GIFGIF+: Collecting Emotional Animated GIFs with Clustered Multi-Task Learning

Weixuan Chen, Ognjen (Oggi) Rudovic, and Rosalind W. Picard
Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
E-mail: {cvx, orudovic, picard}@media.mit.edu

***Abstract*—Animated GIFs are widely used on the Internet to express emotions, but their automatic analysis is largely unexplored. Existing GIF datasets with emotion labels are too small for training contemporary machine learning models, so we propose a semi-automatic method to collect emotional animated GIFs from the Internet with the least amount of human labor. The method trains weak emotion recognizers on labeled data, and uses them to sort a large quantity of unlabeled GIFs. We found that by exploiting the clustered structure of emotions, the number of GIFs a labeler needs to check can be greatly reduced. Using the proposed method, a dataset called GIFGIF+ with 23,544 GIFs over 17 emotions was created, which provides a promising platform for affective computing research.**

1. 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. People often make animated GIFs from scenes of movies, cartoons, and TV shows, and use them on social media, digital forums, message boards and even in emails as an enhanced version of emoticons. As a common means to visually express emotions on the Internet, animated GIFs could be ideal research tools and research objects for affective computing [1]. As research tools, animated GIFs can function as emotional stimuli to induce human emotions in studies. As research objects, they contain a wide variety of facial expressions, gestures and other body language, which lead to questions such as how and why they are perceived as emotional indicators.

Despite animated GIFs' popularity and research value, their information processing and retrieval have been rarely explored in affective computing 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. Thus, it is not trivial to develop artificial intelligence systems specifically for understanding animated GIFs, which would benefit both Internet users and affective computing researchers to use and search them more efficiently.

Emotion recognition is the core problem in GIF analysis, just as the object and scene recognition are in standard image analysis tasks. One potentially powerful tool for emotion recognition from animated GIFs is deep learning [2]. Deep neural networks have the ability to mine massive amounts of visual data, resulting in remarkable success in various tasks such as action recognition and facial expression recognition. However, most deep learning techniques rely on a large quantity of labeled data. Currently, the largest emotion-annotated GIF database is GIFGIF [3], with 6119 GIFs covering 17 human-labeled emotions. Compared with popular datasets for video analysis such as UCF101 [4] (13,000 clips) and Sports-1M [5] (1 million clips), its size is far from adequate for training the latest deep neural network models. On the other hand, there are a considerable number of unlabeled animated GIFs on the Internet that can be accessed easily. For instance, the largest GIF search website Giphy [6] contains around 150 million GIFs in its archive. Labeling these GIFs demands a huge amount of human effort, which can be time-consuming, tedious and error-prone. Therefore, there is a need for methods that can collect animated GIFs and assign them emotion labels in a (semi-)automatic manner, requiring minimal human effort for maximal labeling accuracy.

Figure 1. GIFGIF+ Dataset.

To meet the needs described above, we propose a multi-modal emotional recognizer trained on an existing GIF database with high quality labels, use the trained model to automatically rank a large number of unlabeled GIFs on the Internet, and then manually select target GIFs among those with the highest ranks. We show that by applying multi-task learning based on the clustered structure of emotions, the amount of GIFs a labeler needs to check can be reduced greatly. Furthermore, using the proposed method, we collected a large-scale animated GIF dataset with emotion labels, which we call GIFGIF+.

978-1-5386-0563-9/17/$31.00

To our knowledge, it is the largest dataset of GIFs with annotated emotions.

The rest of the paper is organized as follows: we first review previous works on GIF analysis and multimedia datasets with emotion labels. Then we introduce our semi-automatic pipeline for collecting emotional animated GIFs. After introducing and evaluating several learning methods for emotion recognition, the best one is chosen to create the GIFGIF+ dataset. Finally, we show a qualitative and quantitative analysis of the compared methods and datasets.

2. Related Work

2.1. GIF Analysis

There is surprisingly little scholarly work on GIF analysis. Bakhshi et al. [7] discussed why animated GIFs are more engaging than other media by interviewing Tumblr users and analyzing visual features of GIFs including frame rate, uniformity, and resolution. Cai et al. [8] proposed a spatial-temporal sentiment ontology for GIFs to establish a relationship between visual concepts of GIFs and their sentiment polarity. Gygli et al. [9] trained a visual model on 100K user-generated GIFs and their corresponding video sources to learn to automatically generate animated GIFs from video.

In terms of emotion recognition from GIFs, all previous work we found has been conducted using GIFGIF. Jou et al. [10] compared four different feature representations: color histograms, facial expressions [11], image-based aesthetics [12], and visual sentiment [13] for emotion recognition on GIFGIF. Chen et al. [14] proposed using 3D convolutional neural networks (CNNs) to extract spatiotemporal features from GIFs, which further improved the emotion recognition accuracy on GIFGIF. However, due to the small size and large complexity of the GIFGIF dataset, their accuracies were relatively low and insufficient for practical applications such as reliable and automatic GIF indexing.

2.2. Emotional Multimedia Datasets

There exist several multimedia datasets with emotion-related labels, as shown in Table 3. However, the labels of all these datasets are based on induced emotion, which is different from the perceived emotion labeling of GIFGIF. 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 (induced emotion). According to Jou et al. [10], perceived emotions are more concrete and objective than induced emotions, where labels are less reliable due to their interaction with subjective experience. Specific to animated GIFs, it is their perceived emotions rather than induced emotions that usually determines how GIFS are used. Typically, people post a GIF to express their current emotion instead of to induce a certain emotion from the readers, as with using an emoticon.

