Assistive Image Comment Robot—A Novel Mid-Level Concept-Based Representation

Yan-Ying Chen, Tao Chen, Taikun Liu, Hong-Yuan Mark Liao, *Fellow, IEEE*, and Shih-Fu Chang, *Fellow, IEEE*

Abstract—We present a general framework and working system for predicting likely affective responses of the viewers in the social media environment after an image is posted online. Our approach emphasizes a mid-level concept representation, in which intended affects of the image publisher is characterized by a large pool of visual concepts (termed PACs) detected from image content directly instead of textual metadata, evoked viewer affects are represented by concepts (termed VACs) mined from online comments, and statistical methods are used to model the correlations among these two types of concepts. We demonstrate the utilities of such approaches by developing an end-to-end Assistive Comment Robot application, which further includes components for multi-sentence comment generation, interactive interfaces, and relevance feedback functions. Through user studies, we showed machine suggested comments were accepted by users for online posting in 90 percent of completed user sessions, while very favorable results were also observed in various dimensions (plausibility, preference, and realism) when assessing the quality of the generated image comments.

Index Terms—Visual sentiment, viewer affective concept prediction, comment suggestion, comment robot

1 Introduction

As visual content such as images and video become pervasive on the Web and various forms of social media, there is growing interest in understanding how visual content influences outcomes of social communication online. Visual content is generally considered important in attracting user interest and eliciting responses in social media platforms. Particularly, in order to make messages viral, content conveying strong emotions is often used. For example, the photo shown in the Fig. 1 conveyed a strong positive emotion of compassion and became one of the most widely disseminated Twitter photos in 2012.

Distinction can be made between the affects conveyed in an image intended by the publisher or poster of the image and the affects invoked on the viewer side after viewing the visual content. For example, the Obama picture in Fig. 1 conveying "compassion" and "optimism" is intended to invoke viewer affects of "trust" and "love." Sentiment analysis of textual data (such as social media comments and blogs) has been an active topic of research in recent years [1], [15], [18], [27], [40]. On the visual side, research efforts in recognizing high-level information such as style, aesthetics, and sentiments are just emerging [4], [14], [16], [24]. There are also interesting results in using the biometric response signals of the viewers to classify the emotion categories of the visual stimuli [20], [34]. However, the critical aspect modeling the correlations between the publisher affects and the viewer affects is largely missing.

In [6], we started initial work in developing a Bayesian probabilistic model to address the aforementioned open issue. Recognizing the difficulty in detecting high-level emotion attributes directly from low-level features, we advocate adoption of a mid-level concept based representation for both the intended publisher affects conveyed in the image and the invoked viewer affects. The proposed mid-level representation is general, though in [6] and the evaluation experiments later in this paper we use our prior work SentiBank [4] to extract 1,200 Publisher Affect Concepts (PACs) directly from image content without needing text keywords or description, and use data mining methods to discover 446 Viewer Affect Concepts (VACs) from millions of comments from Web images. Such mid-level concepts serve as concrete vocabularies, over which statistical models such as Bayes and Bernoulli models can be readily applied.

The statistical correlation model offers promising potentials for many interesting applications, such as viewer response prediction, optimal content selection for campaigning, and assistive comment robot. We focus on the latter in this paper to demonstrate the utility of the proposed framework and solutions. Given a new image without any keyword or textual descriptions, we aim to automatically suggest a few comments that consist of multiple sentences, composed of the predicted VACs and correctly reflecting the likely affective responses on the viewer part.

To the best of our knowledge, this is the first work that explicitly distinguishes the two aspects of affects of visual content, Publisher and Viewer, models the correlation

- *Y.-Y. Chen is with the FX Palo Alto Laboratory, Palo Alto, CA, USA. E-mail: yanying@gmail.com.*
- *T. Chen is with the Department of Electrical Engineering, Columbia University, New York, NY, USA. E-mail: taochen@ee.columbia.edu.*
- *T. Liu is with the Department of Computer Science, Columbia University, New York, NY, USA. E-mail: tl2582@columbia.edu.*
- *H.-Y. Mark Liao is with the Institute of Information Science, Academia Sinica, Taipei, Taiwan. E-mail: liao@iis.sinica.edu.tw.*
- *S.-F. Chang is with the Department of Electrical Engineering and the Department of Computer Science, Columbia University, New York, NY, USA. E-mail: sfchang@ee.columbia.edu.*

Manuscript received 19 Aug. 2014; revised 9 Nov. 2104; accepted 12 Nov. 2014. Date of publication 5 Jan. 2015; date of current version 4 Sept. 2015. Recommended for acceptance by M. Soleymani, Y.-H. Yang, G. Irie, and A. Hanjalic.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2014.2388370

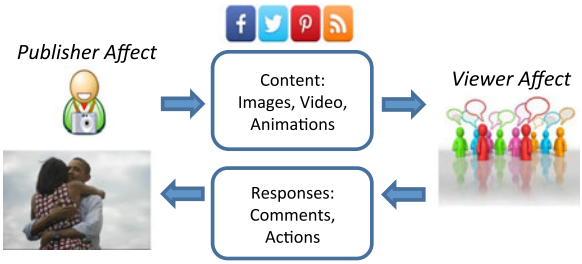


Fig. 1. Visual content with strong emotions plays an important role in influencing the responses of viewers in online social media. This paper studies the correlations between the concepts characterizing the intended affects of the publisher and those of the viewer affects.

between the two, and uses the model to develop image comment robot applications. Other novel contributions of the paper include introduction of the concept-based mid-level representations for both affects, Bayes modeling of the affect correlation, automatic generation of full-sentence comments, and finally an end-to-end system implementation of comment robot applications on popular social media platforms such as FaceBook and Flickr. Extensive experiments and user studies have confirmed the strong performance of the results—in 90 percent of the trial sessions users accept robot suggested comments and post them as the image comments; compared to the manually written comments, the number of likes received by robot suggested comments are only slightly lower than that for the manual comments; and in terms of realism of the comments, more than 50 percent of the robot suggested comments were thought to be manual generated by majority of the evaluators. Though we focus on social image commenting related to affects in this paper, the proposed framework based on mid-level concepts is general and can be extended to address other problems involving high-level information.

In the rest of the paper, we first introduce the proposed framework (Section 2) and review the related studies in (Section 3). The methodology of a mid-level representation (concept) framework and the design of visual-aware comment suggestion are addressed in Section 4, followed by the user study in Section 5 and the evaluation in Section 6. Finally, the open issues and conclusions are discussed in Section 7 and Section 8, respectively.

2 Proposed Framework and System

Fig. 2 shows the architecture of the proposed Assistive Image Comment Robot system. Key to the proposed framework is the statistical correlation model between PACs and VACs, which are discovered from training data offline. In this paper, we use the Flickr images and their associated metadata (keywords, titles, descriptions) and comments as a test domain to illustrate how the proposed framework can be implemented in practice. As shown in Fig. 3, adjective-noun pairs (like "misty woods") with strong sentiment values are discovered and used as PACs, whose automatic classifiers have been made available as SentiBank [4]. Separately, a large pool of image comments from Flickr are used to mine VACs (like "moody"). Further details about PACs and VACs will be provided in Section 4.2. Another component constructed

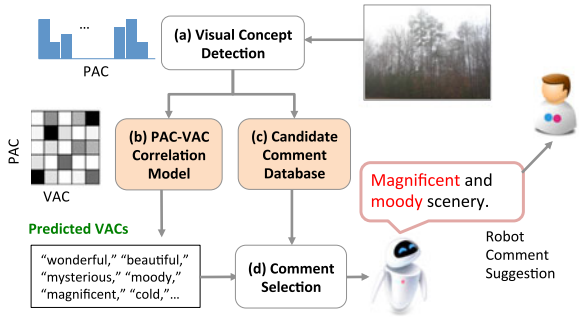


Fig. 2. Proposed assistive image comment robot system. Blocks with the pink filled color are constructed offline and then used with other blocks online to analyze visual content of a new image, detect PAC visual concepts and select a few top comments as suggestions to user.

offline is a database of sentence-length comments that are synthesized based on a large training set of image comments. Each sentence is synthesized according to conditional word occurrence probabilities estimated from the training set. Given a new image without any textual keywords or descriptions, concept classifiers like SentiBank are used to detect PACs and generate a concept score vector, whose elements represent the confidence in detecting corresponding individual concepts (such as "misty woods" or "cute dog"). The detected PAC score vector is then fed to the statistical correlation model to predict the likely VACs that may be evoked on the viewer part. The detected PACs and VACs are then used jointly to select most suitable comments from the pre-synthesized database according to several systematic criteria such as plausibility, relevance, and diversity. The selected comments are then suggested to the user who can further edit the comment before posting. Details of each component mentioned above will be described in Section 4.

