Making use of Semantic Concept Detection for Modelling Human Preferences in Visual Summarization

Stevan Rudinac
ISLA, Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
s.rudinac@uva.nl

Marcel Worring
ISLA, Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

ABSTRACT

In this paper we investigate whether and how the human choice of images for summarizing a visual collection is influenced by the semantic concepts depicted in them. More specifically, by analysing a large collection of human-created visual summaries obtained through crowdsourcing, we aim at automatically identifying the objects, settings, actions and events that make an image a good candidate for inclusion in a visual summary. Informed by the outcomes of this analysis, we show that the distribution of semantic concepts can be successfully utilized for learning to rank the images based on their likelihood of inclusion in the summary by a human, and that it can be easily combined with other features related to image content, context, aesthetic appeal and sentiment. Our experiments demonstrate the promise of using semantic concept detectors for automatically analysing crowdsourced user preferences at a large scale.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*selection process*

General Terms

Algorithms; Human Factors; Experimentation

Keywords

User-informed visual summarization; crowdsourcing; user studies; social media; semantic concepts; learning to rank

1. INTRODUCTION

Due to the increasing popularity of content-sharing and social networking platforms, recently a number of approaches for visual summarization of community-contributed images have been proposed. The approaches presented so far, such as [3, 13, 8], have commonly been guided by the results of well-known user studies [4], suggesting a need for a trade-off between relevance, representativeness and diversity criteria. For example, Kennedy and Naaman present an approach to automatically selecting diverse and representative images of landmarks [8]. In the context of location recommendation, Cao et al. [3] generate visual summaries by first clustering Flickr images based on their geo-coordinates and then represent each location with the most representative tags and images. To a similar end, in [13], the authors mine user-generated travelogues to discover the most representative tags for a particular location and further use them for identifying the most relevant and representative location-specific images. With regard to image search diversification, the research efforts of the community were mainly gathered around ImageCLEFPhoto [10] and, more recently, MediaEval Diverse Images benchmarks [7].

Figure 1: Example images captured in the area around Louvre Museum in Paris. Images in the upper row are more frequently selected for the visual summary by the users. All images are downloaded from Flickr under a Creative Commons license.

For user-centric summarization we conjecture that the notions of relevance, representativeness and diversity are too general to successfully model user information needs in a given summarization scenario. For example, a recent large-scale study on user preferences in visual summarization [14] suggests that a set of criteria that should be taken into account when designing user-centric visual summarization algorithms may be much wider than commonly assumed. Relevant factors revealed by the study include image aesthetic appeal, popularity and the sentiment images evoke in the users. The authors of the study further propose an approach for directly mapping those criteria onto features and deploying them for learning to discriminate between images based on their suitability for visual summarization as judged by the users. Here, we go a step further and investigate to what degree the user's choice of images for a visual summary

CrowdMM'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2660114.2660127.

is influenced by the objects, settings, actions and events (i.e., semantic concepts) they are depicting.

The use of semantic concepts in automatic sentiment analysis from visual content was recently investigated by Borth et al. [1]. Similarly, Khosla et al. investigate the possibilities of deploying content analysis for predicting image popularity on a social media platform [9]. These studies further motivate our assumption about a need for digging into image semantics for better understanding user preferences.

So, how do we obtain information on the relation between concepts and summaries at the scale of tens of thousands of images and hundreds or thousands of potential concepts? Crowdsourcing, which recently emerged as an effective tool for conducting user studies on a large scale would be an option. In case of visual summarization, the crowdworkers may be asked to provide the reasons for selecting individual images using a free-form text input. However, the users are often either unable or not inclined to clearly explain their choice, which makes qualitative and quantitative analysis of such feedback particularly challenging. In this paper we investigate the possibilities of using semantic concept detectors for automatically analysing user preferences at the level of image semantics.

Based on the learned preferences, we analyse whether those can be mapped to a concept-based image representation that can further be used for automatically identifying images likely to be selected for a visual summary by a human (cf. Figure 1).

In Section 2 we describe the dataset used for the experiments. Then, in Section 3 we present the results of our user preference analysis and in Section 4 we describe our approach to preference learning. The experimental results are presented in Section 5.

2. HUMAN-CREATED SUMMARIES

For the study we make use of the large collection of human-created visual summaries introduced in [14], which we briefly describe here. The dataset is created by first selecting 207 geographic locations in Paris, France, output by a location recommender system and then by downloading 100 Creative Commons licensed Flickr images, captured within a radius of 1 km from each of them. The images are downloaded together with various user-generated and automatically captured metadata. To provide a realistic "snapshot" of a social media platform and ensure generalizability of the experiments, images are not pre-selected based on the type or topic and therefore depict a wide spectrum of user interests.

