converted them into a global ranking based on the overall scores.

For numerical evaluation, we use *rank-difference* metric, defined as the average difference in the ranks of images retrieved by AFFNet-CL-P and FECNet-16d embeddings, respectively, divided by the number of retrieved images N. Positive value of this rank-difference metric indicates that FECNet-16d embedding is better than AFFNet-CL-P embedding. The lowest value for this metric is -1, corresponding to the case when all the AFFNet-CL-P retrieval

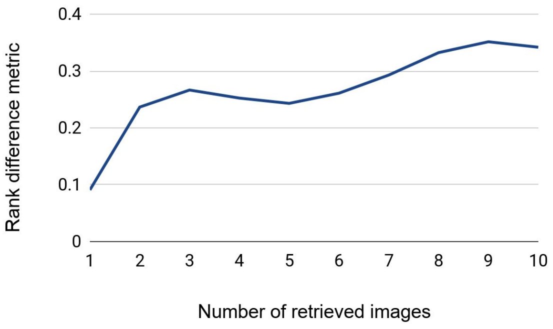


Figure 6: Rank-difference metric as a function of the number of retrieved images N.

results are ranked lower than all the FECNet-16d retrieval results, and the highest value is $+1$, corresponding to the case when all the FECNet-16d retrieval results are ranked lower than all the AFFNet-CL-P retrieval results. Figure 6 shows the rank-difference metric for different values of N. Positive value of the metric for all values of N clearly indicates that the proposed FECNet-16d embedding produces better matches compared to the AFFNet-CL-P embedding.

Figure 8 shows the top-5 retrieved images for some of the queries. The overall results of the proposed FECNet-16d embedding are clearly better than the results of AFFNet-CL-P embedding. Specifically, the FECNet-16d embedding pays attention to finer details such as teeth-not-visible (first query), eyes-closed (second and third queries) and looking straight (fourth query). Results for the full query set are provided in the supplementary material.

Photo album summarization:
In this task, we are interested in summarizing the diverse expression content present in a given photo album using a fixed number of images. Expression embedding can be used for this task by combining it with a clustering algorithm.

For evaluation, we created ten photo albums (100-200 images in each album) by downloading images of ten celebrities using Google image search. For each album, we ran hierarchical agglomerative clustering [7] (10 clusters) with FECNet-16d and AFFNet-CL-P embeddings, and used the images that are closest to cluster centers for generating the summaries. We showed these two summaries to ten human raters and asked them which one is better. Humans preferred the summaries generated by the proposed FECNet-16d embedding for eight out of ten albums. Figure 9 shows the summaries for two photo albums. We can see that the expression content is more diverse in the summaries produced by the FECNet-16d embedding. Please refer to the supplementary material for summaries of the other eight albums and also the number of votes received by both the embeddings for all the ten albums.

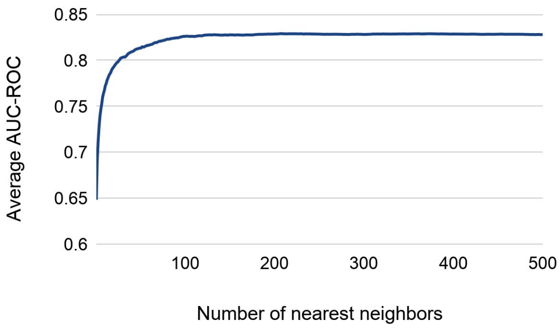


Figure 7: Classification performance of the FECNet-16d embedding on the AffectNet validation set when combined with K-NN classifier.

Emotion classification:
The proposed FECNet-16d embedding can be used for emotion classification by combining it with K-Nearest Neighbor (K-NN) classifier. Figure 7 shows the average AUC-ROC of the FECNet-16d embedding on the AffectNet validation set as a function of the number of neighbors used. The performance increases up to 200 neighbors and then remains stable. Table 4 compares the classification performance of the FECNet-16d embedding (using 200 neighbors) with other approaches. Note that AFFNet-CL and AFFNet-TL have the same architecture as FECNet-16d and are specifically trained for classification using AffectNet training data. Hence, as expected, they perform a bit better than FECNet-16d. However, despite not being trained for classification, the FECNet-16d embedding outperforms AlexNet and VGG-Face based classifiers, demonstrating that it is well-suited for classification.