3. Methods

3.1. GIFGIF Platform

Our goal is to quickly and efficiently collect emotional animated GIFs by expanding an existing dataset with labels. We start with the dataset GIFGIF [3], a crowdsourcing platform enabling users to vote on animated GIFs with their perceived emotions. The GIFs on the platform are imported from the Giphy website [6], 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.

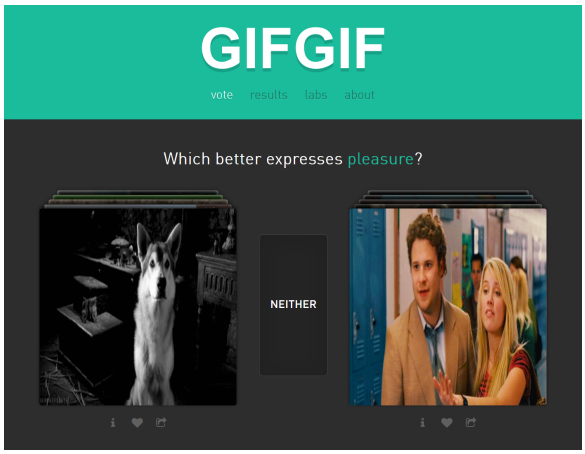

Figure 2. GIFGIF homepage: http://www.gif.gf .

When users enter the homepage of GIFGIF, a pair of random GIFs is presented with a question "which better expresses X?", as shown in Fig. 2, 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 [15]. With all the answers from thousands 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 using the TrueSkill rating algorithm [16], in which the i-th emotion score of the n-th GIF is represented as a normal distribution characterized by a mean $\mu_{n,i}$ and standard deviation $\sigma_{n,i}$. Every GIF is initialized with a prior $\mu_0 = 25$ and $\sigma_0 = 25/3$. When compared with another GIF, it gets a vote or veto, and its $\mu_{n,i}$ will increase or decrease accordingly. As the GIF accumulates more and more votes, we become more confident in its emotion score, as reflected in the decrease of $\sigma_{n,i}$.

Until May 1, 2017, the GIFGIF platform had indexed 6119 animated GIFs with 3,130,780 crowd-sourced annotations. Omitting 6 GIFs with broken links, we downloaded 6113 files with their emotion scores. As an example, the histograms of $\mu_{n,i}$ and $\sigma_{n,i}$ of all GIFs corre-

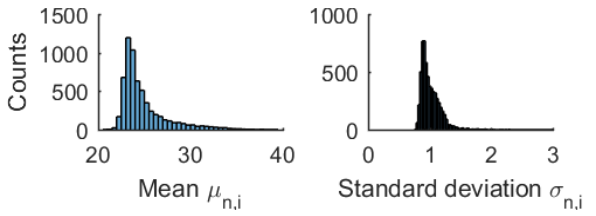

Figure 3. Histograms of the mean $\mu_{n,i}$ and standard deviation $\sigma_{n,i}$ of the emotion scores for all samples on GIFGIF, $i = excitement$.

sponding to the "excitement" emotion are shown in Fig. 3.

There are two main problems with this dataset in its potential use for emotion recognition.

1) Every GIF is annotated with not only an emotional intensity ($\mu_{n,i}$), but also an uncertainty ($\sigma_{n,i}$). However, all previous work [10], [14] only use $\mu_{n,i}$ as their learning labels. As shown in Fig. 3, the means of emotion scores for the off-target samples (GIFs not showing 'excitement' in the example) are within a small range, the scale of which is close to the scores' standard deviations. As a result, to equally treat two GIFs with close means but very different standard deviations would possibly have a negative effect on the training of emotion recognizers.
2) The sample size (6113) is too small for training the latest deep computer vision models.

To solve the first problem, we adopt a different metric introduced by the TrueSkill paper [16]. Instead of using $\mu_{n,i}$, we use the 1% lower quantile $y_{n,i} = \mu_{n,i} - 3\sigma_{n,i}$ as the emotion score, to favor GIFs with both high mean values and low standard deviations. To address 2), we elaborate on our semi-automatic data collection pipeline below.

3.2. Data Collection Pipeline

We decided to collect new animated GIFs from the Giphy website [6], as it is currently the largest GIF search engine, and has a well-documented API[1] for searching and retrieving GIFs. The pipeline of our data collection methodology is depicted in Fig. 4 and introduced as follows.

We first binarized the GIFGIF emotion scores to define positive samples for each emotion. The GIFGIF platform was able to annotate each GIF with continuous scores, because the scores were crowdsouced from thousands of users in the span of more than three years. To greatly expand the dataset in a short time with limited resources without compromising accuracy, it would be better to have more certain labels. Thus we defined a positive sample as a GIF with $y_{n,i} > \mu_0$, which means it has a confidence of more than 99% to be more emotional than the average level in emotion category i.