3 Related Work

Making machines behave like human—not only at the perception level but also the affective level—is of great interest to researchers. Seeing the obvious relationship between human affect and visual perception, many studies in psychology and cognitive science have focused on the analysis of affective responses evoked by different stimuli such as color [11], motion [8] and semantic categories in visual content [22]. Recently, some studies also started to investigate high-level visual aesthetics [7], [9], interestingness of photos

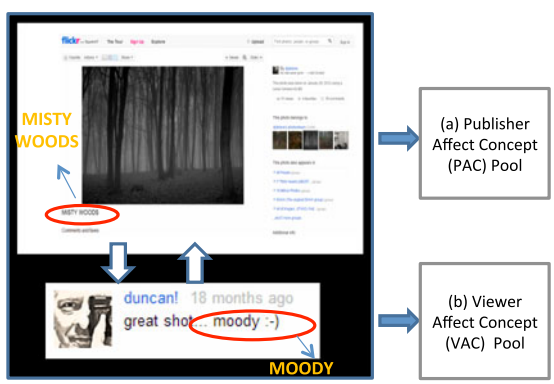


Fig. 3. Publisher affect concepts and viewer affect concepts discovered from Flickr images.

[17], image virality [12] that are potentially useful for applications such as advertising, social media and human-centered interaction.

The computational perspectives on affect and multimedia analysis have been studied in several important works [14], [16], [24], [34]. Hanjalic and Xu [14] first proposed to represent affective content of a given video by the intensity and type of viewer emotion. Irie et al. [16] proposed to classify movie affective scenes and addressed the two critical issues, the features strongly related to viewer emotion and their mapping to emotion categories. Machajdik and Hanbury [24] further exploited the approach driven by psychology and art theory to extract image features for affective image classification. Soleymani et al. [34] incorporated multimodal signals such as electroencephalogram, pupillary response and gaze distance for video emotion recognition. Overall, these studies attempted to map low-level visual features to high-level affect classes and demonstrated that such emotion analysis posed great opportunities for visual search and recommendation. In addition, some related studies of emotion recognition took into account audio cues such as vocal expression [30] and music signals [37] as well as the connection between music affect and user affect [38]. These studies again confirmed the emerging importance of understanding emotion derived from multimedia content. Despite the promising results, direct mapping from low level features remains challenging due to the well-known semantic gap and the emotional gap as discussed in [4], [35].

Facing such challenge, recently a new approach advocates the use of mid-level representations, built upon Visual Sentiment Ontology and SentiBank classifiers [4]. It discovers about 3,000 visual concepts related to eight primary emotions defined at multiple levels in [29]. Each concept in Visual Sentiment Ontology is defined as an adjective-noun pair (e.g., "beautiful flower," "cute dog"), which is specifically chosen to combine the detectability of the noun and the strong sentiment value conveyed in the adjectives. The notion of mid-level representation is also studied in [39], in which visual attributes (e.g., "metal," "rusty") are detected to predict high-level sentiment. However, the work on visual sentiment analysis only focuses on the affect concepts expressed by the content publishers, rather than the evoked emotions of the viewers. For example, a concept "yummy food" expressed in the image by content publishers often evokes the concepts "hungry" and "jealous" on the viewer side.

Social media offers a great resource to analyze the evoked responses of viewers because it carries plentiful comments contributed by users in response to the image content they view. Analysis of comments has been addressed in a broad spectrum of research, including mining opinions in customer reviews. Hu and Liu [15] proposed to summarize comments with specific features of the target product and detect positive or negative sentiments in the customer opinions. Zhuang et al. [40] further incorporated multiple knowledge sources such as a large lexical database and movie casts to specifically improve summaries of movie reviews. Siersdorfer et al. [32] proposed a comment rating predictor for videos by using the dependencies between comments, views, comment ratings and topic categories extracted from the associated metadata. Most of these studies focused on the structures and topics in the comments or metadata without analyzing the media content being viewed.

Online crowdsourcing has been shown promising to collect affective annotations [33]. Analogous to the large concept ontology constructed for describing attributes related to visual sentiment in [4], we propose to mine a large affect concept pool from the viewer comments. Such VACs offer an excellent mid-level abstraction of the viewer emotions and can be used as a suitable platform for mining the correlations between publisher and viewer affects (e.g., "yummy" evokes "hungry," "disastrous" evokes "sad"). Such correlation model was first proposed in our previous work [6], in which a Bayes probabilistic model was developed to analyze the relation between PACs in an image and the likely VACs in the comments. In this paper, we extend to address the issue of automatic generation of multi-sentence comments that are built on the predicted VACs. We further develop an end-to-end system called Assistive Comment Robot, that can suggest comments given a new image without any textual descriptions or keywords.

Generating sentences for an image has been addressed in several studies in image captioning [21], [26]. Kulkarni et al. [21] proposed to generate natural language descriptions from images by recognizing the objects, attributes and prepositional relationships in visual content. Ordonez et al. [26] further demonstrated how large-scale news data can benefit an automatic image captioning system. Different from image captioning, we propose to predict the likely VACs in response to the image content and then exploit the predicted VACs to automatically synthesize and suggest comments to the viewer in social communication. Different from the problem of visual object and attribute recognition, image commenting focuses on predicting the likely viewer responses elicited by the image content and generating comments that are relevant, coherent, and non-redundant.

In summary, this paper includes a set of innovative contributions that clearly differs from prior work in focusing on modeling affect concepts of viewers after viewing images shared in social media, mining of such concepts from online comments, learning their correlation with image content, and developing a comment robot application that automatically suggests comments given an image.

4 Descriptions of Components

4.1 Dataset and Domain

This section introduces the dataset for mining VACs and modeling PAC-VAC correlations. Viewer comments in social media represent an excellent resource for mining VACs. It offers several advantages: (1) the comments are unfiltered and thus preserving the authentic views, (2) there are often a large volume of comments available from major social media, and (3) the comments are continuously updated and thus useful for investigating trending opinions. Since we are primarily interested in affects related to visual content, we adopt Flickr, a semi-professional social media platform to collect the comment data.

a) *Dataset for mining VACs (DVAC).* To ensure that we can get data of rich affects, we first search Flickr with 24 keywords (eight primary dimensions plus three varying

TABLE 1
Our Flickr Training Corpus for Mining VACs Comprises 2 Million Comments Associated with 140,614 Images, Which are Collected by Searching Flickr with the 24 Emotions

emotion keywords (# comments)
ecstasy (30,809), joy (97,467), serenity (123,533)
admiration (53,502), trust (78,435), acceptance (97,987)
terror (44,518), fear (103,998), apprehension (14,389)
amazement (153,365), surprise (131,032), distraction (134,154)
grief (73,746), sadness (222,990), pensiveness (25,379)
loathing (35,860), disgust (83,847), boredom (106,120)
rage (64,128), anger (69,077), annoyance (106,254)
vigilance (60,064), anticipation (105,653), interest (222,990)

strengths) defined in Plutchik's emotion wheel defined in psychology theories [29]. Search results include images from Flickr that contain metadata (tags, titles, or descriptions) matching the emotion keywords. We then crawl the comments associated with these images. The number of comments for each emotion keyword is reported in Table 1, resulting totally around two million comments associated with 140,614 images. To balance the impact from each emotion on the mining results, we sample 14,000 comments for each emotion, resulting in 336,000 comments which will be used to mine VACs.

b) *Dataset for modeling correlations between PAC and VAC (DPVC)*: To collect the training data for modeling the correlations between PAC and VAC, we focus on comments of the images that have PACs related to those defined in our PAC classifier library, namely SentiBank. We use the Visual Sentiment Ontology image dataset [4] in which the associated image metadata (i.e., descriptions, titles and tags) comprises at least one of the 1,200 PACs (cf. Section 4.2.1) defined in the ontology. The comments associated with this image dataset are crawled to form the training data which contain totally around 3 million comments along with 0.3 million images. On the average, an image has about 11 comments, and a comment comprises 15.4 words.

4.2 A Mid-Level Representation Framework Based on Concepts

Our objective is to understand the correlations between intended emotion conveyed by publishers and the evoked emotion on the viewer side. We propose to model these correlations through a mid-level representation framework, that is, presenting the intended and evoked emotion in more fine-grained concepts, i.e., PACs and VACs, respectively. The PAC mined from publisher contributed content is introduced in Section 4.2.1, followed by the VAC discovery from viewer comments in Section 4.2.2 and the correlation model between PACs and VACs in Section 4.2.3.

