Modeling Stylized Character Expressions via Deep Learning

Deepali Aneja[1(✉)], Alex Colburn[2], Gary Faigin[3], Linda Shapiro[1], and Barbara Mones[1]

[1] Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA
{deepalia,shapiro,mones}@cs.washington.edu
[2] Zillow Group, Seattle, WA, USA
alexco@cs.washington.edu
[3] Gage Academy of Art, Seattle, WA, USA
gary@gageacademy.org

Abstract. We propose **DeepExpr**, a novel expression transfer approach from humans to multiple stylized characters. We first train two Convolutional Neural Networks to recognize the expression of humans and stylized characters independently. Then we utilize a transfer learning technique to learn the mapping from humans to characters to create a shared embedding feature space. This embedding also allows human expression-based image retrieval and character expression-based image retrieval. We use our perceptual model to retrieve character expressions corresponding to humans. We evaluate our method on a set of retrieval tasks on our collected stylized character dataset of expressions. We also show that the ranking order predicted by the proposed features is highly correlated with the ranking order provided by a facial expression expert and Mechanical Turk experiments.

1 Introduction

Facial expressions are an important component of almost all human interaction and face-to-face communication. As such, the importance of clear facial expressions in animated movies and illustrations cannot be overstated. Disney and Pixar animators [1,2] have long understood that unambiguous expression of emotions helps convince an audience that an animated character has underlying cognitive processes. The viewer's emotional investment in a character depends on the clear recognition of the character's emotional state [3]. To achieve lifelike emotional complexity, an animator must be able to depict characters with clear, unambiguous expressions, while retaining the fine level control over intensity and expression mix required for nuance and subtlety [4]. However,

The final publication is available at Springer via 10. 1007/978-3-319-54184-6 9.

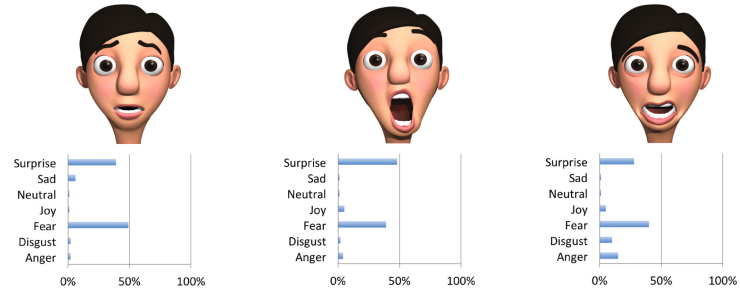

Fig. 1. Expressions are surprisingly difficult to create for professional animators. Three professional animators were asked to make the character appear as surprised as possible. None of the expressions achieved above 50% recognition on Mechanical Turk with 50 test subjects.

explicit expressions are notoriously difficult to create [5], as illustrated in Fig. 1. This difficulty is in part due to animators and automatic systems relying on geometric markers and features modeled for human faces, not stylized character faces.

We focus our efforts on *stylized* 3D characters, defined as characters that no human would mistake for another person, but would still be perceived as having human emotions and thought processes. Our goal is to develop a model of facial expressions that enables accurate retrieval of stylized character expressions given a human expression query.

To achieve this goal, we created DeepExpr, a perceptual model of stylized characters that accurately recognizes human expressions and transfers them to a stylized character without relying on explicit geometric markers. Figure 2 shows an overview of the steps to develop the framework of our model. We created a database of labeled facial expressions for six stylized characters as shown in Fig. 7. This database with expressions is created by facial expression artists and initially labeled via Mechanical Turk (MT) [6]. Images are labeled for each of six cardinal expressions: joy, sadness, anger, surprise, fear, disgust, and neutral. First, we trained a Convolutional Neural Network (CNN) on a large database of human expressions to input a human expression and output the probabilities of each of the seven classes. Second, we trained a similar character model on an artist-created character expression image database. Third, we learned a mapping between the human and character feature space using the transfer learning approach [7]. Finally, we can retrieve character expressions corresponding to a human using perceptual model mapping and human geometry.

We make the following contributions[1]:

1. A data-driven perceptual model of facial expressions.
2. A novel stylized character data set with cardinal expression annotations.
3. A mechanism to accurately retrieve plausible character expressions from human expression queries.

[1] Project page: http://grail.cs.washington.edu/projects/deepexpr/.

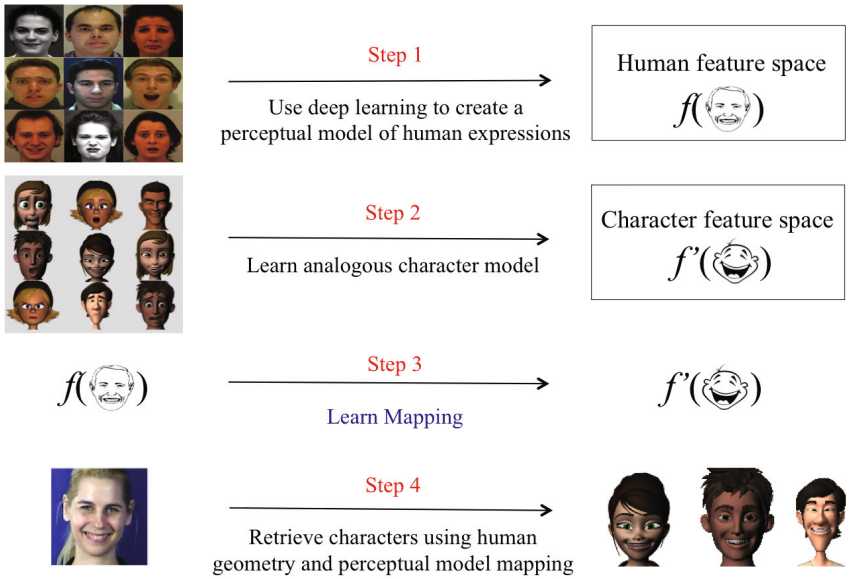

Fig. 2. Overview of our pipeline. Feature extraction using CNNs and transfer learning builds a model of expression mapping.

