

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/335356515>

Sentiment Recognition for Short Annotated GIFs using Visual-Textual Fusion

Article in IEEE Transactions on Multimedia · April 2020

DOI: 10.1109/TMM.2019.2936805

CITATIONS

0

READS

388

6 authors, including:



Tianliang Liu

Nanjing University of Posts and Telecommunications

27 PUBLICATIONS 107 CITATIONS

[SEE PROFILE](#)



Jiebo Luo

University of Rochester

515 PUBLICATIONS 14,178 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



3D reconstruction of large-scale scenes [View project](#)



Machine learning of multi-media social network [View project](#)

Sentiment Recognition for Short Annotated GIFs Using Visual-Textual Fusion

Tianliang Liu , Junwei Wan , Xiubin Dai , Feng Liu , Quanzeng You, and Jiebo Luo, *Fellow, IEEE*

I. INTRODUCTION

Abstract—With the rapid development of social media, visual sentiment analysis from image or video has become a hot spot in visual understanding researches. In this work, we propose an effective approach using visual and textual fusion for sentiment analysis of short GIF videos with textual descriptions. We extract both sequence-level and frame-level visual features for each given GIF video. Next, we build a visual sentiment classifier by using the extracted features. We also define a mapping function, which converts the sentiment probability from the classifier to a sentiment score used in our fusion function. At the same time, for the accompanied textual annotations, we employ the Synset forest to extract the sets of the meaningful sentiment words and utilize the SentiWordNet3.0 model to obtain the textual sentiment score. Then, we design a joint visual-textual sentiment score function weighted with visual sentiment component and textual sentiment one. To make the function more robust, we introduce a noticeable difference threshold to further process the fused sentiment score. Finally, we adopt a grid search technique to obtain relevant model hyper-parameters by optimizing a sentiment aware score function. Experimental results and analysis extensively demonstrate the effectiveness of the proposed sentiment recognition scheme on three benchmark datasets including T-GIF dataset, GSO-2016 dataset and Adjusted-GIFGIF dataset.

Index Terms—GIFs Sentiment, 3-D Convolution, Convolutional Long-Short-Term-Memory, SentiWordNet3.0, Grid Searching.

Manuscript received June 27, 2018; revised April 18, 2019 and July 27, 2019; accepted August 7, 2019. Date of publication August 21, 2019; date of current version March 24, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61001152, Grant 61071091, Grant 31671006, Grant 61572503, Grant 61772286, Grant 61872199, Grant 61872424, and Grant 6193000388, in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK2012437, in part by the China Scholarship Council, and in part by “333” project of Jiangsu Province under Grant BRA2017401. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet. (*Corresponding author: Tianliang Liu.*)

T. Liu, J. Wan, X. Dai, and F. Liu are with the Jiangsu Provincial Key Laboratory of Image Processing and Image Communication, Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Jiangsu Provincial Engineering Research Center for High Performance Computing and Intelligent Information Processing, National Engineering Research Center for Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003 China (e-mail: liutl@njupt.edu.cn; w_j_wan@163.com; daixb@njupt.edu.cn; liuf@njupt.edu.cn).

Q. You is with the Computer Vision Team, Microsoft Cloud + AI, Redmond, WA 98052 USA (e-mail: quanzeng.you@microsoft.com).

J. Luo is with the Department of Computer Science, University of Rochester Rochester, Rochester, NY 14627 USA (e-mail: jluo@cs.rochester.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2019.2936805

AS AN important platform for information exchange, social media has become a main channel for people to communicate with others on almost every single topic of our daily life. In general, social multimedia refers to the multimedia resources posted on online social media [1], which promotes community curation, personal engagement and instant dialogue system. In the era of big data, a huge amount of multimedia data with different modalities such as text, image and video, etc., is generated in all kinds of online social networks per minute. Discovering the knowledge embedded in social multimedia is of great significance. For example, sentiment analysis (known as opinion mining or emotion AI) research has often been carried out on text, e.g., tweets [2]. In this work, we study sentiment analysis of online multimedia contents generated by social network users.

Recently, multimedia contents including images and videos are becoming increasingly popular due to availability of faster and cheaper Fourth Generation (4G) wireless network. The Graphics Interchange Format (GIF) has become the internet's favorite bit-mapped graphics image file type due to its abilities to animate the images on the World Wide Web, CompuServe and BBSs. For instance, since the GIF is a standard of defining a mechanism for the storage and transmission of generalized color raster images or graphics information, the animated GIF video or images (GIFs) have regained huge popularity with their widespread usage in instant messaging, online journalism, social media, online service, among others. Compared with traditional images, the GIFs have better capabilities to show dynamic content, tell stories and convey emotions [14]. Meanwhile, they have obvious advantages, such as silence without sound, and short or small size, which make them more discreet and easily consumable in comparison to long videos that require longer time and larger bandwidth commitment [15]. Recent studies reveal that more than 23 million GIFs emerge everyday in Tumblr, the total number of the short GIFs in Sina or Weibo is over half a billion, and even more than 71% online articles contain short GIFs.

Today, it is common for users to post animated GIFs with relevant textual messages, to express opinions and sentiments on popular social media platforms, such as Facebook, Twitter, Tumblr, Instagram, WeChat and Microblog, etc. Fig. 1 presents several examples of the GIFs with text messages from different representative social media websites with over 100 million registered users including Facebook, Twitter, Tumblr and Instagram. While the GIFs with certain short sentences in textual language description have gained immense popularity, the

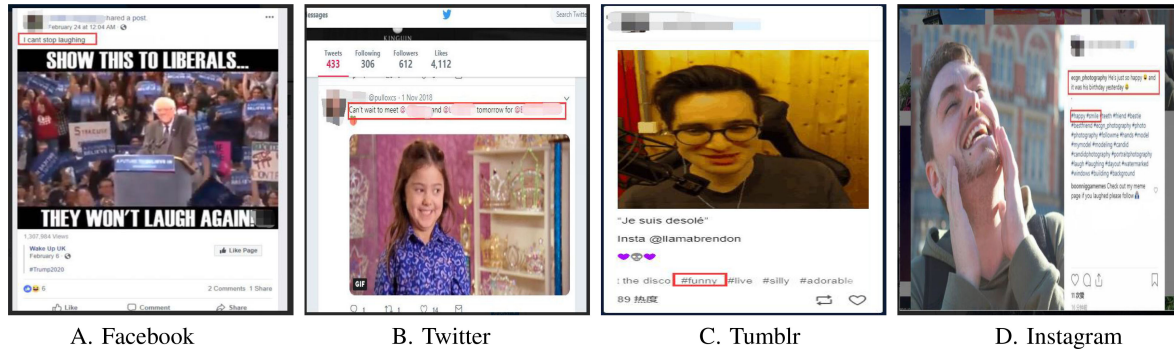


Fig. 1. Examples on daily animated GIFs with short text messages or annotations revealing certain sentiment from the most popular social media including Facebook, Twitter, Tumblr and Instagram, each possessing over 100 million registered users.

implicated sentiment in themselves has not been analyzed together from a computational perspective. In this work, we treat those textual messages as the annotation of the GIFs, which can be used to analyze the sentiment of short annotated GIFs. However, using these text alone is insufficient [5]. For example, in the Fig. 1(B), the user only wrote “Can’t wait to meet A and B tomorrow for something”, which can be considered as neutral sentiment. However, the given GIFs of a smiling girl shows the positive sentiment about the user’s opinion. Furthermore, given the small size and illustrative nature of short annotated GIFs, their low-level statistics will be very different from those of natural images or videos, their behaviour and meaning can be substantially different from that of text and images or videos.

To the best of our knowledge, there are few relevant research projects on analyzing the sentiments of the animated GIFs with textual annotations. The two main challenges in sentiment analysis for the short annotated videos are video understanding and the semantic gap between texts and videos [6]. Video understanding tries to extract the sentiment elements from the short annotated videos, but differs from image sentiment detection. The sequence of images or videos in the short GIFs also contains certain sentiment information, which requires to extract both the spatial and temporal features to learn the integrated sentiment content in the short GIF videos. Similarly, the semantic gap can be also interpreted as a hard problem of semantic comprehension. Without any semantic label measure, the machine can not learn the middle level sentiment semantic elements and their relationship from low level features. Lexicon-based approaches predict the overall sentiment orientation of textual messages based on the words that are annotated by the given polarity or polarity scores [8].

Contributions: In this paper, our contributions are listed to solve the aforementioned challenges as follows:

- We propose an effective and multi-modal sentiment recognition framework based on the visual-textual late fusion to resolve the problem of sentiment classification of short annotated GIFs in social media, which consists of short GIF videos and associated descriptive annotations.
- Our visual understanding model for animated GIF video with sentiment orientations can be constructed firstly by the 3D convolutional neural (C3D) network and VGG-16 network to extract the short-term spatial-temporal features

from the short GIF videos, then exploit a stacked ConvLSTM network to perceive and learn long-term spatial-temporal dependencies of the short annotated whole GIFs.