6. Conclusions and Future Work

In this work, we presented the first large-scale facial expression comparison dataset annotated by human raters, and learned a compact facial expression embedding using this dataset. The embedding learned using this dataset performs better than various other embeddings learned using existing emotion and action units datasets. We experimentally demonstrated the usefulness of the proposed embedding for various applications such as expression retrieval, photo album summarization, and emotion classification.

Since FECNet is trained using human visual preferences, negative samples that are close to the positive samples in the FECNet embedding space can be considered as hard negatives while training a classification model. We plan to explore this further in our future work.

Acknowledgements: We thank Gautam Prasad, Ting Liu, Brendan Jou, Alan Cowen, Florian Schroff and Hartwig Adam from Google for their support and suggestions during the data collection process.

[7] Cosine distance and maximum linkage were used.

Approach	Neutral	Happiness	Sadness	Surprise	Fear	Disgust	Anger	Contempt	Average
AFFNet-CL	84.6	**96.5**	**90.7**	**88.5**	**90.2**	**85.2**	**88.3**	**85.0**	**88.6**
AFFNet-TL	**85.9**	96.0	89.2	**88.5**	89.6	83.7	87.9	82.6	87.9
FECNet-16d + K-NN	83.3	94.9	78.0	83.0	84.5	79.3	78.7	81.2	82.9
AlexNet [42]	-	-	-	-	-	-	-	-	82.0
VGG-Face descriptor [55]	75.9	92.2	80.5	81.4	82.3	81.4	81.2	77.1	81.5
FAb-Net [55]	72.3	90.4	70.9	78.6	77.8	72.5	76.4	72.2	76.4

Table 4: Expression classification results (AUC-ROC) on the AffectNet [42] validation set.