4.2.1 Publisher Affect Concepts

We adopt 1,200 sentiment concepts defined in Visual Sentiment Ontology (VSO) [4] as the PACs in image content. As mentioned earlier, these concepts are explicitly selected based on the typical emotion categories and data mining from images in social media. Each concept combines a sentimental adjective concept and a more detectable noun concept, e.g., "beautiful flower," "stormy clouds." The advantage of an adjective-noun pair is its capability to turn a neutral noun like "dog" into a concept with strong sentiment like "dangerous dog" and make the concept more visually detectable, compared to adjectives only. The concept ontology spreads over 24 different emotions [29] which capture diverse publisher affects to represent the affect content.

PACs can be found in publisher contributed metadata along with an image or detected from the image content itself (Fig. 3a). In the offline (training) stage we use the "pseudo ground truth" labels found in the image metadata. We detect the presence of each PAC in the title, tags, or description of each image. Such ground truth PAC data has been used as training set to learn automatic classifiers for detecting PACs from image content. Here they will be used in the next section to mine the correlation between PACs and VACs. One potential issue with using such metadata is the false miss error—images without explicit labels of a PAC may still contain content of the PAC. We will apply a label smoothing process to partially address this issue.

In the online (testing) stage we exploit visual-based PAC detectors to measure the presence of each PAC in a new image without any publisher contributed metadata. We adopt the SentiBank, which includes 1,200 visual-based PAC detectors, each corresponding to a PAC in VSO. The input to these detectors include low-level visual features (color, texture, local interest points, geometric patterns), object features (face, car, etc.), and aesthetics-related features (composition, color smoothness, etc.). According to the experiment results in [4], all of the 1,200 PAC detectors have F-score greater than 0.6 over a controlled test set.

Given a test image d_i, we apply SentiBank detectors to estimate the probability of the presence of each PAC p_k, denoted as $P(p_k|d_i)$. Such detected scores will be used to perform automatic prediction of VACs. The step will be described in details later.

4.2.2 Viewer Affect Concepts

This section presents how and what VACs are mined from viewer comments (Fig. 3b). We introduce the strategy for crawling observation data, then a post-processing pipeline for cleaning noisy comments and finally the criteria for selecting VACs.

The crawled image comments usually contain rich but noisy text with a small portion of subjective terms. According to the prior study of text subjectivity [5], [36], adjectives usually reveal higher subjectivity which are informative indicators about user opinions and emotions. Following this finding, we apply part-of-speech tagging [3] to extract adjectives. To avoid the confusing sentiment orientation, we exclude the adjectives within a certain neighborhood of negation terms like "not" and "no." Additionally, to reduce the influence by spams, we also remove the hyperlinks and HTML tags contained in the comments.

We focus on sentimental and popular terms which are often used to indicate viewer affective responses. First, we measure the sentiment value of each adjective by SentiWordNet [10]. The sentiment value ranges from -1 (negative sentiment) to $+1$ (positive sentiment). We take the absolute value to represent the sentiment strength of

TABLE 2
The Example VACs of Positive and Negative Sentiment Mined from Viewer Comments

sentiment polarity	VACs
positive	beautiful, wonderful, nice, lovely, awesome, amazing, fantastic, cute, excellent, interesting delicious, lucky, attractive, happy, adorable
negative	sad, bad, sorry, scary, dark, angry, creepy, difficult, poor, sick stupid, dangerous, freaky, ugly, disturbing

a given adjective. To this end, we only keep the adjectives with high sentiment strength (at least 0.125) and high occurrence frequency (at least 20 occurrences). Totally 446 adjectives are selected as VACs. Table 2 presents the example VACs of positive and negative sentiment polarities, respectively.

4.2.3 Correlations Between Expressed And Evoked Concepts

We aim at mining correlations between the intended concepts (PACs) and evoked concepts (VACs). The philosophy behind is searching for PACs in descriptions, titles and tags (provided by publishers) and measuring their co-occurrences of VACs in comments. As mentioned in Section 4.2.1, the intepretability of PACs allows explicit description of attributes in image content related to intended affects of the publisher. Though there remains noisy information in such descriptions, the large scale observation data from social media which can be periodically crawled and updated offers a promising resource for discovering the rich relation between PACs and VACs.

We apply Bayes probabilistic models and the co-occurrence statistics over a training corpus from Flickr (DPVC) to estimate the correlations between PACs and VACs. Given a VAC v_j, we compute its occurrences in the training data and its co-occurrences with each PAC p_k over the training data θ. The conditional probability $P(p_k|v_j)$ can then be determined by,

$$P(p_k|v_j;\theta) = \frac{\sum_{i=1}^{|D|} B_{ik} P(v_j|d_i)}{\sum_{i=1}^{|D|} P(v_j|d_i)}, \tag{1}$$

where B_{ik} is a binary variable indicating the presence/absence of p_k in the publisher provided metadata of image d_i and $|D|$ is the number of images. $P(v_j|d_i)$ is measured by the occurrence counting of v_j in comments of image d_i. Given the correlations $P(p_k|v_j;\theta)$, we can measure the likelihood of an image d_i given VAC v_j by multivariate Bernoulli formulation [25].

$$P(d_i|v_j;\theta) = \prod_{k=1}^{|A|} (P(p_k|d_i)P(p_k|v_j;\theta) + (1 - P(p_k|d_i))(1 - P(p_k|v_j;\theta))). \tag{2}$$

A is the set of PACs in SentiBank. $P(p_k|d_i)$ can be measured by using the scores of SentiBank detectors (cf. Section 4.2.1), which estimate the probability of PAC p_k appearing in image d_i. Here, PACs act as shared attributes between images and VACs, resembling the probabilistic model [25] for content-based recommendation [28]. Then, we can measure the posterior probability of VACs given a test image d_i by Bayes' rule,

$$P(v_j|d_i;\theta) = \frac{P(v_j|\theta)P(d_i|v_j;\theta)}{P(d_i|\theta)}. \tag{3}$$

$P(v_j|\theta)$ can be determined by the frequency of VAC v_j appearing in the training data and $P(d_i|\theta)$ is assumed equal over images. $P(v_j|\theta)$ indicates the popularity of the VAC v_j in social media; for example, the popular VACs (with higher $P(v_j|\theta)$) in Flickr are shown in Fig. 4b. On the other hand, the $P(d_i|v_j;\theta)$ presents relevance of the VAC v_j to the image content in d_i (cf. the VACs ranked by $P(d_i|v_j;\theta)$ in Fig. 4a). Seeing the different characteristics in the predicted probability of VACs, we further include the relevance indicator γ in the measurement of posterior probability to adjust the influence from image content.

$$P(v_j|d_i;\theta) = \frac{P(v_j|\theta)^{1-\gamma} P(d_i|v_j;\theta)^{\gamma}}{P(d_i|\theta)}. \tag{4}$$

The above equation is useful for an interesting application—given an image, we can predict the most possible VACs by the posterior probability. In the evaluation of VAC prediction (Section 6.3.1), γ is set to 0.5 to balance the impact from either side. For comment suggestion, the impact of varying γ value is discussed in Section 6.3.2.

Fig. 4. The VACs ranked by content-based likelihood (a) and prior probability (b). γ value controls the influence of image content on predicting the VACs—the higher it is, the more influence has on the prediction.

4.2.4 Smoothing

In this subsection, we address the issue of the missing associations—unobserved correlations between PACs and VACs. For example, a PAC "muddy dog" will likely trigger the VAC "dirty," but there are no viewer comments comprising this VAC in our data. Intuitively, some publisher affect concepts share similar semantic meaning; for example, "muddy dog" and "dirty dog." To this end, we apply collaborative filtering techniques to fill the potential missing associations. The idea is to use matrix factorization to discover the latent factors of the conditional probability ($P(p_k|v_j)$ defined in Eq. (1)) and use the optimal factor vectors t_j, s_k for smoothing missing associations between PAC p_k and VAC v_j. The matrix factorization formulation can be expressed as $\min_{t,s} \sum_{k,j} (P(p_k|v_j) - t_j^{\top} s_k)^2$. Note that, we specifically use non-negative matrix factorization [23] to guarantee the smoothed associations are all non-negatives which can fit the calculation in the probabilistic model. The approximated associations $P(p_k|v_j)$ between PAC p_k and VAC v_j can then be smoothed by $t_j^{\top} s_k$.