For each of the collection subsets (i.e., geographic locations), 20 unique Mechanical Turk workers were shown a set of 100 images and asked to select 10 of them for the visual summary that should capture the essence of the larger image set. They were further asked to order images according to their importance and provide a reason for selecting each of them using a free-form text field. This additional information about selected images was used for gaining a better insight into user preferences and devising a more effective spam detection mechanism. Finally, the workers were asked several additional questions about the properties of the original image set and their impressions on the task complexity. Since some of the workers completed the task for more than one collection subset, the experiment included 697 unique participants, which created a total of 4140 unique reference visual summaries.

Table 1: Semantic concepts associated with the images most frequently selected by the humans; the percentage of locations for which a particular semantic concept has a higher confidence in case of the positive image examples is reported

Concept	%	Concept	%
Sky	85	Pan_Zoom_Static	77
Clouds	84	Waterscape_Waterfront	76
Daytime_Outdoor	83	Canoe	76
Eukaryotic_Organism	83	Weather	76
House_Of_Worship	81	Religious_Building	75
Flowers	80	Traffic	75
City	79	Rocky_Ground	74
Fields	79	Tent	74
Food	79	Vegetation	74
Urban_Park	79	Church	73
Outdoor	78	Rowboat	73
Cityscape	78	Sunny	73
Fighter_Combat	78	Lakes	73
Hill	78	Raft	73
Landscape	78	Highway	72

3. USER PREFERENCE ANALYSIS

As discussed in Introduction, we conjecture that, from the user's point of view, the presence of specific semantic concepts makes certain images more suitable for visual summarization. To further investigate this hypothesis, we first analyse the results of the crowdsourcing experiment and for each collection subset rank the images based on their frequency of occurrence in human-created reference summaries. Each image in the collection is represented with a *concept vector*, obtained by applying 346 semantic concept detectors from the TRECVID 2012 Semantic Indexing Task [15], where the individual elements of the vector give a confidence of concept presence. Then, the joint representations for "positive" and "negative" image candidates are generated by applying the average pooling [2] on concept vectors associated with the top-N and bottom-N images in the ranked list.

In the following, we investigate which concepts and concept pairs are frequently associated with the images chosen for the visual summary by the humans.

3.1 Individual Concepts

For each collection subset we identify the semantic concepts having higher confidences in case of the positive image candidates as compared to the negative candidates. In Table 1 we show top-30 concepts associated with the positive image examples in the highest percentage of locations.

The results presented in Table 1 clearly suggest users' preference towards "panoramic" images showing the cityscape and individual buildings with the skyline in the background. Further, the users seem to prefer *daytime* to *nighttime* scenes and *outdoor* to *indoor* setting. The images depicting scenes of nature, captured at e.g., city parks and popular resting spots are also frequently chosen for the visual summaries. Additionally, the analysis shows that the images depicting bodies of water (e.g., rivers, canals and lakes) have a higher likelihood of being selected for the visual summary.

We repeat the above-mentioned procedure for identifying concepts typical of images that appear least frequently in the human created reference summaries (i.e., semantic concepts associated with "negative" image examples). Here, some of the concepts populating top-30 of the ranked list are

Event, *Computers*, *Science_Technology*, *Table*, *Chair*, *Furniture* and *Indoor*, but also *Male_News_Subject*, *Advocate*, *Actor*, *Corporate-Leader*, *Civilian_Person*, *Old_People* and *Adult*. Similar to the results presented in Table 1, we can again speak of a strong trend, as the percentage of locations for which the top-30 semantic concepts score higher on the negative examples ranges from 67% to 81%. The results suggest a negative user preference towards images depicting indoor setting and, in particular, a "corporate" environment. Although the collection includes a large number of images depicting people in their everyday activities, it is indicative that they are seldom appearing in human-created reference summaries.

