Predicting Perceived Emotions in Animated GIFs with 3D Convolutional Neural Networks

Weixuan Chen
The Media Laboratory
Massachusetts Institute of Technology
Cambridge, USA
Email: cvx@media.mit.edu

Rosalind W. Picard
The Media Laboratory
Massachusetts Institute of Technology
Cambridge, USA
Email: picard@media.mit.edu

***Abstract*—Animated GIFs are widely used on the Internet to express emotions, but their automatic analysis is largely unexplored before. To help with the search and recommendation of GIFs, we aim to predict their emotions perceived by humans based on their contents. Since previous solutions to this problem only utilize image-based features and lose all the motion information, we propose to use 3D convolutional neural networks (CNNs) to extract spatiotemporal features from GIFs. We evaluate our methodology on a crowd-sourcing platform called GIFGIF with more than 6000 animated GIFs, and achieve a better accuracy then any previous approach in predicting crowd-sourced intensity scores of 17 emotions. It is also found that our trained model can be used to distinguish and cluster emotions in terms of valence and risk perception.**

***Index Terms*—emotion detection; animated GIFs; perceived emotion; 3D convolutional neural network.**

I. Introduction

The Graphics Interchange Format (GIF) is a bitmap image format widespread on the Internet due to its wide compatibility and portability. Different from other popular image formats, GIF supports animations, which makes it a special media form between videos and still images. As a powerful tool for visually expressing emotions online, animated GIFs play an important role in popular culture. People usually make animated GIFs from scenes of movies and TV shows, and use them on social media, digital forums, message boards and even in emails as an enhanced version of emoticons. Despite the format's popularity, its information processing and retrieval have been rarely explored in multimedia and computer vision research. Though similar to videos as spatiotemporal volumes, animated GIFs have a number of unique characteristics such as briefness, looping, silence as well as emotional expressiveness, which bring about particular challenges in their analysis. We believe it is worthwhile to develop artificial intelligence systems specifically for understanding animated GIFs, as this would help the Internet users search, use and recommend them more efficiently, and assist researchers to better understand how humans perceive them.

This paper will focus on predicting perceived emotions in animated GIFs. When a media sample is presented to human subjects, their perceived emotion is the emotion that they think the sample expresses instead of the emotion they feel, which is otherwise called their induced emotion. According to Jou et al. [1], perceived emotions are more concrete and objective than induced emotions, where labels are less reliable due to their subjectivity. Specific to animated GIFs, it is also their perceived emotions rather than induced emotions that usually determines how you use the GIFs, because normally you post a GIF to express your current emotion instead of to induce a certain emotion from the readers. There are many previous works about emotion detection from videos. However, most of them are based on human subjects' emotional responses to videos, i.e. induced emotions [2]–[7]. Pang et al. [8] have realized prediction of perceived emotions from user-generated content including videos, but their emphasis is on multi-modal learning from not only visual but also auditory and textual modalities. There are also several works not mentioning which kind of emotion labels they use [9], [10].

To our knowledge, the only previous study on predicting perceived emotions in animated GIFs is from Jou et al. [1]. On a dataset of over 3800 animated GIFs, they calculated four different feature representations: color histograms, facial expressions recognized by a CNN [11], image-based aesthetics [12], and a mid-level visual representation composed of visual sentiment detectors called SentiBank [13]. After testing three different regression methods, they report a highest prediction accuracy on 17 categories of emotions using the facial expression features. However, a large proportion of GIFs are made from cartoons or anime, in which facial expression recognition can barely work. Hence Jou et al. have to assign average labels to GIFs without a detected face. Moreover, all the features they use are image-based, where all the temporal information related to motion is neglected .

