Affect-based Social Image Search Results Summarization

Eun Yi Kim, Eunjeong Ko, and Yaohui Yu

Visual Information Processing Lab., Dept. of Internet & Multimedia Engineering, Konkuk University
1204 New Millennium Building, Konkuk University, 120 Neungdong-ro, Gwangjin-gu
Seoul, South Korea
{eykim, rritty33, yuyaohui}@konkuk.ac.kr

Abstract

Affect-based social image search results summarization is essential for efficiently visualizing search results and re-ranking them. In this paper, we present the selection of representative images based on human affects. The proposed method consists of three main stages: 1) The images are transformed into the affective space using convolutional neural network, which can automatically find the correlation between visual features and certain affective classes; 2) Images are clustered on affective space and then the resulting clusters are ranked based on the proposed three properties – coverage, affective coherence, and distinctiveness; 3) Representative images are selected from top-ranked clusters. The experiments were conducted on Flickr images and showed the effectiveness of the proposed method.

Keywords-Image summarization; Human affects; Social Images; Ranking model; Convolutional Neural Network

I. Introduction

With the increased availability of devices to capture images and the rapid growth of the social network activities, the number of images on the web and social network services (SNS) is exponentially increasing. Thus, there is urgent needs for these images to be organized and accessed in an easy, fast, and efficient way.

One of the paradigms that can be used to overcome these challenges is to summarize image search results, which can provide a quick overview and accurate impression of what a particular scene looks like. These techniques can help to improve user satisfaction in image retrieval, because it enables the user to select a subset of interests within the query.

Therefore, many researchers and companies have developed cluster-based approaches to summarize image search results, where the clustering can be performed on various aspects [1]. In general, various semantic levels can be extracted from the images [2]. For example, the visual features such as color, texture and shape can be used to describe the images. Furthermore, a high-level of semantics is required to cover both physical semantics, such as named objects and persons, and abstract semantics, such as affective mearnings and moods that are associated with a given image. After clustering the images, the representative images are selected as the top ranked images from the top ranked clusters.

Although the representative images can be defined using their physical contents such as colors, viewpoints, and semantics, abstract semantics such as human affects should be also considered. Because even images that have been categorized as the same theme, it can be interpested differently depending on the user moods. When taking pictures and looking at pictures, there are common feelings caused by and shared by contextual knowledge about how people think, feel, and react. Therefore, considering such feelings can be more meaningful and more important when selecting representative images.

Therefore, we suggest performing the image summarization based on common human feelings. To do this, an image should be first described in the affective space. However, judging such affective qualities of images is not an easy task since there is no direct mapping from the image to the affective meanings. To an end of discovering the correlation between affective classes and visual features, some studies employed well-defined knowledge [3] or learning-based methods [4-5]. Former studies were primarily based on analysis obtained from user studies or theory in the affective computing and psychological studies. Other studies were based on learning methods such as multi-layer perceptron (MLP) [4] and support vector machine (SVM) [5] for estimating such correlations. Recently, a convolutional neural network (CNN) has led to the recent successes in traditional object classification task [6]. In addition, it has been applied for image sentiment analysis problem [7]. Thus, we employ the CNN to extract the meaningful features and then estimate some strong relationships between the feature and human affects.

In the proposed method, image summarization is performed using three steps. First, the images collected over SNS are transformed into an affective feature vector using CNN. Secondly, these images are clustered in the affective space. Then, the selected images should be representative of the images of one theme and they should be distinctive from each other. Therefore, we define three prominent properties that an informative summary should satisfy: coverage, affective coherence, and distinctiveness. Based on these, cluster ranking is performed. Finally, some representative images are selected from the top-ranked clusters by sorting the image based on their variance and density from the center and eliminating the redundant images within cluster.

To verify the effectiveness of the proposed method, the experiments were conducted on images collected from Flickr [7-8]. Then, we performed subjective evaluation to assess the summarization quality on the visual summaries. The results are proven that the proposed method guaranteed high representativeness and diversity.

II. Methods

The goal of this work is to automatically summarize the images over SNS. Here, we propose the selection of these images using the affective space. To do this, two essential techniques are required: one is to automatically describe the images using human affects and the other is to identify a diverse set of representative images.

A. Extracting affective features

In this work, we use a convolutional neural network (CNN), which can automatically predict human affects in images through analyzing the visual features only. To do this, we first choose the affective classes that can represent well social images, and then develop the CNN-based classifier to annotate the image using them.

This work employed the affective corpus defined by Mikels [9]. It is expansion of Ekman's six basic emotions, which have been widely used in image sentiment analysis. The corpus is composed of four positive emotions and four negative ones: {amusement, anger, awe, contentment, disgust, excitement, fear, and sadness}.