4.3 Visual-Aware Comment Suggestion

The viewer response to image content is not limited to a single word or a phrase but is usually conveyed through a sentence or even multiple sentences. Beyond predicting VACs, we target on generating sentence-level comments that are composed of the VACs and reflect the likely evoked affects of the viewer. The proposed automatic comment suggestion process includes two primary steps, (1) synthesizing sentence candidates that have a high probability of occurrence given specific PACs detected in the image, and (2) selecting a small set of comments that consist of the predicted VACs.

4.3.1 Sentence Synthesis (Offline)

Generating sentence-level comments to an image can be formulated as text synthesis problem with the consideration of the likely VACs elicited by the visual content. We adopt a simple text synthesis work [31] that models a sentence as a regular Markov chain. Though more sophisticated sentence generation models may be applied, our current implementation uses the basic one as it is not the main focus of our research. Given a corpus of text, one can compute the probability of occurrence of each word given the previous words in the same sentence, where a word is a state. A plausible sentence can be generated by starting a word seed and repeatedly sampling the following words according to the conditional occurrence probability in the reference corpus as described in Section 4.1. Specially, the future state depends on the past m states, where the order m is finite and less than the current state. By increasing the order, we get a model that looks more like real language with fewer grammar errors but has less flexibility to generate novel sentences in the meantime (m is set to 2 in our implementation).

The reference corpus has a critical impact on the topics of the generated sentences. For example, the reference corpus consisting of sports news has higher prospect for generating a sentence related to sports. In our case, the generated sentences are expected to have higher plausibility if the reference corpus is constructed from images of similar visual content. Thus, we organize the comment reference corpus by grouping image comments to individual distinct PACs (e.g., comments associated with images containing PAC "cute dog" are grouped to a separate corpus.) With Markov chain modeled by such PAC-specific reference corpus, the generated sentences are more likely to follow the topics of the comments that are commonly elicited by the images with the corresponding PACs.

To avoid the online delay in generating PAC-specific corpora, we pre-generate 1,200 pools of sentences in the offline stage, each corresponding to a PAC. The sentences in each PAC-specific pool are generated by the reference corpus consisting of the comments associated with the images containing the specified PAC. In our current implementation, there are about 40 to 30,000 comments associated with each PAC. In the online stage a subset of sentence pools are selected to form the candidate sentence pool S without the overhead of remodeling Markov chain and regenerating sentence candidates. One way to select the subset of pools is based on the detection scores of PAC in the given test image. Only the pools corresponding to the top PACs with the highest detection scores are included.

Automatic PAC detection using visual content classification without relying on textual metadata may result in imperfect results. The false positives could be a PAC with an incorrect adjective or with an incorrect noun. The generated sentences associated with an incorrect noun are more detrimental because the predicted object is actually absent from the image content and thus comments containing such false positive objects are likely to be irrelevant to the image. To exclude the PACs with incorrect nouns, we further aggregate the confidence score of each noun by taking an average of $P(p_k|d_i)$ over all PACs with the same noun. A sentence pool is selected and added to the candidate database S only if its corresponding PAC comprises one of the top 5 nouns with the highest aggregate scores.

This preprocessing is applied to nouns rather than adjectives because adjectives are much more interrelated and subjective than nouns. For example, happy, cute, fluffy, tiny, and adorable are all valid and highly related adjectives often used together with the noun "dog". It is difficult to exclude some of them from others when forming the comment sentence pool. Though the above remedy may not avoid including the sentence pool associated with the PACs comprising an incorrect adjective, such implausible sentences could be filtered out by other methods to be included in the comment selection process (described in the next subsection).

4.3.2 Comment Selection and Suggestion (Online)

A comment could consist of a sentence or multiple sentences. With the pool of sentence candidates S for a given test image, we aim at selecting the most appropriate sentences to form a comment of high quality in terms of two principle criteria: *relevance* and *diversity*. The methods for selecting a single-sentence comment and composing a multi-sentence comment are introduced first, followed by the approach for ranking and suggesting the most appropriate comments.

Relevance. The relevance of a sentence to a given image can be measured by the VACs that appear in the sentence and those predicted to be evoked based on the PAC-VAC correlation model described in Section 4.2.3; for example, for an image containing PAC "yummy food" the sentence containing VAC "tasty" is considered to be more relevant than the sentence containing "handsome," because "yummy food" is more likely to evoke "tasty" rather than "handsome," as predicted by the statistical correlation model. We propose to use VACs V as the shared attributes to measure the relevance of a sentence to an image. Given an image, we first detect the PACs in the image by using SentiBank PAC detectors and then predict the probability of each VAC evoked by the detected PACs by using the Bayes correlation model (cf. Eq. (4)). The given image d_i is represented as a vector, each dimension indicating the probability of evoking a VAC v_j. Each sentence s_q is also represented by a binary indicator vector B_q, each element B_{qj} indicating the presence of v_j in s_q. The relevance between an image d_i and a sentence s_q can be formulated as the likelihood of s_q given d_i,

$$P^{(v)}(s_q|d_i) = \prod_{j=1}^{|V|} B_{qj} P(v_j|d_i) + (1 - B_{qj})\lambda_{ji}. \quad (5)$$

The $P(v_j|d_i)$ can be estimated by the PAC-VAC correlation model as shown in Eq. (4). The first term computes the inner product of the VAC score vector given image d_i and the VAC indicator vector given sentence s_q. The second term provides a smoothing term accounting for other VACs not predicted, with its influence controlled by the parameter λ_{ji}. The value of λ_{ji} is determined as follows.

$$\lambda_{ji} = \frac{\gamma \ddot{P}_{min} + (1-\gamma)\ddot{P}_{max} + \ddot{P}_{avg}}{2}, \tag{6}$$

where $\ddot{P}$ is the set of probability of VACs in the image d_i: $\{P(v_j|d_i)|\forall v_j \in V\}$ and $\ddot{P}_{min}$, $\ddot{P}_{max}$ and $\ddot{P}_{avg}$ are the minimum, maximum and average probability within $\ddot{P}$ respectively. λ_{ji} is controlled by the relevance indicator γ introduced in Eq. (4). The higher γ leads to the lower λ_{ji} and the increasing significance of B_{qj} (the presence of v_j in s_q), thus favoring the s_q that contains v_j likely to be evoked by the image content. γ can be adjusted on demands. The parameter setting is discussed in Section 6.3.2.

It is possible that a sentence may comprise plausible VACs together with implausible keywords other than VACs; for example, the VAC "funny" is relevant to comment on an image with PAC "cute dog," however, the sentence "I love the funny cat" is implausible because of the mismatched noun "cat." To suppress the mismatched nouns, we further consider the noun n_j appearing in the sentence and its probability to appear in the evoked comments for a given image d_i. Similar to the discovery process in Section 4.2.2 we first establish a vocabulary with 1,000 noun concepts, defined as Viewer Noun Concepts (VNC). Likewise, $P(n_j|d_i)$ and $P^{(n)}(s_q|d_i)$ can be measured by using the methods introduced for measuring $P(v_j|d_i)$ and $P^{(v)}(s_q|d_i)$, where the v_j is replaced by n_j. The relevance of a sentence to an image can then be presented in two modalities, $P^{(v)}(s_q|d_i)$ and $P^{(n)}(s_q|d_i)$. The overall relevance score z_{qi} is measured in the log space by a late fusion manner,

$$z_{qi} = \frac{logP^{(v)}(s_q|d_i) + logP^{(n)}(s_q|d_i)}{2|\phi(s_q)|^{0.5}}. \tag{7}$$

$\phi(\cdot)$ is the set of words in the given sentence. $\phi(s_q)$ is a normalization term to favor VAC and VNC words in a sentence. Thus, we can suggest the most relevant sentence s_q with the highest z_{qi} as a single-sentence comment to the given image.

Diversity. To extend the comment beyond a single sentence, we further exhaustively combine μ sentences chosen from the sentence set $\hat{S}$ that have the top sentence scores defined in Eq. (7)[1] to form a multi-sentence comment set $\hat{C}$. To ensure the combined sentences are no redundant, we further adopt a criterion to ensure the diversity of concepts contained in different sentences in the same comment. For example, the comment "I love the funny dog. How cute it is." has more diversity than "I love the funny dog. Very funny." because the first comment includes the VACs "funny" and "cute" while the second comment repeats the same VAC "funny." The comments in $\hat{C}$ are ranked by the summation of relevance scores, $\{z_{qi}|\forall s_q, s_q \in c_l\}$. The diversity δ_l (with value ranging between 0 and 1) of a multi-sentence comment c_l in $\hat{C}$ is measured as follows,

$$\delta_l = 1 - \frac{\left|\bigcap_{s_q \in c_l} \phi'(s_q)\right|}{\left|\bigcup_{s_q \in c_l} \phi'(s_q)\right|}. \tag{8}$$

$\phi'(\cdot)$ means the set of VACs and VNCs in the given text. The most relevant c_l in $\hat{C}$ with δ_l larger than a given threshold[2] is then selected as the suggested comment for the given image.

