Facebook5k: A Novel Evaluation Resource Dataset for Cross-Media Search

Sadaqat ur Rehman[1], Yongfeng Huang[1], Shanshan Tu[2](✉), and Obaid ur Rehman[3]

[1] Tsinghua National Laboratory for Information Science and Technology, Beijing, China
z-sun15@mails.tsinghua.edu.cn, yfhuang@tsinghua.edu.cn

[2] Beijing University of Technology, Beijing, China
sstu@bjut.edu.cn

[3] Sarhad University of Science and IT, Peshawar, Pakistan
obaid.ee@suit.edu.pk

Abstract. Semantic concepts selection for model construction and data collection is an open research question. It is highly demanding to choose good multimedia concepts with small semantic gaps to facilitate the work of cross-media system developers. Since, this work is very scarce therefore; this paper contributes a new real-world web image dataset created by NGN Tsinghua Laboratory students for cross media search. Unlike previous datasets, such as Flicker30k, Wikipedia and NUS have high semantic gap, results in leading to inconsistency with real time applications. To overcome these drawbacks, the proposed Facebook5k dataset includes: (1) 5130 images crawled from Facebook through users feelings; (2) Images are categorized according to users feelings; (3) Facebook5k is independent of tags and language, rather than uses feelings for search. Based on the proposed dataset, we point out key features of social website images and identify some research problems on image annotation and retrieval. The benchmark results show the effectiveness of the proposed dataset to simplify and improve general image retrieval.

Keywords: Cross-media retrieval · Facebook5k dataset
Semantic concepts

1 Introduction

Current era has observed a rapid growth of Multimedia Information Retrieval (MIR). Regardless of constant hard work in the development and construction of new MIR techniques and dataset respectively, the semantic gap is high between images and high-level concepts. In order to reduce the semantic gap, we need a promising paradigm to focus on modeling high-level semantic concepts, either by object recognition or image annotation. This kind of concept-based multimedia search system has been presented into numerous real-world search systems [17–19]. Among various approaches, the first step is dataset selection with high-level

X. Sun et al. (Eds.): ICCCS 2018, LNCS 11063, pp. 512–524, 2018.
https://doi.org/10.1007/978-3-030-00006-6_47

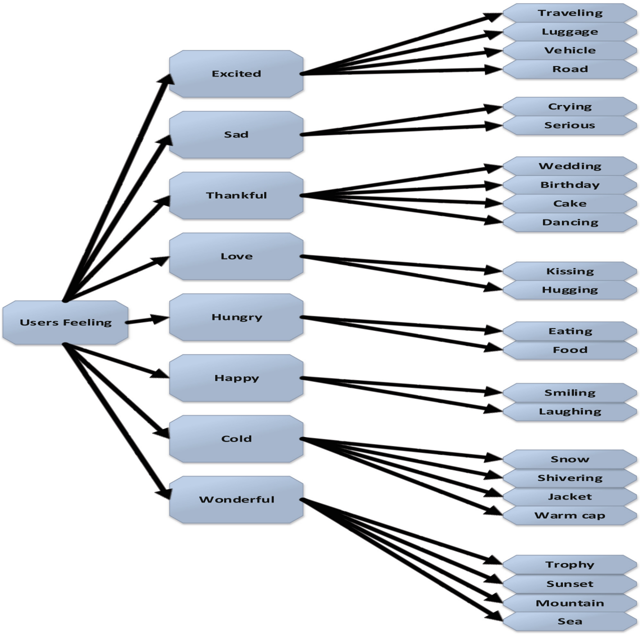

Fig. 1. Concept taxonomy of Facebook5K

concepts and small semantic gaps that is relatively easy for machines understanding and training.

However, existing cross-media datasets totally ignored these issues. For example, firstly, they have shortcoming in media types and categories. Such as, Wiki dataset [2] contains only two types of media (image and text). Similarly, the Pascal VOC 2012 dataset [1] has only 20 different classes. However, cross-media retrieval implicates numerous domains under real-world Internet conditions. Cross-media retrieval systems trained on scanty domain datasets have complications in handling queries from anonymous domain. Second, they lack context information such as link relations. Such context information is quite accurate, and provides significant evidences to improve cross-media retrieval system accuracy. Third, popular cross-media datasets have small sizes, such as Xmedia, IAPR TC-12 dataset [3] (20,000 samples) and the Wiki dataset [2] (2,866 samples). The lack of appropriate data makes difficulties for retrieval systems learning to evaluate the robustness in real-world galleries. Fourth, datasets i.e. ALIPR [4], SML [5], either just used all the image annotation keywords associated with training images, or unenforced any constraint to the annotation vocabulary for example ESP [6], LabelMe [7], and AnnoSearch [8]. Therefore, these datasets essentially disregard the alterations among keywords in terms of semantic gap. A brief summary of some cross-media datasets is provided in Table 1.