3.2 Concept Co-occurrences

We now analyse which concepts frequently co-occur in the images that make good candidates for the visual summary. Similar to the procedure described in the previous section, for each collection subset we identify the pairs of concepts having higher confidences among the positive image examples than among the negative ones. For clarity of presentation, in Figure 2 we select the top-30 semantic concepts identified in the previous section (cf. Table 1) and for each concept-pair display the percentage of locations for which they score higher on images frequently chosen for the visual summary by the humans. We conjecture that some semantic concepts capture similar information, which is caused by, among other factors, the characteristics of the image sets they have been trained on and the features used for training. To account for this effect, considering images from the entire collection, we compute correlation between semantic concepts and then cluster them using the affinity propagation clustering [6] that was proven effective in automatically determining the optimal number of clusters. In Figure 2, the semantic concepts belonging to the same cluster are marked with a "♦" symbol. For reasons of visual neutrality, the blocks on the main diagonal are assigned a mean heat map intensity.

The values reported in Figure 2 range from 58% up to 79%, which suggests that the selected concept co-occurrences tend to be associated with positive image candidates in a particularly high percentage of collection subsets. The results further confirm our assumption about users' preference towards "panoramic" images depicting e.g., cityscapes, individual monumental buildings, landscapes and the bodies of water against the skyline.

4. LEARNING USER PREFERENCES

The results presented in Section 3 reveal a strong relation between user preferences in visual summarization and the image semantics. Therefore, in this section we further investigate the possibilities of utilizing semantic concept detections for learning to rank images according to their likelihood of appearing in the human-created reference summaries.

For each of t subsets (i.e., geographic locations) in the training set, similar to procedure described in Section 3, we rank images according to the number of reference summaries they appear in and then select top-N and bottom-N images as the positive and negative examples respectively. To compensate for the noisiness and imperfection of semantic concept detectors, we reduce dimensionality of concept vectors to 30 components (i.e., less than 10% of the original vector size) using principal component analysis (PCA). We

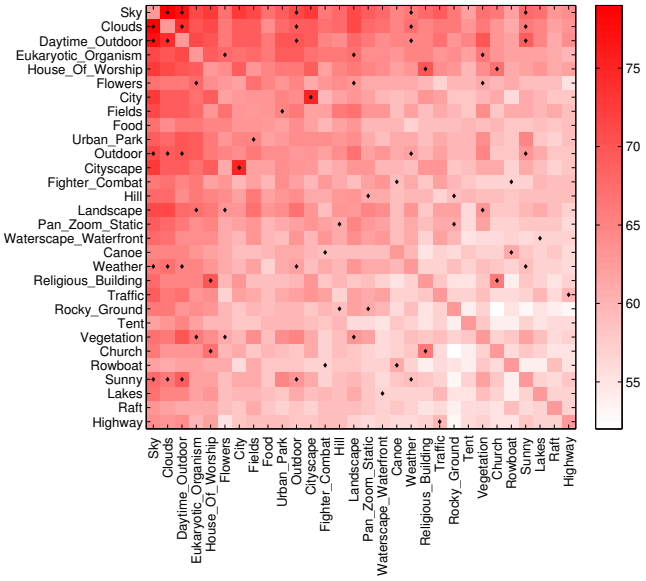

Figure 2: Co-occurrence of semantic concepts in the images frequently selected for the visual summary.

conjecture that the semantic concepts alone are insufficient for making reliable predictions. The reason is that they do not directly incorporate information about image context (e.g., how representative of a given location is a particular image), popularity, aesthetic appeal and sentiment. Therefore, for improved image representation, we combine concept vectors with the heterogeneous set of features proposed in [14]. Further, we feed the N preference pairs, each consisting of a top-ranked and a bottom-ranked image, to a fast pairwise learning to rank algorithm [5]. Since the reference summaries were collected per location and due to the varying underlying distribution of images captured at different locations, the RankSVM algorithm is trained for each location in the training set separately.

Given a location from the test set, we can now apply the trained RankSVM model to produce t ranked lists of images. A single, reinforced results list is then produced using a rank aggregation approach that computes the average rank for an image across all t lists.

5. EXPERIMENTAL RESULTS

To quantify the contribution of image semantics analysis to the overall improvement in automatic image selection, we compare the performance of the approach presented in Section 4 with the one introduced in [14]. For convenience hereafter we refer to them as the *CAS+SC* and *CAS* respectively, to reflect the fact that the approach presented in [14] makes use of features derived from the analysis of image content and context, aesthetic appeal and sentiment, but does not utilize the semantic concepts. Additionally, we show the performance of a "control" baseline denoted as *VC*, which ranks images according to their view count. This is a common ranking strategy in content-sharing websites.