2 Related Work

There is a large body of literature classifying, recognizing, and characterizing human facial expressions. Notably, Paul Ekman's widely adopted Facial Action Coding System (FACS) [8] is used as a common basis for describing and communicating human facial expressions. The FACS system is often used as a basis for designing character animation systems [5,9] and for facial expression recognition on scanned 3D faces [10]. However, despite these advances, creating clear facial animations for 3D characters remains a difficult task.

2.1 Facial Expression Recognition and Perception

FACS for Animation. Though a reliable parameterization of emotion and expression remains elusive, the six cardinal expressions pervade stories and face-to-face interactions, making them a suitable focus for educators and facial expression researchers [11]. To guide and automate the process of expression animation, animators and researchers turn to FACS. For example, FACSGen [9] allows researchers to control action units on realistic 3D synthetic faces. Though Roesch et al. confirmed the tool's perceptual validity settings by asking viewers to rate the presence of emotions in faces developed using action unit combinations found in "real life situations", we found that their faces were unclear as demonstrated in Fig. 3.

HapFACS [5], an alternative to FACSGen, allows users to control facial movement at the level of both action units and whole expressions (EmFACS)

according to Ekman's formulas. The strict use of anatomy-based and constrained motion by these systems limits their generalizability to characters with different anatomy and limits their application, because the most believable animation may require the violation of physical laws [1].

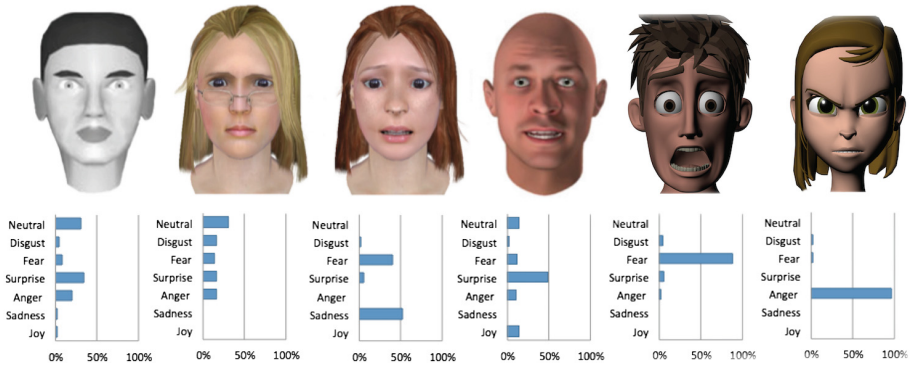

Fig. 3. DeepExpr yields clearer expressions than other approaches when tested on MT. From left to right each generated face was intended to clearly convey an expression: anger from MPEG-4 [12] scored 20% clarity for anger. Anger from HapFACS [5] scored 8% clarity for anger. Fear from HapFACS scored 20% clarity for fear. Fear using FACSGen [9] scored 6% clarity for fear. Anger and fear faces retrieved with our approach, both scored over 85% clarity.

Alternatively, the MPEG-4 standard [12] can describe motion in stylized faces by normalizing feature motion to a standard distance. The MPEG-4 standard provides users with archetypal expression profiles for the six cardinal expressions, but like the FACS-based systems does not give the user feedback on the perceptual validity of their expression, which may lead to unclear faces. As demonstrated in Fig. 3, anatomically valid faces generated by these systems did not consistently yield high recognition rates in MT with 50 test subjects.

Other Perceptual Models. The results shown in Fig. 3 support artists' intuition that anatomy based formulas for expressions must be tailored to each unique face, and necessitate a perceptually guided system to find the optimal configuration for a clear expression. Perceptual models such as Deng and Ma [13] have also been explored for realistic faces with promising results. Deng and Ma polled students' perceptions of the expression of different motion-captured facial configurations and ran Principal Component Analysis (PCA) [14] on the vertices of the meshes of these faces. Using these results, they developed a Support Vector Machine (SVM) [15] model for expression clarity as a function of PCA weights for different areas of the face. They also showed significantly increased expression clarity of generated speech animation by constraining the characters' motion to fit their model. However, the scalability of their procedure is limited

by its reliance on on-site subjects and the size and specificity of the seeding dataset. We addressed these limitations by incorporating MT tests in our character expression data collection and training a deep learning model for expression clarity.

2.2 Feature Extraction and Classification

Facial expression recognition can be broadly categorized into face detection, registration, feature extraction, and classification. In the detection step, landmark points are used to detect a face in an image. In the registration step, the detected faces are geometrically aligned to match a template image. Then the registered image is used to extract numerical feature vectors as the part of the feature extraction step.