Although, there is no doubt these efforts provide significant contribution to cross-media research community in terms of standardization of concept corpus thus open the gateway for researchers to focus ongoing work on a well-defined set of semantics. Nevertheless, we suggest that semantic gap is non-uniform in a low-level feature space realistically and neglecting semantic gap differences is inappropriate. For example, it is well known that modeling broad theme i.e. "Asia" is more challenging than modeling specific theme i.e. "sky" due to absence of significant visual feature that can represent the concept of "Asia". In addition, researchers typically choose local features and color features to model concepts like "sky", and "sunset" respectively.

Motivated by this, this paper focus on the below key issues: Are tags cover the whole scenario in an image? How can cross-media search benefit from users uploading images online? How the contents of image varies according to different users feelings? In other words, how to choose semantic concepts that can be better modeled and simply annotated? To address these problems, this paper makes two major contributions. First, collection of a new large-scale cross-media dataset, named Facebook5k. It contains 5130 image-feelings pairs collected from Facebook[1]. It differentiate itself from the current datasets in two aspects: varied domains and rich context information. Eventually, it provides a more realistic benchmark for cross-media study. Therefore, we construct a standard dataset keeping in mind the research issues to focus research efforts on cross-media retrieval algorithm development rather than the laborious compared methods and results. Second, to the best of our knowledge, it is the first effort to collect a huge dataset of high-level concepts with small semantic gaps on user's semantic descriptions.

Social networking websites i.e. Facebook, Instagram[2], Flicker[3] etc. provides images with rich textual features [9]. Usually these textual features are very close to the semantics of the images i.e. objects name, locations, landmarks, or people present, which helps the users to retrieve relevant images within photo sharing websites using a simple text-based search. According to recent survey [10,20], in every minute more than 2,000 images are uploaded to Flickr, which rises to 12,000 images per second during peak hours. When users upload images to different social websites, they usually categorized the contents of various images through semantic descriptions.

The rest of the paper is organized as follows: Sect. 2 describes the proposed dataset characteristics, collection, potential application and some example images from the proposed dataset. The diversity of the proposed dataset is presented in Sect. 3. Ground truth selection and training set construction are described in Sects. 4 and 5 respectively. Noise level estimation of Facebook5k and its pre-processing, including expert annotations are described in Sect. 6. Section 7 describe some opens discussion. Finally, we conclude the paper and provide some useful future direction in Sect. 8.

[1] facebook.com.

[2] instagram.com.

[3] flickr.com.

2 Proposed Dataset

This section describe a new dataset called Facebook5k, which is comprised of 5130 images collected from Facebook. The complete Facebook5k dataset will be available on NGN[4] soon.

2.1 Dataset Collection

Each step in the dataset collection is briefly explained below.

Seed User Gathering. In order to obtain the real emotion of user associated with the image rather than image contents, we obtain seed users by sending queries to Facebook with various key words, i.e. happy hungry, love, etc.

User Candidate Generation. For this purpose, we develop a web spider to crawl the accounts of the users who are following the seed users. This step repeats a number of times until we get a long list of user candidates.

Feelings Collection. Another web spider collects feelings as a text associated with the corresponding images by visiting the homepages of different users present in the candidate list. We find that about 80% of the users feelings companied images.

Data Pruning. We pruned the data that justify any of the following situations, and is called garbage data.

- Feelings without images;
- Tweets not associated with images or feelings;
- Repeated images with same ID;
- Error images.

In result, we obtain 5130 image-feelings pairs in total. An image and feeling text appearing in one piece of tweet are considered as a pair. Some examples in this new dataset are presented in Fig. 3, 4, 5, 6, 7, 8, 9 and 10.

2.2 Dataset Characteristics

Dataset play a key role in the evaluation of cross-media retrieval methods. The proposed dataset includes a set of images highly associated with users feelings. These images are crawled from Facebook, along with users associated feelings. The Facebook5k dataset is highlighted as: First, since it is collected from social media website, hence it covers a broad range of domains under a single roof of feelings, such as love, hungry, thankful etc.