In Table 2 we compare the performance of the above-mentioned image selection approaches for the various sizes of selected image set N_R. For cross-validation, we adopt a leave-one-out strategy, using at each step the images from

Table 2: Performance of our proposed CAS+SC and CA+SC image selection approaches and the alternatives, expressed in terms of Pyramid score averaged over all 207 locations

Approach	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
VC	0.4135	0.4497	0.4778	0.5015
CA	0.5294	0.5342	0.5602	0.5781
CA+SC	**0.5614**	**0.5803**	**0.6058**	**0.6289**
CAS	0.5660	0.5736	0.5961	0.6216
CAS+SC	**0.5923**	**0.6112**	**0.6323**	**0.6484**

$t = 206$ subsets for training and the remaining subset for validation. The performance is expressed in terms of the averaged Pyramid score [11], inspired by the official evaluation metric of the Text Analysis Conference (TAC) summarization track [12] and adapted to visual summarization domain in [14]. Comparison of results yielded by CAS+SC and CAS approaches confirms that the use of semantic concepts brings a modest, yet consistent improvement to the overall image selection performance.

In many settings (e.g., personal offline image collection), the images are missing annotations and the information about user interactions with them. To investigate the added value of using semantic concept detectors in such cases, here we consider following modifications of the above-mentioned algorithms: *CA* [14], where information about image popularity and the sentiment evoked in the users is not available and the semantic concept detectors are not utilized either; *CA+SC*, a modification of the previous algorithm, extending the feature set with a distribution of semantic concepts.

The benefits of using image semantics analysis are also observed in case of an unannotated image collection, where our proposed CA+SC approach clearly outperforms CA and even emerges as the overall second-best performer (cf. Table 2). The last observation is particularly interesting because it suggests that the use of semantic concepts may in certain use cases, such as the one presented in this paper, compensate for the lack of social annotations and the information about user interactions with the images.

6. CONCLUSION

We have analysed the link between image semantics and user preferences in visual summarization. The experiments conducted in the context of visual summarization of geographic areas indicate that such analysis may be effectively performed on a large scale using automatically detected semantic concepts, despite their noisiness and imperfection. The analysis identified individual semantic concepts and the concept pairs that make certain images good candidates for the visual summary from the user's point of view. Additionally, our experiments suggest that the user preferences can be effectively captured in a distribution of semantic concepts detected in the images. This can be further utilized for improved learning to discriminate between images based on their likelihood of being selected for a visual summary by a human. A performance improvement is observed in both settings of an information-rich collection of user-contributed images and an unannotated, offline image collection.

In our future work, we will further investigate relation between image semantics and various criteria influencing users' choice of images for the visual summary, ranging from topical representativeness and diversity to image aesthetic appeal, sentiment and popularity. Additionally, we will investigate the possibilities of utilizing semantic concepts in evaluation of visual summaries.

7. REFERENCES

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM '13*, pages 223–232, 2013.

[2] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *ICML '10*, pages 111–118, 2010.

[3] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T. Huang. A worldwide tourism recommendation system based on geotagged web photos. IEEE ICASSP '10, pages 2274–2277, 2010.

[4] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR '98*, pages 335–336, 1998.

[5] O. Chapelle and S. Keerthi. Efficient algorithms for ranking with svms. *Inform. Retrieval*, 13(3):201–215, 2010.

[6] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[7] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, and B. Loni. Div400: A social image retrieval result diversification dataset. In *ACM MMSys '14*, pages 29–34, 2014.

[8] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *WWW '08*, pages 297–306, 2008.

[9] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *WWW '14*, pages 867–876, 2014.

[10] M. Lestari Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: Overview of the imageclefphoto task 2009. In *LNCS*, volume 6242, pages 45–59. 2010.

[11] A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2), 2007.

[12] K. Owczarzak and H. T. Dang. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *TAC '11*, 2011.

[13] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang. Summarizing tourist destinations by mining user-generated travelogues and photos. *Comput. Vis. Image Und.*, 115:352–363, 2011.

[14] S. Rudinac, M. Larson, and A. Hanjalic. Learning crowdsourced user preferences for visual summarization of image collections. *IEEE Trans. Multimedia*, 15(6):1231–1243, 2013.

[15] C. G. M. Snoek, K. E. A. van de Sande, A. Habibian, S. Kordumova, Z. Li, M. Mazloom, S. L. Pintea, R. Tao, D. C. Koelma, and A. W. M. Smeulders. The mediamill trecvid 2012 semantic video search engine. In *TRECVID Workshop*, 2012.