To retrieve GIFs matching the positive samples from 150 million entries on Giphy, it is nearly impossible to apply any automatic or semi-automatic filtering direcly, so we did a pre-screening using the tags of GIFs. Most GIFs on Giphy have several tags created by the GIF uploader or website users describing the sources, themes or contents of GIFs. The most intuitive way to retrieve emotion-relevant GIFs would be searching GIFs with the emotion names as their tags, e.g. search "relief" or "relieved" tags to get GIFs perceived as relief. However, most of the emotion names are not common tags on the website, and some of them can lead to confusion with other themes, e.g., searching 'amusement' returns mainly GIFs showing amusement parks, and searching 'pride' returns many results related to pride parades. Hence, we traced the positive samples on GIFGIF to their pages on Giphy, and used their most frequent tags as our search terms. Table 1 shows the top 10 common tags in each emotion group of GIFs on GIFGIF.

With the top 10 tags entered as search terms, Giphy returned on average 50,000 GIF candidates in each emotion category after removing duplicate entries. It is still a huge amount of work to manually assign labels for all these GIFs. Thus in the next step we trained 17 emotion recognizers on the labeled GIFGIF data using both visual and tag features, which will be elaborated in the subsequent sections. Due to the limited size of the labeled data, the trained recognizers are relatively weak in performance, but they are still able to greatly reduce the required human labor. The recognizers are applied to the GIF candidates to re-sort them by the recognizer predictions. Following the new order, human labelers check the GIFs manually to decide if they indeed belong to a specific emotion category or are false positives, until a preset number of positive samples is reached.

3.3. Visual Features

Since the GIFGIF dataset is too small to train a deep vision model from scratch, we adopted the C3D video descriptor [17] as our visual feature representation for transfer learning. C3D is a 3D CNN pre-trained on the Sport1M dataset. It has been shown by Tran et al. [17] that for video analysis volume-based features such as C3D are superior to image-based ones due to their capability of modeling motions. C3D also shows good generalization capability across various video analysis tasks (action recognition, scene classification, and object recognition) without requiring to finetune the model for each task. The details about the architecture of the C3D neural network can be seen in its original paper [17].

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 all the clips were finally averaged and L2-normalized to form a 4096-dim vector for each GIF, which was saved as our visual feature representation.

3.4. Tag Features

To compute tag features, a dictionary was created from the tags of all the 6113 GIFs on GIFGIF. First, all the tags were gathered in one place, among which 11,042 unique

1. GiphyAPI: https://github.com/Giphy/GiphyAPI

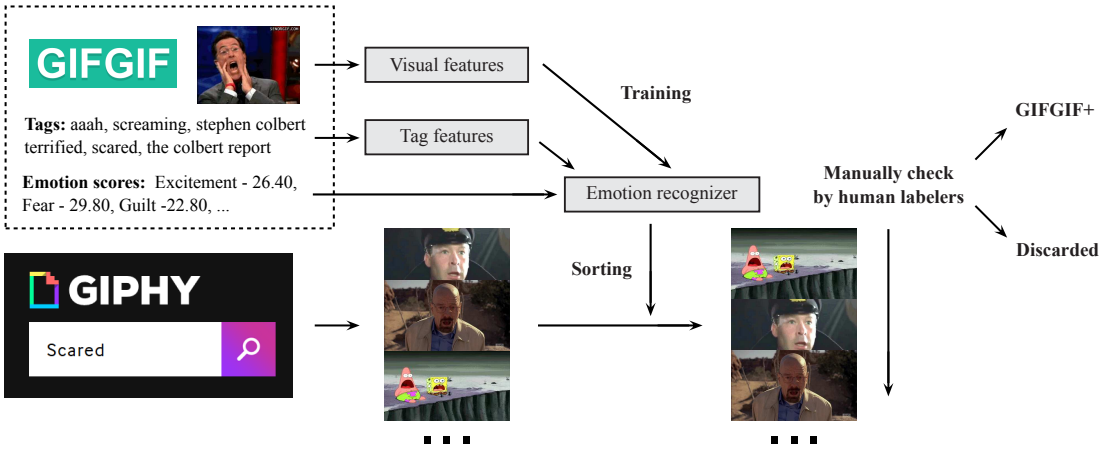

Figure 4. Flowchart describing our data collection pipeline.

TABLE 1. TOP 10 COMMON TAGS FOR EACH EMOTION GROUP OF GIFS ON GIFGIF. GENERIC TAGS SUCH AS "TV" AND "ANIME" ARE REMOVED AND MARKED WITH STRIKETHROUGH.