These features can be *geometry based* such as facial landmarks [16,17], *appearance based* such as Local Binary Patterns (LBP) [18], Gabor filters [19], Haar features [20], Histogram of Oriented Gradients [21], or *motion based* such as optical flow [22] and Volume LBP [23]. Recently, methods have been developed to learn the features by using sparse representations [24,25]. A 3D shape model approach has also been implemented to improve the facial expression recognition rate [26]. A variety of fusion of features has also been utilized to boost up the facial expression recognition performance [27,28]. They are mostly a combination of geometric and appearance based features. In the current practice of facial expression analysis, CNNs have shown the capability to learn the features that statistically allow the network to make the correct classification of the input data in various ways [29,30]. CNN features fused with geometric features for customized expression recognition [31] and Deep Belief Networks have also been utilized to solve the Facial Expression recognition (FER) problem. A recent approach [32] termed "AU (Action Unit)-Aware" Deep Networks demonstrated the effectiveness in classifying the six basic expressions. Joint Fine-Tuning in Deep Neural Networks [33] have also been used to combine temporal appearance features from image sequences and temporal geometry features from temporal facial landmark points to enhance the performance of the facial expression recognition. Along similar lines, we have utilized deep learning techniques as a tool to extract useful features from raw data for both human faces and stylized characters. We then deploy a transfer learning approach, where the weights of the stylized character are initialized with those from a network pre-trained on a human face data set, and then fine-tuned with the target stylized character dataset.

In the last step of classification, the algorithm attempts to classify the given face image into seven different classes of basic emotions using machine learning techniques. SVMs are most commonly used for FER tasks [18,34,35]. As SVMs treat the outputs as scores for each class which are uncalibrated and difficult to interpret, the softmax classifier gives a slightly more intuitive output with normalized class probabilities and also has a probabilistic interpretation. Based on that, we have used a softmax classifier to recognize the expressions in our classification task using the features extracted by the deep CNNs.

3 Methodology

We first describe the data collection approach and design of facial features that can capture the seven expressions: joy, sadness, anger, surprise, fear, disgust, and neutral. Then, we discuss our customized expression recognition and transfer learning framework using deep learning.

3.1 Data Collection and Pre-processing

To learn deep CNN models that generalize well across a wide range of expressions, we need sufficient training data to avoid over-fitting of the model. For human facial expression data collection, we combined publicly available annotated facial expression databases: extended CK+ [36], DISFA [37], KDEF [38] and MMI [39]. We also created a novel database of facial expressions for six stylized characters: the Facial Expression Research Group-Database (**FERG-DB**). Both the databases have labels for the six cardinal expressions and neutral.

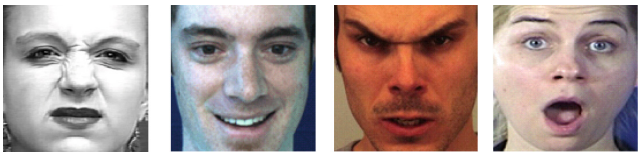

Fig. 4. Examples of registered faces from CK+, DISFA, KDEF, and MMI databases showing disgust, joy, anger, and surprise emotion from left to right.

CK+: The Extended Cohn-Kanade database (CK+) includes 593 video sequences recorded from 123 subjects. Subjects portrayed the six cardinal expressions. We selected only the final frame of each sequence with the peak expression for our method, which resulted in 309 images.

DISFA: Denver Intensity of Spontaneous Facial Actions (DISFA) database consists of 27 subjects, each recorded while watching a four minutes video clip by two cameras. As DISFA is not emotion-specified coded, we used the EMFACS system [5] to convert AU FACS codes to expressions, which resulted in around 50,000 images using the left camera only.

KDEF: The Karolinska Directed Emotional Faces (KDEF) is a set of 4900 images of human facial expressions of emotion. This database consists of 70 individuals, each displaying 7 different emotional expressions. We used only the front facing angle for our method and selected 980 images.

MMI: The MMI database includes expression labeled videos for more than 20 subjects of both genders for which subjects were instructed to display 79 series of facial expressions. We extracted static frames from each corresponding sequence for the six cardinal emotions, resulting in 10,000 images.

We balanced out the final number of samples for each class for training our network to avoid any bias towards a particular expression.

Stylized Character Database. We created a novel database (**FERG-DB**) of labeled facial expressions for six stylized characters. The animator created the key poses for each expression, and they were labeled via MT to populate the database initially. The number of key poses created depends on the complexity of the expression for each character. We only used the expression key poses having 70% MT test agreement among 50 test subjects for the same pose. On average, 150 key poses (15–20 per expression) were created for each character. Interpolating between the key poses resulted in 50,000 images (around 8,000 images per character). The motivation behind the combination of different characters is to have a generalized feature space among various stylized characters.

Data Pre-processing. For our combined human dataset, Intraface [40] was used to extract 49 facial landmarks. We use these points to register faces to an average frontal face via an affine transformation. Then a bounding box around the face is considered to be the face region. Geometric measurements between the points are also taken to produce geometric features for refinement of expression retrieval results as described in Sect. 3.2. Once the faces are cropped and registered, the images are re-sized to 256×256 pixels for analysis. Figure 4 shows examples of registered faces from different databases using this method.

The corresponding 49 landmark points are marked on the neutral expression of the 3D stylized character rig. This supplementary information is saved along with each expression rendering and used later to perform geometric refinement of the result. This step is performed only once per character.