To annotate images using abovementioned emotions, we developed a CNN-based classifier to annotate a given image as 8-D affective vector. We employ transfer learning scheme, which load the weights of pre-trained network based on large dataset, tune the architecture of the network, and then resume the training using small dataset. This learning scheme has been employed to solve overfitting problem [8, 10].

The architecture of CNN used in this work is based on CaffeNet [6], which has been used in recognizing 1000 object classes. Because the social images can contain various objects and scenes, we select the CaffeNet to analyze correlation between these objects containing the images and human affects. The architecture of the CNN is composed of five convolutional layers and three fully-connected layers. We modify the number of node in last output layer as 8 instead of 1000. Except weight parameters of output layer, all parameters are loaded from the pre-trained CaffeNet model.

To learn the CNN, we first collected public available Flickr and Instagram images that is built by You et al [8]. You et al. provided url links with agreements of the emotions obtained from Amazon Mechanical Turks (AMT) workers. Next, after filtering some images with low agreements, we randomly divided the remaining images into training data (80%) and testing data (20%). Thus, 18K images were employed to fine-tune the CNN model and 4.5K images were used for evaluation of the model, respectively.

B. Ranking clusters

For canonical image selection, clustering is first performed on the affective space, and then new criteria are described to select the clusters that have the informative summary. Based on these properties, the clusters are ranked.

Here, we use affective features to discover clusters of images. For the clustering, several algorithms have been considered such as *K*-means clustering and meanshift clustering. By experiments, the *K*-means clustering was selected.

When a set of cluster is given, they are then evaluated according to how relevant to search query, how diverse they are and how much coverage they provide over the query. Generally, the top ranked clusters are likely to include more representative images. Thus, to select the clusters, new criteria needs to be identified. These properties are coverage, affective coherence, and distinctiveness.

- Coverage: The common concepts present in the query are represented by the amount that they cover. Thus, a cluster is included in the summary if it covers a large number of images. Hence, if num() denotes the number of images that belong to the *k*th cluster, then the coverage for that cluster is as follows:

$$Coverage(c_k) = \frac{num(c_k)}{\sum_{all\ i} num(c_i)}$$

- Affective coherence: It is assumed that the images within one cluster share a common feeling. Therefore, the images within a cluster should be cohensive in aspects of affects. To measure this, we use the within variance of cluster (σ), which is used to penalize clusters that have variances that are too large, even if it has a higher coverage. However, there are trade-offs between coverage and affective coherence, because larger clusters are more likely to be sparsely distributed in an affective space than smaller clusters, thus to have larger variances is inevitable.

$$Coherence(c_k) = 1 - \frac{\sigma(c_k)}{max_{all\ i}\,\sigma(c_i)} \cdot (1 - \left(\frac{Coverage(c_k)}{max_{all\ i}\,Coverage(c_i) + 1}\right)^{\frac{1}{\tau}})).$$

- Distinctiveness: The diversity of a summary is a measure of non-redundancy. Thus, clusters that are similar to each other are not contained in the summary. Accordingly, we use a minimum pairwise distance of the clusters to represent the cluster distinctiveness:

$$Distinct.(c_k) = min_{all\ i} Dist(c_i, c_k).$$

Based on three properties, the ranking model can be defined as follows:

$$Rank\ (c_k) = \{\omega_\alpha \cdot Coverage(c_k) + \omega_\beta \cdot Coherence(c_k)\} + \omega_\gamma \cdot 10^3 \cdot Distinct.(c_k).$$

Compared with the terms within a cluster and the term between clusters, scaling is necessary. Thus, 10^3 is multiplied in the third term. This equation provides a score to rank each cluster.

C. Finding canonical images

We generated image summarization through sampling photos according to the ranked order of clusters. For this, we ranked the images within each cluster according to how well they represented the cluster. To measure the representativeness of images, we consider two distance measure together: (1) the distance from the respective image to cluster center μ_k, and (2) the number of images that are adjacent to the selected image within a specified diameter $N_{k-distance(i)}$used in [11]. Thus, the images in a cluster are ranked as follows:

$$Rank(\mathrm{i}) = (|i - \mu_k|)^{-1} \times N_{k-distance(i)}$$

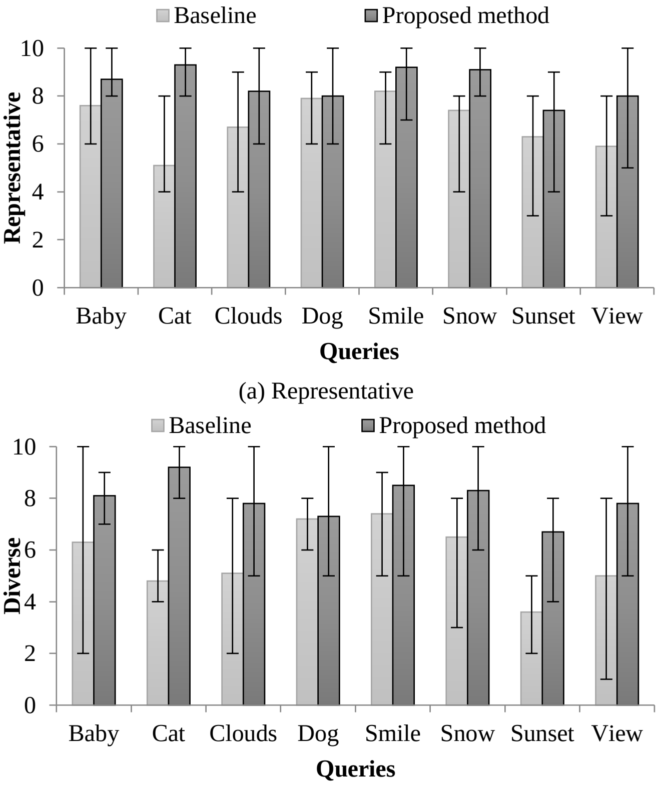
Among the images selected from the respective clusters, we eliminated the redundant images that has small L_2-norm distance between images, and then generate the final summarization.

III. Evaluation

To evaluate the performance of summarizing the images, experiments were conducted on Flickr dataset built by Chen et al. [7] containing 486K images, which are labeled by 1553 adjective-noun pairs (ANPs). These ANPs are pairs of 214 adjectives and 357 nouns. Among these ANPs, we first selected eight nouns as queries: baby, clouds, dog, smile, snow, sunset, and view.

It is difficult to evaluate image summarization method due to the absence of ground truth. Therefore, we evaluated the summarization quality through on user study. In addition, we compared the performance of the proposed method with other method, which uses *K*-means clustering and employs clustering ranking and image evaluation method proposed in [12]. We asked 10 users (average age of the participants is 27) to observe the generated summarization results and rate its quality for every query.

We qualitatively measured the summarization results: How many images in this set are representative of the query (1-10 scale)? And How many images photos in this set are redundant (0-10 scale)? Figure 1 presents a definite subjective evaluation results according to the queries. Although some differences were found according to the queries, the proposed method exhibited the improved performance.



(a) Representative

(b) Diversity

Fig. 1 Subjective evaluation results for the baseline and the proposed method: (a) representativeness and (b) diversity.

Figure 2 shows the examples of summarized images produced using the baseline and the proposed method for five queries 'clouds, 'dogs', 'smiles', 'snows', and 'view'. When visually inspected, the results guaranteed good relevancy; however, the summarized results produced by the baseline contain irrelevant images in 'dogs' and 'clouds'.

Overall, the users who participated in the evaluation gave high representative and diverse scores with less variation between participants.

Acknowledgment

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the University Information Technology Research Center support program (IITP-2016-R2720-16-0004) supervised by the IITP(Institute for Information & communications Technology Promotion).

References

[1] N. Ben-Haim, B. Babenko, and S. Belongie, "Improving web-based image search via content-based clustering," In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop* (CVPRW'06), pp. 1-6, 2006.

[2] A. W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intelli.* Vol. 22(12), pp. 1349-1380, 2000.

[3] M. Solli and L. Reiner, "Emotion related structures in large image databases," In *Proceedings of the ACM international Conference on Image and Video Retrieval* (CIVR'10), pp. 398-405, 2010.

[4] Y. Shin, Y. Kim, and E. Y. Kim, "Automatic textile image annotation by predicting emotional concept from visual features," *Image Vis. Comput.* Vol. 28 (3), pp. 526-537, 2010.

[5] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Studying aesthetics in photographic images using a computational approach," In *Proceedings of the European Conference on Computer Vision* (ECCV'06), pp. 288-301, 2006.

[6] Y. Jia, E. Shelhamer, J. Donahue, and S. Karayev, "Caffe: Convolutional architecture for fast feature embedding," In *Proceedings of the 22nd ACM international Conference on Multimedia* (MM'14), pp. 675-678, 2014.

[7] T. Chen, D. Borth, T. Darrell, and S.F. Chang, "DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks," arXiv preprint arXiv:1410.8586, 2014.