Emotions	Tags sorted by frequency
Amusement	laughing, happy, laugh, excited, smile, lol, ~~tv~~, dancing, exciting, ~~movies~~, funny, smiling
Anger	angry, ~~movies~~, ~~tv~~, frustrated, ~~cartoons & comics~~, funny, hate, anger, ~~anime~~, mad, ~~movie~~, upset, no, rage, annoyed
Contempt	angry, ~~tv~~, eye roll, no, frustrated, smh, ~~movies~~, unimpressed, annoyed, confused, ~~reaction~~, suspicious, smdh
Contentment	happy, smile, dancing, excited, laughing, ~~animals~~, cute, smiling, ~~cartoons & comics~~, ~~movies~~, thumbs up, ~~tv~~, funny, ~~baby~~, laugh
Disgust	no, ~~tv~~, ~~movies~~, angry, disgusted, eye roll, confused, frustrated, shocked, smh, gross, ~~reaction~~, scared
Embarrassment	facepalm, ~~tv~~, awkward, embarrassed, frustrated, nervous, embarassed, funny, ~~movies~~, ~~cartoons & comics~~, sad, annoyed, ~~cartoon~~, disappointed
Excitement	happy, excited, exciting, dancing, ~~tv~~, laughing, funny, ~~cartoons & comics~~, laugh, ~~reaction~~, smile, adventure time, cute
Fear	scared, shocked, ~~movies~~, cat, nervous, ~~cartoons & comics~~, ~~tv~~, surprised, screaming, ~~animals~~, funny, ~~reaction~~, lol, scream, terrified
Guilt	sad, ~~movies~~, crying, ~~tv~~, nervous, ~~cartoons & comics~~, facepalm, embarrassed, sorry, ~~cartoon~~, cry, disappointed, ~~movie~~, awkward, pout
Happiness	happy, laughing, excited, laugh, ~~tv~~, smile, dancing, ~~movies~~, exciting, lol, funny, ~~cartoons & comics~~, smiling
Pleasure	happy, excited, laughing, smile, ~~tv~~, laugh, dancing, exciting, funny, ~~movies~~, cute, love, ~~cartoons & comics~~
Pride	happy, ~~tv~~, excited, yes, dancing, smile, ~~movies~~, smiling, ~~sports~~, exciting, laughing, thumbs up, ~~cartoons & comics~~, celebration
Relief	happy, smile, ~~movies~~, excited, ~~giphytrending~~, laughing, yes, sigh, smiling, exciting, ~~cartoons & comics~~, ~~animals~~, cute, jennifer lawrence
Sadness	sad, crying, ~~movies~~, ~~tv~~, cry, disappointed, upset, sadness, tears, ~~anime~~, ~~cartoons & comics~~, love, pout, ~~movie~~, ~~cartoon~~, ~~disney~~, frustrated
Satisfaction	happy, excited, ~~tv~~, dancing, laughing, smile, exciting, smiling, ~~cartoons & comics~~, funny, ~~movies~~, thumbs up, yes
Shame	sad, facepalm, ~~tv~~, ~~movies~~, crying, frustrated, disappointed, embarrassed, nervous, ~~cartoons & comics~~, embarassed, sorry, awkward
Surprise	shocked, scared, surprised, ~~tv~~, excited, ~~reaction~~, ~~movies~~, happy, funny, omg, exciting, cat, confused

tags were found and sorted by their frequency. Then a common sparsity threshold of 0.995 was applied to the unique tags to only keep those that appear in 0.5% or more of the GIFs, which has proved to help generalization and prevent overfitting. The remaining 139 tags were saved as a dictionary for computing bag-of-words features for the tags of each GIF. After sample-wise L2-normalization, a 139-dim vector counting the appearance of each dictionary entry was finally generated for every GIF as the tag features.

3.5. Learning Methods

To facilitate sorting of the unlabeled GIF candidates, we resort to machine learning approaches that can generalize from a limited number of labeled samples. The learning problem is defined as follows. Specifically, for emotion $i \in (1 \cdots t)$, GIF n has a feature vector $x_{n,i} \in \mathbb{R}^d$ containing the concatenated visual and tag features, and an emotion score $y_{n,i} \in \mathbb{R}$. For N_i GIFs used for training the i-th emotion recognizer, $X_i = (x_{1,i} \cdots x_{N_i,i})$ denotes the feature matrix, and $Y_i = (y_{1,i} \cdots y_{N_i,i})$ denotes the training labels. Our goal is to learn t models to predict Y_i from X_i.

3.5.1. Single-Task Lasso Regression. For every GIF clip, our feature vector has 4235 (4096 visual + 139 tag) features. Because this is comparable to the size of our labeled data, to avoid over-fitting, we used Lasso regression [18] as our single-task learning baseline to train parsimonious models independently for each emotion category. Formally, a linear Lasso regression solves the following problem:

$$\min_{W_i} ||W_i^T X_i - Y_i||_F^2 + \rho||W_i||_1, \ i = 1, 2, \cdots, t \quad (1)$$

where W_i is a linear model for emotion i, $||\cdot||_F$ is the Frobenius norm, $||\cdot||_1$ is the l_1-norm, and ρ is a non-negative regularization parameter optimized via cross-validation.

3.5.2. Gaussian Process Regression. We also consider the Gaussian Process (GP) framework for regression [19]. GPs are particularly fit for the target task due to their abilities to generalize well from a limited amount of data, deal with high dimensional inputs (due to their non-parametric nature), and represent uncertainty in the model's prediction. Formally, given a new test input X_*, the GP for the i-th emotion is defined by its predictive (Normal) distribution with the mean and variance:

$$\mu^{(i)}(X_*^{(i)}) = {k_*^{(i)}}^T (K^{(i)} + \sigma_i^2 I)^{-1} Y^{(i)} \quad (2)$$

$$V^{(i)}(X_*^{(i)}) = k_{**}^{(i)} - {k_*^{(i)}}^T (K^{(i)} + \sigma_i^2 I)^{-1} k_*^{(i)}, \quad (3)$$