3.2 Network Training Using Deep Learning

With approximately 70,000 images of labeled samples of human faces and 50,000 images for stylized character faces, the datasets are smaller in comparison to other image classification datasets that have been trained from scratch in the past. Moreover, since we have to use a portion of this data set for validation, effectively only 80% of the data was available for training. We performed data augmentation techniques to increase the number of training examples. This step helps in reducing overfitting and improving the model's ability to generalize. During the training phase, we extracted 5 crops of 227×227 from the four corners and the center of the image and also used the horizontal mirror images for data augmentation.

Training Human and Character CNN Models. Our human expression network consists of three elements: multiple convolutional layers followed by max-pooling layers and fully connected layers as in [41]. Our character network is analogous to the human CNN architecture and does not require CONV4 for the recognition task as the character images are not very complex. Unlike the human dataset, there are fewer variations in the character dataset (light, pose, accessories, etc.). To avoid overfitting, we limited our model to a fewer number of convolutional parameters (until CONV3). Both networks are trained independently. The details of the network layers are shown in Fig. 5 and network parameters are given in the supplementary material.

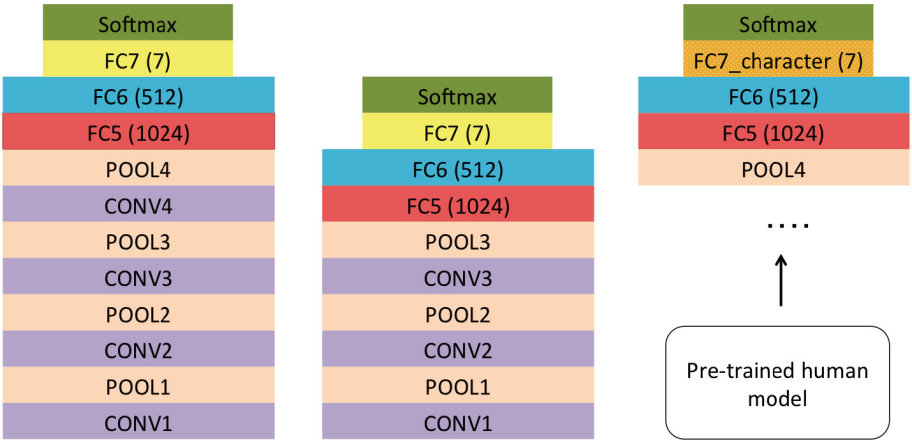

Fig. 5. Outline of the CNN architecture. The convolutional layers, max pooling layers and fully connected layers are denoted as CONV, POOL and FC followed by the layer number. Human expression image trained model (left), Stylized character expression image trained model (middle) and fine-tuned character trained model (right) are shown. In the transfer learning step, the last fully-connected layer (FC7_character) is fine-tuned using stylized character data.

All three color channels are processed directly by the network. Images are first rescaled to 256×256 and a crop of 227×227 is fed to the network. Finally, the output of the last fully connected layer is fed to a softmax layer that assigns a probability for each class. The prediction itself is made by taking the class with the maximal probability for the given test image.

In the forward propagation step, the CONV layer computes the output of neurons that are connected to local regions in the input (resized to 256×256 in the data pre-processing step), each computing a dot product between their weights and a small region they are connected to in the input volume, while the POOL layer performs a downsampling operation along the spatial dimensions. The output of each layer is a linear combination of the inputs mapped by an activation function given as:

$$h^{i+1} = f((W^{i+1})^T h^i) \tag{1}$$

where h^{i+1} is the i^{th} layer output, W^i is the vector of weights that connect to each output node and $f(\cdot)$ is the non-linear activation function which is implemented by the RELU layer given as: $f(x) = max(0, x)$ where x is the input to the neuron. The back-propagation algorithm to used to calculate the gradient with respect to the parameters of the model. The weights of each layer are updated as:

$$\delta^i = (W^i)^T \delta^{i+1} . f'(h^i) \tag{2}$$

where δ^i is the increment of weights at layer i. We train our networks using stochastic gradient descent with hyperparameters (momentum = 0.9, weight

decay = 0.0005, initial learning rate = 0.01). The learning rate is dropped by a factor of 10 following every 10 epochs of training. The proposed network architectures were implemented using the Caffe toolbox [42] on a Tesla k40c GPU.

Transfer Learning. To create a shared embedding feature space, we fine-tuned the CNN pre-trained on the human dataset with the character dataset for every character by continuing the backpropagation step. The last fully connected layer of the human trained model was fine-tuned, and earlier layers were kept fixed to avoid overfitting. We decreased the overall learning rate while increasing the learning rate on the newly initialized FC7_character layer which is highlighted fine-tuned character trained model in Fig. 5. We set an initial learning rate of 0.001, so that the pre-trained weights are not drastically altered. The learning rate is dropped by a factor of 10 following every 10 epochs of training. Our fine-tuned model used 38 K stylized character image samples for training, 6K for validation, and 6 K for test. The proposed architecture was trained for 50 epochs with 40 K iterations on batches of size 50 samples.