[4] http://ngn.ee.tsinghua.edu.cn/.

Second, the relationship between the image and user feeling is often very strong. In the example, first row of Fig. 3, image has a strong knot with the associated feeling. Such is the case in realistic scenario.

Third, Facebook5k is a large-scale dataset, containing 5130 image-feelings pairs, which helps to evade overfitting in system training. Furthermore, it helps the system to test the robustness of cross-media retrieval techniques under a wealth of data.

Fourth, it helps to reduce the semantic gap by providing more accessible visual content descriptors using high-level semantic concepts.

To our knowledge, this is the first dataset collected from Facebook of high-level concepts with small semantic gaps on user's semantic descriptions, and ground-truth of 24 concepts for the whole dataset. Also we believe that this is the only cross-media dataset comprising the above mentioned characteristics.

Table 1. Multimodel datsets summarization

Dataset	Modality	No. of samples	Image features	Text feature	Categories
Wiki	Image-text	2, 866	SIFT+BOW	LDA	10
NUS-WIDE	Image/tags	186,577	6 types	Tag occurrence feature	81
Pascal-VOC	Image/tags	9,963	3 types	Tag occurrence feature	20
Flickr30K	Image/sentences	31,783	-	-	-
Twitter-100K [12]	Image-text	100,000	-	-	-
INRIA-Websearch	Image-text	71,478	-	-	353

2.3 Potential Application Scenario

The proposed dataset provides more practical standard for cross-media retrieval. The potential application scenarios are detailed below.

- Social media website such as Facebook provide predefined emoticons for users to choose when posting a tweet. These emoticons are highly correlated with the posted image and user interpretation of the image. Hence, it is more useful and interesting to link the range of emoticons with users images and recommend suitable image for users according to his/her feelings about the contents of the posts.
- Social network addiction produced huge amount of multimodal data in the internet, which is roughly organized and its annotations are time-consuming and expensive. Apparently, labeling such large-scale multimodal data is challenging. Adding users feeling to images can improve the learning rate of semantic correlations among multimodal data.

2.4 Example Images

The dataset include images of range of feelings i.e. happy (Fig. 3), sad (Fig. 4), wonderful (Fig. 5), cold (Fig. 6), hungry (Fig. 7), love (Fig. 8), excited (Fig. 9) and thankful (Fig. 10).

3 Diversity of the Image Collection

The Facebook5k comprises numerous images of like pose but varying illumination, viewing angle and background. The reason is, most of the images uploaded on social websites with same feeling have similar visual content. For example, "smile" must be common among different people having the feeling of happiness as shown in Fig. 3. However, the viewing angle (Fig. 13), background (Fig. 14) and time schedule (Fig. 15) varies. Therefore, this makes the standard Facebook5k compatible for content-based retrieval tasks as it permits a variety of exemplary quests to explore the efficiency of retrieval systems with these fluctuating settings.

4 Ground Truth for 8 Feelings

In order to analyze the usefulness of research work conducted on Facebook5k, we manually annotate the ground-truth for eight different categories, as described in Fig. 1. Regarding annotations, we have undertaken several rounds of proof analysis by co-authors and external colleagues. Hence, we carefully selected the 8 different feelings in such a way that: (a) they are not inconsistent with the concepts defined in [13–16]. (b) They mainly correspond to the common feelings in Facebook. (c) They give clear evidence of users general perception regarding input image. (d) They belong to different classes comprising happy, sad, excited, wonderful etc.

Since, annotation is challenging task, keeping in mind the following guidelines. If a desire concept exist in the image, label it as positive; if the concept does not exist in the image, or if the annotator is ambiguous regarding the concept, then label it as negative. The number of relevent images for individual users feeling is shown in Fig. 2.

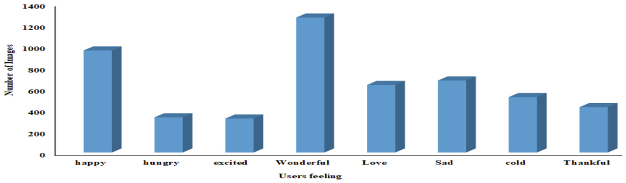

Fig. 2. Facebook5K dataset total images for different users feeling

The key feature about the proposed dataset is that we manually annotate all the images therefore, the chances of error is very less. To ease researchers in system development, we divided the proposed dataset into two parts i.e. testing and training. 4000 images to be used for training and the remaining 1130 images for testing.