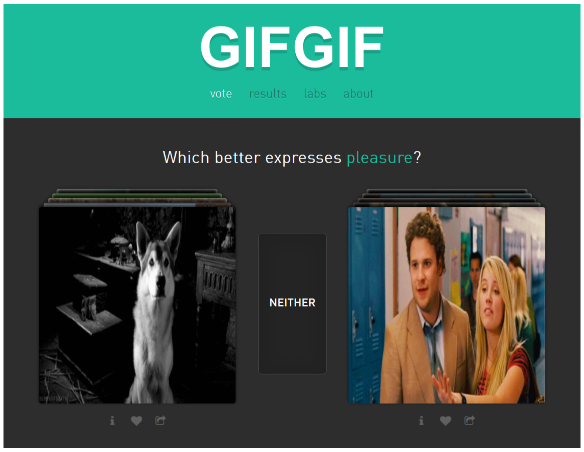
To address these problems, we adopt a 3D CNN for GIF analysis so that spatiotemporal instead of only spatial features can be extracted. It has been shown by Tran et al. [14] that for video analysis volume-based features are superior to image-based ones due to their capability of modeling motions. They develop a video feature representation based on 3D CNN and Sport1M dataset called C3D. It yields good performance on various video analysis tasks (action recognition, scene classification, and object recognition) without requiring to finetune the model for each task. Thus we believe it is also promising to adapt it to the prediction of perceived emotions.

Our contributions in this paper include: (1) We analyze the file structure of GIFs and discuss the main differences from videos; (2) We use 3D CNN features and a Lasso regression to predict perceived emotion intensities of GIFs on a large-scale dataset more accurately then any previous method; (3) We discover that the model we trained reveals useful information for understanding human emotions.

II. Methods

A. Dataset

We collected our data from the GIFGIF website [15], a crowd-sourcing platform enabling users to vote on animated GIFs with their perceived emotions. The GIFs on the platform are imported from a large-scale GIF database called Giphy [16], and cover a wide variety of sources including movies, TV shows, advertisements, sports, cartoons, anime, video games, user-generated content, and user-edited content. As a result, the GIFs span a broad range of resolutions, camera angles, zooming, illumination, grayscale/color, humans/non-humans, numbers of objects, and special effects.



Which better expresses pleasure?

Fig. 1. GIFGIF homepage.

When users enter the homepage of GIFGIF, a pair of random GIFs will be presented with a question ”which better expresses X”, as shown in Fig. 1, where X is one of 17 emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, happiness, pleasure, pride, relief, sadness, satisfaction, shame, and surprise. The users can answer the question by pressing on the GIF that matches the emotion or select ”neither”. The developers of GIFGIF chose the 17 emotion categories based on Paul Ekman’s selection of universal emotions in the 1990s [17]. With all the answers from millions of users, the website is capable of ranking each GIF by its emotion intensities for all the 17 categories. The website API annotates every animated GIF with a 17 x 2 matrix, containing scores between 0 and 50 for each emotion where 25 is neutral and every score’s uncertainty. According to GIFGIF, these scores and uncertainties are generated from users’ votes using the TrueSkill rating algorithm [18]. The [0,50] range is heuristically given by a prior $\mu_0 = 25$ and $\sigma_0 = 25/3$. As soft labels are applied, a GIF can have low emotion intensities in all the 17 categories if it is relatively neutral. Therefore, predicting these labels is better defined as a regression problem rather than a classification problem.

Until May 22, 2016, the GIFGIF platform had indexed 6119 animated GIFs with 3,130,780 crowd-sourced annotations. Skipping 6 GIFs with broken links, we downloaded 6113 files with their corresponding labels as our dataset.

B. Preprocessing

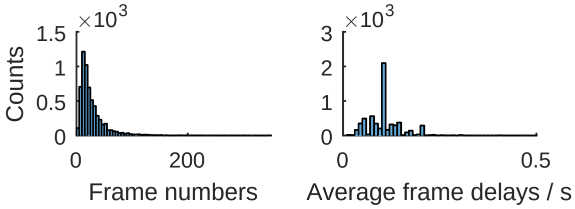


Fig. 2. Histograms of frame numbers and average frame delays in our dataset.