- We design a visual-textual sentiment score function (VT-SSF) to fuse the visual sentiment score extracted from the softmax layer in the ConvLSTM network and the textual sentiment score by the SentiWordNet3.0 model on the short textual annotations, while the suitable and critical threshold for the assumed sentiment richness is introduced to improve the validity of the fused sentiment score for the short annotated GIFs. A grid search technique is applied to learn the assumed model hyper-parameters.
- We implement robust sentiment recognition and conduct extensive experiments with quantitative and qualitative evaluations on three benchmark datasets on short annotated GIFs including T-GIF dataset, GSO-2016 dataset and Adjusted-GIFGIF dataset to verify the effectiveness.

The remainder of this manuscript is structured as follows. Section II presents an overview of related work on our sentiment analysis in multimedia. Section III reviews the whole annotated GIF sentiment recognition framework and the proposed methodology with visual-textual sentiment score function. Experimental results and analysis can be seen in Section IV. Section V draws conclusions and prospects of future work in sentiment analysis for the short annotated GIFs.

II. RELATED WORK

We give a categorized overview of previous works on sentiment analysis of the short annotated GIFs. There are basically two main methodologies in the perspective of computational modality from social multimedia.

A. Unimodal Sentiment Analysis

Unimodal sentiment systems can act as the primary building blocks for a well-performing multimodal framework. In this subsection, we describe the literature of unimodal affect analysis primarily focusing on visual and textual modalities.

1) *Visual Sentiment Analysis:* With the explosive growth of the multimedia datasets (such as image and video) in social media, visual sentiment analysis has become one of the hot topics

in multimedia analysis. According to the types of multimedia contents, visual sentiment analysis can be divided as follows.

Sentiment analysis in image: In visual sentiment analysis, most researchers focus on image sentiment and benefit from the large-scale datasets and the Graphical Processing Units (GPUs). They adopt Convolutional Neural Networks (CNN) as the foundation for special applications. Islam *et al.* built an image sentiment prediction model with CNNs and pre-trained their framework on manually labeled Flickr image dataset to further perform transfer learning [16]. Following works exploited the hyper-parameters learned from a very deep convolutional neural network to improve the effectiveness of image sentiment prediction on the basis of pre-training their model on Twitter image dataset [17]. While increasing the depth of the given neural network, some articles also attempted to adjust the structure of different layers in the CNN network to exploit the proposed framework in affective computing applications. You *et al.* designed a suitable CNN network to make use of noisy machine labeled data and exploited a progressive strategy to fine-tune the deep network to solve the problem of huge number of neurons and connections, while improving the performance on Twitter images by inducing domain transfer with a small number of manually labeled Twitter images [18]. Campos *et al.* added a fully connected layer to the CNN network which is specifically applied for image sentiment analysis [19]. Yang *et al.* connected the CNN network with each candidate region to compute sentiment scores and predict related emotions [20].

Sentiment analysis in video: For video sentiment analysis, especially the short annotated videos such as GIF videos, the researchers started their work from dataset collection and preparation. Li *et al.* collected a new video sentiment dataset, Tumblr GIF (T-GIF), with 100 K animated GIFs from Tumblr website and 120 K natural language descriptions obtained from crowd-sourcing [27]. Jou *et al.* compared the prediction results of color histogram, face expression, aesthetics and SB features in the short GIF videos to find the most useful video features to express meaningful emotions [28]. Based on Jou's work, Chen [29] introduced the short-term temporal sequence features into the related comparison, which were extracted with 3D Convolutional Neural Networks (C3D). Both of their works conducted experiments on GIFGIF dataset built by the MIT Media Lab which contains 6119 GIFs and 17 discrete emotion classes. A spatial-temporal visual mid-level ontology was proposed to construct a semantic tree model to label visual sentiments on video sentiment dataset built by themselves [30].

2) *Textual Sentiment Analysis:* Textual sentiment analysis is one of the most valuable research fields of natural language processing. According to the granularity of the descriptive text, the task of textual sentiment analysis can be divided into three levels: *chapter level*, *sentence level* and *word level*. Chapter level sentiment analysis tries to analyze the positive or negative sentiment conveyed by the whole article. Since textual sentence is composed of limited emotional words, sentiment analysis on the sentence level text generally classifies the sentence into various emotions, such as anger, fear, joy or sadness components of a sentence, then classifies the sentence according to the intensity of different sentiments. According to different ways of

constructing the word bank, word level sentiment analysis can be divided into three kinds: lexicon-based analysis method, network-based analysis method and corpus-based analysis method. With the enrichment of word bank and complexity of network structure, various database of sentiment scores at the word level has been established. Previous researchers started their work from the construction of sentiment word bank and the matching of target text and sentiment word bank. Hu *et al.* constructed a sentiment lexicon with the adjectives in the document and calculated the similarities between the words in the target document and the sentiment lexicon to evaluate the sentiment of the whole document [41]. Wang *et al.* introduced a latent semantic analysis algorithm in textual sentiment analysis, which decomposed the singular value of entry and document matrix. However, both of their works did not group semantically related aspect expressions together [42]. In contrast, the unsupervised knowledge-lean topic modeling approach has been shown to be effective in automatically identifying sentiment aspects and their representative words. Zhang *et al.* used Support Vector Machine (SVM) and an Extreme Learning Machine (ELM) to analyze the emotional tendency of the whole document [43]. Kim *et al.* was the first work which applied the single-layer CNN to textual sentiment analysis and achieved satisfactory results in multiple benchmark [44]. In recent years, more and more data-driven deep learning models like Recursive Neural Networks (RNN), Long-Short Term Memory (LSTM), have greatly improved the performance of textual sentiment analysis.

B. Multimodal Sentiment Analysis

Multimodal analysis has already created a lot of buzz in the field of affective computing. In this subsection, we discuss the approaches to solve the multimodal sentiment recognition problem. For a recent survey on multimodal sentiment analysis, please check the published review by Poria *et al.* in [11]. The convolutional multiple kernel learning was presented to enhance the performance of sentiment analysis and emotion recognition in [12]. The tensor fusion network model to solve the problem of multimodal sentiment analysis was proposed in [3] to learn both intra and inter modality dynamics in an end-to-end framework. The conversational memory network is proposed to train a conversational emotion recognition classifier using dyadic dialogue videos [4], which can be plugged into any dialogue system to generate empathetic responses.

To perceive the sentiment representation from the related multimodal features, some works take both the visual sentiment and the textual content into consideration to work with sentiment analysis. The multimodal features are generally based on the modification of the CNNs structure like increasing the depth of the networks or adjusting the full-connection layers. Image features and text information were combined for sentiment analysis in [21], [22]. The authors in [21] employed vector auto-regression for sentiment selection, while in [22], CNNs were employed for microblog sentiment analysis. Yu *et al.* trained a CNN network on the top of pre-trained word vectors for textual sentiment analysis and employed a Deep Convolutional Neural Network (DNN) with generalized dropout for visual

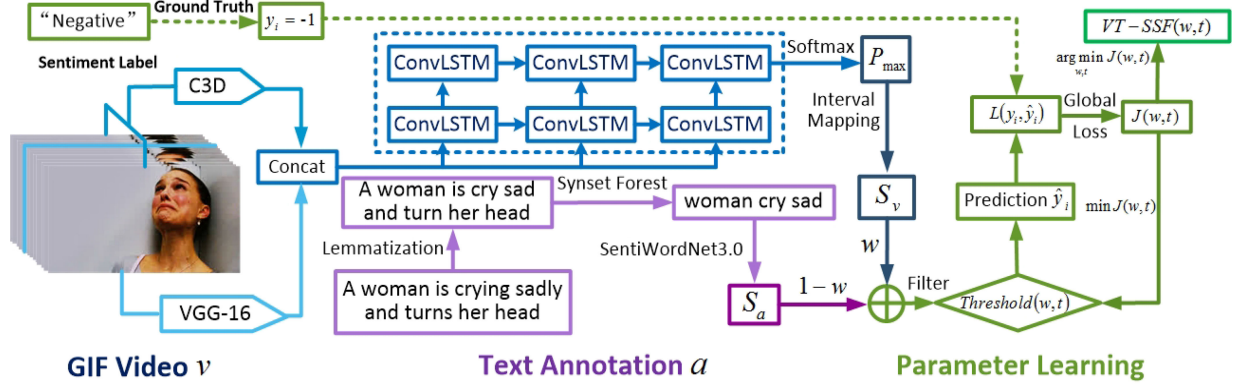


Fig. 2. The sentiment analysis pipeline for short annotated GIFs fused with visual-textual sentiment function. The blue boxes show the calculation of video sentiment score, the purple boxes represent the perception of textual sentiment score and the green boxes denote the model hyper-parameter learning.

sentiment analysis [23]. The mid-level entities or attributes are also employed for image sentiment analysis [24], [25]. In [24], a total of 102 scene attributes were extracted. In [25], a pre-defined 1200 entities with different emotions were employed instead. You *et al.* proposed an attention mechanism with LSTM and an auxiliary semantic learning model to construct a tree-structured LSTM network to analyze the performance of the visual-textual fusion model in visual sentiment prediction in [26].