5 Training Set Construction

This is the most important phase of dataset construction as the efficiency of the learning system totally depend on it. In other words, we need to construct an effective training set for each concept that we want to learn. Training set of each associated target concept must meet the fundamental two properties. (1) The label must be reliable for individual concept of each image. (2) The training samples must possess the properties to cover the entire feature space of the original dataset [11].

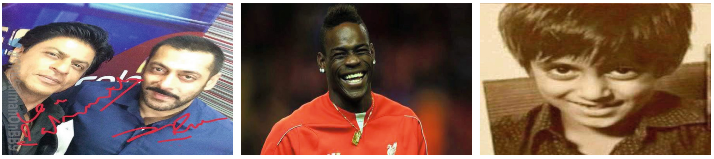

Fig. 3. Feeling happy examples (smiling, laughing)

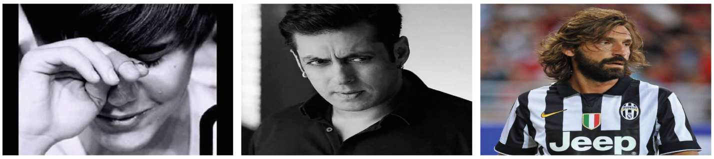

Fig. 4. Feeling sad examples (crying, serious, loose)

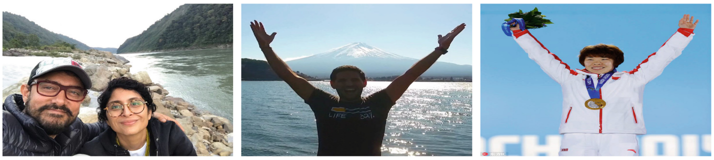

Fig. 5. Feeling wonderful examples (mountain, river, medal)

6 Noise in the Dataset

The key questions arise here are; is the concept of feelings set by the users for associated image possess appropriate features to train intelligent systems for concept detection/classification? What is the quality of associated concept set by the users? Which type of concept can be chosen for accurate detection? To address these questions, we calculate the noise-level of Facebook5k. We simply calculate the precision and recall of the associated feelings in light of ground-truth for 24 different concepts as shown in Fig. 11. It is clear from Fig. 11 that

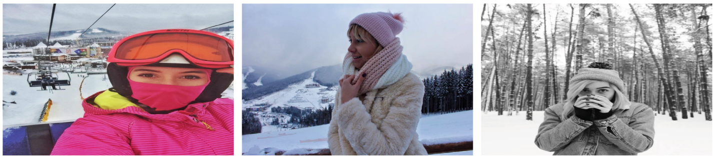

Fig. 6. Feeling cold examples (cap, shivering, jacket, snow)

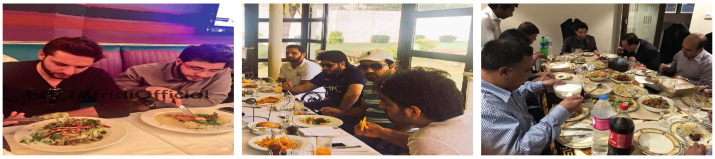

Fig. 7. Feeling hungry examples (food, person)

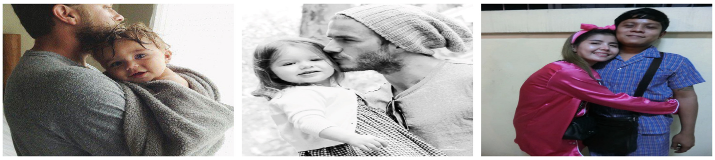

Fig. 8. Feeling love examples (kissing, hugging)

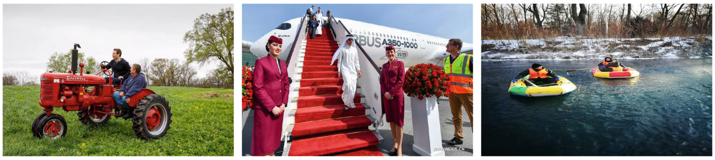

Fig. 9. Feeling exited examples (traveling, luggage, vehicle, road)