Distance Metrics. In order to retrieve the stylized character closest expression match to the human expression, we used the Jensen—Shannon divergence distance [43] for expression clarity and geometric feature distance for expression refinement. It is described by minimizing the distance optimization function in Eq. 3 given as:

$$\phi_d = \alpha\,|\text{JS Distance}| + \beta\,|\text{Geometric Distance}| \tag{3}$$

where JS Distance is given as the Jensen—Shannon divergence distance between FC6 feature vectors of *human* and *character*, and Geometric distance is given as the L^2 norm distance between geometric features of *human* and *character*. Our implementation uses JS Distance as a retrieval parameter and then geometric distance as a sorting parameter to refine the retrieved results with α and β as relative weight parameters. Details of the computation are given as follows:

Expression Distance. For a given human expression query image, FC6 (512 outputs) features are extracted from the query image using the human expression trained model and for the test character images from the shared embedding feature space using the fine-tuned character expression model. The FC7 (7 outputs) layer followed by a softmax can be interpreted as the probability that a particular expression class is predicted for a given input feature vector. By normalizing each element of the feature vector by the softmax weight, the FC6 feature vectors are treated as discrete probability distributions. To measure the similarity between human and character feature probability distributions, we used the Jensen—Shannon divergence [43] which is symmetric and is computed as:

$$JSD(H||C) = \frac{1}{2}D(H||M) + \frac{1}{2}D(C||M) \tag{4}$$

where $M = \frac{1}{2}(H+C)$, $D(H||M)$ and $D(C||M)$ represents the Kullback—Leibler divergence [44] which is given as:

$$D(X||M) = \sum_i X(i) log \frac{X(i)}{M(i)} \tag{5}$$

where X and M are discrete probability distributions.

We used this distance metric to order the retrievals from the closest distance to the farthest in the expression feature space. Our results show that the retrieval ordering matched the query image label, and retrievals were ordered in order of similarity to the query label. To choose the best match out of the multiple retrievals with the same label as shown in Fig. 6, we added a geometric refinement step as described in the next section.

Geometric Distance. The JS Divergence distance results in the correct expression match, but not always the closest geometric match to the expression. Figure 6 shows the retrieval of the correct label (joy). To match the geometry, we extract geometric distance vectors and use them to refine the result.

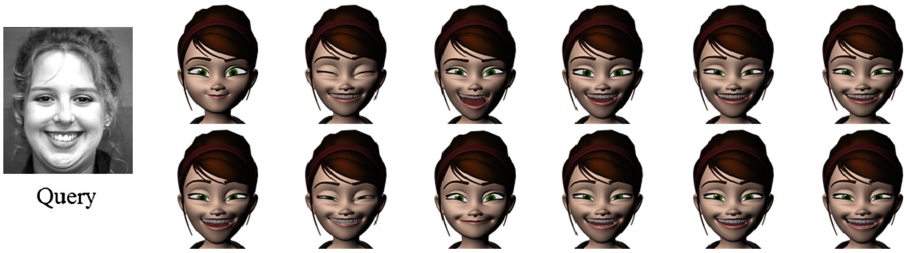

Fig. 6. Multiple retrieval results for the joy query image

We use the facial landmarks as described in Sect. 3.1, to extract the geometric features including the following measurements: the left/right eyebrow height (vertical distance between top of the eyebrow and center of the eye), left/right eyelid height (vertical distance between top of an eye and bottom of the eye), nose width (horizontal distance between leftmost and rightmost nose landmarks), mouth width (left mouth corner to right mouth corner distance), closed mouth measure (vertical distance between the upper and the lower lip), and left/right lip height (vertical distance between the lip corner from respective the lower eyelid). The geometric distance is a normalized space. Each of the distances between landmarks is normalized by the bounding box of the face. After normalization, we compute the L^2 norm distance between the human geometry vector and character geometry vectors with the correct expression label. Finally, we re-order the retrieved images within the matched label based on matched geometry.

4 Experimental Results

The combined DeepExpr features and geometric features produce significant performance enhancement in retrieving the stylized character facial expressions

based on human facial expressions. The top results for all seven expressions on six stylized characters are shown in Fig. 7. Human expression-based image retrievals and character expression-based image retrievals are shown in the supplementary material.

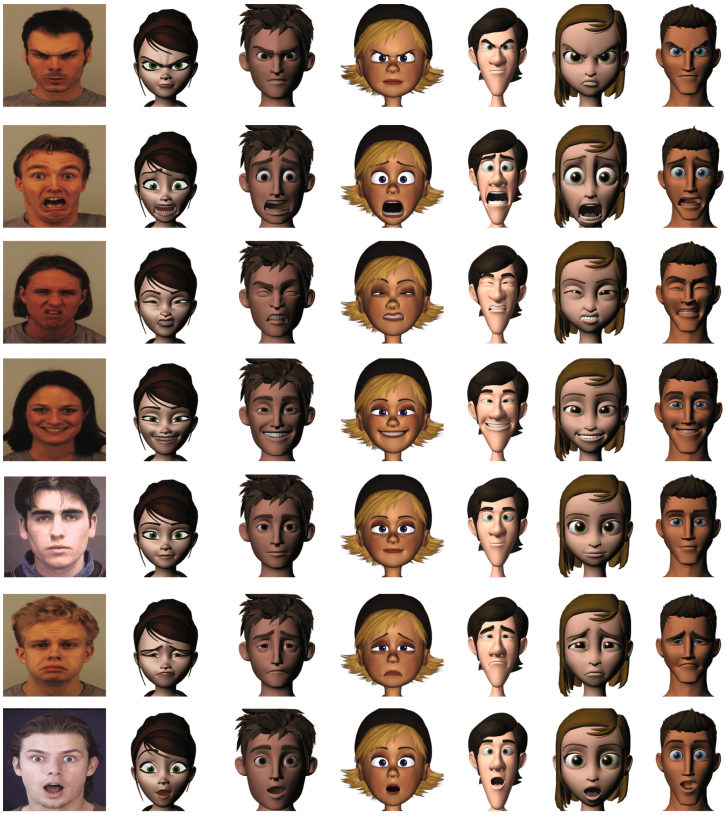

Fig. 7. Results from our combined approach - DeepExpr and geometric features. The leftmost image in each row is the query image and all six characters are shown portraying the top match of the same expression - anger, fear, joy, disgust, neutral, sad and surprise (top to bottom).