Feature extraction from videos for machine learning usually requires a preprocessing to normalize the widths, heights and lengths of videos to be the same. However, different from most videos, animated GIFs not only have varied lengths but also have varied frame rates. Table I shows the structure of an exemplary GIF (#989 on GIFGIF) with 6 frames. As illustrated in the table, every frame has an assigned delay time, and different frames within a file can even have different delays. We read the information from all the GIFs in our dataset, and drew the histograms of their frame numbers and average frame delays (Fig. 2).

As shown in Fig. 2, the longest GIF has 347 frames, while the shortest has only 2 frames. For too short GIFs, we looped all their frames to imitate the way they were usually presented on the Internet and perceived by human eyes. For too long

TABLE I
Structure of an Exemplary Gif File

Frames	Frame Size	Format Version	Left / Pixels	Top / Pixels	Width / Pixels	Height / Pixels	Bit Depth	Background Color	Aspect Ratio	Delay Time / s	Transparent Color
1	31400	89a	1	1	500	492	7	128	0	0.5	128
2	31400	89a	165	178	306	33	7	128	0	0.5	128
3	31400	89a	141	176	325	48	7	128	0	0.2	128
4	31400	89a	136	214	303	69	7	128	0	0.2	128
5	31400	89a	134	274	325	50	7	128	0	0.5	128
6	31400	89a	156	121	331	89	7	128	0	2.0	128

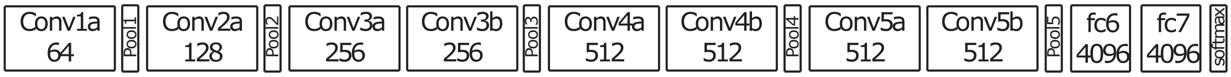


Fig. 3. C3D architecture [14].

GIFs, Jou et al. [1] subsampled ten equally spaced frames within each GIF, because their image-based method does not consider the relationship between frames. On the contrary, in our method using 3D representations, sampling a single frame from tens of frames would destroy the dynamics of any continuous action. Therefore, we chose to always maintain the continuity of consecutive frames by splitting long GIFs into multiple equal-length clips and extracting features from each of them.

Since GIFs have extremely diverse frame rates (Fig. 2), the length of a clip can be defined in two different ways: frame-based or time-based. The frame-based way would directly put the same number of frames into each clip regardless of different frame rates, while the time-based way would interpolate all GIFs to the same frame rate and then do the splitting so that every clip has the same duration. For GIFs, the optimal solution is the latter, because the speed information of actions (e.g. how fast an athlete raise his/her hands) will be maintained, which is possibly useful for distinguishing between emotions. However, as the highest frame rate is about 40 times the lowest in our dataset, it is difficult to realize 40-time motion interpolation without causing visual artifact. Furthermore, some web browsers ignore very short frame delays of GIFs and replace them with a default delay, as a result of which the delay parameters inside a GIF may not necessarily reflect its real playback speed. Therefore, we finally chose the frame-based way to split the raw frames directly.

C. Feature Representations

We used the C3D video descriptor as our feature representation. Fig. 3 shows the architecture of the C3D neural network. It has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. The 3D convolution layers have 3 x 3 x 3 kernels with stride 1 in both spatial and temporal dimensions, and numbers of filters denoted in each box in the figure. The kernels of all the 3D pooling layers are 2 x 2 x 2, except for Pool1 is 1 x 2 x 2. Every fully connected layer has 4096 output units.

Using the same preprocessing parameters as C3D, every GIF was split into 16-frame-long clips with a 8-frame overlap between two consecutive ones. GIFs shorter than 16 frames or not integer multiples of 8 frames were padded via looping first. The clips were then resized to have a frame size of 128 pixels x 171 pixels, and center cropped into 16 frames x 112 pixels x 112 pixels. After all the normalizations, they were passed to the C3D network. As introduced before, the neural network was pre-trained on the Sport1M dataset, and had shown good generalization capability across various datasets. Its fc6 activations formed a 4096-dim vector for each clip, which was finally saved as our feature representation.