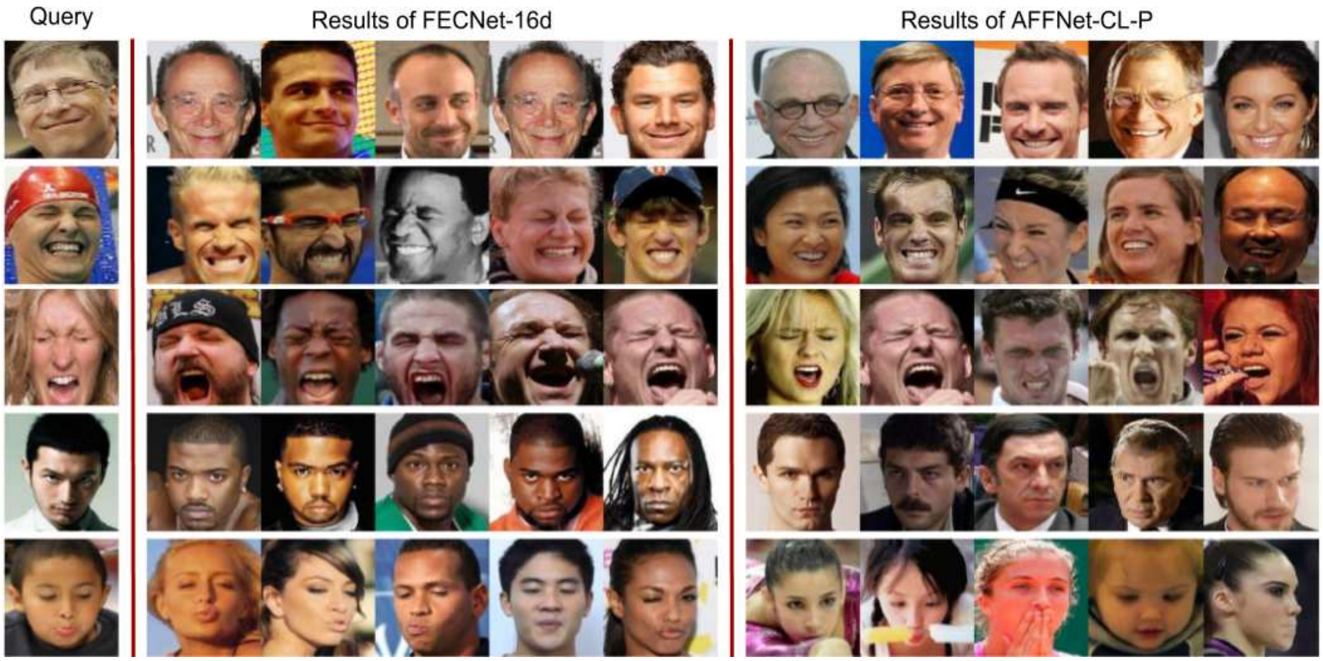


Figure 8: Top-5 images retrieved using FECNet-16d (left) and AFFNet-CL-P (right) embeddings. The overall results of FECNet-16d match the query set better when compared to AFFNet-CL-P.

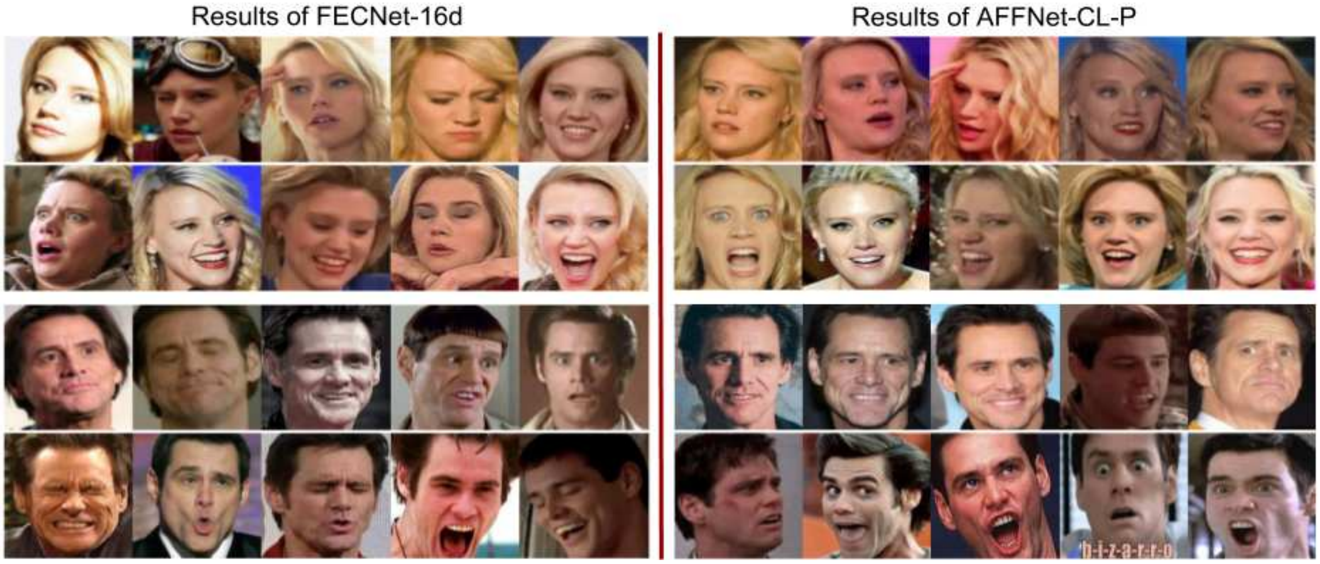


Figure 9: Expression summaries generated for albums (200 images) of Kate McKinnon (top) and Jim Carrey (bottom) using FECNet-16d (left) and AFFNet-CL-P (right) embeddings. The expression content is more diverse in the summaries produced by FECNet-16d embedding (left).