Meanwhile, the multimodel information fusions are also considered in video sentiment analysis. Zhang *et al.* proposed a novel video blog (vlog) management model [31]. The presented management model took both the visual and textual information into consideration, but the proposed algorithm could not deal with large amount of data. More recently, both Adjective Noun Pairs (ANPs) and Verb Noun Pairs (VNPs) were considered for improving sentiment prediction. They also provided a GIF emotion dataset including 1874 GIFs and 1274 SentiPairs. We give examples on this SentiPairs as: “The lovely girl happily takes a cup of coffee and wants to turn around and leave. Then a handsome boy greets the girl and gives a bunch of flowers.”. Here, the words “lovely girl” and “handsome boy” are Adjective Noun Pairs, while the symbol “takes coffee”, “greets girl” and “gives flowers” can be taken as Verb Noun Pairs. However, the assumed textual annotations from co-occurring descriptions or image tags tend in a broad sense to be simple or short and correlative for the annotated GIFs video. For example, they can be meaningful to other available text sources, and comprised of the co-occurring and short sentences, phrases or tags, more than GIFs captions in the strict sense.

III. PROPOSED METHODOLOGY

Sentiment analysis for short annotated GIFs is still in its early stage and more effective methods are needed to bridge the semantic gap in video understanding and text perception. We propose an effective sentiment recognition scheme with visual-textual sentiment score to combine the visual sentiment score with its accompanied descriptive annotations. The proposed framework is shown in Fig. 2. Following [32], a tuple (v, a, y_i) is used as the input of our proposed sentiment recognition,

which consists of short GIF video v , textual annotation a and ground truth of sentiment label y_i .

A. Visual Sentiment Score Calculation

To obtain the visual sentiment score from the short GIF video component, we utilize the 3D Convolutional Neural Networks (C3D) [33] to extract short-term spatial-temporal features, and VGG-16 network [34] to obtain the visual feature of each frame in the GIF video. Then, a new ConvLSTM network [35] is employed to learn the whole long-term spatio-temporal features based on the concatenated visual features and output the sentiment probability of the short GIF video.

1) *Visual Feature Extraction*: As mentioned above, we extract both the *frame-level* and *sequence-level* visual features. To obtain the image features, we exploit the VGG-16 neural network [34] (indicated as green dashed box in Fig. 3). For the sequence features, the C3D network [33] is applied to perceive visual representation of the given short GIF video. The C3D network for a visual feature representation uses 3D convolution and 3D pooling to model the short-term temporal information in the given videos. The C3D network achieves competing results on various video analysis tasks such as action recognition, scene classification and object recognition. Previous work in [29] also employed the C3D network for sentiment prediction. The architecture of the given C3D model network is shown in the blue dashed box in Fig. 3. The kernel size of each Conv3D layer is $3 \times 3 \times 3$ and the sizes of the stride and padding of each Conv3D layer are $1 \times 1 \times 1$. The kernel size of each pooling layer is $2 \times 2 \times 2$ with a stride of 2 and a padding of 2, except that the first pooling layer with $1 \times 2 \times 2$ kernel and stride size. To obtain the early fused representations of the visual modality, we concatenate the sequence-level visual features both from the C3D network and VGG-16 network.

2) *Visual Sentiment Perception*: The above concatenated visual features can only capture short term visual context, which may be insufficient for understanding the sentiment expressed in the whole GIF video. Therefore, we utilize Convolutional Long Short-Term Memory (ConvLSTM) [36] network to further process the concatenated visual features. In such a way, we expect

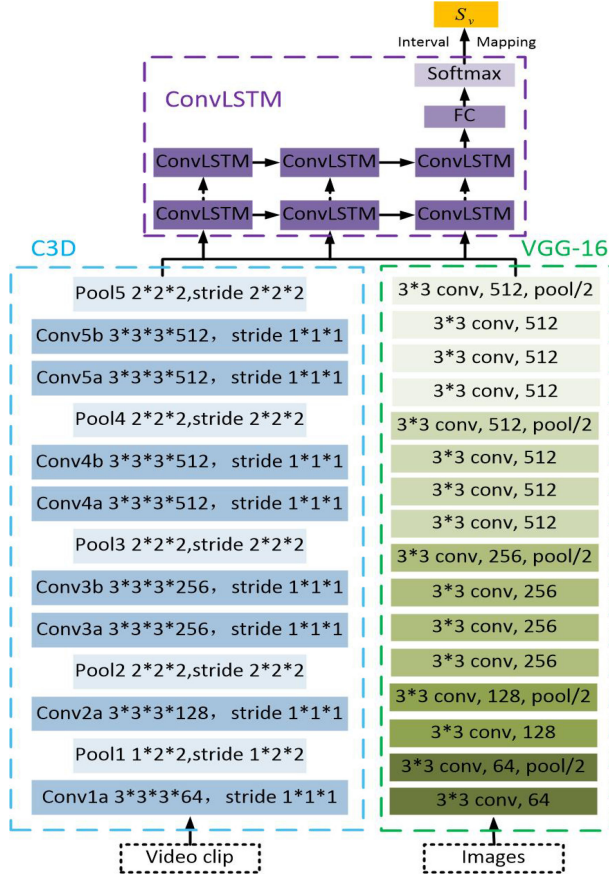


Fig. 3. The architecture for the proposed visual sentiment score. The blue dashed box represents the C3D network, the green dashed box represents the VGG-16 network construction and the purple dashed box denotes the ConvLSTM network. Here, the visual representations obtained from both C3D and VGG-16 network are concatenated before being passed to the given ConvLSTM network. The output of the given whole framework is the proposed visual sentiment score for the given short video component S_v .

our model to incorporate the long-term context for video-level sentiment analysis. The ConvLSTM network can be formulated.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o) \quad (3)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (4)$$

$$H_t = o_t \circ \tanh(c_t) \quad (5)$$

where the symbol '*' represents the convolution operator and the symbol 'o' denotes the Hadamard product.

As illustrated in Fig. 3, a two-level ConvLSTM network is deployed in the visual sentiment score stage. The channels of convolutional layers in our two-level ConvLSTM network are 256 and 384. We adjust the paddings of the layers in ConvLSTM to make the two-level layers produce features with the same spatial size. Until now, we can compute the sentiment probability by using a softmax classifier based on the given features generated by the two-level ConvLSTM network. Next, we map the

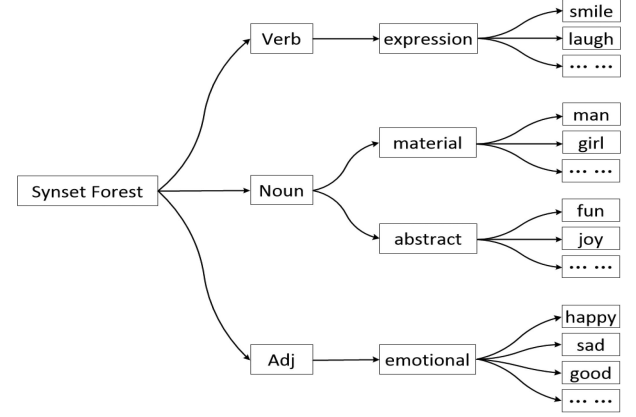


Fig. 4. A rough overview of the Synset Forest in textual sentiment perception.

probabilities into the interval of the visual sentiment score. To obtain visual sentiment score of short GIFs, we first compute the maximum probability P_{\max} :

$$P_{\max} = \max[p_{-1}, p_0, p_1] \quad (6)$$

where p_{-1} , p_0 and p_1 denote the probabilities of negative, neutral and positive sentiment respectively. Then, we remap P_{\max} (belonging to $[\frac{1}{3}, 1]$) to the visual sentiment score as:

$$S_v = \begin{cases} -\frac{P_{\max}-1/3}{1-2/3} & P_{\max} = p_{-1} \\ 0 & P_{\max} = p_0 \\ \frac{P_{\max}-1/3}{1-2/3} & P_{\max} = p_1 \end{cases} \quad (7)$$

where S_v represents the visual sentiment score (belonging to $[-1, 1]$) from visual modality of the annotated GIF video.

B. Textual Sentiment Score Calculation

Here, we discuss how we compute the textual sentiment score of the textual annotations accompanied with the GIF video. Our work applies the proposed SentiWordNet3.0 model [37] on the textual annotations after the Synset Forest procedure.