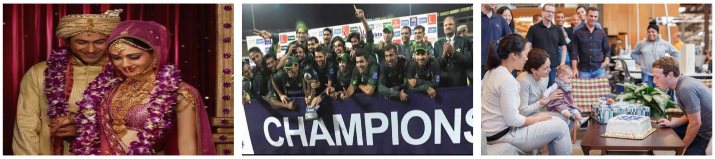

Fig. 10. Feeling thankful examples (wedding, birthday, cake, dancing)

both the average precision and average recall of the original feelings are about 0.75, explicitly, one quarter of the feelings are noisy. Here we define F score as a level of noise measurement:

$$Noise\ level = 1 - F \tag{1}$$

$$F = \frac{2 \times precision \times recall}{precision + recall} \tag{2}$$

An annotated keyword is considered correct subject to its appearance in the ground truth annotation of the target image. We define the Precision and Recall mathematically in Eqs. (3) and (4).

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

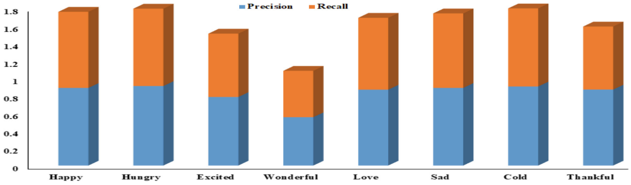

Fig. 11. Precision and recall for 8 different users feelings

where TP (True Positive) represent the total number of positive samples, FP (False Positive) represent the negative samples predictive to be positive, FN (False Negative) represents the number of positive samples predicted to be negative and TN (True Negative) represents the number of negative samples predicted to be negative. Noise of the original feelings for different concepts are shown in Fig. 12.

To improve calculation effects of the number of positive samples and noise level for each concept, we perform the annotation by getting help from expert image scientists as the benchmark for non-tagged image annotation. Hence, we can observe that both the number of positive samples in the dataset and the noise level of the target concept affect the annotation performance.

The number of positive samples influence the results positively. Average precision is directly proportional to the number of positive samples for a firm target concept, i.e. average precision for a target concept is increasing by increasing the positive samples and vice versa. Example of such concepts are “mountain”, “grass”, “car” and “road”. However, the noise level has adverse effects on the outcomes. The noise level increases as the amount of semantic gap of the target concept increases [9].

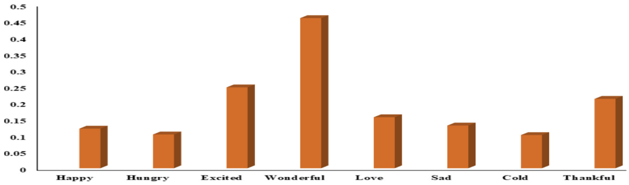

Fig. 12. The noise level for 8 different users feeling

7 Discussion

In this section we discuss that why the proposed dataset is important in cross-media retrieval? We constructed the Facebook5k dataset, keeping in mind the broad-spectrum cover in a single image. Since, a single image unveil thousand words therefore, we feel the need of such dataset, which has a strong knot with the users description. We introduced eight different feelings with 24 concepts, which cover many important aspects of daily life. We pick some key features from users feeling which are described below:

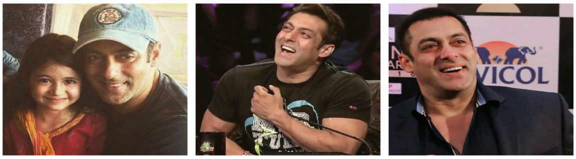

Fig. 13. Same users feeling from different viewing angles

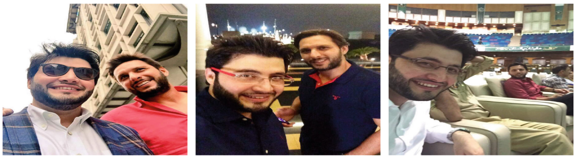

Fig. 14. Same users feeling with different background

Feeling Happy: Users normally share posts with smiling faces with this status. However, there are many cases when the users upload images with a diverse effect.

Feeling Cold: Shivering, warm hat, jacket are the common tags associated with this kind of posts.

Feeling Hungry: Normally, users post this kind of tweets from a restaurant or hotel while taking dinner, breakfast or lunch.

Fig. 15. Users feelings captured in different time span i.e. in the morning, during the day and at night

Feeling Excited: This is a broad feeling however, the data we pruned for this kind of feeling shows users excitement about traveling from one place to another, first time experience and getting into a new place/job. However, many other images have different scenario with the same feeling.