While considering the diversity, a multiple-sentence comment may suffer from the inconsistency problem. That is, the VACs in different sentences in the same comment are not reasonable when used in conjunction in the same comment. An example is, "I love the funny dog. It looks so scary." where the VACs "funny" and "scary" are rarely co-occurred in the same comment for an image. To address this issue, we further restrict the second and later sentences in a comment to be those generated by reference corpora sharing the same PAC nouns as the corpus used in generating the first sentence. This will ensure that all sentences in the same comment are generated from a corpus related to the noun and thus the inconsistency across sentences can be reduced.

To enrich the options for users, this framework can be extended to multi-comment suggestion. Motivated by the previous work for updating users about time critical events [13], we iteratively add an additional comment which adds novel information to the ones already provided. In each iteration, a new comment c^* is selected from the comment set $\hat{C}^{(\tau-1)}$, where c^* should have the fewest VACs and VNCs that are overlapped by the set of suggested comments $\Omega^{(\tau-1)}$ in the previous iteration $\tau - 1$.

$$c^* = \arg\min_{c_l \in \hat{C}^{(\tau-1)}} \frac{|\phi'(\Omega^{(\tau-1)}) \cap \phi'(c_l)|}{|\phi'(\Omega^{(\tau-1)}) \cup \phi'(c_l)|}. \tag{9}$$

The new set of suggested comments $\Omega^{(\tau)}$ is updated as $\Omega^{(\tau-1)} \cup c^*$ and the set of candidate comments $\hat{C}^{(\tau)}$ is updated as $\hat{C}^{(\tau-1)} - c^*$. The initial comment in Ω should follow the criteria aforementioned for single comment selection and each latter comment should satisfy the diversity required for a single comment.

4.4 Prototype and User Interfaces

The proposed comment robot is an assistant tool which can help users comment on photos more efficiently. It can recommend multiple plausible comments relevant to the image content, and users can select any comment based on their own preference. Currently it is available as an extension tool for Chrome browser. Fig. 5 is the user interface of our assistive comment robot. Given an image, it will suggest three comments and offer functions to assist users in finding the preferred comments more efficiently.

1. $|\hat{S}|$ is limited to 50 and μ is set to 2 in our experiments.

2. The threshold is set to 0.8 and iteratively decreased if no δ_l satisfies this criterion.

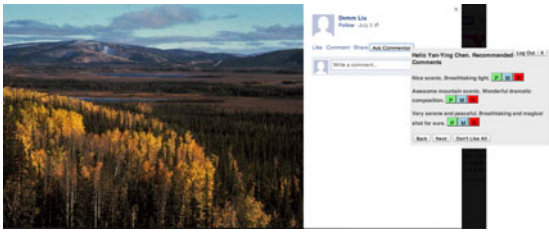


Fig. 5. The user interface of assistive comment robot to recommend comments for Facebook photos.

4.4.1 Functionality

Fig. 6 shows a closeup of the interface. The buttons "Back" and "Next" are designed for returning to the comments in the previous iteration and for requesting more comments in the next iteration, respectively. Once "Next" is clicked, the comments in the current iteration will be logged as *displayed but not selected comments* in our database. The button "Don't Like All" means all the displayed comments in the current iteration are not satisfactory and these comments will be logged as *rejected comments* in our database.

The buttons "R" (red) and "M" (blue) in Fig. 6 are designed for obtaining fine-grain user feedback of each single comment. Clicking "R" leads to a rejection of the corresponding comment that will be then logged as a rejected comment. Clicking "M" activates the request of three more comments related to the corresponding comment, which will be logged as a *preferred comment*. Clicking "P" (green) means the corresponding comment is selected for posting and will be logged as a *posted comment*.

One additional function symbolized as "x" in Fig. 6 means to cancel this session of comment suggestion without any logs being kept. We will discuss how users interacted with the aforementioned functions in Section 6. Furthermore, we have designed tooltips when user's cursor moves over the buttons to give hints of each button. Representing buttons with intuitive icons in a proper size is also potential to improve user interface.

4.4.2 Relevance Feedback

The four types of user interaction logs can be used as informative relevance feedback to further improve comment suggestion. In our proposed mechanism, each type of the comment logs has different impacts on updating the results of VAC prediction and the following comment suggestions. Given an image, the predicted probabilities of VACs $P(v_j|d_i;\theta)$ and VNCs $P(n_j|d_i;\theta)$ in the image are adjusted based on the history of the comments that have been shown to the user and the feedback received so far.

$$P'(v_j|d_i) = (1-\rho_{ji})P(v_j|d_i;\theta) + \rho_{ji}\ddot{P}_{min},$$
$$\rho_{ji} = \max\left(\sum_{c_l \in C^{(v_j,d_i)}} \sigma(c_l), 1\right). \quad (10)$$

$\ddot{P}_{min}$ is the minimum of $P(v_j|d_i;\theta)$ over v_j. ρ_{ji} is the aggregated penalty incurred by the logs in $C^{(v_j,d_i)}$, which is the union of rejected comments and displayed but not selected comments of image d_i that contain v_j. $\sigma(\cdot)$ is an adjustable controlled penalty. If a concept is contained in more comments that have been rejected or not selected, its predicted probability should be reduced and shifted towards the minimal value $\ddot{P}_{min}$.

Fig. 6. The functions of assistive comment robot.

To make the comment suggestion more personalized, the penalty value of $\sigma(\cdot)$ is initially set to 0.1 but will be increased to 1 in subsequent iterations of the same image and user. Moreover, if v_j appeared in the "preferred comments," $P'(v_j|d_i)$ will be set to $\ddot{P}_{max}$, which implies v_j has the highest probability to be included in the following suggested comments.

5 User Study

We conducted two user studies to evaluate the assistive comment robot introduced in Section 4.4. The first aims at investigating how effective the comment robot can help users commenting on images in social media. The second focuses on evaluating the quality of machine-generated comments and its differences from the manually-created comments.

5.1 Assistive Image Commenting

To evaluate how the comment robot can assist users in generating comments for images, we invited 26 Facebook users to join the subjective test. The subjects include 8 females and 18 males aged around 20-35. Mostly are graduate students with a major in computer science or the related field. Before joining the user study, they were not aware of the technical details.

We selected a set of test images with seven topical categories: flower, architecture, scenery, human, vehicle and animal, each comprising 20 images. The seven image categories are selected for our experiments because they are popular topics in consumer photos and commonly appearing in social media. The images in each category are randomly sampled from the Creative Common Licensed photos published by the public website [3] to avoid copyright problems and to make the evaluation dataset publicly available.

The invited users were asked to consider the photos were posted by their friends on social media like Facebook and post comments in the typical manner accordingly. Each user is asked to comment on three images in each topical category by selecting from the machine-suggested comments and another three images without using the robot. The former is

3. Available online at: http://public-domain-image.com/

Fig. 7. The user interface to evaluate quality of comments in terms of plausibility, specificity, preference and the realism. Each metric is given three degrees of options, e.g., "not plausible," "neutral" and "plausible" for plausibility.

defined as machine-assisted comment (machine) and the latter is defined as manually-created comments (manual) hereafter. By "machine suggested comment," we meant the comments automatically suggested by the proposed method given a new image. Such suggested comments are then presented to the user, who may choose any of the suggested comments and post them on FaceBook. We refer to such selected comments as "machine assisted comments," to distinguish them from the comments generated by the user manually, and then evaluate the quality of such "machine assisted comments." Note our method may suggest several comments for an image, but only few of comments are selected and accepted by the user.

For fair comparison, the number of sentences (μ in Section 4.3.2) per comment is set to 2 in our experiments in order to make machine generated comments of similar lengths to those for manually generated comments (on average 6.1 words versus 5.5 words per comment). The proposed system is also capable of generating longer comments with more sentences by adjusting the μ parameter. Compared to shorter comments, there might exist more grammar errors in longer comments. To address this, more sophisticated grammar verification [2], [19] can be used to improve the quality of comments.