1) *Key Information Extraction*: We adopt a robust and practical Lemmatization operation [38] to preprocess the given text annotations. For example, the sentiment score of the word "smile" in the SentiWordNet3.0 model is equal to 0.3, while the sentiment scores of "smile" and "smiling" are both more psychologically close to 1. Without the Lemmatization, we may misidentify the positive sentiment expressed by "smiling".

Then, we build Synset Forest to extract a set of meaningful words from each sentence or short annotation in the textual annotation with respect to the given annotated GIF video. To replace the input of the SentiWordNet3.0 model with the sets of meaningful words in the given text annotations other than the whole sentences or text annotations, there exist two main reasons listed as follows: The set of significant words can be seen as a uniform entity and generalization with respect to Adjective Noun Pair (ANP) [25] and Verb Noun Pair (VNP) [6], but the sets of the meaningful words exploited in our proposed framework could contain the union of ANP [25] and VNP [6], so that the set of the

TABLE I
SOME REPRESENTATIVE EXAMPLES WITH RESPECT TO THE MISJUDGMENT
WORDS BY THE SENTIWORDNET3.0 MODEL

Word	behind	right	long
Sentiment classification by the SentiWordNet3.0 model /Textual sentiment score	Negative /-0.4	Positive /0.2857	Negative /-0.05
Ground Truth w.r.t Sentiment classification	Neutral	Neutral	Neutral

assumed meaningful words can be more general and integrated in describing the corresponding short annotated GIF videos.

In addition, the attention mechanism on the meaningful words from the textual annotations on the given annotated GIF video can avoid effectively some misjudgments of the textual sentiment obtained from the SentiWordNet3.0 model. Table I shows some representative examples of misjudgments cases with respect to the meaningful words.

2) *Textual Sentiment Perception*: We exploit the SentiWordNet3.0 model [37] to evaluate the textual sentiment score of the accompanied annotations. The sentiment score from the short textual annotation can be computed as follows:

$$S_a = \sum_{i=1}^M \text{SentiWordNet}(W_i) \quad (8)$$

where $\text{SentiWordNet}(W_i)$ denotes the sentiment score from the set of the i -th meaningful words extracted from the annotations using SentiWordNet3.0 model, M represents the number of the meaningful words from the given textual annotations.

C. Annotated GIF Video Sentiment Classification

1) *Visual-Textual Sentiment Score Fusion*: Given the visual sentiment score (S_v) and the textual sentiment score of the annotations (S_a), we compute the fused sentiment score S_{va} :

$$S_{va}(w) = w * S_v + (1 - w) * S_a \quad (9)$$

where the coefficient $w \in [0, 1]$ is the balance factor between S_v and S_a . The visual-textual sentiment scores fused with a linear combination are the important evidence criteria for judging sentiment dispositions from socially motivated clues.

Next, we utilize a just noticeable difference threshold $t \in [0, 1]$ to adjust the range of the sentiment score as follows:

$$SR(w, t) = \max(0, |S_{va}(w)| - t) \quad (10)$$

where we use the symbol SR to represent the *Sentiment Richness*, which reflects the social sentiment evidence more than just noticeable difference from observation behavior. In such a way, the sentiment score is insensitive to the noises introduced by either visual sentiment score or textual sentiment score.

2) *Model Parameter Learning*: To estimate the balance parameter w and the difference threshold t , we employ exhaustive grid search to learn them adaptively from a specified subset in the hyper-parameter space in a brute force framework to avoid the arbitrary and capricious behaviour. The global objective loss

for hyper-parameter optimization is defined as:

$$J(w, t) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{y}_i) \quad (11)$$

where N denotes the total number of the training samples and $L(y_i, \hat{y}_i)$ is the loss function on the i -th sample:

$$L(y_i, \hat{y}_i) = \begin{cases} 0, & y_i = \hat{y}_i \\ 1, & y_i \neq \hat{y}_i \end{cases} \quad (12)$$

We discretize both $w \in [0, 1]$ and $t \in [0, 1]$ with the same step size $\delta_w = \delta_t = 10^{-3}$. The optimal value for w and t can be searched from this subset space by minimizing the $J(w, t)$.

3) *Annotated GIF Sentiment Inference*: From $S_{va}(w)$ in Equation (9) and $SR(w, t)$ in Equation (10), final resulting sentiment can be predicted for the short annotated GIFs with certain descriptive and textual language classified as:

$$\hat{y}_i = \begin{cases} 1, & SR(w, t) > 0 \text{ and } S_{va}(w) > \epsilon \\ -1, & SR(w, t) > 0 \text{ and } S_{va}(w) < -\epsilon \\ 0, & SR(w, t) = 0, \text{ otherwise} \end{cases} \quad (13)$$

where ϵ is the positive relaxation coefficient (we set it to 10^{-5}). If the sentiment richness $SR(w, t)$ value is greater than zero, there exist more obvious sentiment inclination or evidence more than just noticeable difference t from observation behavior for the positive or negative one. In the meantime, if the visual-textual sentiment score $S_{va}(w)$ is greater than the relaxation coefficient ϵ , our model outputs the positive sentiment. If the sentiment richness is still positive, but we have a negative sentiment score $S_{va}(w)$, our model outputs the negative sentiment. Otherwise, we expect a neutral sentiment.

IV. EXPERIMENTS AND DISCUSSIONS

In this section, we evaluate our proposed sentiment recognition scheme for the short annotated GIFs with visual-textual fusion in terms of extensive experiments on the Tumblr GIF (T-GIF) dataset, GIF Sentiment Ontology (GSO-2016) dataset and an adjusted GIFGIF (Adjusted-GIFGIF) dataset.

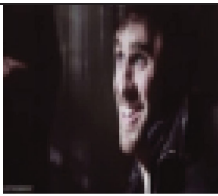



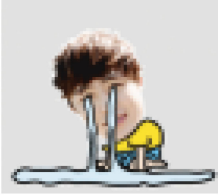




A. Experimental Datasets

1) *T-GIF Dataset*: The T-GIF dataset contains 100K animated GIFs collected from Tumblr website, and 120K natural language sentences annotated via crowdsourcing [32]. We recruited 10 workers who are undergraduate students in our university to give an overall sentiment judgments (positive, negative and neutral) for all the short GIF videos used in the given experiment. For each GIF video, we obtained its sentiment label by using majority voting (Table II shows some representative examples). We labeled 6950 short GIF videos with text annotations including 2320 positive instances, 2310 negative instances and 2320 neutral instances.

2) *GSO-2016 Dataset*: The GSO-2016 dataset was crawled from one of most popular microblog providers [6]. For each GIFs, the ANP and VNP in the SentiPair attribute were chosen as the corresponding textual annotations. The GSO-2016

TABLE II

EXAMPLES ON LABELED T-GIF DATASET, GSO-2016 DATASET AND ADJUSTED-GIFGIF DATASET ACCOMPANIED WITH SHORT TEXTUAL ANNOTATIONS. THE “1”, “-1” AND “0” MEAN THE POSITIVE, NEGATIVE AND NEUTRAL SENTIMENT RESPECTIVELY

T-GIF		GSO-2016		Adjusted-GIFGIF		Label
GIF Thumbnail	Sentence	GIF Thumbnail	SentiPair	GIF Thumbnail	Tag	
	man is looking at something and smiling		funny,man; smile,man		happy; Kristen; siig; celebration; approved	1
	a woman in a black shirt is crying		cry,man; flow,tear		angry; Alan; rickman; flip; furious	-1
	a man is performing skateboarding tricks on ramps		one,cat; dance,cat		man; jacket; play saxophone	0

dataset also has three sentiment labels, e.g., positive, negative and neutral. There are 1874 GIF videos in GSO-2016 dataset labeled with the given SentiPairs. More specifically, we have 1111 positive, 164 negative and 599 neutral samples.

3) *Adjusted-GIFGIF Dataset*: The GIFGIF dataset was built by the MIT media Lab [39]. This dataset classifies 6119 animated GIFs into 17 emotion categories e.g., amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, happiness, pleasure, pride, relief, sadness, satisfaction, shame and surprise. These annotated GIF videos are published publicly on the Internet for sentiment voting. In total, 3,221,504 votes had been collected by April 25, 2018. To use this GIFGIF dataset to evaluate our model, we make several changes. Firstly, we classified the 17 emotions into the positive and negative sentiments using SentiWordNet3.0 model in [37]. The positive sentiment contains four positive emotions: excitement, happiness, satisfaction and surprise. The negative sentiment includes anger, fear, disgust and sadness. Surprise is an example of a complex emotion which can be expressed in both positive and negative sentiment depending on the context. Secondly, we chose the top 300 GIFs video for each of emotion category to form a dataset with 2400 ($=300 \times (4 + 4)$) short GIFs, which had been split into 1200 positive GIFs and 1200 negative GIFs. Thirdly, 1200 neutral GIFs were chosen from the T-GIF dataset as additional neutral samples in this new dataset. In total, we have 3600 ($=2400 + 1200$) annotated GIF videos. We call this Adjusted-GIFGIF dataset.

