



MIT Open Access Articles

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

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation	Chen, Weixuan and Picard, Rosalind W. 2016. "Predicting Perceived Emotions in Animated GIFs with 3D Convolutional Neural Networks."
As Published	10.1109/ism.2016.0081
Publisher	Institute of Electrical and Electronics Engineers (IEEE)
Version	Author's final manuscript
Citable link	https://hdl.handle.net/1721.1/138085
Terms of Use	Creative Commons Attribution-Noncommercial-Share Alike
Detailed Terms	http://creativecommons.org/licenses/by-nc-sa/4.0/

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

Weixuan Chen, Rosalind W. Picard

Media Lab, Massachusetts Institute of Technology, Cambridge, USA

Email: {cvx, 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 than 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.

Keywords—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. 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.

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. 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, image-based aesthetics, and a mid-level visual representation called SentiBank. 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. [2] 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 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.

II. METHODS

We collected our data from the GIFGIF website [3], a crowd-sourcing platform enabling users to vote on animated GIFs with their perceived emotions. When users enter the homepage of GIFGIF, a pair of random GIFs will be presented with a question "which better expresses X", 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 [4]. 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 algorithm [5]. 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.

Animated GIFs usually have varied lengths. In our dataset, 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 GIFs, we chose to always maintain the continuity of consecutive frames by splitting them into multiple equal-length clips without resampling and extracting features from each of the clips.

We used the C3D video descriptor [2] as our feature representation. 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. The fc6 activations of the network formed a 4096-dim vector for each clip, which was finally saved as our feature representation. Since the feature dimension is comparable to the size of our dataset, 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 [6] so that variable selection can be done automatically.

TABLE I
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

III. RESULTS

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 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 clip-level scores over each GIF within the test sets. A metric called normalized mean squared error (nMSE) commonly used before [7] 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.

Table I 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

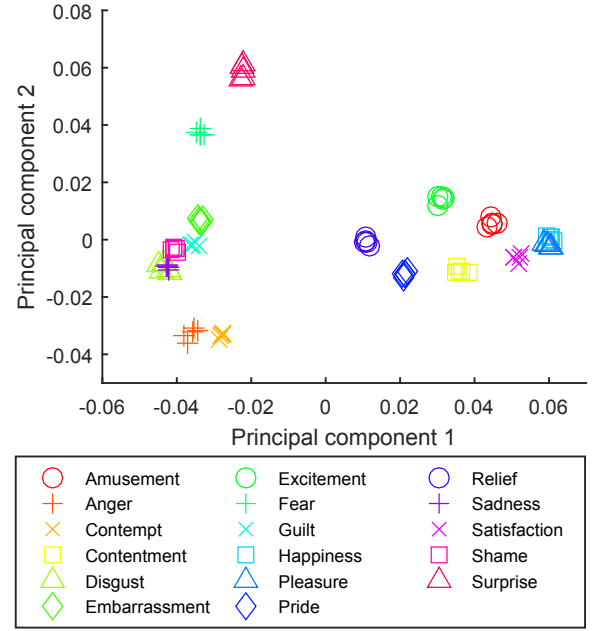


Fig. 1. First and second principal components of our regression coefficients and intercepts.

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 learned from GIFGIF to probe the relationships among emotions. Each pair of coefficients and intercepts 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. 1. 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 [8], on which fear and anger have opposite effects.

REFERENCES

- [1] B. Jou et al., "Predicting Viewer Perceived Emotions in Animated GIFs," in *ACM MM*, 2014, pp. 213–216.
- [2] D. Tran et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," in *IEEE CVPR*, 2014, pp. 675–678.
- [3] T. Rich et al. *GIFGIF* [Online]. Available: <http://www.gif.gf>
- [4] P. Ekman, "All Emotions Are Basic," *The Nature of Emotion: Fundamental Questions*, pp. 15–19, 1994.
- [5] R. Herbrich et al., "TrueSkill: A Bayesian Skill Rating System," in *NIPS*, 2006, pp. 569–576.
- [6] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, pp. 267–288, 1996.
- [7] A. Argyriou et al., "Convex Multi-Task Feature Learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.
- [8] J. S. Lerner and D. Keltner, "Fear, Anger and Risk," *Journal of Personality and Social Psychology*, vol. 81, no. 1, pp. 146–159, 2001.