Feeling Love: This is very special type of feeling as it come with users hugging or kissing. People link this feeling with images when they kiss or hug their love one or pets.

Feeling Wonderful: This kind of feeling engulf a large spectrum due to its generalization. Therefore, we capture more than 1250 picture only for this feeling inorder to make it easy for machine understanding.

Feeling Thankful: Feeling thankful is gratitude. Users can feel grateful for everything and anything. Therefore, normally posts associated with this kind of feeling have marriages, birthday, festivals etc.

Feeling Sad: As the name suggests, images associated with this kind of feelings have crying or serious poses.

8 Conclusion and Future Work

We tried to construct a dataset having high-level concepts with small semantic gap. Among many other multimedia datasets constructed, it is the first efforts, which concentrate on semantic concepts for data collection. The ground-truth annotation of 24 different concepts have many potential applications in concept detection, query optimization and multimedia information retrieval. This dataset can be used for the assessment of users image relationship and multi-label image classification, particularly with the use of visual and text features. Furthermore, we discussed some open research questions and delivered the standard solution.

However, much effort need to be done in future since there is no perfect system ever evolved yet. In the future, we plan to increase the number of images and users feeling for Facebook5k. Also, we plan to design an effective learning method for this dataset.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (No. U1405254, U1536115, U1536207).

References

1. Hwang, S.J., Grauman, K.: Reading between the lines: object localization using implicit cues from image tags. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1145–1158 (2012)
2. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieva. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 251–260 (2010)
3. Grubinger, M., Clough, P., Müller, H., Deselaers, T: The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In: International Workshop Ontoimage, vol. 5 (2006)
4. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 985–1002 (2008)
5. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 394–410 (2007)
6. Von Ahn, L., Dabbish, L: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326. ACM (2004)
7. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**, 157–173 (2008)
8. Wang, X.-J., Zhang, L., Jing, F., Ma, W.-Y.: Annosearch: image auto-annotation by search. In: IEEE computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1483–1490. IEEE Press, New York (2006)
9. Lu, Y., Zhang, L., Tian, Q., Ma, W.-Y.: What are the high-level concepts with small semantic gaps? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Press, New York (2008)
10. Peng, Y., Huang, X., Zhao, Y.: An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. IEEE Trans. Circuits Syst. Video Technol. **28**(9), 2372–2385 (2018)
11. Tang, J., Song, Y., Hua, X.-S., Mei, T., Wu, X.: To construct optimal training set for video annotation. In: Proceedings of the 14th ACM International Conference on Multimedia, pp. 89–92. ACM (2006)
12. Hu, Y., Zheng, L., Yang, Y., Huang, Y.: Twitter100k: a real-world dataset for weakly supervised cross-media retrieval. IEEE Trans. Multimed. **20**, 927–938 (2017)
13. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. J. Mach. Learn. Res. **3**, 1107–1135 (2003)
14. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. **106**, 59–70 (2007)
15. Naphade, M., et al.: Large-scale concept ontology for multimedia. IEEE Multimed. **13**, 86–91 (2006)
16. Snoek, C.G.M., Worring, M., Van Gemert, J.C., Geusebroek, J.-M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th ACM International Conference on Multimedia, pp. 421–430. ACM Press (2006)

17. Lu, Y.-J., Nguyen, P.A., Zhang, H., Ngo, C.-W.: Concept-based interactive search system. In: Amsaleg, L., Guðmundsson, G.Þ., Gurrin, C., Jónsson, B.Þ., Satoh, S. (eds.) MMM 2017. LNCS, vol. 10133, pp. 463–468. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51814-5_42
18. Kambau, R.A., Hasibuan, Z.A.: Concept-based multimedia information retrieval system using ontology search in cultural heritage. In: Second International Conference on Informatics and Computing (ICIC), pp. 1–6. IEEE Press, New York (2017)
19. Kambau, R.A., Hasibuan, Z.A.: Evolution of information retrieval system: critical review of multimedia information retrieval system based on content, context, and concept. In: 11th International Conference on Information & Communication Technology and System (ICTS), pp. 91–98. IEEE Press, New York (2017)
20. Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C.G.M., Bimbo, A.D.: Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval. ACM Comput. Surv. (CSUR) **49** (2016)