Table III shows the differences of three datasets in which the given visual contents focus on “real-life scene”, “cartoon” and “movie”, respectively. To preview visually the GIF contents with short annotations in late display, the GIF thumbnail image is created by using the powerful command line *ffmpeg* package and

TABLE III
DIFFERENCES AMONG THE THREE EXPERIMENTAL DATASETS

Modality \ Dataset	T-GIF	GSO-2016	Adjusted-GIFGIF
Visual content	real life scene	cartoon	movie
Visual source	web scraping	man made	web scraping
Textual content	sentences	two phrases	several tags
Textual source	man made	man made	web scraping

PHP’s ability to execute server commands through *shell_exec* from the given GIF source file. In addition, the textual annotations are either sentences or word phrases.

B. Experimental Results and Analysis

The proposed framework is implemented by using the Tensorflow and Tensorlayer platforms [35]. All our experiments are carried out on a Linux server with Intel(R) Core(TM) i7-4790K CPU@4.00 GHz and 32 GB RAM memory as well as one NVIDIA TITAN BLACK GPU. For all the experiments, we split the datasets into 80% as training and 20% as testing. For all the GIFs used in our experiments, we can treat each 16 frames as certain temporal fragment. At the same time, two rules are adopted in this scheme: 1) If the length of a GIFs video is less than 16 frames, the first and the last frame will be repeated until we have a 16 frame segment. 2) If the length of a GIFs video is longer than 64 frames, we sample the frame with a step size of two and then generate the segments.

We resize all the images in the given short GIF videos to 112*112 and train the C3D network and ConvLSTM network with a batch size of 16 and the learning rate is initialized to be 0.01. We finetune the VGG-16 network, C3D network and ConvLSTM network with the given pre-trained models on the T-GIF dataset, GSO-2016 dataset and Adjusted-GIFGIF dataset,

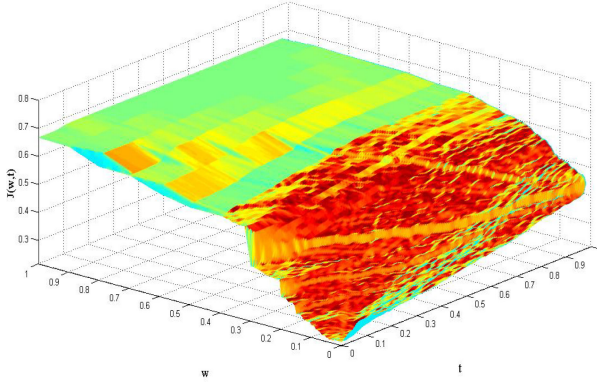


Fig. 5. The road map of the grid search procedure to sweep the given model parameters in the global loss function on the T-GIF dataset.

TABLE IV
PERFORMANCES OF THE PROPOSED SENTIMENT RECOGNITION USING VISUAL-TEXTUAL FUSION COMPARED WITH THAT OF USING SOME PARTS OF THE RELATED MODEL COMPONENTS ON THE T-GIF DATASET

Model		Precision	Recall	F1	Accuracy
Visual Only	VGG-16	0.4708	0.4692	0.4702	0.4688
	C3D	0.4911	0.4897	0.4868	0.4845
	VGG-16 + C3D + ConvLSTM	0.5346	0.5313	0.5329	0.5315
Textual Only	Raw Sentence	0.6462	0.6121	0.6287	0.6112
	Lemmatization + Synset Forset + SentiWordNet3.0	0.7588	0.7612	0.7600	0.7630
Visual + Textual	Proposed Fusion	0.7819	0.7826	0.7822	0.7839

respectively. We conduct the Lemmatization preprocessing step by using the Natural Language Toolkit (NLTK) [38]. We also manually fix some missing cases, where the words have rich meanings, such as the word “lying”. After another preprocessing using the Synset Forest, we send the meaningful words to the SentiWordNet 3.0 model [37] and obtain the sentiment scores of the textual annotations.

1) *Results on T-GIF Dataset:* In Fig. 5, we visualize the objective function $J(w, t)$ with different w and t values in our grid search. The loss objective function slopes stably down and the optimal model parameters are $w = 0.068$ and $t = 0.028$ on the T-GIF dataset. We report Precision, Recall, F1 Score and Accuracy to compare the performance of different models. All hyper-parameters are the same for different variants of our model. Table IV shows the experimental results on the T-GIF dataset. The sentiment recognition with only visual input (VGG-16 + C3D + ConvLSTM) achieves better performance compared with that of using single network model such as VGG-16 network and C3D network. The results also suggest that visual sentiment analysis on the GIF videos can benefit from the sequence features. Compared with single-frame model (VGG-16 only), the short-term context (C3D) model can improve the accuracy from 46.88% to 48.45%. With long-term context (VGG-16 + C3D + ConvLSTM), the accuracy can be further improved to 53.15%. The sentiment recognition with only textual annotations using Lemmatization and Synset Forest can increase by more than 10%, compared with that of using raw sentence in all four quantitative measures, and especially the related increment can even reach 15.18% with respect to the given accuracy. These results indicate the Lemmatization and Synset

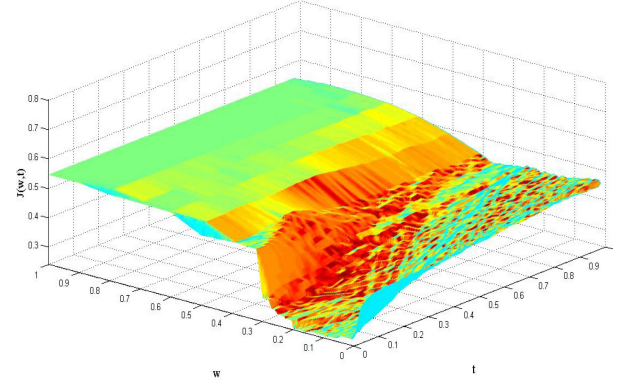


Fig. 6. The road map of the grid search to sweep the optimal model parameters with respect to the global loss function on the GSO-2016 dataset.

TABLE V
PERFORMANCES OF THE PROPOSED SENTIMENT RECOGNITION FUSED WITH VISUAL-TEXTUAL SENTIMENT SCORE COMPARED WITH THAT OF USING SOME PARTS OF THE RELATED MODELS ON GSO-2016 DATASET

Model		Precision	Recall	F1	Accuracy
Visual Only	VGG-16	0.4439	0.4402	0.4420	0.4613
	C3D	0.4509	0.4487	0.4498	0.4776
	VGG-16 + C3D + ConvLSTM	0.4595	0.4510	0.4552	0.5163
Textual Only	Raw Sentence	0.5283	0.5063	0.5171	0.5658
	Lemmatization + Synset Forset + SentiWordNet3.0	0.7474	0.6900	0.7176	0.7366
Visual + Textual	Proposed Fusion	0.7208	0.6848	0.7023	0.7513

Forest are necessary for textual sentiment analysis on GIF video annotations. The joint visual and textual model achieves the best performance compared with visual-only or textual-only models. This result also coincides with the findings in [22], [26], [40].

2) *Results on GSO-2016 Dataset:* We also show the objective function $J(w, t)$ on this dataset in Fig. 6. It has similar pattern with Fig. 5 when w is larger than 0.5. However, the shape of $J(w, t)$ becomes quite different with a smaller w . We believe that the differences shown in Table III can lead to the differences of $J(w, t)$ on these two datasets. The optimal model parameters are $w = 0.039$ and $t = 0.151$ on the GSO-2016 dataset. Table V summarizes the experimental results. Similarly, we can see that the presented technique with the Lemmatization and Synset Forest operations and the C3D network and ConvLSTM models have improved the performance compared with that of using only textual sentiment and only visual sentiment. We find that 16.53% of the SentiPairs from the textual annotations in the GSO-2016 dataset contain significant words which are not positively correlated with the sentiment of the GIFs, but have strong influence on the results of classification. After removing those uncorrelated words with the Lemmatization and Synset Forest, the sentiment classification accuracy can be increased from 56.58% to 73.66%. We also achieve more than 17% performance increase in terms of Precision, Recall and F1.

Almost all of the related sentiment models have slightly poor performance on the visual sentiment classification, and both of the accuracies are lower than 50% in using the VGG-16 (46.13%) network and the C3D (47.76%) network. This may be caused by the small size of the GSO 2016 dataset with only 1874 GIF videos and the highly unbalanced distributions of different

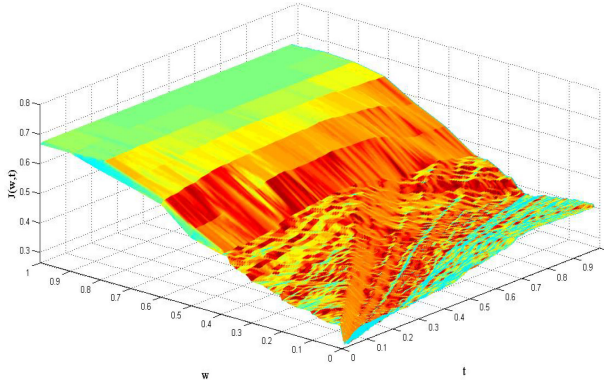


Fig. 7. The road map of the grid search to sweep the optimal model parameters w.r.t the global loss function on Adjusted-GIFGIF dataset.