The robot-assisted comments and manually-created comments are mixed in the displayed Facebook page after they are posted. Namely, there is no markers to indicate which ones are generated with assistance of robot. Then, we invite the users to review the posted comments and indicate which comments they like in the normal manner they do when interacting with the images on the social media. Note that, in such setting users may see their own comments and have known these comments were either machine-assisted or manually-created. That may considerably influence their preferences. To reduce such ambiguity, we conduct a Turing test where the subjects were completely unaware of how the comments were generated. More details are discussed in the next section.

5.2 Comment Quality

We invited another 10 subjects to evaluate the quality of the machine-assisted comments (selected by the users) and manually-created comments gathered in the previous user study. As the interface shown in Fig. 7, each single test includes an image and a comment either machine-assisted or manually-created. The subjects were asked to evaluate the given comment in terms of (1) plausibility: how plausible the comment is to the given image, (2) specificity: is the comment specific (specific to the given image content) or generic, (3) preference: how much the subject likes the given comment and (4) realism: whether the subject can tell the comment was synthesized and suggested by a comment robot. Each of the 140 test image-comment pairs was evaluated by three subjects, totally 420 evaluation results.

6 Evaluation

We present the evaluation at different levels, (1) the overall usability of the assistive comment robot, (2) the quality of the comments suggested by assistive comment robot and (3) evaluation of component algorithms included in the system, including PAC-VAC correlation model, comment synthesis and relevance feedback.

6.1 Overall Usability

We evaluate the overall usability of assistive comment robot through the user study described in Section 5.1. The invited users contributed 405 test sessions. Each test session was finished either by posting a selected comment or by rejecting all suggested comments. As reported in Table 3, on average the users finished a session within 3.43 iterations, each iteration comprising 3 suggested comments. The # posts means the number of sessions that the users accepted one of the suggested comments and posted it at the end of the session. It's very encouraging to see the accept rate can reach around 90 percent. Among the seven image classes, the accept rate of the classes "flowers" and "scenary" are the highest. Both classes are of outdoor scenes or close-up objects that might occupy the whole image, perhaps resulting in the more accurate PAC detection from visual content[4]. The class "human" has the lowest acceptance rate (still as high as 81 percent) probably because commenting on human requires more familiarity with the subjects.

We further evaluate the difference in preference between comments produced by humans and assistive robots. The number of "likes" for a Facebook comment is an intuitive indicator of preference. Fig. 8 reports the average number of likes per machine-assisted/manually-created comment in each photo class. The average like of machine-assisted comments (0.37) is understandably lower than that of manually-created comments (0.45), and the results are consistent in the comments for images of different classes.

In a small number of sessions, users used "cancel" by clicking on the "x" button in Fig. 6 without accepting any suggested comment or explicitly rejecting all suggested comments. Through additional survey, we found in such cases, users did not have strong evaluations of the suggested comments. Very often users found the suggested comments reasonable but desired to look for even better comments by canceling the session and started a new session again. We did not include such canceled sessions when computing the accept rates reported above.

6.2 Quality of Comments

In the second user study (cf. Section 5.2), we aim at evaluating the quality of the comments produced by humans with

4. The PAC detectors we used are based on SentiBak version 1.1 that relied on the visual features of a whole image rather than localized objects.

TABLE 3
Usability of Assistive Comment Robot for Suggesting Social Comments for Images

image class	food	flowers	architecture	scenery	human	vehicle	animal	all
# sessions	50	52	60	51	57	74	60	405
# posts	45	51	53	50	47	63	54	363
# iterations	2.36	2.58	3.15	2.29	3.48	3.99	5.57	3.43
# clicks (next)	1.68	2.17	2.69	1.94	2.90	3.47	5.00	2.92
# clicks (more)	0.68	0.40	0.46	0.35	0.59	0.51	0.57	0.51
# clicks (reject)	1.56	0.71	2.20	0.29	2.24	2	2.83	1.75
accept rate	0.90	0.98	0.88	0.98	0.81	0.85	0.90	0.90

or without assistive robots (machine versus human). The three degrees of each metric (cf. Fig. 7) are given different scores, 0, 0.5 and 1, from left to right. For each metric, the score of each image-comment pair is the average of the scores given by three subjects. Fig. 9 reports the average scores of four quality metrics, plausibility, specificity, preference and realism. Note that, the preference is different from that measured by the likes on Facebook in Fig. 8.

Among the four evaluation metrics in Fig. 9a, the manually-created comments and machine-assisted comments have nearly the same specificity and less 5 percent difference in plausibility. The difference in realism is larger than the other three metrics; therefore we investigate more details of the realism in Fig. 9b. It shows the number of users who "guessed" the given comment (either machine or manually generated) has been generated by robot or human. More than 50 percent $(0.43 + 0.11)$ of machine-generated comments fooled the evaluators and were thought to be manually-created by the majority of the subjects (at least 2 of the 3 subjects for a test). Through these numbers we can see that although the machine-generated comments do not perfectly resemble manually-created ones, they still look very convincing.

Fig. 10 shows some examples of image-comment pairs that are considered to be"real" (i.e., manually-created) by all the three subjects. The comments in the upper bar are machine-generated and those in the lower bar are manually-created. All of them present high plausibility and some of them mention the specific details in the given image (e.g., (a)-1 and (b)-2). It is worth noticing that several comments that are considered as manually-created comprise question sentences, e.g., (b)-1 and (b)-3, which provides a distinct style not implemented by the robot program yet. This certainly points out an interesting direction for future expansion of the system.

Besides, the design of the user study may have encouraged the subjects to beat the system because in the tests we asked them whether the given comment is machine-generated or manually-created. That implies some of the comments were machine-generated and the subjects may have felt "challenged" and thus "motivated" to defeat the system. The phenomenon can be found in the correlation between realism and the other metric, preference. As shown in Table 4, the score of realism has positive correlations with all the three metrics and with the highest one with "preference." That suggests the subjects might tend to dislike a comment if they think it is machine-generated. Although in this setting the subjects can differentiate machine-assisted comments from the manually-created ones considerably (69 percent of real comments were guessed to be generated manually, while 54 percent of robot generated ones were considered so), the disparity is expected to be reduced in real applications if users do not proactively investigate the subtle difference.

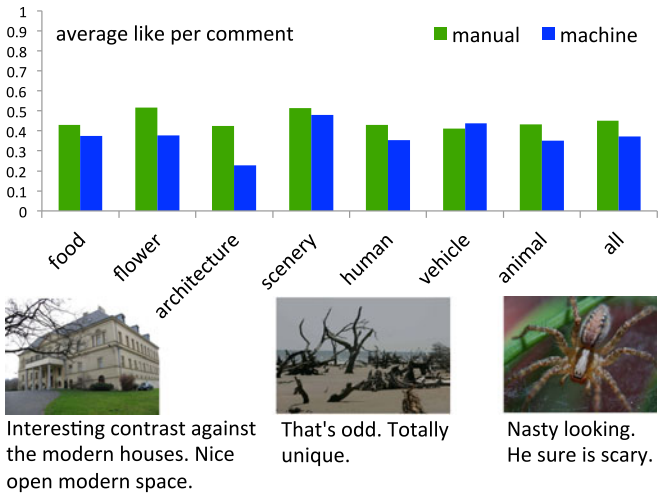


Fig. 8. The average number of likes per comment over different photo classes and the example image-comment pairs accepted by the subjects.

6.3 Component Evaluation

We further provide evaluation of individual algorithm components included in the whole system in this subsection.

6.3.1 Evaluation of PAC-VAC Correlation Model

Table 5 reports the top PAC-VAC correlated pairs ranked by $P(p_k|v_j)$ (cf. Eq. (1)) and filtered by statistical significance

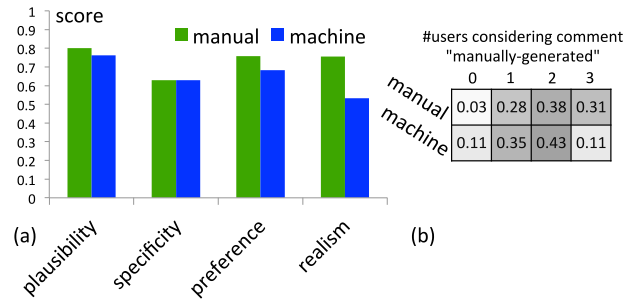


Fig. 9. The results of subjective evaluation of image comments. (a) scores of the plausibility, specificity, preference and realism for the manually-created (manual) and machine-assisted (machine) comments, respectively. (b) # user considering the comment generated by humans manually. 54 percent (43% + 11%) of machine-assisted comments are thought to be generated by humans by the majority of users.