where $k_*^{(i)} = k^{(i)}(X^{(i)}, X_*^{(i)})$, $k_{**}^{(i)} = k^{(i)}(X_*^{(i)}, X_*^{(i)})$ and $K^{(i)}$ are kernel functions computed on train-test, test-test, and train-train data, respectively. Typically, a sum of Radial Basis Function (RBF) and noise term (σ_i^2) is used in the kernel function, and we adopted the same. Parameter estimation in a GP is easy as it does not require lengthy cross-validation procedures, and it consists of finding the kernel hyper-parameters (in our case, length scale and noise term stored in $\theta^{(i)}$) that maximize the log-marginal likelihood:

$$\log p(Y^{(i)}|X^{(i)}, \theta^{(i)}) = -\frac{1}{2}\text{tr}\left[(K^{(i)} + \sigma_i^2 I)^{-1} Y^{(i)} {Y^{(i)}}^T\right] - \frac{C}{2}\log|K^{(i)} + \sigma_i^2 I| + \text{const.} \quad (4)$$

To solve the maximization problem, gradient ascent is used (based on conjugate gradients [19]). Finally, to leverage the confidence information provided by GPs, we sort the target GIFs according to the following (probability) score:

$$p(GIF_* \in i|X_*^{(i)}) \sim exp(-\frac{50 - \mu^{(i)}(X_*^{(i)})}{2V^{(i)}(X_*^{(i)})}), \quad (5)$$

where we assumed that the GIF_* is more likely to belong to emotion i if its mean is closer to the maximum (i.e., 50).

3.5.3. Multi-task Regression with Trace-norm Regularization. The two regression methods introduced above assume that the 17 emotion recognition tasks are independent. However, the emotion classes in the GIFGIF dataset are in fact highly related. For example, positive emotions such as "Happiness," "Pleasure" and "Excitement" share similar visual and tag features. To account for this, we adopt the use of multi-task learning, in which related tasks are learned simultaneously by leveraging information shared across tasks. In this way, the parameter regularization is achieved rendering a model more robust to overfitting. Formally, we denote the learned models for all emotions as $W = (W_1 \cdots W_t)$. To capture the emotion relatedness, we assume that different emotions share a low-dimensional subspace, captured by a low-rank projection matrix W. This can be posed as the following rank minimization problem:

$$\min_W \sum_{i=1}^{t} ||W_i^T X_i - Y_i||_F^2 + \rho Rank(W). \quad (6)$$

Solving for W is NP-hard in general, so a popular substitute [20] is to replace the rank function with a trace norm:

$$\min_W \sum_{i=1}^{t} ||W_i^T X_i - Y_i||_F^2 + \rho \sum_{j=1}^{min(d,t)} \sigma_j(W), \quad (7)$$

Where $\sigma_j(W)$ are the successive singular values of W. The regularization factor ρ is found via cross-validation.

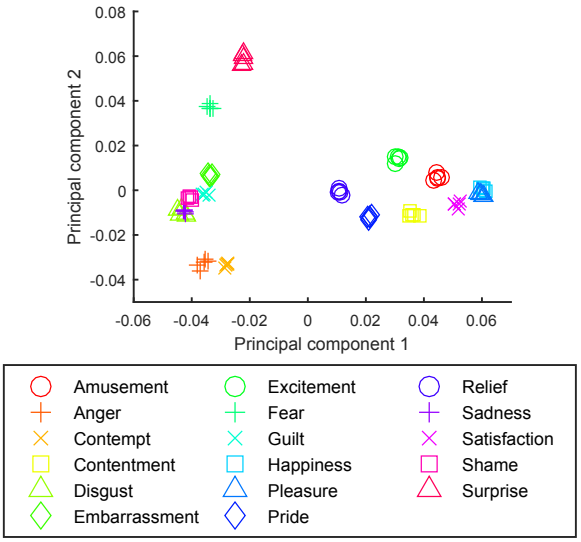

Figure 5. First and second principal components of visual-feature-based emotion recognizers trained on 5 folds of GIFGIF using Lasso regression. (This figure is reproduced with permission from [14].)

3.5.4. Clustered Multi-task Regression. The trace-norm regularization assumes that all learning tasks are related, so that all the emotion recognition models share a common low-dimensional subspace. This assumption is restrictive, as emotion pairs like "happiness-sadness" and "happiness-pleasure" likely do not share information to the same level. According to our observation, the emotions exhibit a more sophisticated group structure where the models of emotion recognizers from the same group are closer to each other than those from a different group. Fig. 5, reproduced from [14], shows the principal components of emotion recognizers trained on GIFGIF using only visual features, which imply clustered patterns related to the valence and risk perception [21] of emotions.

To make use of the clustered structure, clustered multi-task learning [22] is a viable solution. Assuming emotions can be clustered into $k < t$ groups, the cluster assignment can be represented by a $t \times k$ binary matrix E, in which $E_{i,m} = 1$ if emotion i is in cluster m. For easier expression, define $M = E(E^T E)^{-1} E^T$ and U a $t \times t$ projection matrix whose entries are all equal to $1/t$. A general framework for clustered multi-task learning includes three penalties: (i) a global penalty on the elements of the weight matrix, (ii) a measure of between-cluster variance (the difference between the clusters), and (iii) a measure of within-cluster variance (the compactness of the clusters). To make the learning problem tractable, a relaxed convex

solution was proposed in [23]:

$$\min_{W} \sum_{i=1}^{t} ||W_i^T X_i - Y_i||_F^2 + \rho_1 \eta(1+\eta) tr(W(\eta I + M)^{-1} W^T) \quad s.t.: \ tr(M) = k, M \prec I, M \in S_+^t, \eta = \frac{\rho_2}{\rho_1} \quad (8)$$

where ρ_1 and ρ_2 are non-negative regularization parameters optimized via cross-validation.