5 Evaluation

5.1 Expression Recognition Accuracy

For human facial expression recognition accuracy, we performed the subject independent evaluation, where the classifier is trained on the training set and evaluated on images in the same database (validation and test set) using K-fold cross-validation with $K = 5$. On average, we used 56 K samples for training in

batches of 50 samples, 10 K samples for validation and 10 K for testing. The overall accuracy of human facial expression recognition was 85.27%. Similarly, for stylized character expression, we used 38 K character images for training in batches of 50 samples, 6 K for validation, and 6 K for testing, and achieved the recognition accuracy of 89.02%. Our aim with human expression accuracy was to achieve a good score on the expression recognition which is close to the state-of-the-art results in order to extract relevant features corresponding to a facial expression. The details of human expression recognition accuracy for each expression are given in the supplementary material.

5.2 Expression Retrieval Accuracy

We analyze our retrieval results by computing the retrieval score to measure how close is the retrieved character expression label is to the human query expression label. We also compare our results with a facial expression expert by choosing 5 random samples from the retrieved results with the same label and rank order them based on their similarity to the query image. The details of analysis are discussed as follows:

Retrieval Score. We measured the retrieval performance of our method by calculating the average normalized rank of relevant results (same expression label) [45]. The evaluation score for a query human expression image was calculated as:

$$score(q) = \frac{1}{1 - N \cdot N_{rel}} \left(\sum_{k=1}^{N_{rel}} R_k - \frac{N_{rel}\left(N_{rel} + 1\right)}{2} \right) \tag{6}$$

where N is the number of images in the database, N_{rel} the number of database images that are relevant to the query expression label q (all images in the character database that have the same expression label as the human query expression label), and R_k is the rank assigned to the k^{th} relevant image. The evaluation score ranges from 0 to 1, where 0 is the best score as it indicates that all the relevant database images are retrieved before all other images in the database. A score that is greater than 0 denotes that some irrelevant images (false positives) are retrieved before all relevant images.

The retrieval performance was measured over all the images in the human test dataset using each test image in turn as a query image. The average retrieval score for each expression class was calculated by averaging the retrieval score for all test images in the same class. Table 1 shows the final class retrieval score, which was calculated by averaging the retrieval scores across all characters for each expression class using only geometry and DeepExpr expression features. The best match results in Fig. 8 confirm that the geometric measure is not sufficient to match the human query expression with clarity.

Comparison. In order to judge the effectiveness of our system, we compared DeepExpr to a human expert and MT test subjects. We asked the expert and

Table 1. Average retrieval score for each expression across all characters using only geometry and DeepExpr features.

Expression	Geometry	DeepExpr
Anger	0.384	0.213
Disgust	0.386	0.171
Fear	0.419	0.228
Joy	0.276	0.106
Neutral	0.429	0.314
Sad	0.271	0.149
Surprise	0.322	0.125

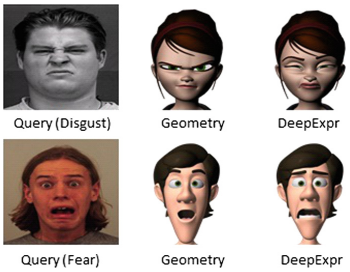

Fig. 8. Best match results from our Deep-Expr approach compared to only geometric feature based retrieval for Disgust (top) and Fear (bottom).

the MT subjects to rank five stylized character expressions in order of decreasing expression similarity to a human query image. The facial expression expert, 50 MT test subjects and DeepExpr ranked the same 30 validation test sets. We aggregated the MT results into a single ranking using a voting scheme. We then compared the DeepExpr ranking to the results, measuring similarity with two measures. Both measures found a high correlation between DeepExpr ranking compared with the expert and the MT ranking results. The details of the ranking comparison tests are given in the supplementary material.

The **Spearman rank correlation coefficient** ρ measures the strength and direction of the association between two ranked variables [46]. The closer the ρ coefficient is to 1, the better the two ranks are correlated.

The average ρ coefficient for the expert rank orderings is 0.773 ± 0.336 and for MT tests is 0.793 ± 0.3561. The most relevant correlation coefficient is between the first rank chosen by the expert and the first rank chosen by DeepExpr as they represent the best match with the query image. The Spearman correlation with expert best rank is 0.934 and with MT best rank is 0.942, which confirms the agreement on selection of the closest match to the human expression.

The **Kendall τ test** is a non-parametric hypothesis test for statistical dependence based on the τ coefficient [47]. It is a pairwise error that represents how many pairs are ranked discordant. The best matching ranks receive a τ value of 1. The average τ coefficient for expert validation rank orderings is 0.706 ± 0.355, and the best rank correlation is 0.910. For the MT ranking, the average Kendall correlation coefficient is 0.716 ± 0.343 and 0.927 is the best rank correlation.