Fig. 10. Image-comment pairs that are agreed by all three users to be "real" (i.e., manually generated). (a) Three pairs generated by machine and (b) three pairs created by users.

value (p-value), e.g., "hilarious" for "crazy cat," "delicate" for "pretty flower" and "hungry" for "sweet cake." It is interesting to note that sometimes the adjectives in the PACs and VACs could be quite different, e.g., "cute" for "weird dog" and "scary" for "happy halloween."

We further demonstrate how PAC-VAC correlations benefit VAC prediction. Given a test image d_i, we aim at predicting the most likely VACs stimulated by this image. We measure the posterior probability of each VAC v_j by the probabilistic model in Eq. (3). The higher posterior probability means that the VAC v_j is more likely to be evoked by the given image d_i. In addition, we compare our method (Corr) with the baseline using PACs [4] only. Given a test image, the baseline method (PAC-only) chooses all the VACs appearing in the comments associated with the training images which comprises the PACs with the highest detection scores in the test image. In contrast, our method (Corr) considers the soft detection scores of all PACs and use the PAC-VAC correlations described in Eq. (3) to rank VACs based on $P(v_j|d_i;\theta)$. The predicted VACs are with probabilities higher than a threshold. For fair comparison without being affected by sensitivity of threshold setting, the threshold is set to include the same number of VACs predicted by the baseline method.

The test images are downloaded from the public dataset [4] (separated from (DPVC) in Section 4.1). Each test image has comments comprising at least one VAC. Totally 2,571 test images are evaluated by the two performance metrics, overlap ratio and hit rate. Overlap ratio indicates how many predicted VACs are covered by the ground truth VACs, normalized by the union of predicted VACs and ground truth VACs. As shown in Table 6, the overlap of our approach (Corr) outperforms the baseline approach by 20.1 percent. The higher overlap indicates higher consistency between the predicted VACs and the ground truth VACs given by real users.

Considering the sparsity in comments, the false positives in the predicted VACs may be missing but actually correct. To address this issue, we further evaluate hit rate, that is, the percentage of the test images that have at least one predicted VAC hitting the ground truth VACs. Hit rate is similar to overlap ratio but deemphasizes the penalty of false positives in the predicted VACs. As shown in Table 6, our approach achieves 19.0 percent improvement in overall hit rate compared to the baseline. The gain is even higher (22.9 percent) if the hit rate is computed only for the top 3 predicted VACs (hit rate (3)). All the results confirm the apparent contribution from the PAC-VAC correlation model for VAC prediction that is a critical component for comment suggestion.

TABLE 4
The Pearson's Correlation Coefficient of Realism with the Other Three Metrics Plausibility, Specificity and Preference

Pearson's r	plausibility	specificity	preference
realism	0.3126	0.3871	0.4579

TABLE 5
PAC-VAC Pairs with Strong Correlations

PAC	#1 VAC	#2 VAC	#3 VAC
tiny dog	cute	adorable	little
weird dog	weird	funny	cute
crazy cat	hysterical	crazy	hilarious
cloudy morning	ominous	serene	dramatic
dark woods	mysterious	spooky	moody
wild water	dangerous	dynamic	wild
terrible accident	terrible	tragic	awful
broken wings	fragile	poignant	poor
bright autumn	bright	delightful	lovely
happy halloween	spooky	festive	scary
pretty flowers	delicate	joyful	lush
wild horse	wild	majestic	healthy
silly girls	sick	funny	cute
mad face	mad	funny	cute
beautiful eyes	expressive	intimate	confident
sweet cake	yummy	hungry	delicious
nutritious food	healthy	yummy	delicious
colorful building	colourful	vivid	vibrant
haunted castle	spooky	mysterious	scary

6.3.2 *Evaluation of Comment Synthesis*

The proposed comment suggestion considers the relevance between a sentence and the given image content as well as the diversity among multiple sentences in a comment. Fig. 11 reveals the influence of relevance indicator γ (cf. Eq. (4) and Eq. (6)). The higher γ leads to selecting the more content-relevant sentence (as (+), $\gamma = 1$) for the given image while the lower γ leads to selecting the more generic sentence (as (–), $\gamma = 0.1$), though both are plausible.

Furthermore, the generated comments without/with enriching diversity (cf. Eq. (8)) are shown in Fig. 12 (–) and (+), respectively. Obviously, many repetitive VAC words (underline) appear in the comments generated without considering the diversity, e.g., "dramatic," "yummy" and "floral" in the comments of (–). Compared to the comments composed with the consideration of diversity in (+), the comments in (–) present redundant information that might decrease the quality.

Overall, increasing relevance and diversity can enrich the information in a comment. However, the subjective quality

TABLE 6
Accuracy of VAC Prediction Given a New Image

method	PAC-only [4]	Corr
overlap	0.2295	0.4306 (+20.1%)
hit rate	0.4333	0.6231 (+19.0%)
hit rate (3)	0.3106	0.5395 (+22.9%)

Our method outperforms the baseline by 20.1 percent in predicted concept overlap, 19 to 22.9 percent in hit rates.

Fig. 11. Influence of the relevance control parameter, γ. Higher (+), Lower (–).

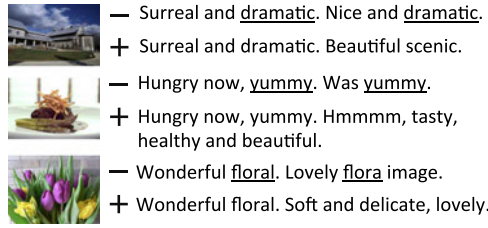


Fig. 12. Example comments when the diversity metric is incorporated (+), compared to the baseline (–).

of the comments still depends much on the personal and social context. User studies for quantitatively assessing such effects will be part of our future work.

6.3.3 Evaluation of Relevance Feedback

As introduced in Section 4.4, the assistive comment robot offers several functions to gather relevance feedback from users including "M" (requesting more comments related to the reference one) and "R" (rejecting a specific comment). As reported in Table 3, "M" (#more) and "R" (#reject) were clicked 0.51 and 1.75 times per session before the users accepted a comment. This suggests there might exist some comments that particularly interest users or look implausible to users. Relevance feedback returned in such critical cases can be used to further improve the performance. Moreover, the function "Next" can also be used as a form of implicit relevance feedback. As shown in Eq. (4), it can be used to iteratively reduce the probabilities of VACs that have appeared in the comments of the previous iterations. Averagely, the users made a post after clicking 2.92 "Next". That suggests the relevance feedback can moderately help the comment suggestion task in the online assistive comment robot.

7 Open Issues and Future Work

We briefly discuss the limitations of the current system, open issues, and future work in the following.

7.1 Domain Differences

Though the proposed concept-based mid-level representation is general, several components need to be adapted when extending to other domains, e.g., product review, news, film critiques, and advertising. The image data sets, frequent concepts, user types, and user behaviors will vary and thus the current tools and models will need to be updated. Moreover, our approach could be extended to more diverse sentence types, e.g., question sentences, by collecting corpora for specific sentence types. However, we expect the same framework and methodologies for concept discovery, correlation modeling, and comment recommendation to be generalizable.

7.2 Modeling Of Users And Social Networks

Our current approaches have not considered the variations among individual users in terms of their demographics, interests, and other attributes. Such personalized factors have been used in most recommendation systems and will likely contribute to further improvement of the proposed system, especially in modeling correlation between image content and viewer affects and customizing the preferred comments in response to shared images.

In addition, evoked viewer affects are expected to be influenced by context in which the image is shared and the social relations between the publisher and the viewers. The same image content may evoke different affective responses when it is framed in different social or cultural contexts or embedded in different conversation threads. Moreover, responses of individual users are likely to be influenced by the opinion leaders in the community. Nonetheless, the framework presented in this paper offers a sound system to which additional models for users and networks can be incorporated.

8 Conclusions

This paper presents an innovative and systematic effort in understanding how visual content is used in conveying affects intended by publishers of images in online social media and how such content evokes affective responses on the viewer side. We propose a concept-based mid-level representation and most importantly statistical methods like Bayes models are developed to characterize the correlations between Publisher Affect Concepts and Viewer Affect Concepts. We show that the PAC-VAC correlation model can be used to predict the likely responses of the image viewers, and select the most appropriate comments from a comment database made of candidate comments automatically synthesized in advance. To demonstrate the utility of the framework, we developed and tested an end-to-end social media application, called Assistive Image Robot, which automatically suggest comments to users with 90 percent accept rate and comparable quality compared to manually generated ones from users. Future work includes extension for predicting affective attributes or other high-level information associated with visual content in other applications domains as viewer responses to videos, films, and animations. In addition, incorporation of personalized user models and social network relations will be important in order to understand the influence of personalized attributes, user-user interaction context, and social network structures of the community.