D. Lasso Regression

For every GIF clip, our feature representation exhibits 4096 features, which are comparable to the size of our dataset. Consequentially, an ordinary linear regression without regularization would likely give poor results because of over-fitting. To address the problem, we trained parsimonious models using a Lasso regression [19] so that variable selection can be done automatically.

Consider a sample consisting of N observations, each of which consists of p features and a single outcome. Let y_i be the outcome and $x_i = (x_1, x_2, \cdots, x_p)^T$ be the feature vector for the i-th observation. A Lasso regression solves the problem:

$$\min_{\beta_0, \beta} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \qquad (1)$$

where β and β_0 are fitted coefficients and intercept, and λ is a nonnegative regularization parameter optimized via cross-validation.

III. Results

TABLE II
The Normalized Mean Squared Errors (nMSE) for Emotion Prediction on GIFGIF. Lower nMSE indicates better performance.

Methods	nMSE
3858 GIFs	
Color histograms + Trace-norm regularized multi-task regression	1.4641 ± 0.1935
Face expression + Ordinary least squares linear regression	0.8925 ± 0.0036
Image-based aesthetics + Trace-norm regularized multi-task regression	1.0361 ± 0.0093
SentiBank + Logistic regression	1.4944 ± 0.0593
C3D + Lasso regression	0.6652 ± 0.0545
6113 GIFs	
C3D + Lasso regression	0.7161 ± 0.0519

To compare with previous methods, we used the same approach as Jou et al. [1] to train and test our model. The emotion intensity scores from the TrueSkill rating algorithm were normalized to [-1,1], and applied to each GIF clip as weakly supervised labels. Since Jou et al. only tested 3858 GIFs on GIFGIF up to April 29, 2014, we assessed our method on two different set sizes: the first 3858 GIFs and the whole 6113 GIFs. For either size, a 5-fold cross-validation was employed with regression results reported by averaging

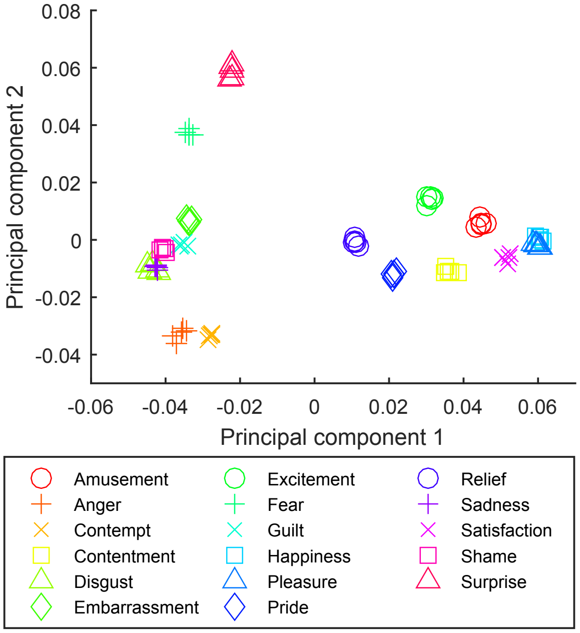


Fig. 4. First and second principal components of our regression coefficients and intercepts.

clip-level scores over each GIF within the test sets. A metric called normalized mean squared error (nMSE) commonly used before [20], [21] was applied to our predicted scores and the ground truth to evaluate the prediction accuracy. It is defined as the mean squared error (MSE) divided by the variance of the target vector to assure that the error is not over-weighted by those high-variance emotions.

Table II lists our final results and the best nMSEs reported before using other feature representations. The mean and standard deviation values in the table were calculated across 17 emotions over five test sets of each. On the small set with 3858 GIFs, our method achieved a performance better than all the previous. On the whole dataset with 6113 GIFs, the nMSE becomes a little higher, which is probably because the later a GIF was posted on GIFGIF the fewer votes it would usually get. The number of votes a GIF received affects the reliability of its emotion intensities, partially quantified as the uncertainties of the scores. In our dataset, the first 3858 GIFs had an average TrueSkill uncertainty of 1.0286, while the latter 2255 GIFs' was 1.1259.