References

[1] MegaFace challenge leaderboard: http://megaface.cs.washington.edu/results/facescrub.html. 4

[2] MS-Celeb-1M challenge leaderboard: http://www.msceleb.org/leaderboard/iccvworkshop-c1. 4

[3] Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 6

[4] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martínez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016. 2

[5] Y. Chang, C. Hu, R. S. Feris, and M. Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006. 2

[6] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010. 2

[7] C. A. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *ECCV*, 2018. 5

[8] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900E7909, 2017. 3

[9] A. S. Cowen and D. Keltner. Clarifying the conceptualization, dimensionality, and structure of emotion. *Trends in Cognitive Sciences*, 22(4):274–276, 2018. 3

[10] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *ICMI*, 2017. 2

[11] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *ICMI*, 2013. 2

[12] A. Dhall, O. V. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *ICMI*, 2015. 2

[13] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System - Manual*. A Human Face, 2002. 1, 2

[14] J. Fiss, A. Agarwala, and B. Curless. Candid portrait selection from video. *ACM Transactions on Graphics*, 30(6):128:1–128:8, 2011. 1

[15] W. Ge, W. Huang, D. Dong, and M. R. Scott. Deep metric learning with hierarchical triplet loss. In *ECCV*, 2018. 2

[16] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 5

[17] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. C. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. T. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015. 2

[18] R. Gross, I. A. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. 2

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *ECCV*, 2016. 4

[20] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai. Triplet-center loss for multi-view 3D object retrieval. In *CVPR*, 2018. 2

[21] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 2

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 4

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[24] W. B. Johnson and J. Lindenstrauss. Extension of lipschitz mapping into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. 2

[25] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 4

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[27] S. Koelstra, C. Mühl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. 2

[28] A. S. Koepke, O. Wiles, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. In *BMVC*, 2018. 2

[29] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. W. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *CoRR*, abs/1804.10938, 2018. 2

[30] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 2

[31] A. Krizhevsky. Convolutional deep belief networks on CIFAR-10. *Unpublished manuscript*, 2010. 4

[32] S. Li and W. Deng. Deep facial expression recognition: A survey. *CoRR*, abs/1804.08348, 2018. 1, 2

[33] S. Li and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 2

[34] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, 2016. 2

[35] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 6

[36] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, 2010. 2

[37] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017. 1, 2

[38] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2, 5

[39] D. McDuff, R. E. Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. W. Picard. Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild. In *CVPR Workshops*, 2013. 2

[40] A. Mehrabian. *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*. Oelgeschlager, Gunn & Hain, 1980. 2

[41] H. Meng, T. Lin, X. Jiang, Y. Lu, and J. Wen. LSTM-based facial performance capture using embedding between expressions. *CoRR*, abs/1805.03874, 2018. 2

[42] A. Mollahosseini, B. Hassani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *CoRR*, abs/1708.03985, 2017. 2, 3, 5, 8

[43] A. Mollahosseini, B. Hassani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor. Facial expression recognition from world wild web. In *CVPR Workshops*, 2016. 2

[44] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *CVPR*, 2017. 4

[45] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *ICME*, 2005. 2

[46] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *FG*, 2013. 2

[47] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. 2

[48] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 2, 3, 4, 5

[49] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *ECCV*, 2018. 5

[50] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 2

[51] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 4

[52] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 2

[53] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep metric learning with angular loss. In *ICCV*, 2017. 2

[54] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009. 3

[55] O. Wiles, A. S. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. In *BMVC*, 2018. 5, 8

[56] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 2

[57] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 2

[58] B. Zhuang, G. Lin, C. Shen, and I. D. Reid. Fast training of triplet-based deep binary embedding networks. In *CVPR*, 2016. 2