TABLE VI
PERFORMANCE OF THE PROPOSED SENTIMENT RECOGNITION JOINTLY WITH VISUAL-TEXTUAL SENTIMENT SCORE AGAINST THAT OF USING SOME PARTS OF THE RELATED MODELS ON ADJUSTED-GIFGIF DATASET

Model		Precision	Recall	F1	Accuracy
Visual Only	VGG-16	0.5673	0.5549	0.5610	0.5353
	C3D	0.6647	0.6395	0.6518	0.5621
	VGG-16 + C3D + ConvLSTM	0.6705	0.6524	0.6613	0.5847
Textual Only	Raw Sentence	0.5673	0.5588	0.5630	0.5583
	Lemmatization + Synset Forset + SentiWordNet3.0	0.7048	0.6610	0.6822	0.6597
	Proposed Fusion	0.7441	0.7404	0.7422	0.7403

classes, which leads to the obtained threshold t being pretty high. Although we try to make this dataset balanced by using an oversampling technique, the experimental results are below expectation. If the scale of total number of the GSO-2016 dataset further increase, the critical threshold value t would be decreased predictively. Although there exists some negative side effects, final whole sentiment recognition using the given visual-textual fusion model could be slightly inferior (about 2% decline) and similar or equivalent to that of using the textual sentiment classification in terms of the precision, recall and F1 measures, while being superior to that of using only textual annotations (1.47% increase) and that of using only visual sentiment (23.50% rise) in terms of the accuracy for the annotated GIFs.

3) *Results on Adjusted-GIFGIF Dataset:* Again, we demonstrate the objective function $J(w, t)$ on this dataset in Fig. 7. The function map is quite different from both Fig. 5 and Fig. 6. This could be due to the construction of this dataset, which includes samples from T-GIF dataset as our neutral samples. The optimal model parameters $w = 0.086$ and $t = 0.023$ are obtained by the exhaustive grid search technique.

Table VI summarizes the experimental results of our proposed model on the Adjusted-GIFGIF dataset. Again, the results suggest that visual-textual fusion model have the best performance among all the baselines. Compared with the T-GIF dataset and GSO-2016 dataset, a great majority of the textual annotations of the GIFGIF dataset were collected from the original websites, which are weakly correlated with the content of the given short GIFs video. This could lead to the accuracy of the sentiment classification with only textual annotations after Lemmatization and Synset Forest reaching only 65.97%. This may explain why the performance of the given models using only textual

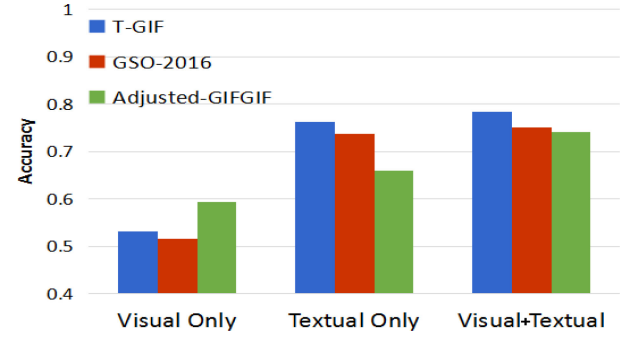


Fig. 8. The accuracies of the sentiment classification on the T-GIF dataset, GSO-2016 dataset and Adjusted-GIFGIF dataset with only visual sentiment score, only textual sentiment score and the visual-textual sentiment score.





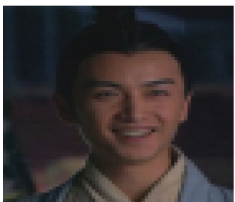

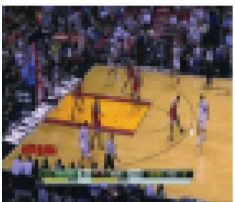

annotations on this dataset is worse than that of the assumed modes on the other two datasets. However, different with the first two datasets labeled with no more than 10 workers, the ground truth of the sentiment classifications on GIFGIF dataset were voted by numerous internet users in a crowdsourcing platform and each of the short annotated GIF video in Adjusted-GIFGIF ranks in top 300 in their emotions. We think that this dataset with higher quality may enhance the performance of the proposed sentiment classification with only visual sentiment in terms of the accuracy to achieve 58.47% and the precision to reach 67.05%, the recall to hit 65.24% and the F1 measure to approach 66.13%, which are the highest among the three short annotated GIF datasets. In view of model performances, the accuracy of the proposed sentiment recognition technique with the textual and visual sentiment score can increase by over 8% than that of using only single modal sentiment score and perform excellently in other quantitative evaluations.

4) *Results Analysis:* We compared the sentiment classification with different types of multimedia contents on the three GIF datasets in Fig. 8 in terms of the accuracy while keeping the same experiment setting. From the exhibited histogram in Fig. 8, we can conclude as follows. The accuracy of visual sentiment classification can be dependent on the quantity and distribution of training samples. This may be the main reasons why the GSO-2016 dataset with unbalanced distribution obtains the lowest accuracy while the Adjusted-GIFGIF performs best in visual sentiment classification. At the same time, we believe that the quantitative performance of the related sentiment classification could be highly related to the reliability of the way in which ground truth labels are collected. The comparison between the contributions of using “Textual Only” and “Visual + Textual” parts in Fig. 8 shows that the accuracy of textual sentiment classification could be the determining factor in the visual-textual fusion model, since the textual sentiment factors make dominating contribution and to certain extents this textual component can predominate and limit the floor of the accuracy of final classification using visual-textual fusion. The final accuracy can be certainly higher than that of the textual sentiment.

The proposed annotated GIF video sentiment technique can increase the performance of the final results by that of using visual-textual fusion model on all the three short annotated GIF video datasets, such as 2.09% on T-GIF dataset, 1.47% on GSO-2016 dataset and 8.06% on Adjusted-GIFGIF dataset

TABLE VII

EXPERIMENTS ON SOME EXAMPLES W.R.T THE ANNOTATED GIF VIDEOS ON THE T-GIF DATASET AND ADJUSTED-GIFGIF DATASET, THE WORDS IN RED MEAN THE INCORRECTLY CLASSIFIED RESULTS BY THE PROPOSED SENTIMENT RECOGNITION APPROACH

Dataset	T-GIF			Adjusted-GIFGIF
GIF Thumbnail				
Annotation	a man is laughing and singing with a mic.	a boy is beating a guy and the guy is trembling.	a person is playing with a small reptile on the sand.	happy; excited; minions applause; cheering
Sentiment (Ground Truth)	Positive	Negative	Neutral	Positive
Sentiment (Visual Only)	Negative	Negative	Negative	Negative
Sentiment (Textual Only)	Positive	Neutral	Positive	Positive
Sentiment (V-T Fusion)	Positive	Negative	Neutral	Positive
Dataset	T-GIF			Adjusted-GIFGIF
GIF Thumbnail				
Annotation	cool man smile man	old woman wave hand	shoot basketball jump high	Tv; movies gordon ramsay; hells kitchen;
Sentiment (Ground Truth)	Positive	Neutral	Neutral	Negative
Sentiment (Visual Only)	Negative	Neutral	Negative	Negative
Sentiment (Textual Only)	Positive	Positive	Positive	Neutral
Sentiment (V-T Fusion)	Positive	Neutral	Neutral	Negative

respectively, compared with the results of using only textual models after the Lemmatization and Synset Forest. These improvements in quantitative evaluation performances have the same or similar trendy with the rankings in “Visual Only” part in Figure 8). This reflects that the introduction with visual feature representations for visual sentiment could achieve more excellent performance than the presented sentiment recognition scheme without visual contribution, and the textual effect for sentiment classification may determine the ceiling of final accuracy of the proposed sentiment recognition.

C. Case Studies

For further researches in the sentiment classification with visual-textual sentiment score function and the effectiveness of *sentiment richness*, which can be calculated from the formula of sentiment richness, we show some cases gathered from those three datasets in Table VII and VIII. Table VII shows that when one kind of single model in visual or textual component results in the wrong results of the sentiment classification, another kind of multi model will achieve certain correct results of the sentiment recognition, so that the effective result of final classification can reverse the contribution of the sentiment recognition by the visual-textual sentiment score function. There even exist

certain phenomena that both the visual model and textual model may obtain incorrect result, but the final recognition result can be calculated to be right and corrected by the whole sentiment classification in terms of the offset between positive score and negative score in Table VII.