3.6. Evaluation

To evaluate the aforementioned methods before choosing the one for data collection, we separated a part of the GIFGIF dataset to form test sets. For each emotion category, the human-labeled GIFs with the top 10 common tags in Table 1 were chosen to resemble the distribution of the unlabeled GIF candidates. From them, 33% were randomly selected as test sets, and all the remaining labeled GIFs were used for training the models.

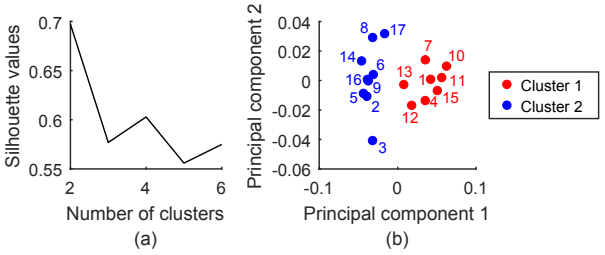

Figure 6. (a) Silhouette values w.r.t. the chosen number of clusters in k-means. (b) First and second principal components of W with k-means partition ($k = 2$). The numbers indicate 1: amusement, 2: anger, 3: contempt, 4: contentment, 5: disgust, 6: embarrassment, 7: excitement, 8: fear, 9: guilt, 10: happiness, 11: pleasure, 12: pride, 13: relief, 14: sadness, 15: satisfaction, 16: shame, and 17: surprise.

In clustered multi-task regression, there is an extra hyper-parameter k, the number of emotion clusters. To find an appropriate k, k-means clustering was performed on the model W trained by the trace-norm regularized multi-task regression. Different choices of k were compared using the Silhouette criterion [24]. As shown in Fig. 6 (a), $k = 2$ gives the highest Silhouette value, which indicates the best cluster partition. We draw the k-means partition results using $k = 2$ along the first two principal components of W in Fig. 6 (b), which shows the two clusters respectively correspond to positive emotions and the other emotions. Note that the distribution of emotions becomes different from Fig. 5, probably due to the introduction of tag features, but the presence of a positive cluster is robust.

After all the regularization parameters were optimized via 5-fold cross-validation, the emotion recognizers were re-trained on the whole training sets. With the test sets sorted by the trained recognizers, precision and recall values can be computed for different thresholds. Fig. 7 illustrates the precision-recall curves of all the learning methods, averaged among 17 emotions. As shown in the figure, without sorting, randomly checking the GIF candidates would only give a precision of 0.25. By introducing learning on visual and tag features, the efficiency

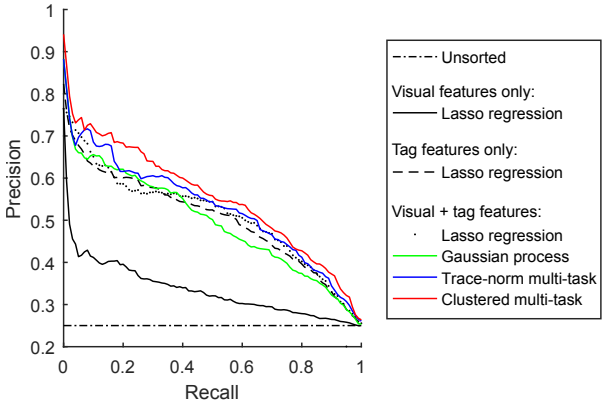

Figure 7. Average precision-recall curves for all the tested learning methods.

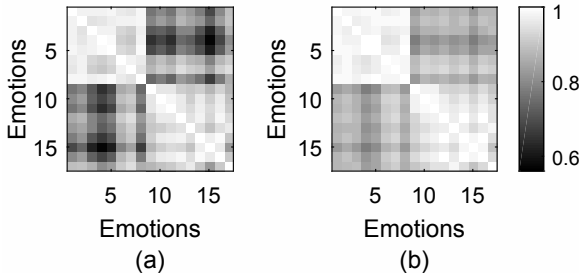

Figure 8. Model correlations among 17 emotions of (a) clustered multi-task regression and (b) trace-norm regularized multi-task regression. For better visualization of the two clusters, the emotions are reordered as follows: 1: amusement, 2: contentment, 3: excitement, 4: happiness, 5: pleasure, 6: pride, 7: satisfaction, 8: relief, 9: anger, 10: contempt, 11: disgust, 12: embarrassment, 13: fear, 14: guilt, 15: sadness, 16: shame, and 17: surprise.

can be greatly improved. Compared with using only the tag features, Lasso regression on visual features gives much lower precisions, probably because learning was not conducted on the raw GIFs directly and the performance of using the C3D representation for transfer learning was just passable. Nonetheless, combining the visual and the tag features still produce better results than using only the tag features. The curves also demonstrate the superiority of multi-task learning over single-task learning, and clustered multi-task regression displays overall the best performance. To explain why clustered multi-task regression could beat trace-norm regularized multi-task regression, we drew the correlation coefficients between the learned emotion models W_i, $i = 1 \cdots 17$ in Fig. 8, which shows that the clustered multi-task method better captured the clustered structure of positive emotions and non-positive emotions.