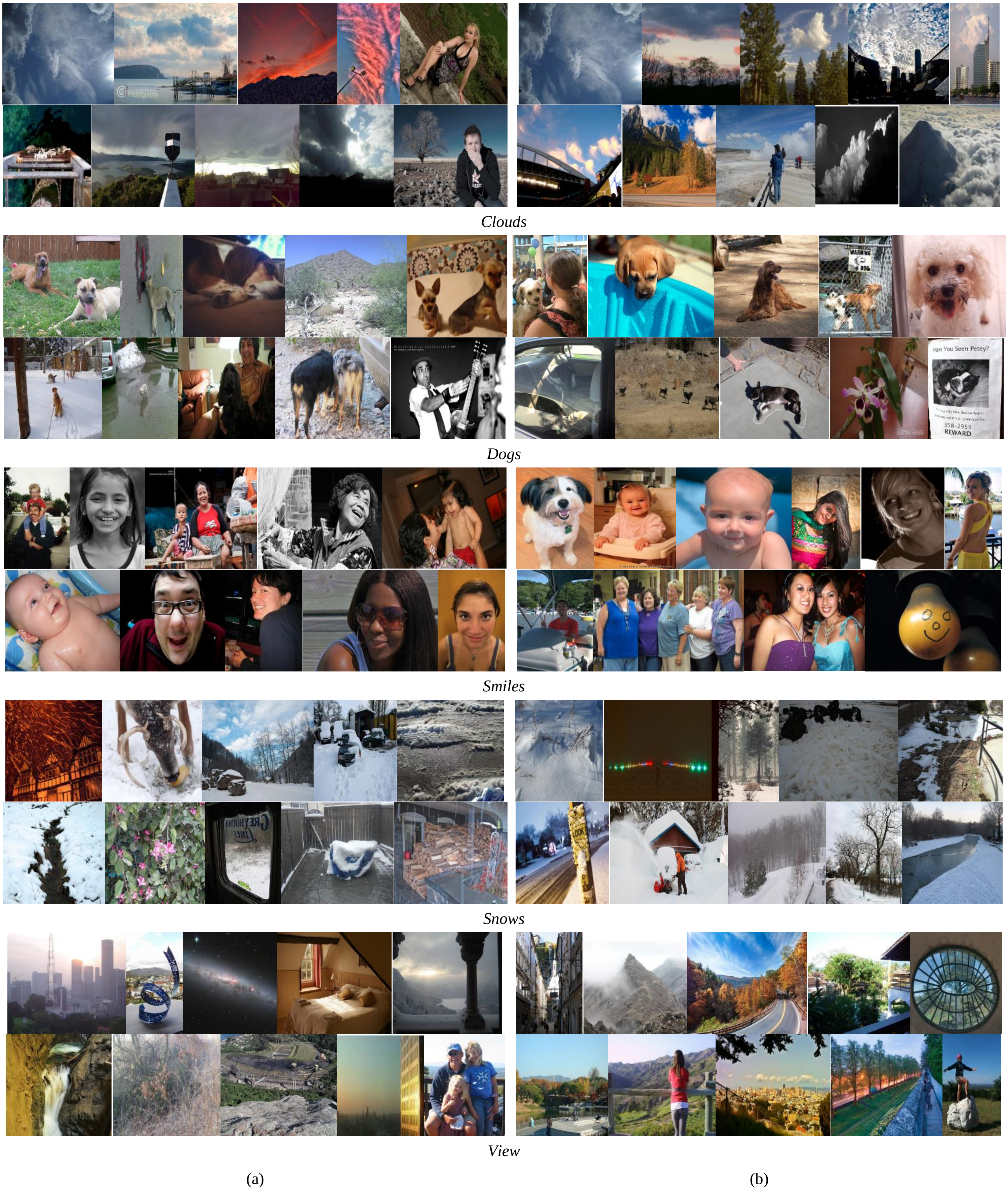
[8] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: the fine print and the benchmark,". In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI'16), pp. 308-314, 2016.

[9] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior Research Methods* vol. 37, pp. 626-630, 2005.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," In *Proceeding of Advances in Neural Information Processing Systems* (NIPS'14), pp. 3320-3328, 2014.

[11] M. M. Breuning, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," In *Proceedings of the ACM SIGMOD international Conference on Management of Data* (SIGMOD/PODS'00), pp. 93-104, 2000.

[12] Y. Jing, S. Baluja, and H. Rowley, "Canonical image selection from the web," In *Proceedings of the 6th ACM international Conference on Image and Video Retrieval* (CIVR'07), pp. 280-287, 2007.



Clouds

Dogs

Smiles

Snows

View

(a)

(b)

Fig. 2 Sets of representative images produced using (a) baseline and (b) proposed method for five queries.