Table VIII shows four cases of sentiment richness with respect to positive sentiment and negative sentiment respectively. In terms of visual contents in the given short annotated GIFs, different levels of sentiment can be achieved with the presence of facial expression, while the corresponding textual sentiment progress can be shown in the change of the meaningful words. For example, from “smile” to “laugh”, from “annoyed” to “angry”, the richest sentiments of the textual annotation basically contain several types of significant sentiment words such as the fourth GIF thumbnail in positive and negative conditions on the given short annotated GIF videos.

D. Computational Efficiency Analysis

Table IX shows the computational time and model network with the numbers of network parameters and multiadds of computation (Flops) in the inference stage by the proposed sentiment recognition approach fused with the assumed visual-textual sentiment fusion score with the visual part, textual part and

TABLE VIII
EXPERIMENTS OF SOME EXAMPLES WITH RESPECT TO THE GIVEN SHORT GIF VIDEO WITH TEXTUAL ANNOTATIONS AND THE CALCULATED SENTIMENT RICHNESS ON THE T-GIF DATASET, GSO-2016 DATASET AND ADJUSTED-GIFGIF DATASET






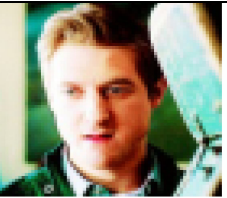

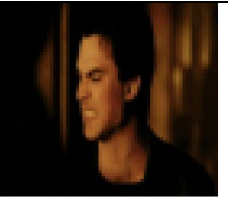
Label	Positive Sentiment			
Dataset	T-GIF	Adjust-GIFGIF	T-GIF	T-GIF
GIF Thumbnail				
Annotation	an Asian man is talking in front of a microphone and smiling .	reaction;Disney;animation; excited ; funny pics; princess and the frog	a woman is talking with beautiful smile .	the two men are smiling and laughing .
Sentiment Richness	0.1515	0.3251	0.5688	0.9715
Label	Negative Sentiment			
Dataset	Adjust-GIFGIF	T-GIF	Adjust-GIFGIF	T-GIF
GIF Thumbnail				
Annotation	black and white; Halloween; Hocus Pocus; witch	a man in a green jacket looking annoyed .	angry ; Tom and Jerry	a man is showing a very disgusted look and turns away in anger .
Sentiment Richness	0.0953	0.2675	0.5698	0.8247

TABLE IX
RUNNING TIME AND MODEL NETWORK BY THE SENTIMENT RECOGNITION WITH DIFFERENT COMPONENTS IN INFERENCE STAGE ON THREE DATASETS

Dataset	T-GIF		GSO-2016		Adjusted-GIFGIF		Model Network (Million/units)	
Running time (s)	Total	Average	Total	Average	Total	Average	Parameters (/Numbers)	MultAdds (/Flops)
VGG-16	121	0.087	53	0.080	73	0.091	14.7	3837
ConvLSTM	125	0.090	52	0.078	74	0.095	63.7	20537
C3D + ConvLSTM	187	0.135	78	0.117	110	0.143	95.6	30805
Visual	308	0.242	131	0.197	183	0.234	110.3	34642
Textual	3.8	—	3.8	—	3.8	—	—	—
Visual + Textual	311.8	—	134.8	—	186.8	—	110.3	34642

the overall model, including total testing time and average testing time of each of short annotated GIFs in the three datasets such as T-GIF dataset, GSO-2016 dataset and Adjusted-GIFGIF dataset. It is noted that, the number of testing samples in T-GIF, GSO-2016 and Adjusted-GIFGIF is 1390, 665 and 720 respectively. It is obvious that, the computational time costs shown in Table IX are proportional to the size or number of the related testing samples. In terms of the average time in different components, the computational time in the C3D network component increased by 49.7% compared with that of the VGG-16 network model, while the calculation time in the whole visual sentiment analysis increased by 11.07% compared with that of the C3D network model. The reason is that the dimension of the visual feature extracted by the C3D network is higher than that of the feature extracted from the 2D CNN network, while the given ConvLSTM model can just enhance the understanding of the obtained visual features. At the same time, the average time of the related models is the lowest in the GSO-2016 dataset, which could be caused by the oversampling of the neural and negative samples in the GSO-2016 dataset. Moreover, the short GIFs in the GSO-2016 dataset mainly consist of cartoons with

relatively simple visual features compared with those of T-GIF dataset and Adjusted-GIFGIF dataset. Finally, the improvement of the proposed sentiment recognition approach in terms of four quantitative measures can be attributed to the fusion of visual and textual sentiment score. To some extent, it is valuable to enhance the performance of the proposed sentiment recognition approach at the cost of time.

Additionally, the computational times in the training procedure will be directly proportional to the number of the training samples and inversely proportional to the intervals of model parameters δ_w and δ_t to be learnt in the grid searching stage.

E. Comparison With Similar Works

To the best of our knowledge, the researches on sentiment analysis for the short annotated animated GIFs with textual annotations are still in the beginning stages. And currently there exist few previous work to resolve simultaneously the problem of sentiment recognition for the short annotated GIFs with textual description from the visual and textual modalities. Hence, in perspective of the modality of similar and implicated sentiment,

TABLE X

PERFORMANCE OF THE PROPOSED SENTIMENT ANALYSIS APPROACH AND OTHER RELATED WORKS ON T-GIF DATASET AND GSO-2016 DATASET

Dataset	Method	Precision	Recall	F1	Accuracy
T-GIF	DNN + Word2Vec [23]	0.7143	0.7189	0.7166	0.7212
	Our proposed approach	0.7819	0.7826	0.7822	0.7839
	ANP + VNP [6]	0.7009	0.6508	0.6749	0.7140
GSO-2016	Our proposed approach	0.7208	0.6848	0.7023	0.7513

we make some performance comparison between our proposed whole approach and two public related works [6], [23] on the T-GIF and GSO-2016 dataset shown in Table X. The “DNN + Word2Vec” method in [23] standing at static imagery and text modality of view, was designed to classify the sentiment of the images with long sentence labels. As we can be seen in the T-GIF dataset in Table X, our proposed technique with dynamic video and static imagery and text modalities can obtain better performance in terms of precision, recall, F1 measure and accuracy in comparison to the “DNN + Word2Vec” method in [23]. This may benefit from the superiority of particularly 3D CNN network and stacked ConvLSTM with VGG-16 network from the visual understanding of the short GIF video in the visual perception component and textual perception component in our proposed sentiment recognition approach.

Furthermore, as is shown in the GSO-2016 dataset in Table X, our proposed whole approach for the short annotated GIFs with the textual tags can also obtain much better performance in terms of precision, recall, F1 measure and accuracy, compared with the SentiPair (ANP + VNP) model by the provider of public GSO-2016 dataset [6] from visual and textual modalities such as dynamic video and static imagery and text. The superiority of VGG-16 network, particularly 3D CNN network and stacked ConvLSTM from the visual understanding for the short annotated GIFs, will be propitious to improve the related quantitative performance, particularly in terms of the related feature representation with temporal information and the understanding of the sentiment semantics.

V. CONCLUSIONS

In this work, we propose an effective sentiment analysis approach for short annotated GIFs with visual-textual sentiment score. Firstly, from the given GIF video, we perceive the visual sentiment score with recent C3D network, VGG-16 network and ConvLSTM model. Then, we extract the textual sentiment score with the SentiWordNet3.0 model from the results of Synset Forests on the short text annotations w.r.t the corresponding GIF video. And then, we design an affective and multimodal fusion function integrated with the visual and textual sentiment score and achieve the hyper-parameters of the affective fusion with grid search. Later, the extensive experiments involving both quantitative and qualitative evaluations verify the effectiveness of the proposed GIF video sentiment analysis system. In future works, we will consider how to enhance the visual sentiment representation and how to learn efficiently and robustly the complicated parameters with the visual-textual sentiment function for huge amounts of emerging short annotated GIF videos.