Comparison of the evaluations is summarized in Table 2. First, the area-under-curve (AUC) was computed for each precision-recall curve in Fig. 7. Then, the number of GIF candidates a human labeler needs to check to get enough positive samples for each emotion category was estimated as

$$\tilde{N} = \frac{N_p}{Precision}, \ s.t. \ Recall = \frac{N_p}{N_{all}} \quad (9)$$

in which N_p is the targeted number of positive samples, and N_{all} is the number of all GIF candidates. We report the numbers for $N_p = 3,000$ and $N_{all} = 50,000$. Table

TABLE 2. AREA UNDER CURVE (AUC) OF THE PRECISION-RECALL CURVES, AND $\bar{N}$ THE EXPECTED NUMBER OF GIFS TO CHECK FOR EACH EMOTION CATEGORY ON AVERAGE.

Methods	AUC	Expected numbers
Unsorted	0.250	11,996
Visual features only		
Lasso regression	0.336	7,196
Tag features only		
Lasso regression	0.504	4,486
Visual + tag features		
Lasso regression	0.511	4,302
Gaussian process regression	0.497	4,550
Trace-norm multi-task regression	0.530	4,247
Clustered multi-task regression	0.555	4,033

2 demonstrates that the clustered multi-task regression achieves the highest AUC, while requiring the fewest GIFs to be manually examined.

4. GIFGIF+ Dataset

Based on the evaluation, the best emotion recognizer was clustered multi-task regression using both visual and tag features. The recognizer was then applied to the GIF candidates we collected from Giphy to sort them by the predicted emotion scores. In the last step, two labelers manually checked the GIF candidates following the new order, and assigned GIFs to emotion categories only when consensuses were reached. In this way, 3,000 GIFs were collected with associated tags for each of the 17 emotions. Many GIFs were assigned to have multiple emotion labels. In sum, a total of 23,544 GIFs collected. We call this expanded dataset GIFGIF+[2]. The comparison of this dataset with previous emotion-annotated multimedia datasets is summarized in Table 3.

TABLE 3. COMPARISON OF GIFGIF+ WITH PREVIOUS EMOTION-ANNOTATED MULTIMEDIA DATASETS.

Study	Dataset size	Modalities
Wang and Cheong [25]	36 full-length popular Hollywood movies (2040 scenes)	7 emotions
Arifin and Cheung [26]	43 videos (10970 shots and 762 video segments)	6 emotions
Zhang et al. [27]	552 music videos in different languages and different styles	Arousal and valence
Soleymani et al. [28]	8 famous Hollywood movies (64 movie scenes)	Arousal and valence
Yan et al. [29]	4 films (112 scenes)	4 emotions
Baveye et al. [30]	160 movies (9800 video clips)	Valence
GIFGIF+	23,544 GIFs	17 emotions

The main difference between animated GIFs and videos is that GIFs usually have shorter lengths and much more varied frame rates. Fig. 9 shows the histograms of the frame numbers and the average frame delays of GIFGIF+. According to the figure, the longest GIF has 347 frames, while the shortest has only 2 frames. Also, the highest frame rate is about 40 times the lowest in the dataset.

Another characteristic of the dataset is that a single GIF can belong to multiple emotion categories. Fig. 10

2. Available at http://affect.media.mit.edu/share-data.php

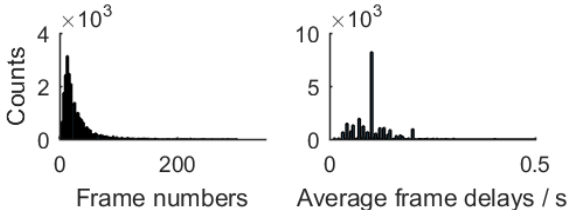

Figure 9. Histograms of frame numbers and average frame delays in GIFGIF+.

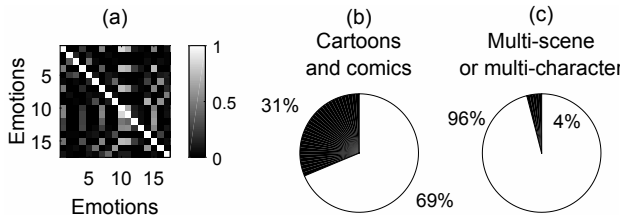

Figure 10. (a) Overlapping rates between 17 emotions in GIFGIF+. The intensity of a pixel indicates the percentage of GIFs of the row emotion that also belong to the column emotion. The emotions are in the same order as Fig. 6. (b)(c) Pie charts showing the percentage of GIFs made from cartoons or comics, and showing multiple scenes and/or multiple characters.

shows the overlapping rates between all the emotions. To help with the analysis of the dataset, we also provided two useful flags for each GIF: one indicates if a GIF is made from cartoons or comics, and the other indicates if a GIF includes multiple scenes and/or multiple characters. The two flags were created, because the mixture of drawings and real-world scenes, and the existence of multiple scenes/characters are the main difficulties for emotion recognition from GIFs. With the flags, users can easily choose a subset of GIFs for more simplified learning.