To further verify the effectiveness of our method, we analyzed the 85 sets (17 emotions x 5 test repetitions) of regression coefficients and intercepts β and β_0 learned from GIFGIF to probe the relationships among emotions. Each pair of β and β_0 was concatenated into a 4097-dim vector, and fed into principal component analysis (PCA). The first and second principal components of all the sets were visualized in Fig. 4. According to the figure, the first principal component clearly indicates the valence of emotions, as positive emotions including happiness, pleasure, amusement and contentment are clustered at high values, and negative feelings such as disgust, sadness and shame appeared on the far left of the figure. On the other hand, the second principal component appears to reflect the risk perception of emotions [22], on which fear and anger have opposite effects. According to Lerner and Keltner, emotions like fear and surprise relate to pessimistic risk estimates, while emotions like anger and contempt exhibit optimistic risk estimates. This discrimination perfectly matches our y-axis in Fig. 4.

IV. Conclusion

We have described, implemented and evaluated a novel methodology for predicting perceived emotions from animated GIFs using 3D CNNs. Our method not only predicts emotion intensities of GIFs more accurately than any previous approach, but also proves to be a promising tool for emotion taxonomy. One potential future direction of this work will be considering both the emotion intensity scores and their uncertainties as the training labels.

References

[1] B. Jou *et al.*, "Predicting Viewer Perceived Emotions in Animated GIFs," in *ACM MM*, 2014, pp. 213–216.

[2] C. Chamaret *et al.*, "LIRIS-ACCEDE : A Video Database for Affective Content Analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 43–55, 2015.

[3] M. Soleymani *et al.*, "A Multimodal Database for Affect Recognition and Implicit Tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[4] A. Schaefer *et al.*, "Assessing the Effectiveness of a Large Database of Emotion-Eliciting Films: a New Tool for Emotion Researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[5] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[6] H. L. Wang and L.-F. Cheong, "Affective Understanding in Film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, 2006.

[7] S. E. Kahou *et al.*, "Emonets: Multimodal Deep Learning Approaches for Emotion Recognition in Video," *Journal on Multimodal User Interfaces*, no. 1, pp. 1–13, 2015.

[8] L. Pang *et al.*, "Deep Multimodal Learning for Affective Analysis and Retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[9] Y.-G. Jiang *et al.*, "Predicting Emotions in User-Generated Videos," in *ICAI*, 2014, pp. 73–79.

[10] B. Xu and Y. Fu, "Video Emotion Recognition with Transferred Deep Feature Encodings," in *ACM ICMR*, 2016, pp. 15–22.

[11] Y. Tang, "Deep Learning using Linear Support Vector Machines," in *ICML*, 2013.

[12] S. Bhattacharya *et al.*, "Towards a Comprehensive Computational Model for Aesthetic Assessment of Videos," in *ACM MM*, 2013, pp. 3–6.

[13] D. Borth *et al.*, "Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs," in *ACM MM*, 2013, pp. 223–232.

[14] D. Tran *et al.*, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *IEEE CVPR*, 2014, pp. 675–678.

[15] T. Rich *et al. GIFGIF* [Online]. Available: http://www.gif.gf

[16] Giphy, Inc. *Giphy* [Online]. Available: http://giphy.com

[17] P. Ekman, "All Emotions Are Basic," *The Nature of Emotion: Fundamental Questions*, pp. 15–19, 1994.

[18] R. Herbrich *et al.*, "TrueSkill: A Bayesian Skill Rating System," in *NIPS*, 2006, pp. 569–576.

[19] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, pp. 267–288, 1996.

[20] J. Chen *et al.*, "Integrating Low-Rank and Group-Sparse Structures for Robust Multi-Task Learning," in *ACM SIGKDD KDD*, 2011, pp. 42–50.

[21] A. Argyriou *et al.*, "Convex Multi-Task Feature Learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[22] J. S. Lerner and D. Keltner, "Fear, Anger and Risk," *Journal of Personality and Social Psychology*, vol. 81, no. 1, pp. 146–159, 2001.