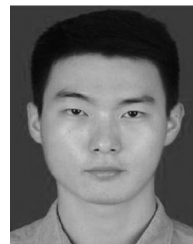
REFERENCES

- [1] M. Naaman, “Social multimedia: Highlighting opportunities for search and mining of multimedia data in social media applications,” *Multimedia Tools Appl.*, vol. 56, no. 1, pp. 9–34, Jan. 2012.
- [2] P. Chikersal, S. Poria, E. Cambria, A. Gelbukh, and C. E. Siong, “Modelling public sentiment in twitter: Using linguistic patterns to enhance supervised learning,” in *Computational Linguistics and Intelligent Text Processing*. New York, NY, USA: Springer, 2015, pp. 49–65.
- [3] A. Zadeh, M. H. Chen, S. Poria, E. Cambria, and P. H. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 1114–1125.
- [4] D. Hazarika *et al.*, “Conversational memory network for emotion recognition in dyadic dialogue videos,” in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Jan. 2018, vol. 1, pp. 2122–2132.
- [5] R. R. Ji, D. Cao, and Y. Zhou, “Survey of visual sentiment prediction for social media,” *Frontiers Comput. Sci.*, vol. 10, no. 4, pp. 602–611, Aug. 2016.
- [6] D. Z. Lin, D. L. Cao, and Y. P. Lv, “GIF video sentiment detection using semantic sequence,” *Math. Problems Eng.*, vol. 2, pp. 1–11, May 2017.
- [7] M. Soleymani *et al.*, “A survey of multimodal sentiment analysis,” *Image Vision Comput.*, vol. 65, pp. 3–14, Aug. 2017.
- [8] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, “A survey on sentiment analysis and opinion mining for social multimedia,” *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 6939–6967, Mar. 2019.
- [9] Z. Li, Y. Fan, W. Liu, and F. Wang, “Image sentiment prediction based on textual descriptions with adjective noun pairs,” *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1115–1132, Jan. 2018.
- [10] S. Poria *et al.*, “Multimodal sentiment analysis: Addressing key issues and setting up baselines,” *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov. 2018.
- [11] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Inf. Fusion*, vol. 37, pp. 98–125, Feb. 2017.
- [12] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, “Convolutional MKL based multimodal emotion recognition and sentiment analysis,” in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.
- [13] H. Moiss *et al.*, “Fusing audio, textual and visual features for sentiment analysis of news videos,” in *Proc. 10th Int. AAAI Conf. Web Social Media*, 2016, pp. 659–662.
- [14] G. Michael and S. Mohammad, “Analyzing and predicting GIF interest-ness,” in *Proc. ACM Multimedia Conf.*, 2016, pp. 122–126.
- [15] S. Bakhshi *et al.*, “Fast, cheap, and good: Why animated GIFs engage us,” in *Proc. 34th Annu. ACM Conf. Human Factors Comput. Syst.*, 2016, pp. 575–586.
- [16] J. Islam and Y. Q. Zhang, “Visual sentiment analysis for social images using transfer learning approach,” in *Proc. IEEE Int. Conf. Big Data Cloud Comput.*, 2016, pp. 124–130.
- [17] S. Jindal and S. Singh, “Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning,” in *Proc. Int. Conf. Inf. Process.*, Pune, India, 2015, pp. 447–451.
- [18] Q. Z. You, J. B. Liu, H. L. Jin, and J. C. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.
- [19] V. Campos, B. Jou, and X. Gir-I-Nieto, “From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction,” *Image Vision Comput.*, vol. 65, pp. 15–22, Sep. 2017.
- [20] J. F. Yang, D. Y. She, and M. Sun, “Visual sentiment prediction based on automatic discovery of affective regions,” *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2513–2525, Sep. 2018.
- [21] Q. Z. You, L. L. Cao, Y. Cong, X. C. Zhang, and J. B. Luo, “A multifaceted approach to social multimedia-based prediction of elections,” *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2271–2280, Dec. 2015.
- [22] F. H. Chen, R. R. Ji, J. S. Su, D. L. Cao, and Y. Gao, “Predicting microblog sentiments via weakly supervised multimodal deep learning,” *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 997–1007, Apr. 2018.
- [23] Y. H. Yu, H. F. Lin, and J. N. Meng, “Visual and textual sentiment analysis of a microblog using deep convolutional neural networks,” *Algorithms*, vol. 9, no. 2, pp. 1–11, 2016.
- [24] J. B. Yuan, S. McDonough, and Q. Z. You, “Stribute: Image sentiment analysis from a mid-level perspective,” in *Proc. 2nd Int. Workshop Issues Sentiment Discovery Opinion Mining*, 2013, pp. 1–8.
- [25] D. Borth, R. R. Ji, and T. Chen, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 223–232.

- [26] Q. Z. You, L. L. Cao, H. Jin, and J. B. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1008–1017.
- [27] Y. C. Li et al., "TGIF: A new dataset and benchmark on animated GIF description," in *Proc. IEEE Comput. Vision Pattern Recognit.*, 2016, pp. 4641–4650.
- [28] B. Jou, S. Bhattacharya, and S. Chang, "Predicting viewer perceived emotions in animated GIFs," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 213–216.
- [29] W. X. Chen and R. W. Picard, "Predicting perceived emotions in animated GIFs with 3D convolutional neural networks," in *Proc. IEEE Int. Symp. Multimedia*, 2017, pp. 367–368.
- [30] Z. Cai, D. L. Cao, and D. Z. Lin, "A spatial-temporal visual mid-level ontology for GIF sentiment analysis," in *Proc. IEEE Congr. Evol. Comput.*, 2016, pp. 4860–4865.
- [31] X. Y. Zhang, C. S. Xu, J. Cheng, H. Q. Lu, and S. D. Ma, "Annotation and search for video blogs with integration of context and content analysis," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 272–285, Feb. 2009.
- [32] Y. Jang, Y. L. Song, Y. Yu, Y. Kim, and G. Kim, "TGIF-QA: Toward spatio-temporal reasoning in visual question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1359–1367.
- [33] D. Tran, L. Bourdev, and R. Fergus, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 4489–4497.
- [34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sept. 2014, *arXiv:1409.1556*.
- [35] G. M. Zhu, L. Zhang, and P. Y. Shen, "Multimodal gesture recognition using 3D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.
- [36] X. J. Shi, Z. R. Chen, and H. Wang, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, MIT Press, 2015, pp. 802–810.
- [37] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2010, pp. 83–90.
- [38] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python*. Newton, MA, USA, O'Reilly Media, 2009, pp. 581–592.
- [39] W. X. Chen, O. O. Rudovic, and R. W. Picard, "GIFGIF+: Collecting emotional animated GIFs with clustered multi-task learning," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 410–417.
- [40] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 1556–1566.
- [41] M. Q. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 1556–1566.
- [42] X. Wang and Q. Zheng, "Text emotion classification research based on improved latent semantic analysis algorithm," in *Proc. Int. Conf. Comput. Sci. Electron. Eng.*, 2013, pp. 210–213.
- [43] X. Zhang and X. Zheng, "Comparison of text sentiment analysis based on machine learning," in *Proc. IEEE Int. Symp. Parallel Distrib. Comput.*, 2017, pp. 230–233.
- [44] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.



Tianliang Liu received the Ph.D. degree in biology and medical engineering from the School of Biology and Medical Engineering, Southeast University, Nanjing, China, in 2010. From 2013 to 2014, he was a visiting scholar with the Department of Computer Science, the University of Rochester, NY, USA. He is currently an Associate Professor with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, China. His current research interests include computer vision and pattern recognition.



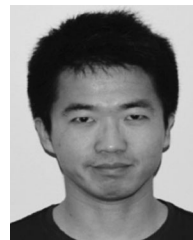
Junwei Wan received the B.E degree in telecommunications engineering from Nanjing University of Information Science and Technology, Nanjing, China, in 2016. He is currently a graduate student in Signal and Information Processing with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include image processing, computer vision, action recognition, and machine learning.



Xiubin Dai received the Ph.D. degree in biology and medical engineering from the Department of Biology and Medical Engineering, Southeast University, Nanjing, China, in 2009. From 2013 to 2014, he was a visiting scholar with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He is currently an Associate Professor with the School of Geography and Biological Information, Nanjing University of Posts and Telecommunications, China. His research interests include pattern recognition and image processing.



Feng Liu received the Ph.D. degree in electronic and optical engineering from the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China, in 1997. He is currently a Professor with the School of Education Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include image processing and multimedia communication.



Quanzeng You received the B.E. and M.E. degrees from the Dalian University of Technology, Dalian, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Department of Computer Science, University of Rochester, Rochester, NY, USA, in 2017. He is currently a Researcher with the Microsoft AI Perception and Mixed-Reality. His research interests include social multimedia, social networks, data mining, high-level visual understanding including image captioning, and visual sentiment analysis. He is interested in developing effective machine learning algorithms to help us understand the data.



Jiebo Luo was with the University of Rochester in Fall 2011 after more than 15 prolific years with Kodak Research Laboratories, where he was a Senior Principal Scientist leading research and advanced development. His research interests include image processing, computer vision, machine learning, data mining, social media, biomedical informatics, and ubiquitous computing. He was involved in numerous technical conferences, including serving as the program Co-Chair of ACM Multimedia 2010, IEEE CVPR 2012 and IEEE ICIP 2017. He was with the editorial boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *ACM Transactions on Intelligent Systems and Technology*, *Pattern Recognition*, *Machine Vision and Applications*, and *Journal of Electronic Imaging*. He is a Fellow of the SPIE, IAPR, ACM, and AAAI. In addition, he is a board member of the Greater Rochester Data Science Industry Consortium. He is a pioneer for contextual inference in semantic understanding of visual data, and social multimedia data mining. He has published extensively in these fields with nearly 400 peer-reviewed technical papers and more than 90 U.S. patents.