5. Potential Usage of the Database

The most obvious usage of the expanded database would be training visual emotion recognizers on GIFs. As each GIF can appear in multiple emotion categories, this is a multi-label classification problem. Using C3D visual features and linear support vector machines (SVM) with 20% hold-out testing, we give a binary relevance baseline in Table 4 by averaging the results of 17 independently trained classifiers. Potential directions to improve the result include training recognizers using the raw GIFs, and using the frame rate information to re-sample the GIF frames.

TABLE 4. MEAN AND STANDARD DEVIATION OF PRECISION, RECALL, AND F1 SCORES FOR EMOTION CLASSIFIERS TRAINED ON GIFGIF+.

Methods	Precision	Recall	F1
C3D + linear SVM	0.20 ± 0.11	0.55 ± 0.12	0.29 ± 0.13

It is also interesting to apply unsupervised learning to each category of GIFs to learn representative actions. The animated GIFs in GIFGIF+ contain not only a wide variety of facial expressions but also different gestures and other body language. A great number of GIFs within each emotion category share similar actions even between comic characters and real actors. Learning these actions

from the dataset would help advance human emotion analysis beyond facial expression recognition.

6. Conclusion

We have proposed a novel clustered multi-task learning approach for predicting perceived emotions from a diverse set of animated GIFs. This approach combines 3D CNNs and transfer learning to enable an efficient labeling of a large set of target GIFS in terms of 17 emotion categories (i.e., tasks) and their intensity. We showed that the proposed method outperforms previous approaches for emotion prediction from GIFs, and also provides the GIF representations that map onto intuitively interpretable clusters (e.g., the cluster of positive discrete emotions). Using this approach, combined with human labeling in a way that maximizes precision-recall while also minimizing the effort required to label the data, we were able to speed up the development of a large database containing more than 20,000 emotion-labeled GIFs. This database is labelled in terms of 17 emotion categories and will be made publicly available for research and educational purposes.

Acknowledgements

This research was supported by the SDSC Global Foundation, and the MIT Media Lab Consortium. The work of O. Rudovic has been funded by the European Community Horizon 2020 under grant agreement no. 701236 (EngageMe - Marie Curie Individual Fellowship).

References

[1] R. W. Picard, *Affective computing*. MIT Press Cambridge, 1997.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

[3] T. Rich, K. Hu, and B. Tome, "GIFGIF." [Online]. Available: http://www.gif.gf/

[4] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1725–1732, 2014.

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

[7] S. Bakhshi, D. A. Shamma, L. Kennedy, Y. Song, P. de Juan, and J. J. Kaye, "Fast, Cheap, and Good: Why Animated GIFs Engage Us," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 575–586, 2016.

[8] Z. Cai, D. Cao, D. Lin, and R. Ji, "A Spatial-Temporal Visual Mid-Level Ontology for GIF Sentiment Analysis," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pp. 4860–4865, 2016.

[9] M. Gygli, Y. Song, and L. Cao, "Video2GIF: Automatic Generation of Animated GIFs from Video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1001–1009, 2016.

[10] B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting Viewer Perceived Emotions in Animated GIFs," in *Proceedings of the ACM International Conference on Multimedia*, pp. 213–216, 2014.

[11] Y. Tang, "Deep Learning using Linear Support Vector Machines," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[12] S. Bhattacharya, B. Nojavanasghari, and T. Chen, "Towards a Comprehensive Computational Model for Aesthetic Assessment of Videos," in *Proceedings of the ACM International Conference on Multimedia*, pp. 3–6, 2013.

[13] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs," in *Proceedings of the ACM International Conference on Multimedia*, pp. 223–232, 2013.

[14] W. Chen and R. W. Picard, "Predicting Perceived Emotions in Animated GIFs with 3D Convolutional Neural Networks," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pp. 367–368, 2016.

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

[16] R. Herbrich, T. Minka, and T. Graepel, "TrueSkill: A Bayesian Skill Rating System," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 569–576, 2016.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 675–678, 2014.

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

[19] C. E. Rasmussen, *Gaussian processes for machine learning*. Citeseer, 2006.

[20] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, PhD thesis, Stanford University, 2002.

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

[22] L. Jacob, J.-p. Vert, and F. R. Bach, "Clustered multi-task learning: A convex formulation," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 745–752, 2009.

[23] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 702–710, 2011.

[24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[25] 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.

[26] S. Arifin and P. Y. K. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Trans. Multimed.*, vol. 10, no. 7, pp. 1325–1341, 2008.

[27] S. Zhang, Q. Huang, Q. Tian, S. Jiang, and W. Gao, "Personalized MTV affective analysis using user profile," in *Advances in Multimedia Information Processing-PCM*, pp. 327–337, 2008.

[28] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *IEEE International Symposium on Multimedia (ISM)*, pp. 228–235, 2008.

[29] L. Yan, X. Wen, and Z. Wei, "Study on Unascertained Clustering for Video Affective Recognition," in *Journal of Information and Computational Science*, vol. 8, no. 13, pp. 2865–2873, 2011.

[30] Y. Baveye, J. N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret, "A large video data base for computational models of induced emotion," in *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 13–18, 2013.