The Spearman and Kendall correlation coefficients of DeepExpr ranking with the expert ranking and MT test ranking for 30 validation experiments are shown in Fig. 9. Note that more than half the rankings are perfectly correlated, and most of them are above 0.8. Only two of the rankings had (small) negative correlations in both correlation experiments: the order was confusing because of very subtle difference in expressions (see supplementary material for details).

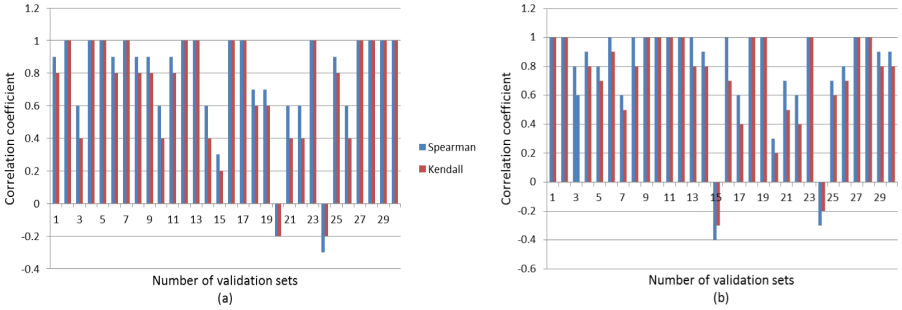

Fig. 9. Correlation rank order result charts with (a) expert and (b) MT tests

6 Comparison with Character Animator

Currently to our knowledge, no other system performs stylized character retrieval based on a learned feature set. The closest match to DeepExpr tool is Adobe Character Animator (Ch) [48] which creates 2.5-D animations for characters. We conducted an expression recognition experiment by creating a similar character in Ch with different expressions as layers. We queried three human expression images for each of the seven expressions. Then, we asked 50 MT test subjects to recognize the expression for best matches from DeepExpr retrieved images and Ch results. The results of the experiment are shown in Table 2. On an average, joy, neutral and surprise had comparable recognition performance. DeepExpr showed great improvement in recognition of fear and disgust. In Ch, fear was confused with surprise due to the dependence on geometric landmarks of the face showing an open mouth and disgust was most confused with anger. For anger and sad, the closed mouth was most confused with neutral in Ch. An example of a fear expression MT test is shown in Fig. 10. DeepExpr achieved higher (83%) expression recognition accuracy as compared to the Ch animator tool (41%).

Table 2. Average expression recognition accuracy (%) for each expression across all characters using Ch animator and DeepExpr.

Expression	Ch animator	DeepExpr
Anger	60	85
Disgust	47	86
Fear	42	81
Neutral	87	88
Joy	95	97
Sad	43	89
Surprise	93	95

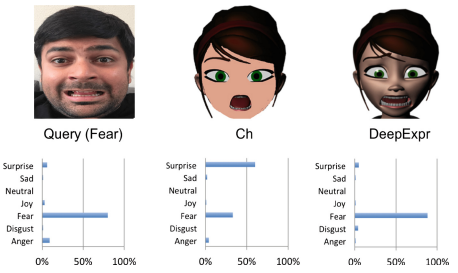

Fig. 10. Expression matching results for fear query image. Ch [48] result scored 41% clarity for fear and DeepExpr result scored 83% clarity for fear.

7 Conclusions and Future Work

We have demonstrated a perceptual model of facial expression clarity and geometry using a deep learning approach combined with artistic input and crowdsourced perceptual data. Our results are highly correlated with a facial expression expert, in addition to MT subjects and have a higher expression recognition accuracy as compared to Character Animator.

DeepExpr has several practical applications in the field of storytelling, puppeteering and animation content development. For example, the system could assist animators during the initial blocking stage for 3D characters in any production pipeline. When there are multiple animators working on the same character during a production, using our expression recognition system will help enable a consistent approach to the personality for that character. More importantly, our approach provides a foundation for future facial expression studies. For example, our perceptual model could be used to evaluate existing FACS.

Our system demonstrates a perceptual model of facial expressions that provides insight into facial expressions displayed by stylized characters. It can be used to automatically create desired character expressions driven by human facial expressions. The model can also be incorporated into the animation pipeline to help animators and artists to better understand expressions, communicate how to create expressions to others, transfer expressions from humans to characters, and to provide a mechanism for animators/storytellers to more quickly and accurately create the expressions they intend.

Acknowledgements. We would like to thank Jamie Austad for creating our stylized character database. We would also like to thank the creators of the rigs we used in our project: Mery (www.meryproject.com), Ray (*CGTarian Online School*), Malcolm (www.animSchool.com), Aia & Jules (www.animationmentor.com), and Bonnie (*Josh Sobel Rigs*).