Acknowledgments

This work was sponsored in part by the U.S. Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-C-0028. The views and conclusions contained in this document are those of the

author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Defense Advanced Research Projects Agency or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. This work was performed at Columbia University. Yan-Ying Chen was the corresponding author of the article.

REFERENCES

[1] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. Int. Conf. Weblogs Soc. Media*, 2010, pp. 122–129.

[2] R. Barzilay, K. R. McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in *ACL Proc. Annu. Meet. Assoc. Comput. Linguistics Comput. Linguistics*, 1999, pp. 550–557.

[3] S. Bird, E. Klein, and E. Loper, "Natural language processing with python: Analyzing text with the natural language toolkit," 2009, http://www.nltk.org/book

[4] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Multimedia*, 2013, pp. 223–232.

[5] R. F. Bruce and J. M. Wiebe, "Recognizing subjectivity: A case study of manual tagging," *Nat. Lan. Eng.*, vol. 5, 1999, pp. 187–205.

[6] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang, "Predicting viewer affective comments based on image content in social media," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2014, p. 233.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 288–301.

[8] B. H. Detenber, R. F. Simons, and G. G. Bennett, Jr., "Roll 'em!: The effects of picture motion on emotional responses," *J. Broadcasting Electron. Media*, vol. 42, no. 1, pp. 113–127, 1998.

[9] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1657–1664.

[10] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2006, pp. 417–422.

[11] L. Fitzfibbons and R. F. Simons, "Affective response to color-slide stimuli in subjects with physical anhedonia: A three-systems analysis," *Psychophysiol.*, vol. 29, no. 6, pp. 613–620, Nov. 1992.

[12] M. Guerini, J. Staiano, and D. Albanese, "Exploring image virality in Google Plus," in *Proc. Int. Conf. Soc. Comput.*, 2013, pp. 671–678.

[13] Q. Guo, F. Diaz, and E. Yom-Tov, "Updating users about time critical events," in *Proc. Eur. Conf. Informat. Retrieval*, 2013, pp. 483–494.

[14] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[15] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proc. AAAI Conf. Artif. Intell.*, 2004, pp. 755–760.

[16] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010.

[17] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 145–152.

[18] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proc. Int. Conf. Comput. Linguistics*, 2004.

[19] K. Knight and D. Marcu, "Statistics-based summarization—step one: Sentence compression," in *Proc. Nat. Conf. Am. Assoc. Artif. Intell.*, 2000, pp. 703–710.

[20] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classication of eeg and peripheral physiological signals for recognition of emotions induced by music videos," in *Proc. Int. Conf. Brain Inform.*, 2010, pp. 89–100.

[21] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011.

[22] P. Lang, M. Bradley, and B. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," *Tech. Rep. A-8. Univ. Florida, Gainesville, FL, USA*, 2008.

[23] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Annu. Conf. Neural Informat. Process. Syst.*, 2001, pp. 556–562.

[24] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Multimedia*, 2010, pp. 83–92.

[25] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *Proc. AAAI Workshop Learn. Text Categorization*, 1998.

[26] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. Annu. Conf. Neural Inform. Process. Syst.*, 2011, pp. 1143–1151.

[27] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empir. Methods Nat. Lang. Process.*, 2002.

[28] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web: Methods and Strategies of Web Personalization. Volume 4321 of Lecture Notes in Computer Science*. New York, NY, USA: Springer, 2007.

[29] R. Plutchik, Emotion: A *Psychoevolutionary Synthesis*. New York, NY, USA: Harper & Row, 1980.

[30] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 1136–1139.

[31] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[32] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro, "How useful are your comments?: Analyzing and predicting Youtube comments and comment ratings," in *Proc. Int. World Wide Web Conf.*, 2010, pp. 891–900.

[33] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Proc. ACM SIGIR, Workshop Crowdsourcing Search Eval.*, 2010, pp. 4–8.

[34] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr.–Jun. 2012.

[35] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 117–120.

[36] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, 1999, pp. 246–253.

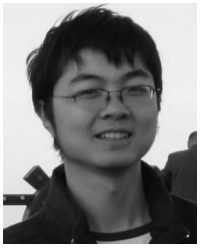
[37] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[38] Y.-H. Yang and J.-Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1304–1315, Oct. 2013.

[39] J. Yuan, Q. You, S. McDonough, and J. Luo, "Sentribute: Image sentiment analysis from a mid-level perspective," in *Proc. Workshop Sentiment Discov. Opin. Min.*, 2013.

[40] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in *Proc. ACM Int. Conf. Inform. Knowl. Manage.*, 2006, pp. 43–50.

Yan-Ying Chen received the PhD degree in computer science at National Taiwan University, Taipei, Taiwan, in 2014. She is currently a research scientist at FX Palo Alto Laboratory, Palo Alto, CA. Previously, she was a visiting researcher in the DVMM Lab at Columbia University, New York, NY, and a postdoc researcher in the Institute of Information Science at Academia Sinica, Taiwan. Her research interests include social media analytics, affective computing, multimedia data mining and retrieval.



Tao Chen received the BS degree in fundamental science class and the PhD degree in computer science from Tsinghua University, Beijing, China in 2005 and 2011, respectively. He is currently a postdoctoral researcher in Department of Electrical Engineering, Columbia University, New York, NY. His research interests include multimedia, computer graphics and computer vision. He received the Netexplorateur Internet Invention Award of the World in 2009, and China Computer Federation Best Dissertation Award in 2011.

Taikun Liu received the BE degree in software engineering from Fudan University, Shanghai, China, in 2013, and the MSc degree in computer science from Columbia University, New York, NY, in 2014. She is currently working toward the Master's degree at Columbia University, New York, NY. Her research interests include image processing, computer vision.

Hong-Yuan Mark Liao received the PhD degree in electrical engineering from Northwestern University in 1990. In July 1991, he joined the Institute of Information Science, Academia Sinica, Taiwan, where he is currently a distinguished research fellow. He is jointly appointed as a professor in the Computer Science and Information Engineering Department of National Chiao-Tung University, Hsinchu, Taiwan. From 2009 to 2012, he was jointly appointed as the Multimedia Information Chair Professor at National Chung Hsing University. Since August 2010, he has been an adjunct chair professor of Chung Yuan Christian University, Zhongli, Taiwan. He received the Young Investigators Award from Academia Sinica in 1998, the Distinguished Research Award from the National Science Council of Taiwan in 2003 and 2010, the National Invention Award of Taiwan in 2004, the Distinguished Scholar Research Project Award from National Science Council of Taiwan in 2008, and the Academia Sinica Investigator Award in 2010. His professional activities include: cochair, 2004 International Conference on Multimedia and Exposition (ICME), technical cochair, 2007 ICME, Editorial Board member, IEEE Signal Processing Magazine (2010-present), associate editor of the *IEEE Transactions on Image Processing* (2009-present), *IEEE Transactions on Information Forensics and Security* (2009-2012), and *IEEE Transactions on Multimedia* (1998-2001). He has been a fellow of the IEEE since 2013.

Shih-Fu Chang received the BS degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1985, and the MS and PhD degrees in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, in 1991 and 1993, respectively. He is the Richard Dicker Professor in the Departments of Electrical Engineering and Computer Science, Senior Vice Dean of School of Engineering and Applied Science, and Director of Digital Video and Multimedia Lab at Columbia University. He has made significant contributions to multimedia search, computer vision, media forensics, and machine learning. He has been recognized with ACM SIGMM Technical Achievement Award, IEEE Kiyo Tomiyasu Award, Navy ONR Young Investigator Award, IBM Faculty Award, Recognition of Service Awards from ACM and IEEE Signal Processing Society, and NSF CAREER Award. He and his students have received several Best Paper Awards, including the Most Cited Paper of the Decade Award from *Journal of Visual Communication and Image Representation*. He is a fellow of the IEEE and a fellow of the American Association for the Advancement of Science. He has also served as the Editor-in-Chief for *IEEE Signal Processing Magazine* (2006-2008), and Chair of Department of Electrical Engineering, Columbia University, New York, NY, from 2007 to 2010.