References

1. Lasseter, J.: Principles of traditional animation applied to 3D computer animation. SIGGRAPH Comput. Graph. **21**, 35–44 (1987)
2. Porter, T., Susman, G.: On site: creating lifelike characters in pixar movies. Commun. ACM **43**, 25 (2000)
3. Bates, J.: The role of emotion in believable agents. Commun. ACM **37**, 122–125 (1994)
4. Pelachaud, C., Poggi, I.: Subtleties of facial expressions in embodied agents. J. Vis. Comput. Anim. **13**, 301–312 (2002)
5. Amini, R., Lisetti, C.: HapFACS: an open source API/Software to generate FACS-based expressions for ECAs animation and for corpus generation. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 270–275 (2013)
6. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. **6**, 3–5 (2011)

7. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724 (2014)
8. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
9. Roesch, E.B., Tamarit, L., Reveret, L., Grandjean, D., Sander, D., Scherer, K.R.: FACSGen: a tool to synthesize emotional facial expressions through systematic manipulation of facial action units. J. Nonverbal Behav. **35**, 1–16 (2011)
10. Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L.: Static and dynamic 3D facial expression recognition: a comprehensive survey. Image Vis. Comput. **30**, 683–697 (2012)
11. Adolphs, R.: Recognizing emotion from facial expressions: psychological and neurological mechanisms. Behav. Cogn. Neurosci. Rev. **1**(1), 21–62 (2002)
12. Pereira, F.C., Ebrahimi, T.: The MPEG-4 Book. Prentice Hall PTR, Upper Saddle River (2002)
13. Deng, Z., Ma, X.: Perceptually guided expressive facial animation. In: Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2008 Eurographics Association (2008)
14. Jolliffe, I.: Principal Component Analysis. Wiley, Hoboken (2002)
15. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
16. Kobayashi, H., Hara, F.: Facial interaction between animated 3D face robot and human beings. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, vol. 4, pp. 3732–3737. IEEE (1997)
17. Dibeklioglu, H., Salah, A., Gevers, T.: Like father, like son: facial expression dynamics for kinship verification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1497–1504 (2013)
18. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**, 803–816 (2009)
19. Liu, C., Wechsler, H.: Independent component analysis of Gabor features for face recognition. IEEE Trans. Neural Netw. **14**, 919–928 (2003)
20. Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., Movellan, J.: Toward practical smile detection. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 2106–2111 (2009)
21. Shu, C., Ding, X., Fang, C.: Histogram of the oriented gradient for face recognition. Tsinghua Sci. Technol. **16**, 216–224 (2011)
22. Kenji, M.: Recognition of facial expression from optical flow. IEICE Trans. Inf. Syst. **74**, 3474–3483 (1991)
23. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 915–928 (2007)
24. Mahoor, M.H., Zhou, M., Veon, K.L., Mavadati, S.M., Cohn, J.F.: Facial action unit recognition with sparse representation. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), pp. 336–342. IEEE (2011)
25. Lin, Y., Song, M., Quynh, D.T.P., He, Y., Chen, C.: Sparse coding for flexible, robust 3D facial-expression synthesis. IEEE Comput. Graph. Appl. **32**, 76–88 (2012)

26. Jeni, L.A., Lőrincz, A., Szabó, Z., Cohn, J.F., Kanade, T.: Spatio-temporal event classification using time-series kernel based structured sparsity. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 135–150. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_10
27. Tan, X., Triggs, B.: Fusing Gabor and LBP feature sets for kernel-based face recognition. In: Zhou, S.K., Zhao, W., Tang, X., Gong, S. (eds.) AMFG 2007. LNCS, vol. 4778, pp. 235–249. Springer, Heidelberg (2007). doi:10.1007/978-3-540-75690-3_18
28. Ying, Z.-L., Wang, Z.-W., Huang, M.-W.: Facial expression recognition based on fusion of sparse representation. In: Huang, D.-S., Zhang, X., Reyes García, C.A., Zhang, L. (eds.) ICIC 2010. LNCS (LNAI), vol. 6216, pp. 457–464. Springer, Heidelberg (2010). doi:10.1007/978-3-642-14932-0_57
29. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016)
30. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442. ACM (2015)
31. Yu, X., Yang, J., Luo, L., Li, W., Brandt, J., Metaxas, D.: Customized expression recognition for performance-driven cutout character animation. In: Winter Conference on Computer Vision (2016)
32. Liu, M., Li, S., Shan, S., Chen, X.: Au-aware deep networks for facial expression recognition. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–6. IEEE (2013)
33. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)
34. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2562–2569. IEEE (2012)
35. Dumas, M.: Emotional expression recognition using support vector machines. In: Proceedings of International Conference on Multimodal Interfaces. Citeseer (2001)
36. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101. IEEE (2010)
37. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: a spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**, 151–160 (2013)
38. Lundqvist, D., Flykt, A., Öhman, A.: The Karolinska directed emotional faces-KDEF. CD-ROM from department of clinical neuroscience, psychology section, Karolinska Institutet, Stockholm, Sweden. Technical report (1998). ISBN 91-630-7164-9
39. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, p. 5. IEEE (2005)
40. Xiong, X., Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539 (2013)

41. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
42. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
43. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theor. **37**, 145–151 (1991)
44. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
45. Müller, H., Marchand-Maillet, S., Pun, T.: The truth about corel - evaluation in image retrieval. In: Lew, M.S., Sebe, N., Eakins, J.P. (eds.) CIVR 2002. LNCS, vol. 2383, pp. 38–49. Springer, Heidelberg (2002). doi:10.1007/3-540-45479-9_5
46. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. **15**, 72–101 (1904)
47. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**, 81–93 (1938)
48. Character Animator: Adobe After Effects CC 2016. Adobe Systems Incorporated, San Jose, CA 95110–2704 (2016)